Handwritten Arabic Documents Indexation using HOG Feature

Y. Elfakir
LIPI/ ENS FES, MAROC

G. Khaissidi
LIPI/ ENS FES, MAROC

M. Mrabti
LIPI/ ENS FES, MAROC

D. Chenouni
LIPI/ ENS FES, MAROC

ABSTRACT
The old manuscripts are a part of the richest cultural heritage and legacy of civilizations where the digitalization is a solution for the preservation of these manuscripts. The conception of handwriting recognition system knows today a great expansion and appears as a necessity in order to exploit the wealth of information contained in ancient manuscripts. In this paper, a holistic approach for spotting and searching query, especially, for images documents in handwritten Arabic is proposed. These operations need a lot of time and effort to do manual work. For this, we use in the first time text line segmentation of handwritten document image based on partial projection, where a sliding-window approach is used to locate the document regions that are most similar to the query. Histograms of Oriented Gradients (HOGs) are used as the feature vectors to represent the query and documents image, then Support Vector Machines (SVM) is used to produce a better representation of the query and to classify feature vectors. Finally, the application of the reclassification technique at the indexation stage, leads to better results.

General Terms
Pattern Recognition.

Keywords
Indexation, Classification, SVM, Segmentation, Arabic handwritten documents, Histograms of Oriented Gradients (HOG).

1. INTRODUCTION
Many digitization projects have been developed such as manuscripts d'Oc and d'Oïl in the Vatican Library (MOOV) [1], Saint-Omer [2] and Better Access to Manuscripts and Browsing of Images (BAMBI) [3] …treat Latin scripts.

The ancient Arabic manuscripts are a treasure priceless on the textual planes, intellectual and artistic, manual manipulation repetitive of fragile documents could destroy them. In order to exploit this wealth, we are led to scan them and creating electronic libraries. For the reasons above, and in order to develop a complete system for recognition Arabic handwriting, the first step for the creation of this system is presented in this article, namely indexation. That, through two steps: a segmentation phase based on the lines extraction by partial projection method. The second phase is the matching between the query and the scanned document using Histograms of Oriented Gradients descriptor. From literature survey of handwritten word recognition techniques, it is found that, in general, several approaches of textual indexation handwritten documents images are inspired by one of two following categories:

Heuristic approaches (see [4]): they are based on successive segmentation methods, which starting from the text image, allow arriving to the characters.

Holistic approaches (see [5]): address the indexing problem at words, most of the authors prefer a word-level rather than character-level approach.

Many works treat a Latin's manuscripts documents we quote:

Zhang and Tan [6] compute local key-points over the document images for segmentation level and use features based on the Heat Kernel Signature (HKS), while Leydier et al. [7] represent the document images with gradient-based features. The main drawback is that they use a costly distance computation, which is not scalable to large datasets.

Howe [8] uses a generative word appearance from a single positive sample, resulting in an example of a one-shot learning approach and then uses the model to retrieve similar words from arbitrary documents.

Kamble and Hegadi [9] use the Rectangle Histogram Oriented Gradient for extraction the features of handwritten Marathi characters. Feed-Forward Artificial Neural Network (FFANN) classification technique is used in this algorithm.

Rath et al. [10] extracted discrete feature vectors that describe word images, which are then used to estimate similarity between word images after training step of the probabilistic classifier.

Jon et al. [11] propose a sliding-window approach for word spotting in document images where the documents are represented with a grid of HOG descriptors and the retrieval step is performed by using an exemplar support vector machine framework.

Liang et al. [12], described a novel approach to overcome the problem of the lack of existing large data sets for training which uses a character-based modeling for training. A word modeling technique is used for enabling the retrieval of keywords that have not explicitly been seen in the training set.

The million documents were written in Arabic in various disciplines between the seventh and fourteenth centuries. The digitalization is a solution for the preservation of these manuscripts with the advances in digital scanning and electronic storage. However, few works treat an Arabic manuscripts document.

Mohamad et al.[13] use the Hidden Markov Models (HMM) vertical window and two HMM slanted windows, one window is slanted to the left, and one to the right. They proposed this method to remedy with the problem of writing inclination, overlapping and diacritical marks. The approach combined the three slanted windows HMM-based classifiers at the decision step. All classifiers have the same topology as the reference system and vary only in the orientation of the window (query).

Kessentini et al. [14] use multi-stream hidden Markov Models for off-line handwritten word recognition. The proposed approach combines low level feature streams namely, density

based features of sliding windows, and contour based features. In this technique, several feature representations are modeled separately by HMM classifiers. Kundu et al. [15] use Variable Duration Hidden Markov Models (VDHMM) where all Arabic words are modeled by one HMM. Each character is a state in VDHMM and has a variable duration to model a character model of multiple segments.

The paper is organized as follows. Section 2 gives an overview of indexation system process. Section 3 deals with the overall text line segmentation of handwritten document image based on partial projection. Afterwards, section 4 describes the R-HOG approach for word-spotting. In section 5, classification using SVM classifier is presented. Section 6 discuss experimental results. Finally, conclusions are summarized in section 7.

computation. The document images have been preprocessed to enhance them. For this purpose, a model for the restoration of the degradations [16], which uses a series of multi-level classifiers [17] is applied to document images. Then, the text lines are extracted using histogram partial projections method which consists to decompose, classify and search text lines to facilitate research in these manuscripts. Later, a window applied on the segmented lines; reduce the time of the descriptor computation and matching. Histograms of Oriented Gradients (HOGs) are used to represent and to compare the query with the region of the document. A better representation of the query is obtained by suing the SVM training set. Identical positive set is produced by slightly the window around the query and sample negative set is obtained by taking a sampler random regions. The regions with high similarity will be used in reranking step. The general process of the proposed system is shown in "Figure 1".

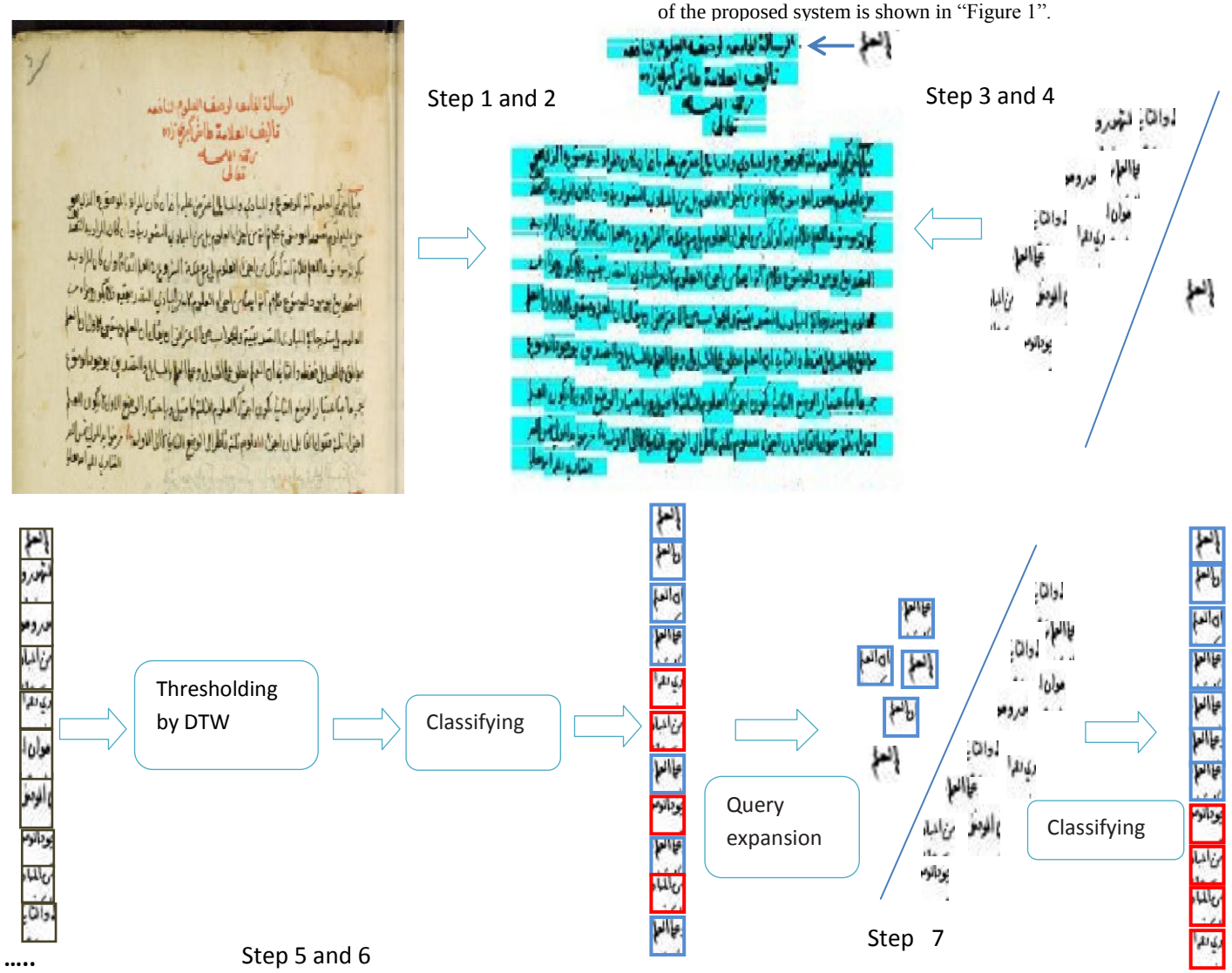

Fig 1: Proposed system process

2. PROPOSED SYSTEM

In this work, we concentrate on historical Arabic documents. The application of HOG descriptor and line segmentation in the context of indexation for Arabic handwriting provides better performance in average precision and time of descriptor

The proposed indexation system is achieved in the following steps:

1. Image preprocessing based on the multi-level classifiers

2. Text line segmentation based on partial projection method.
3. Characterization of the query and the regions in documents using the HOGs descriptor.
4. SVM training set
5. Thresholding by DTW
6. SVM classifying set
7. Reranking

3. TEXT LINE SEGMENSTATION

In this section, the text line segmentation method of Arabic manuscripts is presented. The method is based on the histogram of the partial projections which consists to decompose, classify and search the text lines to facilitate research in these manuscripts.

- Image decomposition

The document image is divided into columns, the width of the column is that of the query, then; we determine the histogram of the horizontal projections for each column. The minimum of these histograms provide the text blocks and the height of each block h (i) "Figure 2".

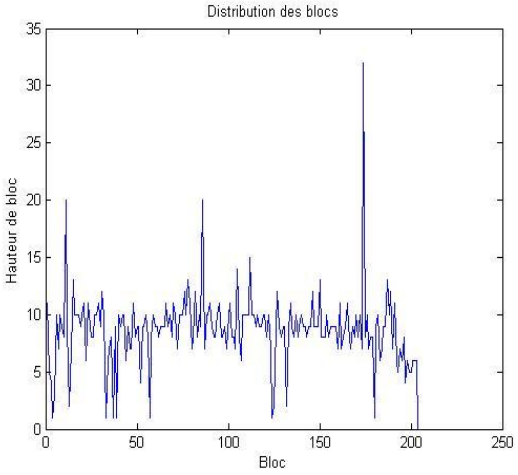

Fig 2:The height of the blocks

- blocks classification

A text document image can contain three types of blocks according to their height: small, medium and large blocks, represent successively diacritics symbols and also components generated by the division of the image in columns, the main path of words and finally overlapping ascending and descending characters well as characters glued with neighboring lines. Each block type represents a class, and the detection of block types is made using the automatic k-means classification algorithm which provides three outputs (classes). The description of the algorithm is given by [18].

- Searching the text lines

When the segmentation blocks are achieved, we pass to the research lines of text. A matching between blocks of different columns is done. For this, we use the Euclidean distances between the lower ordinates blocks. We compare the blocks of column i with those of columns i-1 and i + 1, except for extreme columns which the comparison is respectively with column 2 and n-1 column. The blocks where the distance between their ordinates lower is are paired together. The separators lines correspond to the ordinates lower of these blocks.

4. CONCEPT OF HOG DESCRIPTOR

HOG feature descriptors are used in computer vision and image processing for the object recognition purpose. The main idea behind the HOG descriptors is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. The implementation of these descriptors can be achieved by dividing the image into small connected regions, called cells, and for each cell computing a histogram of gradient directions. The combination of these histograms represents the descriptor.

4.1 Feature Extraction

The main contribution of this paper is the application of HOG feature descriptors for word spotting in handwritten Arabic documents and using text line segmentation. The feature extraction is a most important part of word spotting system, that mean transforming the input query into the set of features. Rectangle Histogram Oriented Gradient (R-HOG) is used to detect and extract feature of Arabic handwritten documents "Figure 3". Initially, we remove noise for sliding-window and region of documents with Gaussian filter. After smoothing, the Sobel kernel is used to calculate the horizontal and vertical components of the gradients.

Let, I_s the smoothed image and the horizontal and vertical components of image gradient is $I_x(x,y)$ and $I_y(x,y)$ respectively.

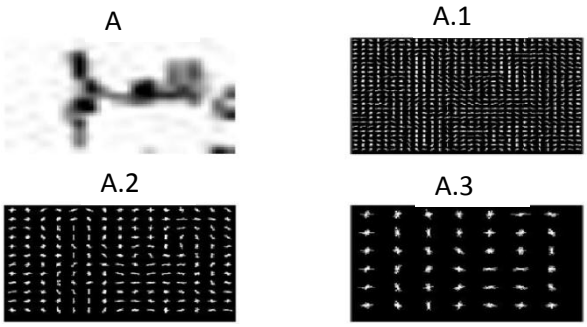

Fig 3: Rectangle HOG of 2*2, 4*4 and 8*8 block size of Ibn Sina (A1, A2, A3) dataset.

$$I_x(x,y) = I_s * \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \text{ and } I_y(x,y) = I_s * \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

The magnitude M(x, y) and direction D(x, y) of the gradient at pixel (x, y) in the smoothed image are computed as follows:

$$M(x,y) = \sqrt{I_x^2(x,y) + I_y^2(x,y)}$$

$$D(x,y) = \tan^{-1}\left(\frac{I_x(x,y)}{I_y(x,y)}\right)$$

Then, histogram of all blocks can be computed using the block size of character; each pixel is assigned in certain category according to its gradient direction, "Figure 4" shows the feature extraction process.

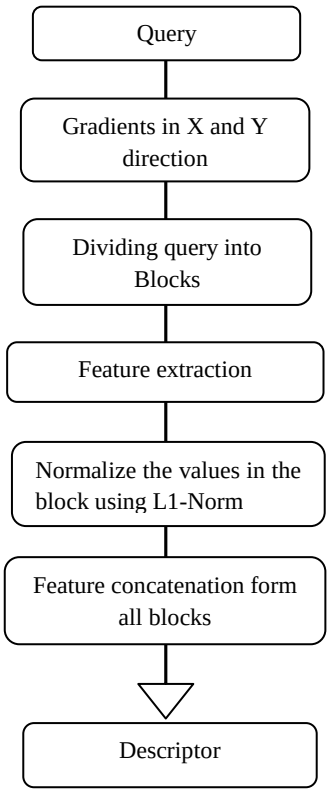

Fig 4: feature extraction process

5. CLASSIFICATION

Support Vector Machines (SVM) are a group of supervised learning methods with associated learning algorithms that analyze and recognize data. We have used SVM classifier with linear function for the recognition documents in handwritten Arabic. The Width of the margin between the classes is the major optimization criterion, the empty area around the decision boundary, defined by the distance to the nearest training pattern. These patterns called support vectors, which finally define the function for classification "Figure 5".

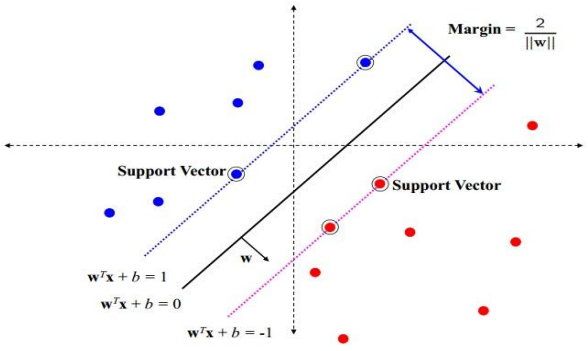

fig5: linear model classification

Following are the four basic kernels used in SVM classifications:

- Linear $K(x_i, x_j) = x_i^T \cdot x_j$
- Polynomial $K(x_i, x_j) = (\gamma x_i^T . x_j + r)^d, \gamma > 0$
- RBF $K(x_i, x_j) = e^{(-\gamma |x_i - x_j|^2)}$
- Sigmoid $K(x_i, x_j) = \tanh(\gamma x_i^T \cdot x_j + r)$

In a linear model, separating hyper-plane has equation

$$w^T \cdot x + b = 0$$

Considering a binary classification problem with training data $\{(x_1, y_1).....(x_i, y_i)\}$ where $x_i \in (\mathrm{N}, \mathrm{P})$ and $y_i \in \{+1, -1\}$

The SVM attempts to find the hyper-plane < w, b > that maximizes the margin.

The positive and negative support vectors respectively is $w^T \cdot x_+ + b = +1$ and $w^T \cdot x_- + b = -1$ so

$$\frac{w}{\|w\|} \cdot (x_+ - x_-) = \frac{w^T (x_+ - x_-)}{\|w\|} = \frac{2}{\|w\|}$$

So we can deduce that maximize the margin amounts to minimizing. This can be casted as an optimization problem as

$$\arg_w \min \frac{1}{2} \|w\|^2 + c_1 \sum_{(x_+, y_+) \in \mathrm{P}} L(y_+ w^T x_+) + c_2 \sum_{(x_-, y_-) \in \mathrm{N}} L(y_- w^T x_-)$$

$c_{1,2}$ Is a regularization parameter, $y_+ = +1$ and $y_- = -1$

$x_+ \in \mathrm{P}$ is constructed by deforming the query, To produce the negative set , the sample random regions over all the documents $x_- \in \mathrm{N}$

6. EXPIREMENTS

In this section, we present the result of our approach for searching query on public datasets of the Lord Byron (LB) and Ibn Sina handwritten manuscripts. A comparison with [11] is given. MATLAB is used to measure all score and running times of the different sections (computing the HOG descriptors, calculating the scores with query and training the SVM).

"Table 1" shows the mean average precision of the approach proposed, Vinciarelli+DTW and HOG+SVM applied to the Lord Byron dataset.

"Table 2" shows the time of descriptor computation with the approach proposed and SVM+HOG approach

Table 1. Mean Average Precision

Dataset	HOG+SVM [11]	HOG	Vinciarelli + DTW	Our approach
LB	83.04	75.37	83.47	85.63

Table 2. Time of descriptor computation:

Dataset	HOG + SVM	Our approach
LB	700s	267s

As can be seen in Table 1 and 2, our approach provides better performance at mAP and time of descriptor computation. In order to evaluate the performance of the techniques in handwritten Arabic documents in Ibn Sina, the cell and block sizes is changed. "Figure 6", shows that the best mean average precision (75 %) is obtained for 2×2 blocks of 2×2 pixel cells.

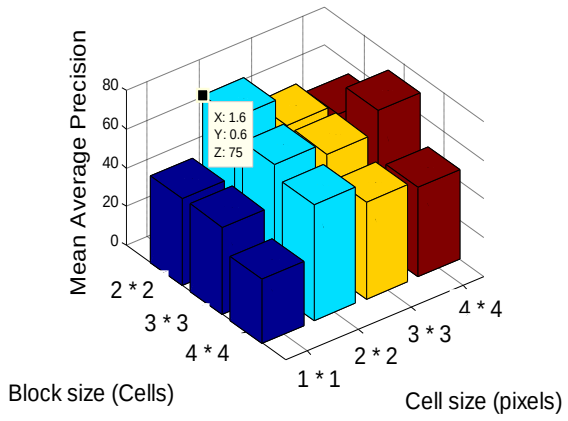

Fig 6: The mAP when the cell and block sizes change

7. CONCLUSION

This paper presents a system for Handwritten Arabic documents indexation using HOG Feature. The system has been evaluated on a large amount of handwritten Ibn Sina and datasets of the Lord Byron. The experimental results show that the used of HOG based feature extraction method and SVM classifier with linear function provides good results in mAP and the time of descriptor computation. In future work, the conception and realization of handwriting recognition system application will be developed, also, reducing the time of descriptor computation by improving the preprocessing step, and also feature descriptors extraction.

8. REFERENCES

[1] IRHT, coord. Maria Careri (Université de Chiet - membre associé à l'IRHT), Anne-Françoise Leurquin et Marie-Laure Savoye (tt://jonas.irht.cnrs.fr/ 2011 – 2021.

[2] IRHT, coord. Dominique Stutzmann (IRHT) htt://form-tei.irht.cnrs.fr/ 2011 – 2018.

[3] CALABRETTO, Sylvie; BOZZI, Andrea; PINON, Jean-Marie, décembre 1999. " Numérisation des manuscrits médiévaux ": le projet européen BAMBI, in: Actes du colloque Vers une nouvelle érudition: numérisation et recherche en histoire du livre, Rencontres Jacques Cartier, Lyon.

[4] A.Zahour,B.Taconet,S .Ramdane,2004." Contribution à la segmentation de textes manuscrits anciens," Conférence Internationale Francophone sur l'Ecrit et le document,CIFED'04.

[5] K. Khurshid, C. Faure, and N. Vincent, 2008. "Recherche de mots dans les images de documents par appariements de caractères", Proceedings of the 10ème Colloque International Francophone sur l'Ecrit et le Document (CIFED08), Rouen, France , p. 91-96.

[6] X. Zhang, C.L. Tan, 2013. "Segmentation-free keyword spotting for handwritten documents based on heat kernel signature", in: International Conference on Document Analysis and Recognition, pp. 827–831.

[7] Y. Leydier, A. Ouji, F. Lebourgeois, H. Emptoz, 2009. "Towards an omnilingual word retrieval system for ancient manuscripts", Pattern Recognit. 42 (2009) 2089–2105.

[8] N.R. Howe, 2013. "Part-structured inkball models for one-shot handwritten word spotting", in: International Conference on Document Analysis and Recognition, pp. 582–586.

[9] M.Kamble, S.Hegadi, 2015. "Handwritten Marathi character recognition using R-HOG Feature", in: International Conference on Advanced Computing Technologies and Applications (ICACTA), Procedia Computer Science 45 (2015) 266 – 274015.

[10] T. Rath, V. Lavrenko, and R. Manmatha, 2003. "Retrieving historical manuscripts using shape", Technical Report, Center for Intelligent Information Retrieval Univ. of Massachusetts, Amherst.

[11] J. Almazán, A. Gordo, A. Fornés, E. Valveny, 2014. "Segmentation-free word spotting with exemplar SVMs", Pattern Recognition, 47 (12), pp. 3967–3978.

[12] Y. Liang, M. C. Fairhurst, and R. M. Guest, 2012. "A synthesised word approach to word retrieval in handwritten documents", Pattern Recognition. 45(12), 4225 –4236.

[13] R.A. Mohamad, L. Likforman-Sulem, C. Mokbel, 2009. "Combining slanted-frame classifiers for improved HMM-based Arabic handwriting recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (7) (2009) 1165–1177.

[14] Y. Kessentini, T. Paquet, A. Ben Hamadou, 2010. "Off-line handwritten word recognition using multi-stream hidden Markov models", Pattern Recognition Letters 31 (1) (2010) 60–70.

[15] A. Kundu, T. Hines, J. Phillips, B. Huyck, L. Van Guilder, 2007. "Arabic handwriting recognition using variable duration HMM", in: 9th International Conference on Document Analysis and Recognition (ICDAR), pp. 644–648.

[16] M .Cheriet, R. Farrahi Moghaddam, R. Hedjam, 2013. "A learning framework for automation and optimization of document binarization methods", Computer Vision and Image Understanding 117(3): 269-280.

[17] Y.Elfakir, G. Khaissidi, M. Mrabti, 2014. "Traitement des documents anciens par les classificateurs multi-niveaux", Colloque International sur le Monitoring des Systèmes Industriels, CIMSI14.

[18] Y.Elfakir, G. Khaissidi, M. Mrabti, Z. Lakhliai, D. Chenouni, M.Elyacoubi, 2015. "Contribution à l'indexation des documents manuscrits arabes scannés", Mediterranean Telecommunication Journal Vol. 5, N° 2.