Neural Networks and Support Vector Machines Classifiers for Writer Identification Using Arabic Script

Sami Gazzah[1] and Najoua Ben Amara[2]
[1]National School of Engineers of Sfax, Tunisia [2]National School of Engineers of Sousse, Tunisia

Abstract: *In this paper, we present an approach for writer identification carried out using off-line Arabic handwriting. Our proposed method is based on the combination of global and structural features. We used genetic algorithm for feature subset selection in order to eliminate the redundant and irrelevant ones. A comparative evaluation between two classifiers is done using Support Vector Machines and Multilayer Perceptron (MLP). The best results have been achieved using optimal feature subset and MLP with an average rate of 94%. Experiments have been carried out on a database of 120 text samples. The choice of the text samples was made to ensure the involvement of the various internal shapes and letter locations within a subword.*

Keywords: *Writer identification, off-line Arabic handwriting, genetic algorithm, support vector machines, multilayer perceptron.*

Received June 16, 2006; accepted July 23, 2006

1. Introduction

The off-line handwriting is the most difficult problem to solve in the field of OCR-Optical Character Recognition due to the wide variability of handwriting styles, existence of within /between writer variations and the absence of an additional source of information (compared to the on-line system). In this case the recognition rate drops with the use of a large vocabulary and multi-writer context. However most of the research effort has been devoted to arming the recognition system by an additional improving step. This additional step can be held:

- After the OCR (a posterior or post-recognition step) by matching the recognized word against a lexicon to measure the similarity and using this information to make a suitable correction in order to improve the recognition rate. The disadvantage of this method is that the running time increases linearly with the number of words in the lexicon.
- Before the OCR (a priori) by extracting from image text the writer's style identity. Writer identity offers a valuable additional source of information to the recognition system that allows transforming the complexity of the multi-writers system into mono-writer ones, as shown in Figure 1.

In spite of its usefulness and meaningful improvement in recognition for off-line handwriting, personal identification has been almost neglected especially in Arabic OCR [9]. The identification task includes a comparison of a set of handwriting samples provided from an unknown writer with each class reference pattern, and decides which reference pattern matches the unknown writer features best.

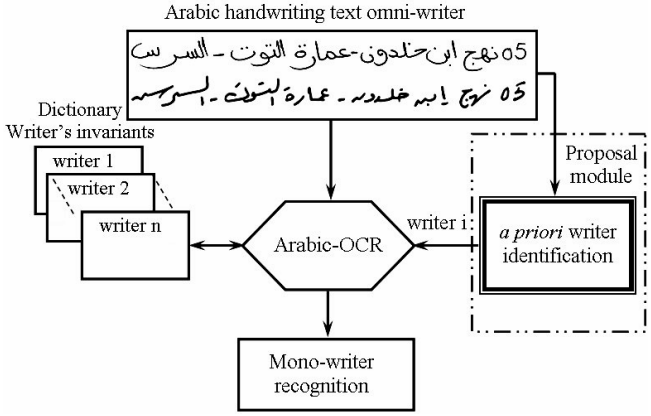


Figure 1. Recognition system improved by adding a priori writer identification module.

As a signature, a person's handwriting has long been considered as an important behavioural characteristic and was accepted as a form of human identification. In spite of the existence of within-writer variation, we can notice that this variation is less significant than the between-writer variation, as shown in Figure 2. Each person writes in his own style. This problem can be treated in two ways: the verification and the identification of the writer [1].

Writer verification is a process to compare questioned handwriting with samples of handwriting obtained from sources for the purpose of determining whether the two samples were written by the same writer or by two different writers. This is a two class categorization problem: authorship or non-authorship [3, 20, 24].

(a) Written 6 times by the same writer.

(b) Written by 6 different writers.

Figure 2. Illustration of within-writer and between-writer variation by the superposition of a same word.

The investigations in the field of writer verification are attributed basically in the biometric axis of research. This axis still remains a very active research topic because it aims to replace the intuitive methods (made by experts) with scientific rigorous ones, whose effectiveness can easily be established with precision.

According to our bibliographic research, the problem of writer identification has been carried out only for Latin [1, 3, 18, 19, 20, 24] and Chinese scripts [23, 25]. However, such studies will be a valuable contribution in the case of Arabic. In this paper, we present a contribution to the writer identification using off-line Arabic handwriting. Our approach is based on the combination of global and structural features. Our various works in recognition of the off-line Arabic handwriting [6] shows the complexity of the task particularly in a multi-writer context. The variations within and between-writers are accentuated by the calligraphic nature of Arabic script.

In this paper, we present in the next section some problems related to the morphological variation of this script. In section 3, we give a brief description of the Multilayer Perception classifier, SVM classifier and features selection problem. In section 4, we give an overview of our system design. In section 5, the experimental results are given for two feature sets. Conclusions and future work are drawn in the last section.

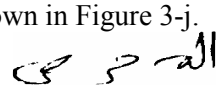
2. Problems Related to the off-Line Arabic Handwriting

The recognition of the off-line multi-writers Arabic handwriting is a non-trivial task due to the morphological characteristics of this script. It is assumed to be among the most difficult styles [13]. To analyze the related problems, we undertook a study on handwritten Arabic variations on a significant number of Tunisian town/village names extracted from the IFN/ENIT database [17].

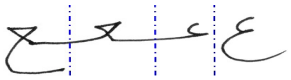
We retained significant morphological variations at different levels:

- For Arabic script, the word is composed of one or several subwords. This concept often loses its significance in the case of the handwriting because of the very frequent discontinuity of the writings or overlapping in consecutive subwords, as shown in Figure 3-a [4].
- A character assumes up to four shapes according to its location within a subword: in isolation, beginning, middle, or final, as shown in Figure 3-b.
- The characters layouts present a wide morphological variability, as shown in Figure 3-c.
- The writing presents different inclinations, as shown in Figure 3-d.
- The thickness of the writing depends on the device of writing, the way of writing and the pressure applied on the pen, as shown in Figure 3-e. A fine layout leads to a better shape of the letters, but may cause disconnections in the continuity of the shape. A thick writing is likely to cause the obstruction of the loops [25].
- Several characters can be written in a combined way according to the style of the writer. One subword can be written differently with one, two or three connected characters (vertical ligatures). A ligature occurs when 2 or more letters overlap vertically and touch, as shown in Figure 3-f.
- For aesthetic reasons, certain writers involve a sense of beauty and "lengthen" the horizontal ligatures. The density of the writings varies from one writer to another and the same subwords will not have the same length in the context of the handwriting. A too dense writing can involve an anomalous structure of the letters, while a too weakened writing can involve an abnormal extension of the letters being able to degenerate into an excessive elongation at the end of the word, as shown in Figure 3-g.
- Certain handwriting present ambiguous characters, which induce confusions in the stage of recognition by gathering characters belonging to different classes, as shown in Figure 3-h.
- Overlaps between two or more subwords (not necessary between succeeding subwords) are very frequent and pose challenging problems. An overlap occurs when two or more subwords overlap vertically without touching, as shown in Figure 3-i.
- For Arabic script, sixteen characters include in their initial shape diacritic dots which can be at the number of one, two or three. These dots can be either above or below the primary stroke of the character. These dots can take abnormal forms or locations compared to the layout shape of the character. They can be connected together or with

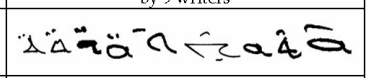
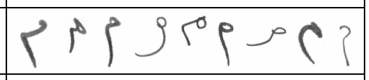
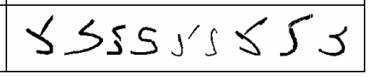
the body of the character. Sometimes, they can be missed or take a wrong position. Certain diacritics (such as kachida "ˇ") can have forms which can be confused with diacritic dots or even with low-size characters, as shown in Figure 3-j.



(a) Example of a discontinuous writings.

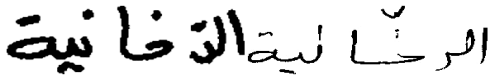


(b) Illustration of the different shapes of the letter "ع" (ain), from right to left: initial, middle, end and isolated.

Printed from	Same character form written by 9 writers
ة	
م	
ك	

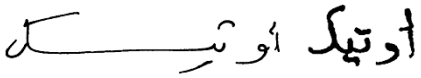
(c) Example of morphological variability of the character's shape written by nine writers compared to the printed form of the same character.

(d) Inclination of the writing.

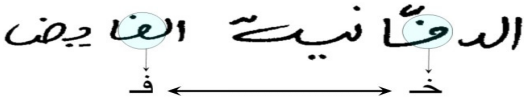


(e) Examples of thickness variation.

(f) Several characters can be written in a combined way according to the style of the writer.



(g) Writing density variations, from left to right: slackened writing & dense writing.



(h) Ambiguous characters.

(i) Examples of overlaps and ligatures.

(j) Dots taking different position or shape.

Figure 3. Illustration of some problems related to the off line Arabic handwriting.

3. Preliminaries

In this section, we give a brief overview of the classification used methods.

3.1. Overview of Multilayer Perceptron

Neural networks have been intensively used in the area of pattern recognition and have increasingly received considerable attention in various areas such as signal processing, pattern recognition and automatic control [5].

The most popular network is the multilayer perceptron architecture which is trained using the backpropagation algorithm. It consists of at least three layers: an input layer, an output layer and one or more hidden layers, as shown in Figure 4.

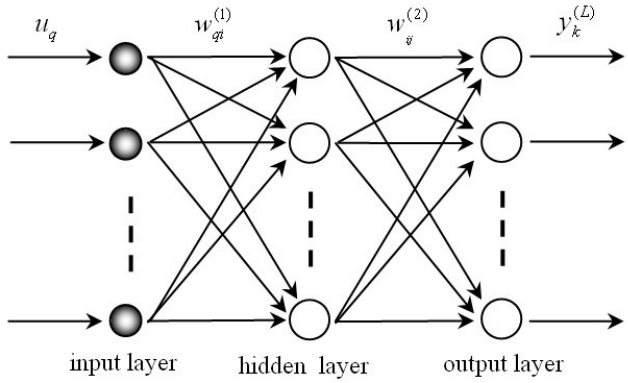


Figure 4. An example of two layers feed forward network architecture.

Suppose we have *n* hidden layers and *u* input vector, the output *y* will be [15]:

$$y=\xi_n[w_n\xi_{n-1}[w_{n-1}+...+w_2\xi_1[w_1u+b_1]+b_2+...+b_{n-1}]+b_n] \quad (1)$$

Where $(w_i)_{1\le i\le n}$ is the weight matrix of the layer *i*, b_i is the bias vector of the neuron and ξ_i is a non-linear operator defined by:

$$\xi_i(x)=[\gamma_{i1}(x_1),\gamma_{i2}(x_2),...,\gamma_{ik}(x_k)]^T \quad (2)$$

Where $\gamma_{iv}(x_v)$ is the activation function of neuron *i* from the layer *v*. The transfer function is in general a sigmoid:

$$f(x)=1/(1+e^{-cx}) \quad (3)$$

MLP is trained by a gradient descent using the backpropagation algorithm to optimise the cost function. For example, the most common cost function is the mean square error criteria which summed the squared error between the desired and actual output vectors.

$$E(w,b)=\frac{1}{2}\sum_i^l (Y_i-O_i)^2 \quad (4)$$

Where *Y* is the true target value and *O* is the output of the network (function of *w* and *b*). The purpose of

neural-network training is to adjust w in such way that the error function is minimized. The backpropagation is a gradient descent on this cost function and the backpropagation weight updates are equivalent to:

$$\Delta w_{ij}(n) = -\eta(\delta E / \delta w_{ij}) \quad (5)$$

Where η is a positive step size known as the learning rate and w_{ij} is the connection strength of the wire between nodes i and j. In (5), if η is small enough, the above weight updating will decrease the error between the desired output and the actual network output.

$$\Delta w_{ij} = -\eta \delta_i E_j \quad (6)$$

Output layer neurons are:

$$\delta_i = f'(e_i)({}_iO_i - Y) \quad (7)$$

And the hidden layer neurons are:

$$\delta_i = \sum_{k=1}^{n\_nodes} \delta_k w_{ki} \quad (8)$$

Where k describes all the nodes in the next layer.

The number of hidden neurons depends upon the complexity of the problem to be solved. The precise number of hidden neurons required for a given modelling task remains an open research problem.

3.2. Overview of Support Vector Machines

In the last years, Support Vector Machines have been massively used by machine learning and pattern recognition communities. They have successfully been applied to several different areas ranging from face recognition and verification, speaker verification, text categorization, prediction, image retrieval, and handwriting recognition [16]. A detailed description of SVM classifiers can be found in [2, 14, 22].

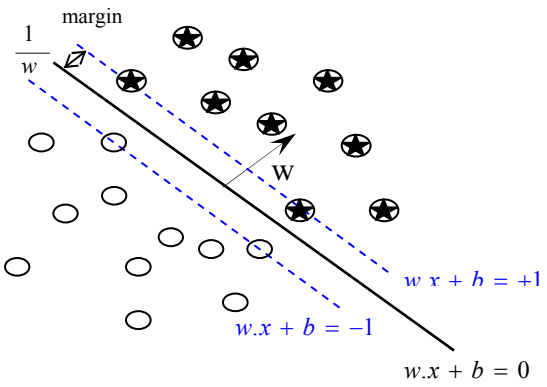


Figure 5 . SVM linear binary classifier.

Suppose we have n training data points $(x_i, y_i)_{1\le i\le n}$, where $x_i \in \mathcal{R}^d$ and $y_i = \{\pm 1\}$. We want to classify x among two classes.

First, we consider the case of linear support vector machines to discriminate two classes (positive and negative classes), which are separated by a hyperplane.

$$y = sign((w.x) + b) \quad (9)$$

Where w is normal to the separating hyperplane, b is an offset and $(w.x)$ is the scalar product between w and x, as shown in Figure 5.

Our target is to try to maximize the margin of the classification on the training data set. We consider the margin as the minimal distance between any training points to the decision boundary.

For separable data belonging to respectively positive and negative classes, we have:

$$\begin{cases} x_{i.}.w + b \ge +1 & for\ y_i = +1 \\ x_{i.}.w + b \le -1 & for\ y_i = -1 \end{cases} \quad (10)$$

The goal of maximizing the margin can be formulated as a quadratic optimization problem:

$$min_{w,b} \frac{1}{2}\|w\|^2$$

$$\text{subject to} \quad y_i(x_i.w + b) - 1 \ge 0 \quad (11)$$

Thus, by solving this optimization problem with inequality constraint, we get the following Lagrangian:

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i (y_i.((w.x_i) + b) - 1) \quad (12)$$

Where α_i are non-negative Lagrange multipliers.

We have to minimize $L(w,b,\alpha)$ with respect to primal variables w, b, and to maximize it with respect to α_i. We have to consider the conditions:

$$\frac{\partial L}{\partial b} = 0 \quad and \quad \frac{\partial L}{\partial w} = 0 \quad \text{which implies:}$$

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (13)$$

substituting them into $L(w,b,\alpha)$, we have the following equation in which the primal variables are eliminated:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i.x_j \quad (14)$$

From this equation we can solve the lagrange multiplier α_i and replace it in $w = \sum_{i=1}^{n} a_i y_i x_i$. Finally, the classification occurs using the following hyperplane decision function:

$$\begin{aligned} f(x) &= sign((w.x) + b) \\ &= sign((\sum_{i=1}^{n} \alpha_i y_i x_i).x + b) \\ &= sign(\sum_{i=1}^{n} \alpha_i y_i (x_i . x) + b) \end{aligned} \quad (15)$$

Due to the fact that classification problems can't be solved easily by a linear classifier, it is more practical to give the classifier the ability to solve non-linear problems. Therefore it is necessary to map the input space into a high dimensional features space F through non-linear map:

$$\Phi : \mathbb{R}^n \rightarrow \mathbb{F} \tag{16}$$

The training algorithm will use the data set through dot products in F. We substitute $\Phi(x_i)$ for each training example x_i :

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \Phi(x_i).\Phi(x_j) \tag{17}$$

Thus, when we replace $(\Phi(x_i).\Phi(x_j))$ with suitable kernel function $K(x_i, x_j)$, the dual quadratic optimization problem becomes [12]:

$$\max_{\alpha} \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i,j} a_i a_j y_i y_j K(x_i x_j)$$

$$\text{subject to } \sum_{i=1}^{n} a_i y_i \qquad a_i \geq 0, (1 \leq i \leq n) \tag{18}$$

Thus, by solving this optimization problem we get the following non-linear function:

$$f(x) = sign(\sum_{i=1}^{n} a_i y_i . K(x_i . x_j) + b) \tag{19}$$

The following table shows the most common Kernels:

Table 1. Most common kernels.

Kernel	Function
Linear	$K(x, y) = (x.y)$
Gaussian (radial basis function)	$K(x, y) = \exp(-\frac{\|x - x_i\|^2}{2\sigma^2})$
Polynomial	$K(x, y) = (x.y)^p$
Tangent Hyperbolic	$K(x, y) = \tanh(x.y - \theta)$

Table 2. provides a summary of the main differences between MLP and SVM classifiers

Table 2. Comparison between MLP and SVM.

	MLP	SVM
Architecture choice	• Number of hidden layers. • appropriate initial weights • Number of neurons per hidden layers. • Activation function	• Kernel function • Correspondent parameter to the kernel function
Performance	• May converge to local minimum • Risk of overtraining	• Converge to global minimum • No risk of overtraining

4. Proposed Approach

The proposed system is composed mainly of four modules: text-image acquisition, pre-processing, features extraction and the writer style classification, as shown in Figure 6.

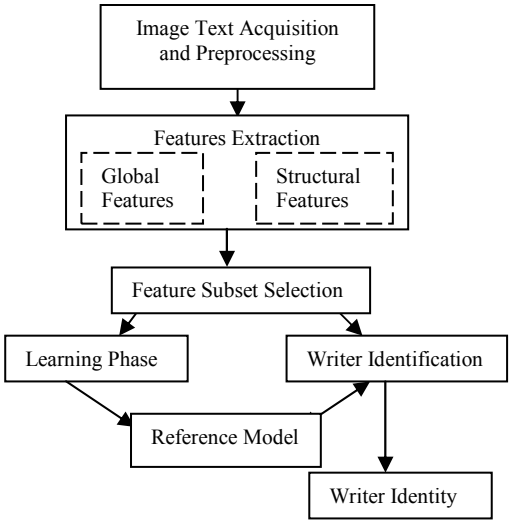


Figure 6. The flow chart of the proposed writer identification system.

Text images are digitized in grayscale at a resolution of 300 dpi, a median filter is used then to remove salt and pepper noise from each page.
In the following, we describe in more details feature extraction and classification modules.

4.1. Features Extraction

Each selected feature should morphologically characterize the style of writing of a script writer (authorship) to dissociate it from the other script writers (non-authorship). We combined two types of features: structural and global ones. Feature vectors are extracted from handwriting at three levels: ascender, subwords and text line, as shown in Table 3.

Table 3 . Features extraction levels.

Feature	Extraction Level		
	Ascenders	Subwords	Text line
Line height			x
Spaces between subwords			x
Inclination	x		
Dots boldness and shape		x	
Wavelet transforms		x	
Entropy		x	

We retained four structural features:

a. *Average line height:* In a previous step, we fixed space between the text lines to avoid the overlaps between succeeding ones. Under this condition, the extraction of the text lines of writings becomes easier by the method of the horizontal projection profile. The height 'h' of the line is measured directly on the histogram.
b. *Spaces between subwords:* An algorithm allows to cross the baseline and to calculate the average of white spaces between subwords.
c. *Inclination of the ascenders:* First, we extract the zone delimiting the ascenders then we measure the angle of inclination 'γ' for each detected ascender, as shown in Figure 7.

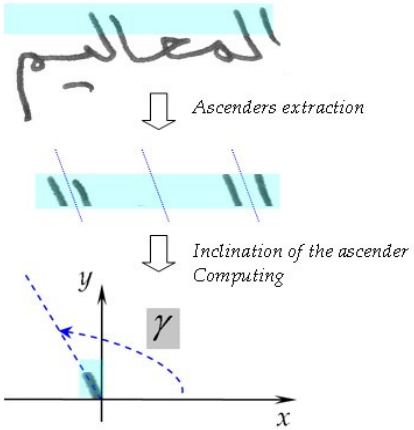


Figure 7. Ascenders inclination of the feature extraction.

d. *Features extracted from dots:* These features are extracted from isolated diacritic dots. The height and the width of each diacritic dot (isolated dots) are measured on bounding box. The ratio of the sum of black pixels is also calculated. Then we compute:

$$R_{h/l} = \frac{a}{b} \quad \text{and} \quad R_{pixels} = \frac{P_n}{P_t} \times 100 \qquad (20)$$

Where b is the width, a is the height, P_n is the sum of the black pixels and P_t is the sum of the total pixels in the bounding box.

We used wavelet transform and entropy to extract textural features as follows:

a. *Wavelet transforms*: We used Daubechies wavelet transforms to describe cursive Arabic subwords. The same feature set was used in the case of Arabic Optical Font Recognition and shown promising performances [11]. The 2D wavelet decomposition is applied until level three. We retain the mean and the standard deviation from each matrix subband (approximation, horizontal details, vertical details and diagonal details) as parameters describing the handwriting feature.

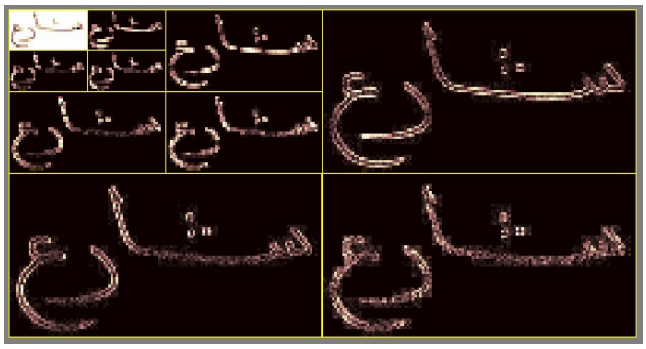


Figure 8. Decomposition of the word شارع (Street) at level three.

b. *Entropy:* The entropy is a statistical measure of randomness (probability distribution) used to characterize the texture of the input image. It is defined as:

$$-\sum p(i) \log\ p(i) \qquad (21)$$

where $p(i)$ is the probability of the i^{th} bins level being used and computed from the grayscale histogram.

The feature vector is composed of twenty parameters corresponding to each line text. Table 4 gives the whole parameters set.

Table 4. Extracted features.

f_1	Line height.
f_2	Average of white spaces between subwords
f_3	Mean of ascender's inclination.
f_4	Standard deviation of ascender's inclination.
f_5	Mean of average ascender's width.
f_6	Standard deviation of average ascender's width.
f_7	Mean $R_{h/l}$ computed from each text line.
f_8	Standard deviation $R_{h/l}$.
f_9	Mean R_{pixels}.
f_{10}	Standard deviation R_{pixels}.
f_{11}	Mean of approximation image matrix.
f_{12}	Standard deviation of approximation image matrix.
f_{13}	Mean absolute deviation of approximation image matrix.
f_{14}	Mean horizontal details matrix.
f_{15}	Standard deviation horizontal details matrix.
f_{16}	Mean vertical details matrix.
f_{17}	Standard deviation vertical details matrix.
f_{18}	Mean diagonal details matrix.
f_{19}	Standard deviation diagonal details matrix.
f_{20}	entropy.

The Mean Absolute Deviation (MAD) is defined by:

$$MAD = \frac{1}{N} \sum_{i=1}^{N} \left| x_i - \bar{x} \right| \qquad (22)$$

where the features parameters size is N, the parameters have values x_i and the mean is:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (23)$$

4.2 Feature Selection

The main idea of feature selection is to select a subset of input variables by cutout features with weakly or non predictive information while maintaining or performing classification accuracy. In this context, genetic algorithms have been successfully implemented to select feature subsets in classification problems. Genetic algorithm is a form of inductive learning strategy inspired by the process of biological evolution [21]. In this work, we use the genetic algorithm to feature subset selection for writer identification using Arabic script.

Consider the original feature set extracted in the previous step, as shown in Table 4, $F = \{f_1, f_2, ..., f_{20}\}$ we used genetic algorithm for feature subset selection in order to eliminate redundant and irrelevant features. Experiments were conducted using the following parameter settings:

Population size: 30
Number of generation: 1000
Probability of crossover: 0.8
Probability of mutation: 0.002
Selection function: roulette wheel
Fitness function: Mean Square Error (we minimize this function). The initial population is generated randomly.

The different experiments lead to the selection of twelve parameters among the original feature set.

$$F_{subset} = \{f_1, f_3, f_4, f_5, f_7, f_9, f_{11}, f_{12}, f_{14}, f_{17}, f_{18}, f_{20}\}$$

5. Experimentation and Results

In this section, we first describe our database design and then we present the different results achieved using Multi-Layer Perceptron and Support Vector Machines.

3.2. Database Description

For the experiments reported in this paper, we have designed an Arabic letter source document which contains 505 characters, 15 numerals and 6 punctuations, as shown in Figure 9-a. The choice of the letter contents was made to ensure the involvement of the various internal shapes of the letter within a subword (isolated, initial, middle and end). Handwriting samples of 60 persons were collected. Each person was required to copy the same letter three times: two samples were used for the training and the other for the tests which makes a total of 180 A4 format sample pages. A black pen, a white sheet and a plain were provided. These text images are initially digitized in grayscale at a resolution of 300 dpi.

5.2. Results

In this section, we briefly describe the different techniques that we have used to implement a modular network. In fact, it has been shown that the modular architecture offers superiority in terms of convergence and recognition capability over non-modular network [7, 11, 12]. In our case, the classification problem of the sixty writers is decomposed into two-classified sixty sub-problems: authorship and non-authorship. During the phase of training, a sub-network learns the features of the authorship and non-authorship. Thus, each sub-network learns the good examples as well as the bad ones (among the whole training base). In this stage we faced the problem of the non-proportionality between the numbers of true examples compared to the false ones: each subnetwork learns two true examples and 118 false.

This problem may bias the results because in the training stage the system learns more non-authorship examples. To solve this problem we used for each non-authorship writer text only one text line chosen randomly from the whole document. Finally, the pages of test are presented at each of the sixty trained sub-networks.

Two sets of classification experiments were carried on the same feature vectors using two classifiers which are support vector machines and multi-layer perceptron.

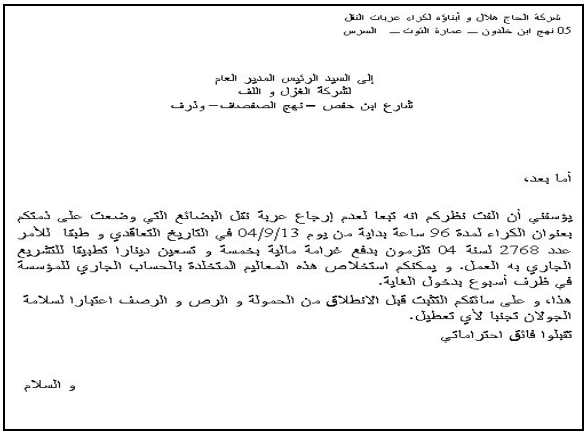
a. *Classification using support vector machines*

We use a support vector machines classifier with radial base function core. Class-modular classification architecture was adopted. It consists of 60 different classes SVM, each one representing the style of one writer which is trained separately. During the training phase, a sub-network learns the features of the authorship and non-authorship. We recorded an average rate of 93.76 %. Table 5 reports the results obtained by each subnetwork of the ensemble for optimal feature subset (obtained after using a genetic algorithm) [10].

b. *Classification using multilayer perceptron*

In the second set of experiments, MLP classifiers were applied in the recognition of the same set of classes. The same set of handwriting attributes (features) for both training and test stages was used as in SVM classifiers. A modular architecture network classifier was adopted. It consisted of an ensemble of 60 MLP sub-networks. The architecture of each subnetwork is composed by two-layers of neurons interconnected by weights; the number of neurons in the input layer is equal to the length of features' vectors. One output neuron takes Boolean values [8]. We recorded an average rate of 94.7 %. Table 6 reports the results

obtained by each subnetwork of the ensemble for optimal feature subset.



(a) The printed form of the document.

(b) An extract of three handwriting samples written by three writers.

Figure 9. Test document.

Table 5. Experimental results obtained with SVM modular network.

Writer	Rate	Writer	Rate	Writer	Rate	Writer	Rate
1	93.12	16	97.74	31	97.57	46	90.93
2	92,05	17	92.01	32	96.48	47	92.17
3	99.19	18	93.22	33	94.88	48	92.95
4	93.44	19	97.74	34	86.95	49	94.17
5	93.93	20	96.51	35	96.27	50	94.85
6	94.48	21	94.63	36	92.54	51	98.14
7	96.26	22	91.74	37	93.78	52	92.31
8	89.8	23	91.12	38	95.57	53	90.89
9	91.12	24	98.6	39	96.73	54	93.71
10	91.56	25	91.61	40	93.49	55	96.5
11	97.67	26	88.08	41	94.88	56	91.45
12	93.93	27	91.84	42	92.68	57	95.66
13	90.19	28	91.59	43	98.37	58	95.11
14	89.72	29	91.06	44	91.08	59	93.44
15	93.22	30	89.16	45	96.2	60	99.76

Table 6. Experimental results obtained with MLP modular network.

Writer	Rate	Writer	Rate	Writer	Rate	Writer	Rate
1	94.16	16	97.66	31	93.47	46	95.12
2	96.73	17	96.04	32	96.04	47	91.16
3	96.74	18	96.96	33	97.44	48	94.63
4	97.20	19	94.63	34	96.25	49	93.01
5	93.26	20	96.03	35	96.87	50	92.06
6	94.19	21	92.29	36	95.10	51	98.14
7	97.20	22	91.59	37	93.94	52	95.10
8	92.99	23	92.29	38	96.97	53	96.03
9	92.99	24	97.67	39	99.07	54	93.47
10	92.74	25	98.37	40	95.37	55	92.54
11	96.27	26	93.24	41	95.35	56	90.44
12	96.73	27	92.54	42	96.28	57	94.63
13	96.03	28	93.46	43	97.90	58	93.72
14	88.08	29	91.36	44	92.77	59	94.87
15	91.36	30	94.63	45	90.42	60	99.30

6. Conclusion and Future Works

In this paper, we have described an approach for writer identification using off-line Arabic handwriting. Our proposed method is based on the combination of optimal local and global feature subset, which was optimised by using a genetic algorithm. Two machine learning algorithms were used: support vector machines and multilayer perceptron. Through experiments, we have found that the MLP provides slightly better results than SVM classifier in the writer identification task.

The study of the performances per author shows that for some writers the MLP classifier achieved better results whereas for the others the SVM gives better performances. A hybrid approach is under development in order to profit from the advantages offered by each classifier.

References

[1] Bensefia A., Paquet T., and Heutte L., "Identification Et Vérification Du Scripteur: 2 Approches Complémentaires Pour L'analyse Quantitative Des Documents Manuscrits," *14e Congés Francophone Sur La Reconnaissance Des Formes Et Intelligence Artificielle (RFIA'04)*, pp. 215-223, Toulouse, France, 2004.

[2] Burges C., "Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and knowledge Discovery,* vol. 2, no. 2, pp. 121–167, 1998.

[3] Cha S., and Srihari S., "Multiple Feature Integration for Writer Verification," *Proceedings of the 7th International Workshop on Frontiers in Handwriting Recognition (IWFHR'00)*, pp. 333-342, Amsterdam, 2000.

[4] Crettez J., "L'écrit Et Le Document: Premier Degré De Caractérisation Des Ecritures Manuscrites," *Traitement du Signal*, vol. 12, no. 6, pp. 567-573, 1995.

[5] Egmont-Petersen M., de Ridder D., and Handels H., "Image Processing with Neural Networks: A Review," *Pattern Recognition Journal*, vol. 35, no. 10, pp. 2279-2301, 2002.

[6] Essoukri Ben Amara N. and Ellouze N., "Overview and Advances in Arabic Optical Character Recognition," *Asian Journal on Information Technology*, vol. 3, no. 4, pp. 217-240, 2004.

[7] Essoukri Ben Amara N., and Gazzah S., "Une Approche d'Identification des Fontes Arabes," *in Huitième Colloque International Francophone sur l'Ecrit et le Document (CIFED'04)*, La Rochelle, France, 2004.

[8] Gazzah S. and Essoukri Ben Amara N., "Writer Identification Using Modular MLP Classifier and Genetic Algorithm for Optimal Features Selection," *Lecture Notes in Computer Science*, *Advances in Neural Networks*, *The Third International Symposium on Neural Networks (ISNN2006)*, vol. 3972, pp. 271-276, Chengdu, China, 2006.

[9] Gazzah S. and Essoukri Ben Amara N., "Neural Networks and Support Vector Machines Classifiers for Writer Identification using Arabic Script," *The second International Conference on Machine Intelligence (ACIDCA-ICMI'2005)*, pp. 1001-1005, Tozeur, Tunisia, 2005.

[10] Gazzah S. and Essoukri Ben Amara N., "Writer Identification Using SVM Classifier and Genetic Algorithm for Optimal Futures Selection," *International Arab Conference on Information Technology (ACIT'2005)*, pp. 461-466, Amman, Jordan, 2005.

[11] Gazzah S. and Essoukri Ben Amara N., "Utilisation des Ondelettes et des Réseaux de Neurones en Identification des Fontes Arabes," *Conférence Internationale Signaux, Circuits et Systèmes (SCS'04)*, pp. 498-502, Monastir, Tunisia, 2004.

[12] Il-Seok OH. and Suen C., "A Class-modular Feedforward Neural Network for Handwriting Recognition," *Pattern Recognition*, vol. 35, no. 1, pp. 229-244, January 2002.

[13] Khorsheed M., "Automatic Recognition of Words in Arabic Manuscripts," *PhD Thesis*, Churchill College, University of Cambridge, 2000.

[14] Müller K., Mika S., Rätsch G., Tsuda K., and Schölkopf B., "An Introduction to Kernel-Based Learning Algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181-201, 2001.

[15] Njah M., "Contribution à la Synthèse des Réseaux de Neurones Multicouches Application Industrielle," *PhD Thesis*, Ecole Nationale d'Ingénieurs de, Tunis, 2003.

[16] Oliveira L. and Sabourin R., "Support Vector Machines for Handwritten Numerical String Recognition," *9th International Workshop on Frontiers in Handwriting Recognition (IWFHR-9)*, pp. 39-44, IEEE CS Press, Tokyo, Japan 2004.

[17] Pachwitz M., Maddouri S., Märgner V., Ellouze N., and Amiri H., "IFN / ENIT Database of Handwritten Arabic Words," *Colloque International Francophone sur l'Ecrit et de document (CIFED'02),* pp. 129-136, Hammamet, Tunisie, 2002.

[18] Said H., Tan T., and Baker K., "Personal Identification Based on Handwriting," *Pattern Recognition*, vol. 33, pp. 149-160, 2000.

[19] Schomaker L. and Bulacu M., "Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 787- 798, 2004.

[20] Srihari S., Cha S., Arora H., and Lee S., "Individuality of Handwriting," *Journal of Forensic Sciences*, vol. 47, no. 4, pp. 856-872, July, 2002.

[21] Vafaie H. and De Jong K., "Genetic Algorithms as a Tool for Feature Selection in Machine Learning," *4th International Conference on Tools with Artificial Intelligence*, pp. 200-203, Arlington, VA, USA, 1992.

[22] Vapnik V., *The Nature of Statistical learning Theory*, Springer, 1995.

[23] Wang X., Ding X., and Liu H., "Writer Identification Using Directional Element Features and Linear Transform," *7th International Conference on Document Analysis and Recognition*, *(ICDAR'03)*, pp. 942-945, Edinburgh, Scotland, 2003.

[24] Zhang B., Srihari S., and Lee S., "Individually of Handwriting Characters," *7th International Conference on Document Analysis and Recognition (ICDAR'03)*, pp. 1086-1090, Edinburgh, Scotland, 2003.

[25] Zhu Y., Tan T., and Wang Y., "Biometric Personal Identification Based on Handwriting," *in Proceedings of the ICPR'00*, vol. 2, pp. 2797-2800, Barcelona, Spain, 2000.

Sami Gazzah was born in Tunis/Tunisia. He received the Master degree in automatic control and industrial computer science from National School of Engineers of Sfax/Tunisia in 2004 and Engineer in electrical engineering

from the same university. He is currently a PhD student in automatic control and industrial computer science in National School of Engineering of Sfax/Tunisia. His current research interests include image document analysis, wavelets, and machine learning,

Najoua Ben Amara received the BSc, MSc, PhD, and HDR degrees in electrical engineering, signal processing and system analysis, from the National School of Engineers of Tunis, Tunisia, in 1985, 1986, 1999, 2004 respectively. From 1985 to 1989, she was a researcher at the Regional Institute of Informatics Sciences and Telecommunications, Tunis, Tunisia. In September 1989, she joined the Electrical Engineering Department of the National School of Engineers of Monastir, Tunisia, as an assistant professor. In July 2004, she becomes a senior lecturer at the National School of Engineers of Sousse, Tunisia. Her research interests include mainly optical character recognition applied to arabic documents, image processing, compression, ancient document processing, biometric and the use of stochastic models and hybrid approaches in the above domains.