Kamel BOUMAZA, Mokhtar SELLAMI BADJI
Mokhtar Annaba University Research Laboratory on computer sciences kamel_boumaza@hotmail.com
sellami@lri-annaba.net

ARABIC AND LATIN WRITING DISCRIMINATION SYSTEM

ABSTRACT

In this paper, we propose a method of an Arabic and Latin discrimination writings, based on morphological, geometrical and statistical analysis. We focused on feature extractions in order to curry out a system able to differentiate between printed or handwritten Arabic and Latin scripts. First, an image acquisition is made by means of a scanner, and then come some pre-treatment steps. After that, we extract the different features constituting an input vector of multi-layer perceptron (MLP) with back-propagation algorithm. Finally and thanks to the network outputs we have been able to undertake a discrimination of each script (Printed or handwritten). The aim of our work is to constitute a postal sorting system in French-Arabic speaking countries.

Key words: Arabic and Latin writing discrimination, morphological, geometrical, statistical features, postal sorting system. MLP, back-propagation algorithm, printed and handwritten scripts.

1. INTRODUCTION:

Pattern recognition is a very vast domain; hence, the difficulty of writing is until now implying several researches. Nowadays, computer scientists are seeking a practical solution to solve their problems, among these problems writing printed or handwritten recognition. The recognition of some printed or handwritten languages is a major problem for the establishment of an automatic system able to identify the different language words [19] .

However, when it comes to several languages at the same time, the complexity of the problem is increased as well as it's currying out [18].

On the most of time and when countries are Multilanguage, a single document page may contain several language scripts. In Algeria for example, there are two kinds of language: the Arabic language and the Latin one, these languages are written in two different scripts: the Arabic scripts and the Latin ones.

In Algeria, postal checks, bank checks, postal addresses and some pre-printed documents are written in Arabic scripts, in Latin scripts or simultaneously in both scripts. In the case of India, there are eighty official languages. Two or more of these languages may be written in one script. Twelve different scripts are used for writing these languages. Under the three- language formula, many of the Indian documents are written in three languages namely, English, Hindi and the state official language [1].

To develop a multi-script optical character recognition (OCR) system for these countries, it is necessary to separate different script forms before feeding them to the individual script recognizers. This is so because development of hybrid OCR for multiple scripts is more difficult than separate OCRs of individual scripts [1].

There are many fields, in which one may apply writing system recognition, the latter seems very important when replacing manual human efforts. Such systems could be found in bank cheques, postal cheques as well as when extracting texts from images and video sequences, in the recognition of postal addresses... etc [5] [3] [13]. Our research is relating to the postal addresses recognition; in French-Arabic speaking countries postal addresses writing is made via two languages (Arabic and French).

In order to set a system able to recognize postal addresses, it is obvious to curry out a Latin writing Recognition system and an other one for the recognition of Arabic writing. The problem we first faced was to determine when exactly we may distinguish or separate the Arabic addresses from the Latin ones? Discriminating these two types of addresses is a particular case of Arabic and Latin writing. It's obvious that the discrimination between two patterns must be made by means of an extraction of significant features for each pattern .These features must be as least as possible, the most identifying and more powerful to put each pattern in its class.

In spite of much efforts and many works have been effectuated aiming to discriminate between languages, there is no efficient system yet.

In this paper we propose an Arabic and Latin writing discrimination method in order to constitute a system able to differentiate and recognize each writing

(Arabic and Latin printed or handwritten).We start by the different methods proposed in the literature then we

describe each script (Arabic and Latin) their nature (printed or handwritten), also neural network

(MLP) with back-propagation algorithm after that we give more details concerning our method and finally, we achieve this paper by a conclusion.

2. LITERATURE PROPOSED METHODS

According to the different existing works differentiating between different writings, we can regroup four principal existing method classes, according to the information analysis level, in order to make decision of identification, we distinguish:

- the methods based on the text block analysis
- the methods based on the text line and word analysis
- the methods based on the connected component analysis
- the methods based on mixed analysis

2.1 Bloc text-analysis-based methods

These methods use the totality of text bloc to make process, they consider that the text bloc as a single entity, they do not make another analysis in text line or connected entity, they suppose that the text to identify is normalized, uniform (interline and inter-word space are constant) and homogeneous (only writing)[5-7][17].

2.2 Text line-analysis-based methods

These methods are based essentially on statistical analysis of text line (one or more words) [1]. We mention the Elgammal and Ismail methods to separate between Arabic and English printed writing [2] and the Fan and al method to discriminate between English, Japanese and Chinese printed and handwritten writing [4].

2.3 connected entity-analysis-based methods

Bloc text analysis based methods and text line analysis based methods have sometimes poor results to identify similar writings or to extract morphological and intrinsic features from some characters. In this case it is necessary to use the connected entity analysis. These entities can be of course segmented when writing is unconnected like Latin, Asian, etc or they can be obtained after an explicit segmentation when writing is connected like Arabic, Bangala, Devnagari, etc[8-11][14].

2.4 Mixed analysis based methods

These methods attempt to develop a technique to differentiate between writings using different available information in different levels of textual entity to identify (bloc, line or word, connected entity). In this field, proposed methods are all based on connected entities analysis and another analysis level which can be text line or text bloc [12-13].

In this frame of methods, text line analysis is used generally in first time to separate between group of writings, then connected entities analysis are exploited to distinguish between different writings of the single group [13]; another strategies seek to take identification decision about independent writings for each analysis level then, they try to make combination and adopt them in order to curry out a system able to resolve crisis.

3. LATIN AND ARABIC SCRIPTS

3.1. Latin scripts

In the Latin alphabet, there are twenty-six letters. Each letter has two different forms: small or capital. Moreover, Latin alphabet is low in diacritic. Two letters “i” and “j” have each one diacritic in up. No diacritic in bottom in Latin alphabet. Therefore, Latin script is semi-cursive in the hand written documents

[3] [12]. On the other hand printed Latin script is a sequence of separated letters [2]. Both following figures show printed and hand-written Latin text:

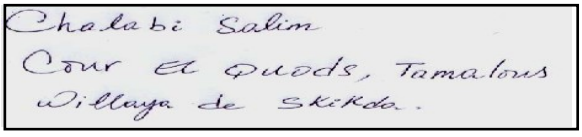


Figure1: Latin handwritten writing

Figure2: Latin printed writing

3.2. Arabic script

Contrary to the Latin script, the Arabic alphabet is rich and complex in characteristics. In the Arabic alphabet there are twenty-eight basic letters [3][16][20,21]. These letters change a shape according to their position in the word. That it is at beginning, medium, end or isolated. So, each character can correspond up to four different forms. In addition, more of half Arabic characters have one, two or three diacritics. These diacritics can be at up or bottom of characters, but never in up and bottom at the same time. The knowledge of presence of these diacritics and their position allows classifying the characters belonging to the same forms family [2]. Arabic scripts are cursive of nature. The letters are connected in the printed and hand-written documents

[1,2]. Both following figures show printed and hand- written Arabic text:

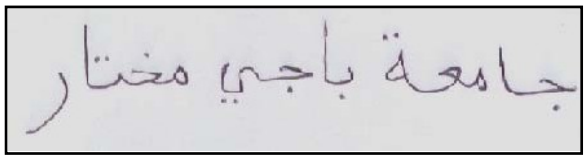


Figure3: Arabic handwritten writing

Figure4: Arabic printed writing

4.NEURAL NETWORK

4.1 Physiological argument

The formal neuron network idea comes from the survey of the human brain. This last is composed of a set of cells called neurons. The human brain is estimated to have around 10 billion neurons each connected on average to 10,000 other neurons. A neuron is composed of a core, of incoming connections (dendrites) and a connection leaves called the axon. The nervous impulse always moves of dendrites toward the core, and of core toward the axon. The impulse transmitted in the axon is function of the value of the impulse in each dendrite. Some dendrites can have a motor effect encouraging the transmission of information in the axon; others have an inhibitory effect that blocks the transmission of the impulse in the axon on the contrary. It seems that the core acted like a summer (adder) of impulses coming from dendrites, while affecting a weight (that can be negative) to these dendrites. If the sum of impulses is superior to a threshold, an impulse is transmitted in the axon. If the sum of impulses is lower to this threshold, no signal is transmitted. An axon can, thereafter, either been subdivided in several filaments that will serve each entry to the others neurons while connecting to dendrites of these neurons via synapses, either to attack a motor element directly (muscle for example) [22][23]. The figure below shows the physiological neuron of human brain.

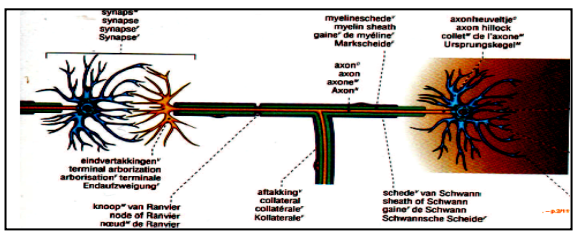


Figure5: physiological neuron of human brain.

4.2 Perceptron Neuron Model

A perceptron neuron, which uses the transfer function ***f***, is shown below.

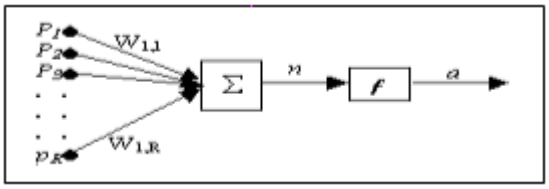


Figure6: Perceptron neuron with transfer function *f*

Each external input is weighted with an appropriate weight w_{1j}, and the sum of the weighted inputs is sent to the transfer function ***f***. The perceptron neuron produces a signal if the net input into the transfer function is equal to or greater than certain threshold, otherwise no signal is produced [24]. The transfer function gives a perceptron the ability to classify input vectors by dividing the input space into regions [25].

5. SYSTEM ARCHITECTURE

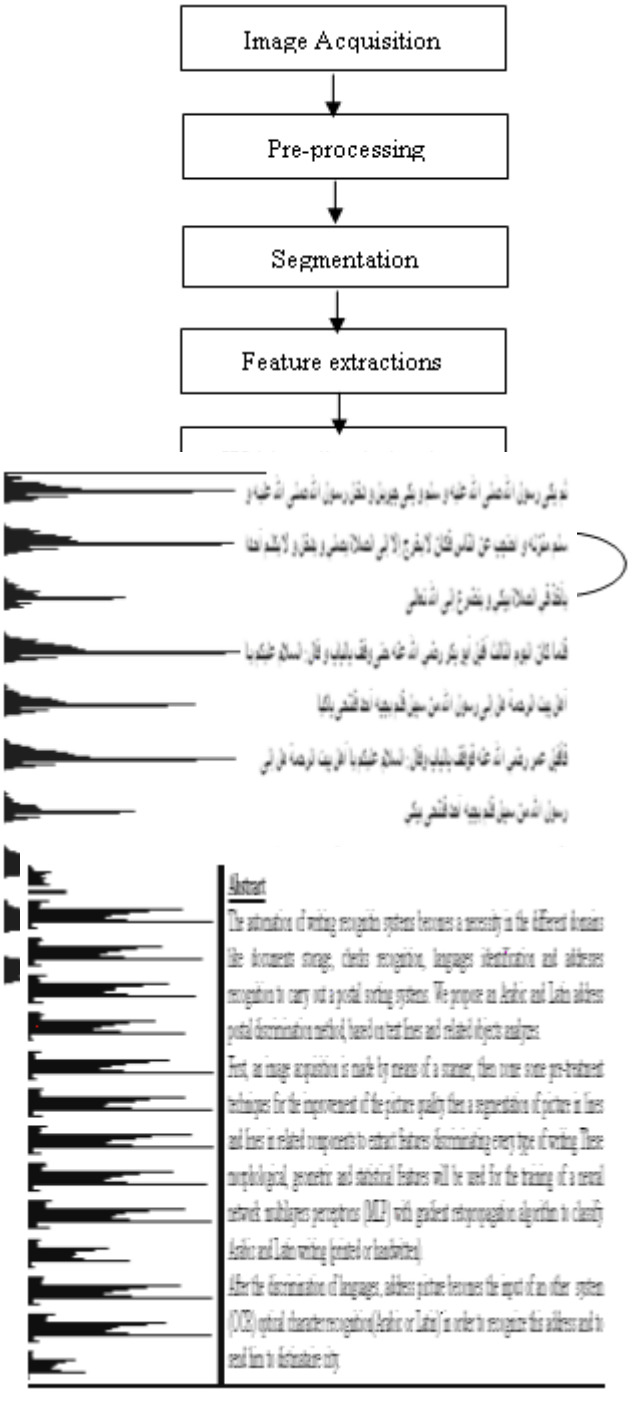


Figure8: horizontal projection profiles of Arabic text lines (on the left) and English text lines (on the right).

Our method is based on an extraction of morphological, geometrical and statistical features. First of all, we suppose that each text line is written in either Arabic or Latin writing.

The architecture of our method is shown on the figure7 above.

First, an image acquisition is made by means of a scanner, and then come some pre-treatment steppes. After that, we extract the different features in order to constitute an input vector for a multilayer perceptron with hidden layer using the back-propagation algorithm. Finally and thanks to the network outputs we have been able to undertake a discrimination of each script.

First, images obtained after scanning must be converted in white and black images. Background colours are white-converted and writing-colours are black-converted. The obtained image is white and black. After finishing this, we undertake a binarization step by which the image becomes a pixel matrix. White pixels are represented by one whereas black ones are represented by zero. After that, a regulation of slant image is effectuated, then a normalization step which consists to resize image with those of the training base is carried out. Once the image is likely to be treated, the image is segmented into lines. Since we supposed that each line is written in one language (Arabic or Latin). In order to carry out this segmentation, a horizontal projection of the image was made (see figure9).

After segmentation of the image into text line, we extract the baseline and the low and high reference lines and another segmentation of line into connected entities is effectuated. After these stages, various

Features discriminating each writing in order to put it in its class are extracted.

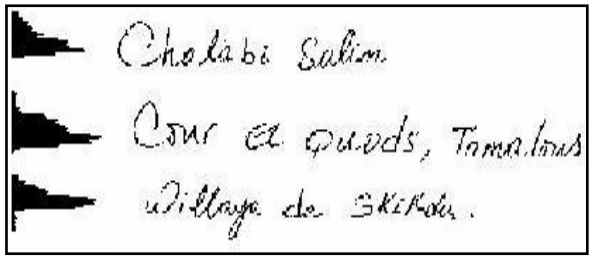


Figure9: horizontal projection of Latin handwriting

The first feature to extract is the number of peaks descended of the horizontal projection profiles of text line[1], so The horizontal projection profiles of an Arabic printed text line have a single peak around the middle of the text line[2][3]. This peak corresponds to the base line of the Arabic writing where characters are connected together. By contrast, projections of Latin printed text line have two major peaks [2] the figure8 above presents different peaks resulting by horizontal projection profiles of various Arabic and English text lines.

After that, we determine number of ascendants and descendants and diacritic points from which we distinguish several types:

- One point on up.
- Two points on up
- Three points on up
- One point in bottom
- Two points in bottom

The figure 10 shows the diacritic points of an Arabic handwriting text line.

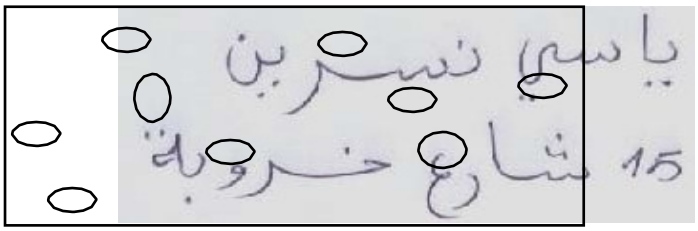


Figure10: Arabic handwriting diacritic points.

The Alif character is more used in the case of Arabic writing so, we extract this character in the phase of connected-entities extractions, and consequently the number of Alif detection is used as a feature characterizing Arabic writing (see figure11).

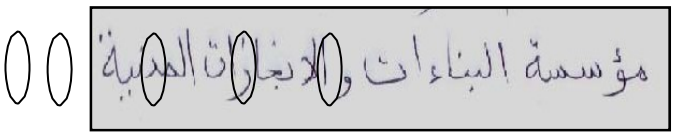


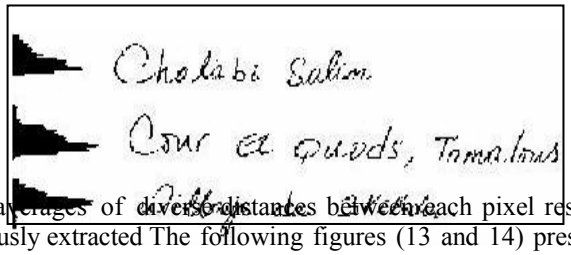
Figure11: Arabic handwriting Alif character.

After extraction of the various connected entities in each text line, we limit each connected component by box. We distinguish that dimensions of boxes including Arabic connected components are different from Latin ones, thus we calculate the average of box width by box height of various connected components in the text line. In addition, the density of black pixels of Arabic connected entities is commonly smaller than the one of Latin ones. Therefore we calculate the average of different densities for each connected entity in the text line.

In the case of Arabic handwritten writing, boxes including connected entities are strongly interconnected between them and sometimes some connected entities include one or more of another ones, by contrast the intersections and inclusions in the case of Latin writing is infrequent . We use the number of intersections between connected entities and number of inclusions as two features for the input vector.

In the case of Latin or Arabic printed writing the interconnected component spaces are regular, thus, we compute the number of equal spaces in order to differentiate between printed and handwritten writing, we used the vertical projection of text line to calculate this number(see figure 13 and 14).In additionin in the case of Latin printed writing , most connected entities have the same height, thing that doesn't exist in the case of printed Arabic, therefore we calculate the number of these connected entities to differentiate between printed Latin and Arabic writing. Another feature discriminating between both printed writing is the concept of water overflow for a reservoir shown below [1] [15].

Figure12: concept of water overflows for a reservoir



The two other features are the averages of diverse distances between each pixel resulting from the high and low profile and the baseline previously extracted The following figures (13 and 14) present a vertical projection, horizontal projection, high profile, low profile and the Sobel edge from top to bottom of Arabic and English printed text line.

Figure13: Arabic printed text line.

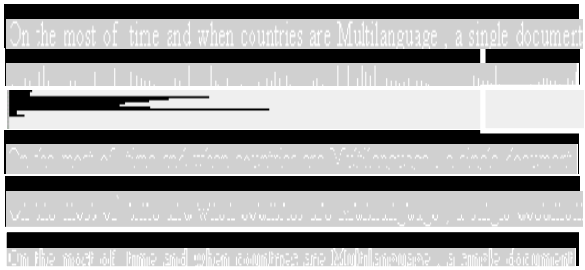


Figure14: English printed text line.

As it is illustrated in both figures above, the sum of distances of each pixels rusulting

from low profile of an Arabic text line is very smaller than Latin text line by contrast; the sum of distances of each pixels rusulting from high profile of an Latin text line is very greater than Arabic text line.

The denstity of pixels resulting from Sobel edge detection is calculated and used as feature discriminating between Arabic and Latin text lines.

Finally vector of twenty features is constituted; this last will be used as an input vector for multilayer perceptron with back-propagation algorithm. The network architecture is presented on the following figure.

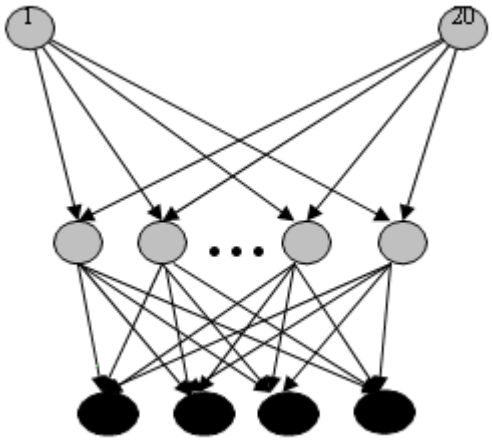


Figure15: network architecture

Network outputs are Printed Latin (PL), Printed Arabic (PA), Handwritten Latin (HL) or Handwritten Arabic (HA)

6. CONCLUSION

This paper presents a method discriminating Latin and Arabic Writing (printed or handwritten) based on morphological, statistical and geometrical features in order to curry out postal sorting system in French-Arabic speaking countries.

The main features of our approach include line text detection, middle zone to separate the upward and downward and diacritic points and connected entities in order to extract the different features which allow identifying each scripts.

Beginning results are encouraging, so, in order to improve our method discriminating between both writings (printed and handwritten) a base of text lines with different writing by different writers with various categories and various old is in development.

7 REFERENCES

[1] U. Pal, S. Sinha and B. B. Chaudhuri "Multi- Script Line identification from Indian Documents" Proc, IEEE, ICDAR 2003.

[2] A. M. Elgammal and M. A. Ismail," Techniques for Language Identification for Hybrid Arabic- English Document Images", 2001.

[3] S.Kanoun, A. ENNAJI, Y LECOURTIER, "Script And Nature Differentiation for Arabic and Latin Text Images", IEEE, 2002.

[4] K. Fan, L. Wang, Y. Tu, "Classification of machine-printedand handwritten texts using character block layout variance", IJPR, Volume31, number9, pp. 1275-1284, 1998.

[5] L Wood, X.Yao, K. Krishnamurthi, Dang, "Language Identification for Printed Text Independent of Segmentation", Proc.IEEE ICIP'95, pp. 428-431, 1995.

[6] T.N. Tan, "Rotation Invariant Texture Features and Their Use in Automatic Script Identification", Proc. IEEE, PAMI, vol.20, no. 7, pp.751-756, 1998.

[7] Y. Tao, Y.Y. Tang, "Discrimination of Oriental and Euro-American Scripts Using Fractal Feature", ICDAR'01, pp.1115-1119, 2001.

[8] A. L Spitz, "Determination of the Scripts and Languages content of document images", PAMI, vol.19, no. 3, pp. 235-245, 1997

[9] D.S. Lee, C.R. Nohl, H.S. Baird, "Language Identification in Complex, Unoriented and Degraded Document Images", DAS'1996, pp. 63-79, 1996.

[10] J. Hochberg, P. Kelly, T. Thomas, L. Kerns, "Automatic Scripts Identification From Document Images Using Cluster-Based Templates", PAMI, vol. 19, no. 2, pp.176-181, 1997.

[11] J. Hochberg, K. bowers, M. Canon, P. Kelly, "Script and Language Identification for Hundwritten Document Images", IJDAR, vol. 2, pp. 45-52, 1999.

[12] L. Lam, J. Ding, and C.Y. Suen," Differentiating between oriental and European scriptsby statisticl features", IJPRAI, Volume 12, Number 1, pp. 63-79, 1998.

[13] A. Bennasri, A. Zahour, B. Taconet, "Arabic Script Preprocessing and Application to Postal Addresses", ACIDCA'2000, March 22-24, Monistir, Tunisia, pp.74-79, 2000

[14] V. Ablavsky and M.R. Stevens, "Automatic Feature Selection with Applications to Script Identification of Degraded Documents", Proc IEEE ICDAR 2003.

[15] U. Pal and B. B. Chaudhuri, "Automatic Identification of English, Chinese, Arabic, Devnagari and Bangla Script line", ICDAR'01, pp. 790-794,2001.

[16] A. Sehad and al,"Détection de l'inclinaison des documents arabes imprimés", 2003.

[17] C. L. Tan and al, "Language Identification in Multilingual Documents", 1998.

[18] R. Manthalkar and P. K. Biswas," An Automatic script identification scheme for IndianLanguages", 2002.

[19] M. Morita and al, HMM-MLP Hybrid System to Recognize Handwritten Dates, IEEE 2002.

[20]Y El-Ohali and al," Data bases for Recognition of Handwritten Arabic Cheques", Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, pp 601-606, September 11-13 2000.

[21] Y. Al-Ohali and al," Databases for recognition of handwritten Arabic cheques", CENPARMI Computer Science, 2002.

[22] J.M. Alliot, T. Schiex, " intelligence artificielle et informatique téorique", CEPADUES-éditions 1994.

[23] J. Vreeken, "Spiking neural networks, an introduction", 2003.

[24]C. Touzet," les réseaux de neurones artificiels, cours exercices et travaux pratiques", édition 1992.

[25] "Neural Network Toolbox for Use with MATLAB".