A semi-automatic approach of old Arabic documents indexing

Abderrahmane KEFALI[(1)], Chaouki CHEMMAM[(1)]
[(1)]Laboratoire de Gestion Electronique de Documents (LabGED)
University Badji Mokhtar-Annaba, Algeria
kefali@labged.net chemame@labged.net

***Abstract*—indexing is a largely used technique in retrieval systems. It has as goal to extract and to represent the meaning of a document so that it can be found by the user. We can cite two types of indexing: manual indexing, and automatic indexing. The automatic indexing requires to use character and words recognition engines which work only over the texts of contemporary documents. In this paper, we propose a semi-automatic approach of old Arabic documents images indexing and searching without resorting to recognize their contents in order to deal with the incapacity of the recognition techniques to understand the contents of old documents. The proposed approach repose on the representation of the documents according to the structural features of their indexes chosen manually from each document by an expert. The approach is tested on a sample of approximately 1100 envelopes and shows good results.**

Keywords-component; indexing, old documents, structural features, documents analysis

I. INTRODUCTION

The libraries, museums and other institutions in pedagogic or sociopolitic matters contain considerable collections of documents, mostly handwritten. Historical documents of old civilizations and the public archives are the typical example of such richnesses which represent the patrimony and the nation's history. The old Arabic handwritten documents form a good part of this patrimony. Indeed, these documents incur a progressive degradation because they are not preserved in good conditions and they are accessible only by a small number of people.

The archives have thus an important challenge to rise: how to preserve the original documents and to make available to the public these millions of pages containing handwritten information for which there are not yet retrieval tools built by the archivists? For a few years, the libraries have started to digitize the collections of historical documents which have much interest for general public. However, this simple digitalization is not sufficient to answer the needs of the numerical libraries users because the difficulties of access are the same as that on paper or microfilm. The content is generally not structured what makes always necessary to leaf through an enormous quantity of pages before finding the pages containing the desired information.

Various solutions to this problem are possible: a simple method to structure the collections of historical documents is to order them chronologically. Documents annotations with principal topics allow to refine the granularity. A very great level of detail in the contents annotation can be achieved by the transcription. It allows a search of complete text by using a traditional textual search engine. Because the cost of the electronic contents annotation increases considerably with the desired level of details and the size of the annotated collection, a compromise is usually preferred.

Automatic approaches of contents annotation and research are thus desirable in order to reduce the enormous cost of human transcription. Several systems of indexing were already proposed but the majority of them are effective only over images of recent documents. The automatic recognition of handwritten historical documents can appear as an obvious choice, but the manuscripts recognition reached a good level of exactitude only in two fields: the recognition *online*, and the out-*line* applications with very limited vocabularies, such as for example the checks processing or automatic mails sorting. For the old documents which include wide vocabularies, inconsistent orthography, in addition to the presence of several types of degradations, the recognition results remain far from being satisfactory. It is thus essential to define new tools giving access to the old handwritten documents without recognition of their contents.

In this paper, we propose a semi-automatic approach of old Arabic documents images indexing and searching without resorting to a recognition of the contents in order to deal with the incapacity of the recognition techniques to understand the contents of the old documents. The proposed approach repose on the representation of the documents according to the structural features of their indexes chosen manually from each document by an expert.

The paper is divided into sections and starts with presenting some work in the same field. It is followed by a detailed description of the proposed approach and the various stages implied in the processes of indexing and research. Lastly, we show the obtained results before concluding.

II. PREVIOUS WORKS

The indexing and search of words in the Latin documents images aroused a considerable attention recently. One of the first works on this field is that of Spitz [4] in which the author

proposed to code the characters of the printed texts according to their forms. The method extracts for each word of the document the features of its characters basing on the connected component count of each character, and on their position in relation to two base lines. The characters are then coded according to their features to form a *word shape token* (WST). Query words are mapped in the same way to WSTs. Indexing and retrieval of documents can now be done as usual, but using WSTs instead of words. Later, Smeaton and Spitz [5] showed that this technique is useful only if the images are of bad quality implying a failure of the OCR.

Chen and *al.*, [6] proposed an approach based on the information of the words forms instead of the characters forms. Firstly they identify upper and downer contours of each word. From these contours, they extract the form information basing on the pixels location. Then, the Viterbi decoding is used to match the word image with the query.

Keyword spotting is even more difficult for handwritten text. One of the first works on the handwritten documents retrieval is that of Manmatha and *al.* [7] in which they proposed a semi-automatic approach to index handwriting documents. In this work, the similar images of words are grouped into equivalence classes. The most frequently occurring classes are eliminated and the most remain classes are manually coded in ASCII and used as index.

Several works was also carried out over historical document retrieval. [8] for example proposed an approach for searching in the cards of military incorporation of the XIXe century. The idea of this approach is to index automatically the old forms by an ordered chain of graphemes associated to the case of the handwritten patronym, and to also transform the alphabetical query of the user into a chain of graphemes. The comparison is made by using the traditional edit-distance.

In [9], manuscripts in Telugu language were characterized with representations by wavelets of the words. The representation by wavelets provides information on the contents of the image to various scales. It exploits the characteristics inherent in the characters of Telugu. The application of this representation by wavelets on the Latin characters does not give good results [9].

Rath and Manmatha [3][10] presented an holistic approach of words searching in historical handwritten documents. This approach consists to group the words images in similar groups. Then it seeks the set of the most representative groups and labels them. Each labelled group is used as index. In [3], the authors propose to represent the words images by four features of profile which are then matched by using various methods. [10] used the correspondences between the angular points to classify the words images in historical manuscripts.

Motivated by the work of Rath and Manmatha, Adamek and *al.* [12] proposed to compare words contours instead of the whole words for the holistic recognition in historical manuscripts. The search of a word in a document is implemented then by a comparison between words contours.

Another recent work of indexing and searching of old documents is that of Ramel and al. [13]. In this work, the authors proposed an approach searching in documents of the Higher Study Center of the Rebirth without preliminary recognition of the documents model. They made a study of the structural characteristics of these documents. Then, they applied a hybrid analysis which benefits simultaneously the advantages of the ascending and downward methods.

In [2], the authors are interested by searching words in the images of old printed documents. The authors proposed here to work with the characters and not with the words, and to represent each character by a set of features. At the time of research, the query is treated in a way similar to the total document and the features of the query characters are matched with the features of the words characters already stored in index files by using the DTW algorithm.

Bai et al. [14] proposed an approach based on the coding of the words according to their forms. For each word, we extracts a set of 7 features and each word is coded thereafter by a sequence of codes. For searching, the formulated query is coded by the same manner in a sequence of codes and the query code is matched with the codes of each word in each document by using the DTW algorithm.

For the old Arab documents retrieval, we can cite the work of Sari and Kefali [15] in which the authors proposed a method of searching in old Arab manuscripts images. They represent each document by the structural features of its sub-words. Then, at the time of research, the textual query of the user is coded in the same manner, and the comparison is performed by using the DTW algorithm.

Another work is that of Benmohamed and al. [1]. In this work the authors proposed a semi-automatic approach of document indexing and retrieval. After the choice of the indexes manually, the method detects their contours then parameter these contours by representing them in the form of "chain of codes". The latter will be refined in order to decrease its size, and finally arranged in a index table. At the time of research, the sought word is coded in the same manner proceeded in the indexing, and the optimal code obtained is compared thereafter with the optimized codes of the indexes arranged in the table using the DTW algorithm [1].

III. PROPOSED APPROACH

The objective of our work is to facilitate the access to the old Arabic documents by proposing a semi-automatic approach of indexing and retrieval of these documents without complete recognition of their contents. The proposed method repose on the representation of the documents according to the structural features of their indexes chosen manually from each document by an expert (specialist). After the manual selection of the indexes (images of one or more words), the features are extracted automatically from each index by applying a series of treatments resulting from the documents analysis field. The method starts with the transformation of the image into gray levels, and the resulting image will be binarized then smoothed. Then, it segments the index image into sub-words, and extracts finally their features. These last will be coded in ASCII and will be recorded in an indexes file associated to the treated document. All these treatments are summarized by the figure 1.

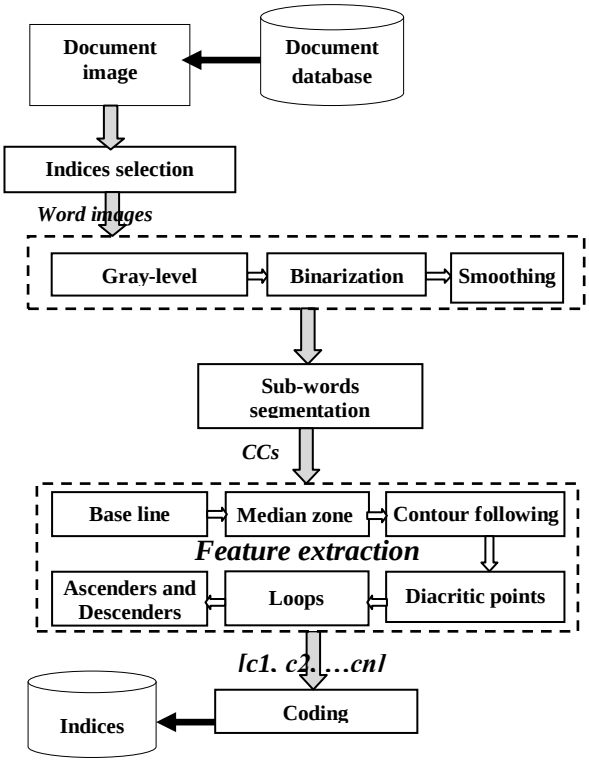


Figure 1. General outline of the indexing

At the time of the search of a query, the system compares its features with the features of the indexes already recorded on disc. The words for which the distance is lower than a threshold are the accepted words (figure 2).

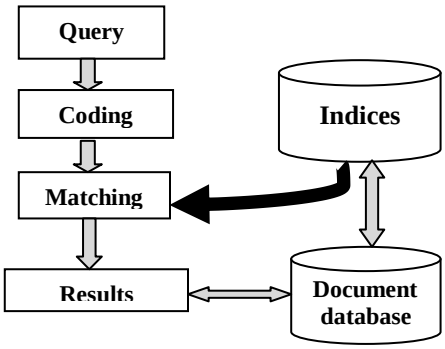


Figure 2. General outline of the query retrieval

IV. INDEXING

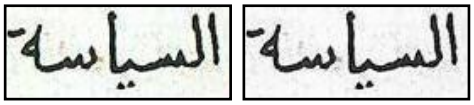
We detail in this section the various stages of the documents indexing.

A. Indexes selection

This stage is very important because it influences the final result of our system. The indexes are the key words which describe the best our document. For our case (contrary to [3]) the indexes selection is performed manually by an expert or a specialist. This last chooses the most representative words in the document. This stage allows us to obtain a whole of word images. The following stages will be applied to each one of these images separately.

B. Gray level transformation

The gray level transformation is necessary because the employed binarization method is performed over gray level images (256 gray levels) and not over colors images. This transformation can be carried out simply by assigning to each pixel of the image the average of its values of the colors red, green, and blue (Figure 3).



(a) ***(b)***

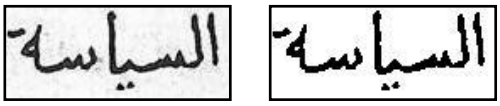
Figure 3. (a) color image, (b) gray level image

C. Binarization

The binarization is one of the most important stages in the images processing and analysis process. it is an irreversible processing which allows to transform a gray level or color image into a binary image (in black and white) according to a definite threshold (Figure 4). To choose a method to be applied in our approach, we based on a comparative study of Kefali and al. [16] in which the authors showed that the NICK method [17] is the best for the binarization of the degraded documents. The NICK method is an adaptive local method in which the binarisation threshold T is calculated for each pixel of the image by using the following formula:

$$T = m + k\sqrt{\frac{(\sum p_i^2 - m^2)}{NP}}$$

Such as: k varies between -0.1 and -0.2, m: the average gray level on a window centred over the current pixel, p_i: the gray level of the pixel i and NP is the total number of pixels.



(a) ***(b)***

Figure 4. (a) gray level image, (b) binary image

D. Smoothing

In certain cases, the processes of acquisition or binarization can introduce noise to the image, which results in particular in the presence of irregularities along the characters tracing, and which can thus degrade the performances of our system. To palliate this problem, we apply a smoothing by using the algorithm of [18] which reduces the noise of a binary image by eliminating the isolated pixels on the one hand and by stopping the empty holes on the other hand (Figure 5).

(a) ***(b)***

Figure 5. (a) binary image, (b) smoothed image

E. Segmentation into sub-words

As we said previously, we chose to work with the sub-words and not with the characters, because the sub-words seems to us to offer a better compromise between the complexity of the segmentation of the handwritten Arabic words in letters and the global solution of recognition which handles the word in its totality. In addition, the characters are not always easily separable (even for a human). This difficulty was clearly evoked by Sayre in 1973 and can be summarized by the following dilemma: "to recognize the letters, we must segment the tracing and to segment the tracing, we must recognize the letters", see also [21].

The sub-words extraction consists in labeling the various connected components of the image by gathering the neighbor black pixels in a distinct unit, and we use for that the pixels aggregation method (figure 6). This stage can be summarized by the following pseudo-code:

1- Find a non-visited black pixel
2- Find all its neighbours: if one of the neighbours is a black pixel we gather it with the first and we reiterate recursively the operation for all the neighbours.
3- We stop when all the black pixels are visited.
4- Return to stage 1.

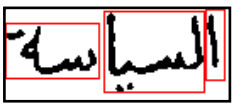


Figure 6. Sub-words extraction

F. Features extraction

One of the fundamental problems of the image analysis is to determinate which features a use for obtaining good results. According to [11] [19] the structural primitives coming from the human perception which are connected to the writing shape are considered as relevant features for the discrimination of the handwriting characters. In our system, we chose to extract from each sub-word four structural features: ascenders, descenders, loops and diacritics points. In order to extract these features, we proceed as follow:

1) Base line detection

The base-line is the line on which rests the characters which do not have descenders. The most used method for detecting the base-lines is the horizontal projection of the image. The base-line corresponds to the line whose projection contains the greatest number of black pixels (red line in the figure 7).

Figure 7. Base-line and median zone detected on a word image

2) Median zone localization

The median zone is the zone which includes the body of the Arabic words (part of letters without ascenders or descenders). In our system, we detect the median zone by tracing two borders, high and low relatively to the base line. The median zone will be thus the space between these two borders (blues lines in figure 7).

3) Contour following

The contour following is commonly used for the structural features extraction. The sub-word contour is the set of points delimiting the tracing (Figure 8). Each contour is coded by specifying a starting point followed by a codes chain or sequence of the Freeman chain.

Figure 8. Contours of a word image

4) Extraction of the diacritic points

The Arabic writing is rich in diacritic points, they are simples or multiples points which appear above or below the principal body of the Arabic letters, and allow to differentiate between the letters having a similar body, it is for that that the detection of these points is important in our system.

A connected component is considered as a diacritic point if its size is lower than a certain threshold. The threshold is chosen as the average of the various connected components sizes.

5) Extraction of the loops

Certainly the loops are one of the most important and used structural features for the Arabic writing, and allow the detection of certain letters. For our system, we detect the presence of a loop when all the coordinates of a contour include all the coordinates of another.

6) Extraction of the ascenders and descenders

An ascender corresponds to a rise which exceeds the median zone, in the same way a descender corresponds to a descent which finishes outside the median zone i.e. located in the lower zone. To extract the ascenders, we traverses the higher zone of the image, all the components which are in this zone and which are not diacritic points are regarded as ascenders. For the descnders, we proceed in the same way, but we traverse the lower zone of the image.

G. ***Coding***

After, the image will be coded according to their features extracted previously by corresponding to each feature an ASCII code (ascenders whith h, descenders whith j, loops with b, high diacritic points with p, and lower diacritic points with q). We adds another character "#", which represents the space inter sub-words.

The word السياسة for example is coded by the string: **h#hqqh#hbp**.

The obtained codes are stored in an index file corresponding to the treated document.

V. SEARCH

As we said, the objective of the user is to find all the documents in the base comprising certain words. In our

system, the user query is a string in Arabic language, which can be a simple or composed word.

The query is coded in the same manner as the coding of the documents indexes in the first phase, i.e. according to the structural features of its letters. Coding is performed by corresponding to each letter of the query a precise code describing its structural features, which give us a code string. We established a correspondence table which summarizes the codes of the various Arabic letters (Table. 1).

TABLE 1 .CHARACTER TRANSCRIPTION IN ASCII CODES

Character	*Code*	*Designation*
ا - ا - ك - ل	h	Ascender
إ	hq	Ascender +Down Diacritic
أ - أ	ph	Up Diacritic +Ascender
ل	hj	Ascender +Descender
ط	bh	Loop +Ascender
ظ	bph	Loop +Up Diacritic +Ascender
لا	hbh	Ascender +Loop+ Ascender
ك	hp	Ascender +Up Diacritic
ي	jq	Descender +Down Diacritic
غ - خ - ذ - ت - ئ - ن	p	Up Diacritic
غ	jp	Descender +Up Diacritic
ش - ث	pp	Up Daicritic +Up Diacritic
ن - ز - خ - ئ	jp	Descender +Up Diacritic
ش	ppj	Up Diacritic + Up Diacritic + Descender
ض	bpj	Loop + Up Diacritic + Descender
ض - ف - ة - غ - ق	bp	Loop +Up Diacritic
ق	pbj	Up Diacritic + Loop + Descender
ب - ج - ب	q	Down Diacritic
ح-ع- س - ر - ى	j	Descender
ج	jq	Descender + Down Diacritic
ه - م - ص - ه - ه	b	Loop
هـ	bb	Loop +Loop
ع - و - ص - م	bj	Loop +Descender
ﻻ	hh	Ascender +Ascender
ة	pb	Up Diacritic +Loop
غ	pbj	Up Diacritic + Loop + Descender
ؤ	bjp	Loop + Descender + Up Diacritic
لأ	hbhp	Ascender + Loop + Ascender + Up Diacritic
لإ	hbqh	Ascender+Loop+Down Diacritic+ Ascender

After, and contrary to [1], the codes string obtained will be compared with the indexes of all the documents already saved in the indexes base, and not only with the indexes which have the same number of sub-words as the query. The comparison between the codes string and the indexes is done by using the DTW algorithm in order to take account of possible errors. The advantage of employing the DTW is that it allows to take account of the non-linear stretching and the compression of the words. In this manner, two identical words which differ by their sizes will be put in correspondence correctly, unlike the correlation where the words must have the same dimension to be matched [2]. Noting also that the use of the DTW algorithm allows to return occurrences close to the query but also increases the number of false occurrences. With my personal opinion, turn over false occurrences is better than to unaware relevant occurrences.

The comparison by using the DTW algorithm can be summarized by the following pseudo-code:

n = size of the first string *X*
m = size of the second string *Y*
C(*a*, *b*) = transformation cost to pass from de *a* à *b*
lambda = void
D [*m*, *n*] = edit distance from *X* to *Y*
For *i* = 0...*n* **do** *D*[0, *i*]=*i*
For *j* = 0...*m* **do** *D*[*j*, 0]=*j*

For *i* = 1...*m* **do**
  For *j* = 1...*n* **do**
    D [*i* ,*j*] = Min {
    D [*i*-1, *j*-1] + *C* (X_i, Y_j), // substitution (0 if the characters X_i and Y_j are equals)
  D [*i*-1, *j*] + *C* (X_i, *lambda*), // effacement
  D [*i*, *j*-1] + *C* (*lambda*, Y_i)} // insertion
    }
  End for
End for

Finally, the distance between two words *X* and *Y* is contained in the last cell of matrix *D*. A document is considered as relevant for the user if the distance between one of its indexes and the query is lower than a certain threshold.

VI. EXPERIMENTS AND RESULTS

In view of the absence of a benchmark data base of the old Arabic documents to use it in the validation, we tested our approach (as first draft) on a sample of Algerian postal envelopes (approximately 1100 images of envelopes). The data base used is available in our research laboratory (LabGED), and the envelopes composing this base is written by various writers and scanned under the same conditions.

To each image of this base, we applied the stages of the first phase of indexing, by choosing 3 indexes manually corresponding to: the family name, the first name, and the name of Wilaya (Figure 9).

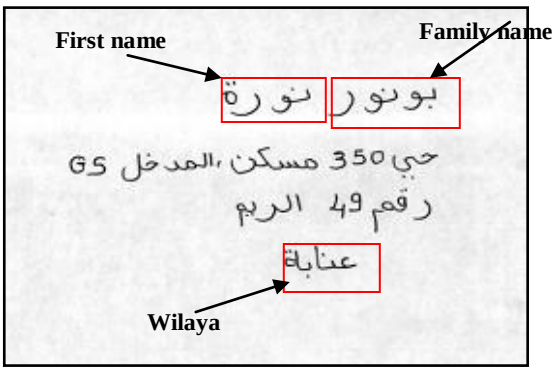


Figure 9. Indexes selected from an envelope image

Afterwards, we examine the base by a set of 300 textual queries in Arabic language of various lengths (from 4 to 20

characters), for which we established the list of the relevant documents manually. The evaluation of the obtained results is done in terms of recall and precision.

We tested in our experiments several values of the threshold *K* (allowed error) in order to choose the best. The table.2 recapitulates the average recall and the precision obtained for various values of the threshold *K*.

TABLE.2. OBTAINED RESULTS FOR DIFFERENT VALUES OF THE THRESHOLD

Threshold value (*k*)	**Recall**	**Precision**
0	0.25312	0.99885
1	0.49143	0.8411
2	0.7215	0.6745
3	0.9148	0.4122
4	1	0.2125

The obtained results show that the search recall and the precision depend on the value of the threshold *K* and vary in an opposite way. The best value of the threshold is thus that which presents the best compromise between the recall and the precision. The various tests performed, show that the best compromise between the recall and the precision is not constant but it differs according to the size of the queries codes. To take account of this remark, we proposed to adapt the threshold according to the queries sizes (the attribution of the values of *K* is done by experiments). The threshold is thus allotted as follows:

If the query size <= 7 then *k*=1;

Else, if 7 < the query size < 12 then *k*=2;
Else *k*=3;

Considering optimum threshold values for various word lengths, our system carries out an accuracy of 75% while obtaining a recall of 87%.

VII. CONCLUSION

In this paper we proposed a semi-automatic method of old Arabic documents images indexing. The manual aspect of this method consists in the indexes selection, while the automatic aspect consists in their processing and representation. Over the indexes chosen manually (words images) we applies a series of processing resulting mainly from the field of the documents analysis (binarisation, skew correction, segmentation, etc) aiming to extract the structural features of the sub-words composing the indexes. These features are then coded and stored under a string form in an indexes file associated to the treated document. At the time of the search of a new query, the system extracts its structural features, and represents them by a character string (codes string). Then the search will be translated by a comparison between the codes string and the indexes of all the documents already stored in the base by using the DTW algorithm. The tests are carried out over a sample of 1100 images of Algerian postal envelopes and the obtained results are very encouraging. Improvements are still possible by refining the various stages, and by adding other features.

REFERENCES

[1] A. Benmohamed, T. Sari, M. Sellami, « Une approche semi automatique pour la recherche de documents anciens », Journées Gestion Electronique de Documents & Réseaux de Recherche en sciences et Technologies d'information, pp. 156-163, Annaba-Algérie, Mai 2009.

[2] K. Khurshid, C. Faure, N. Vincent, « Recherche de mots dans des images de documents par appariement de caractères », 10ème Colloque International Francophone sur l'Écrit et le Document, France, pp. 91-96, 2008.

[3] T.M. Rath, R. Manmatha, « Word Spotting for historical documents », International Journal on Document Analysis and Recognition, vol. 9, No. pp. 139-152, 2007.

[4] A. Spitz, "Using character shape codes for word spotting in document images", Dori D. and Bruckstein A. (Eds.), Shape, Structure and Pattern Recognition, World Scientific, Singapore, pp.382–389, 1995.

[5] Smeaton, A. Spitz, "Using character shape coding for information retrieval", in the 4th ICDAR, pp.974–978, 1997.

[6] F. Chen, D. Bloomberg, "Summarization of imaged documents without OCR", Computer Vision and Image understanding, vol.70, No.3, 1998.

[7] R. Manmatha, C. Han, E. Risemen, "word spotting: a new approach to indexing handwriting", IEEE Conference on Computer Vision and Pattern Recognition, pp.631-637, 1996.

[8] J. Camillerapp, L. Pasquer, B. Coüasnon, " Indexation automatique de formulaires anciens par reconnaissance du patronyme manuscrit", in RFIA, pp. 1493-1502, France, 2004.

[9] A. K. Pujari, C.D. Naidu, B.C. Jinaga, "An adaptive character recogniser for telugu scripts using multiresolution analysis and associative memory", ICVGIP, 2002

[10] T.M. Rath, R. Manmatha, « Features for Word Spotting in Historical Manuscripts », 7th International Conference on Document Analysis and Recognition, 2003.

[11] M. Cheriet, H. Miled, C. Olivier, Y. Lecourtier, « Visual Aspect of Cursive Arabic Handwriting Recognition », Vision Interface, pp. 262-270, 1998.

[12] T. Adamek, N.E. O'connor, A.F. Smeaton, « Word matching using single closed contours for indexing handwritten historical documents », International Journal on Document Analysis and Recognition, vol. 9, No. 2-4, pp. 153-165, avril 2007.

[13] J.Y. Ramel, « User driven page layout analysis of historical printed books », International Journal on Document Analysis and Recognition, 2007.

[14] S. Bai, L. Li, C.L. Tan, "Keyword Spotting in Document Images through Word Shape Coding", in the10th ICDAR, 2009.

[15] T. Sari, A. Kefali, "A search engine for Arabic documents", Dixième Colloque International Francophone sur l'Écrit et le Document (CIFED), Octobre 2008

[16] A. Kefali, T. Sari, M. Sellami, « Evaluation de plusieurs techniques de seuillage d'images de documents arabes anciens », 5éme symposium international Images Multimédias Applications Graphiques et Environnements, pp. 123-134. Biskra, Algérie, Novembre 2009.

[17] K. Khurshid, I. Siddiqi, C. Faure, N. Vincent, « Comparison of Niblack inspired Binarization methods for ancient documents », 16th Document Recognition and Retrieval Conference, USA, Jan 2009.

[18] A.S. Mahmoud, « Arabic Character Recognition Using Fourier Descriptors and Character Contour Encoding », International Conference on Pattern Recognition, vol. 27, No. 6, pp. 815-824, 1994.

[19] A. Amin, J. F. Mari, « Machine Recognition and Correction of Printed Arabic Text », IEEE Transactions on Systems, Man and Cybernetics, vol. 19, No. 5, pp.1300-1306, 1989.

[20] P.A. Devijver, J. Kittler, « Pattern recognition, a statistical approach », Englewood Cliffs, London, 1982.

[21] T. Sari, M. Sellami, « State of the art of line Arabic handwriting segmentation », International Journal of Computer Processing of Oriental Languages, vol. 20, No. 1, pp.53-73, 2007.