A search engine for Arabic documents

T. Sari, A. Kefali

To cite this version:

T. Sari, A. Kefali. A search engine for Arabic documents. Colloque International Francophone sur l'Ecrit et le Document, Oct 2008, France. pp.97-102. hal-00334402

HAL Id: hal-00334402
https://hal.science/hal-00334402

Submitted on 26 Oct 2008

A search engine for Arabic documents

Toufik SARI [1] and Abderrahmane KEFALI[1]

[1] Laboratoire de Gestion Electronique de Documents (LabGED)
Université Badji Mokhtar – Annaba

email : {sari, kefali}@lri-annaba.net

***Abstract** : This paper is an attempt for indexing and searching degraded document images without recognizing the textual patterns and so to circumvent the cost and the laborious effort of OCR technology. The proposed approach deal with textual-dominant documents either handwritten or printed. From preprocessing and segmentation stages, all the connected components (CC) of the text are extracted applying a bottom-up approach. Each CC is then represented with global indices such as loops, ascenders, etc. Each document will be associated an ASCII file of the codes from the extracted features. Since there is no feature extraction technique reliable enough to locate all the discriminant global indices modelling handwriting or degraded prints, we apply an approximate string matching technique based on Levenshtein distance. As a result, the search module can efficiently cope with imprecise and incomplete pattern descriptions. The test was performed on some Arabic historical documents and shown good performances.*

Keywords: Document retrieval, Arabic handwriting recognition, handwriting segmentation.

1 Introduction

Digital libraries are becoming a sound area of information technology merging most other disciplines [BAI04]. Since '*digital'* refers to digital information which could be of any kind : sound, text, image or video with the respective processing tools; and *'library'* employs the technologies for database, information systems, client-server, information retrieval, datamining, etc. To date, most libraries hold a huge amount of digitized documents from various resources: historical, newspapers, legacy materials, etc. To efficiently access to such collections on CDs or over the Internet requires an effective indexing and searching strategies. Such indexes are usually created manually. While this approach may be feasible for small numbers of documents, the cost will be prohibitive for large collections.

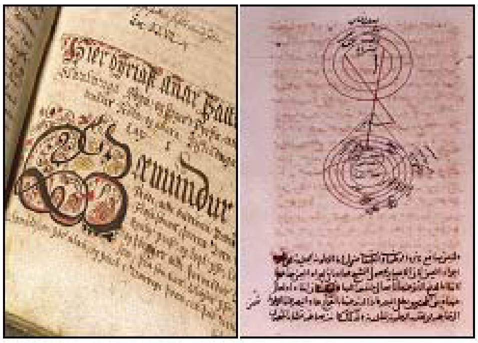


FIG. 1 – Examples of complexe document images.

Recognizing the textual content of the documents may be a promising alternative, since this technology has advanced enough to make commercial applications possible. However, OCR techniques have only been successful in the online domain, in offline applications with very limited lexicons for handwriting, and for machine printed. For general documents with open/large lexicons, the usually degraded image quality, with multilingual text, and so on, traditional OCR techniques are not adequate [BAI04], [LEE02] (see Fig.1). We should point here that the ultimate goal of a multimedia IR system is not to recognize the patterns, textual or graphical, in the documents rather than to retrieve the original materials. Several commercial systems have been developed using page segmentation and layout analysis techniques following OCR, Heinz Electronic Library Interactive Online System (HELIOS), Excalibur EFS and PageKeeper from Caere, are some examples.

Since one of the crucial steps in text-based IR is the identification of keywords in a piece of text, researchers have attempted to locate keywords in text images using properties like word-shapes. This is known as *keyword spotting*.

As we can retrace the literature, the work of A. Spitz et *al.* seems to be the most closest to the current investigation. A. Spitz in [SPI95] proposed the use of *character shape codes*. First, words are identified, and then each character in a word is mapped to a character shape code. For example, letters with ascenders may be mapped to the code *A*. The sequence of obtained codes form a *word shape token* (WST). Thus, the wordbook may be mapped to the WST *AxxA*. Query words can be similarly mapped to WSTs. Indexing and retrieval of documents can now be done as usual, using WSTs instead of words. Note that this method only determines the general shape of a character rather than trying to identify individual characters. The hope here is that shape information can be obtained more accurately and at a lower cost than full OCR. Later, A. Smeaton and A. Spitz [SME97] show that this may not be a very useful tool except when OCR quality is very low. It was noted that WST-based retrieval performed worse than conventional word-based retrieval even on poor quality OCR output. Chen et *al.*, [CHE98] proposed a segmentation free approach using word shape information. They first identify upper and lower contours of each word using morphology and then extract shape information based on pixel location among these contours. Next, Viterbi decoding of the encoded word shape is used to map the word image with the given keyword.

If keyword spotting is a non-trivial problem for images of printed text, it is even more difficult for handwritten text. Fortunately, Manmatha et *al.* [MAN96], proposed a technique for indexing handwritten documents. The page is first segmented into words, and word images are grouped into equivalence classes. Obvious mismatches are eliminated using like area and ratio features. The most frequently occurring classes are eliminated since they represent function words, and the most remain equivalence classes are used as index.

As so many efforts have been devoted to OCR-free processing and retrieval of document images, no work dealt with Arabic handwritten documents even within the TREC framework.

To deal with the defectiveness of pattern recognition techniques to understand the complex structures and the heterogeneous content of the documents [SAR07] and to keep in mind the goal of IR, we propose a free-recognition approach for retrieving handwritten Arabic document images from huge collections. The search strategy proposed is based on *approximate string matching* as opposed to *exact string matching* [NAV01]. We do not perform any recognition of words or characters. Another important innovation in our current work is that end-users can search document images by textual queries, *i.e.* a noisy edition of the text already present in documents.

The rest of the paper is organized as follows. Section 2 gives some basic definitions of the edit distance problem; section 3 presents our approach illustrated by preliminary investigations and the experiments in section 4. The conclusion is set on section 5.

2 Basic problem of edit distance

The first theoretical work which constitutes a formal model for handling noisy data is based on the theory developed by Shanon in 1948 at AT&T Bell laboratories modelling a noisy communication channel such as a telephone line. *The Research Group of Yorktown Heights in New York applied Noisy Channel Model (NCM), also known as source-channel model, for the first time* in a continuous speech recognition system. Since, this model was carried to the fields of machine translation, spellchecking and several other applications as part-of-speech tagging, OCR and handwriting recognition, and information retrieval.

In this approach, a sequence of text, *A*, is sent in a communication channel and a noisy text, *T*, occurs. In spellchecking, for example, the noisy text corresponds to the text, containing errors, keyed by a typist or resulting from a speech/written text recognition system; and in machine translation it corresponds to a text in another language. Thus, the solution to the problem is to find/recover the original text from the generated text. Some basic edition operations transform the generated data from the original one.

We are interested in a distance that enables to transform a string x into a string y using three kinds of basic operations: the substitution of a letter of x by a letter of y and the deletion of a letter of x or the insertion of a letter of y. A cost is associated to each of these operations and for each letter of the alphabet:

- Sub (a, b) is the cost of the substitution of the letter a by the letter b.
- Del (a) is the cost of the deletion of the letter a.
- Ins (b) is the cost of the insertion of the letter b.

The general problem consists of finding a sequence of such basic operations to transform x into y minimizing the total cost of the operations used. The total cost is equal to the sum of the costs of each of the basic operations. This cost is a distance on the words.

The recurrence relations imply an obvious ternary-recursive routine. This is not a good idea because it is exponentially slow, and impractical for strings of more than a very few characters.

3 The proposed approach

Among the various types of documents, the old manuscripts are probably those whose digitalization is most imperative. Indeed, the historical materials are damaged by time and, when they do not deteriorate naturally, suffer to be too often handled for consultation. In order to overcome this problem, digitalization was considered. With the huge numerical tools now available, it is possible to extract the useful data from digital images. In this paper, we focus on Arabic historical manuscripts. In order to preserve such heritage two approaches are possible. The first approach relates on the development of algorithms for document content recognition and their transcription in text or only their annotation. The second approach will make users capable for searching, visualizing and browsing into the image database of ancient manuscripts without a complete identification of the content. This is due mainly because of their complexity and diversity of the structure and the spatial disposal of the objects even textual or graphical.

Our main objective is the development of an information retrieval system for content-based image search using descriptors extracted from handwritten words actually present on old Arabic manuscripts. The manuscripts are initially books, which contain one or more interesting works by their contents. Their reproduction, their distribution, their owning by a library or an institution, also confirm the importance of the knowledge and its circulation at one historical period and in a given region. Moreover, the manuscript is an archaeological document, an object that can be studied for its aesthetic production, its material characteristics and its artistic decoration (Fig.2).

FIG. 2 - A decorative Arabic historical document

Before looking further into the proposed approach, let us first introduce Arabic writing characteristics. Arabic is cursive, which transcribes from right to left. It involves 28 basic characters in addition of the HAMZA (ء) which can be considered as an entire character or grouped with other letters to yield to additional shapes (Fig.3.a). Each character can have up to four different shapes according to its position within the word: in the beginning of the word BW (only connected by the left), in the middle of the word MW (two-sides connected) in ending of the word EW (only connected by the right) or isolated IS (two-sides unconnected) (Fig.3.b). The letter TA (ت) has two additional shapes TA-MARBOUTA (ة ending: ـة isolated and:). Six Arabic basic characters cannot be linked by the left. They have only two shapes: isolated and linked by the right. An Arabic word containing at least one of these last letters will be cut up in several parts (Fig.3.c) known as pseudo-words, sub-words or even PAW for Pieces of Arabic Word. In the following of the paper, we will use one of the three appellations indifferently. Several Arabic characters have dots, which can be disposed over or below the main body of the letters. They are called secondary parts or diacritics; the remaining and most important part is the primary part. No letter in Arabic language has high and low diacritics together. Several other characters have a loop, an ascending or a descending part according to a middle line, the baseline of the writing, which contains the most ligatures between letters. Two, or more, characters can be superposed vertically constituting vertical ligatures (Fig.3.d). The most vertical ligatures are not obligatory, they appear for the aesthetic reasons excepting the LAM-ALIF ligature (لا) as for is obligatory. The combination of the letters LAM (ل) and ALIF (ا) cannot be transcribed as (لـــا). Phonetically, Arabic alphabet involves only consonants and long vowels. The short vowels, representing the different sounds associated to letters, are the horizontal or slanted strokes, the denture CHEDDA (ّ) or the MEDDA. Thus, an Arabic word can have several meanings according to used short vowels [SAR07] (Fig.3.e).

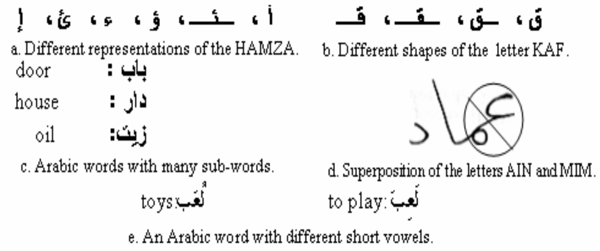

إ ، ئ، ء ، ؤ ، ـئـ ، أ
a. Different representations of the HAMZA.
ق ، ـق ، ـقـ ، قـ
b. Different shapes of the letter KAF.
door : باب
house : دار
oil :زيت
c. Arabic words with many sub-words.
d. Superposition of the letters AIN and MIM.
toys:لُعَب
to play: لَعِب
e. An Arabic word with different short vowels.

FIG. 3 - Arabic writing special features.

Our system works in two phases: The first one aims representing each document in the database by its words' descriptions, while in the second phase users can search the collection by keying some keywords in the user interface.

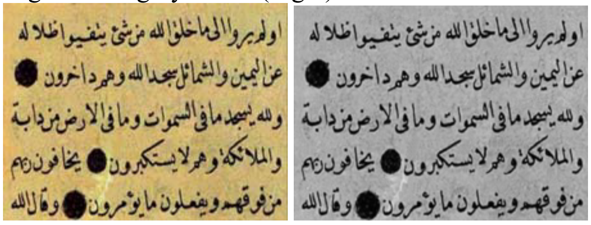
3.1 Document representation

After four steps, words are detected, isolated and coded in ASCII transcriptions. Henceforth, each document, *i.e.* the text of the document, corresponds to an ASCII file of coded text and employed for indexing and searching.

a. Preprocessing

Images of Arabic historical documents are highly noised. This is due to the degenerated quality of the original media and for technical/numerical reasons related to the computerizing. The images, which we could have from our partners and those collected from the Internet, are degraded and come in different format of colours.

- *Gray level transformation:* First, we transform each image in 256-gray levels (Fig.4):



(a) Coloured image (b) Gray image

FIG. 4 - Image transformation to gray.

- *Binarization:* It consists of comparing the pixels' gray levels with a threshold calculated from the histogram of gray levels of the image. First, we compute the histogram of gray levels by setting a table of 256 cells corresponding to the 256 level of gray to zero. Then, we traverse the entire image and count the number of pixels having each level of gray. The following stage determines the threshold for separating pixels in blacks and whites.

In our experiments, the level of gray with the greatest value used as threshold gave the best results.

- *Smoothing:* we gave to each pixel the dominant colour in its neighbouring (Fig.5).

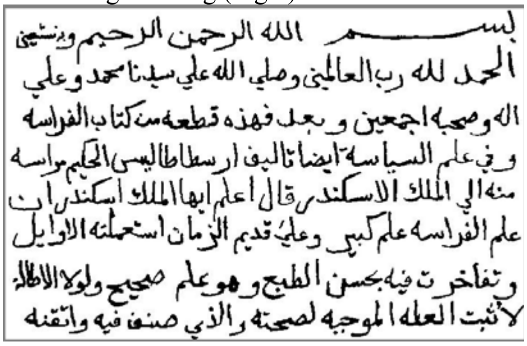
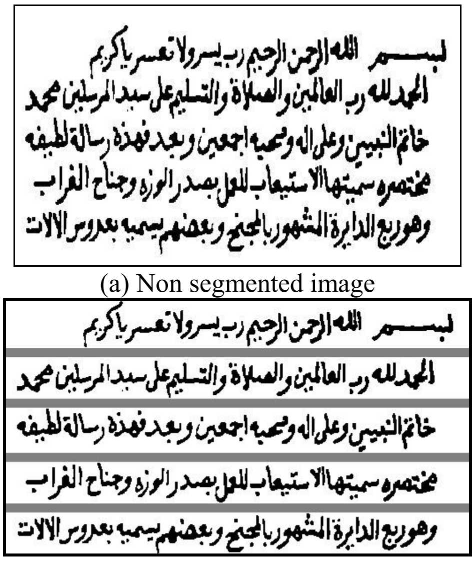


FIG. 5 - Binarization and smoothing.

b. Line segmentation

In this phase, we separate the lines from the text using the measure of the density of white lines in the horizontal projections. Local minima in the horizontal projections are analysed to localize lie separation.

Finally, we assign each pixel in the image to the area within successive red lines according to a distance. Hence, the text lines are detected (Fig.6).



(a) Non segmented image

(b) Line detection

FIG. 6 - Line detection and segmentation by horizontal projection analysis.

c. Subword decomposition

This stage consists in detecting the various connected components (CC) in an image, i.e., gathering the neighboring pixels in one unit called a connected component. At this level, we consider the diacritics as CC (Fig.7).

The following algorithm performs the line segmentation in CC:

Input: Line images

Output: the connected components

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. To frame this CC, the four extreme points are determined (high, low, left, right) which define the bounding box.
5. Return to stage 1.

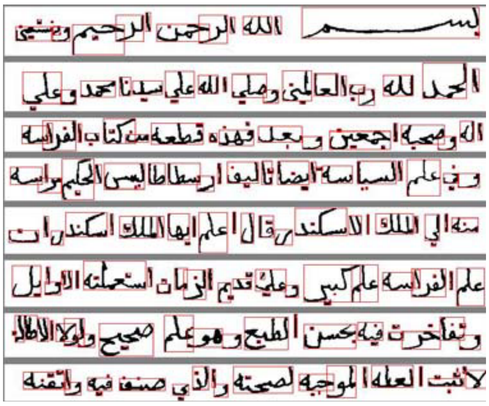


FIG. 7 - Connected components extraction.

d. Feature extraction and coding

This phase must guarantee a maximum of reliability, because the subsequently processing will not handle the image any more, but rather the results provided by this module. We run the below steps.

- *Base line detection:* The base line in the Arabic text carries significant information on the orientation of the text and the position of the diacritics. The most used method for base lines detection is the horizontal projections of the line. The base line corresponds to that whose projection contains the greatest number of black pixels. It is computed in the line segmentation phase (subsection b of 3.1).

- *Median zone detection:* The median zone consists in representing the body of the words or characters. To obtain this zone, high and low, base lines were located according to the principal base line; the space between these two lines is the median zone (Fig.8).

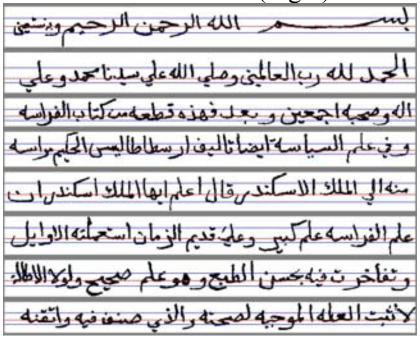


FIG. 8 - Base line and median zones detection.

- *Contour following:* The extraction of contours is an economic function that reduces the object recognition process and the three-dimensional localization. A contour is a sequence of points where each is coded by its direction according to the Freeman code.

After localizing the base line, the median zone of a line and the contours are extracted (Fig.9):

- *Diacritics:* The diacritics are simple or multiples points, which vary from one to three points, these points can be written up or down the primary part of characters.
- *Descenders:* Certainly, the descenders are the most used primitives in the recognition of handwriting. They are detected by a stroke lasting down of the median zone i.e. located in the lower zone.
- *Ascenders:* As opposite to descenders, ascenders are detected by a growth in the writing exceeding the median zone; therefore, we check the existence of ascenders in the higher zone.
- *Loops or holes:* In Arabic writing, loops are generally located in the median zone; they are as being a contour in another.

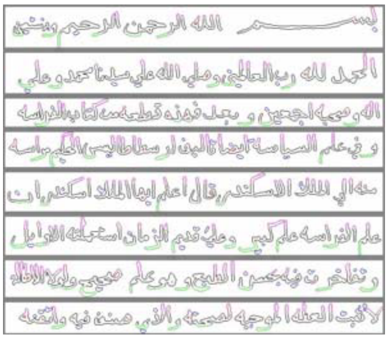


FIG. 9 - Feature extraction.

Each feature described above is coded in ASCII. For example, the word in the Fig.10 will have the code "qhhbj" which means the features (Down diacritic, Ascender, Ascender, Loop, Descender).

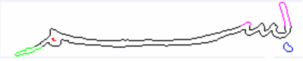
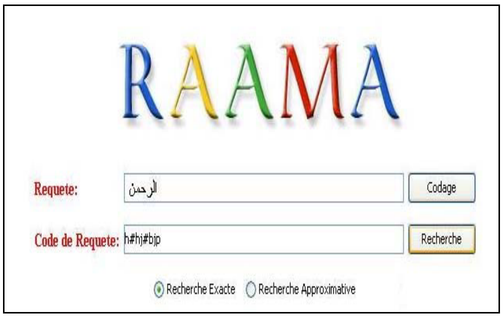


FIG. 10 - An Arabic word with its features.

The role of document representation is assigning to each image in the database a symbolic representation. Thus, a parallel database of file codes will be available for searching and indexing instead of image database. The transcription of images to ASCII representation allows simple, fast and already existent searching algorithms to be used.

3.2 User Interface

The user users can search document images by textual queries in Arabic natural language (Fig.12). To do this Arabic texts should be also coded as done for images. We use the TAB.1 for transcription.



RAAMA

Requete: الرحمن [Codage]

Code de Requete: h#hj#bjp [Recherche]

● Recherche Exacte ○ Recherche Approximative

FIG. 12 - The user interface

In this stage of the process, the character string (user query) is coded making correspondence to each letter a precise code, using Table.1, resulting in code strings. This latter will be compared with the file codes inputs.

By clicking the "Recherche" button, the results are displayed as in the Fig.13 ordered by their edit distance.

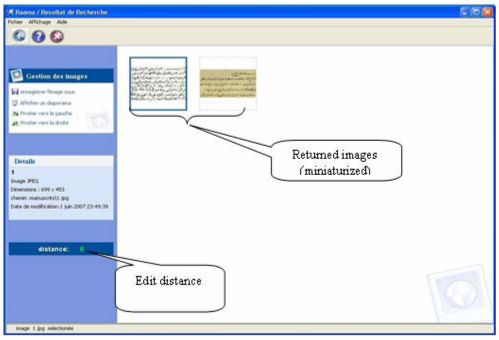
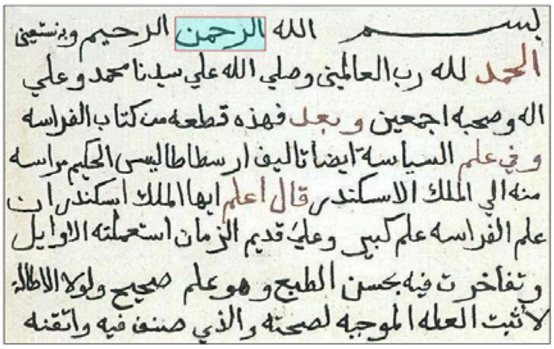


FIG. 13 - Search results

Character	Code	Designation
ـل - ـكـ - ـا - ا	h	Ascender
إ	hq	Ascender+ Down Daicritic
ـإ - أ	ph	Up Daicritic+Ascender
ل	hj	Ascender+Descender
ط	bh	Loop+Ascender
ظ	bph	Loop+Up Daicritic+Ascender
لا	hbh	Ascender+Loop +Ascender
ك	hp	Ascender+ Up Daicritic
ي	jq	Descender+ Down Daicritic
ة - ـضـ - ـخـ - - ـتـ ـنـ - نـ	p	Up Daicritic
غ	jp	Descender+ Up Daicritic
ث - شـ	pp	Up Daicritic+ Up Daicritic
ق - خ - ز - ن	jp	Descender+ Up Daicritic
ش	ppj	Up Daicritic+Up Daicritic+ Descender
ض	bpj	Loop+Up Daicritic+ Descender
ـقـ - ـغـ - ـة - ـفـ - ضـ	bp	Loop+ Up Daicritic
ق	pbj	Up Daicritic+Loop+ Descender
ب - جـ - يـ	q	Down Daicritic
ى - ر - س - ع - ح	j	Descender
ج	jq	Descender
ه - ـه - ـصـ - ـمـ - ـهـ	b	Loop
هـ	bb	Loop+Loop
م - ص - و - ع	bj	Loop+Descender
ـلا	hh	Ascender+Ascender
ـة	pb	Up Daicritic+Loop
ـغ	pbj	Up Daicritic+Loop+ Descender
ف	bjp	Loop+Descender + Up Daicritic
لأ	hbhp	Ascender+Loop +Ascender+ Up Daicritic
لإ	hbqh	Ascender+Loop + Down Daicritic+ Ascender

TAB. 1 – Character transcription in ASCII codes.

The user can increase the image size that he wants by a double click, to better see the result. In the window that is displayed the word nearest to the user query will be framed (Fig.14).

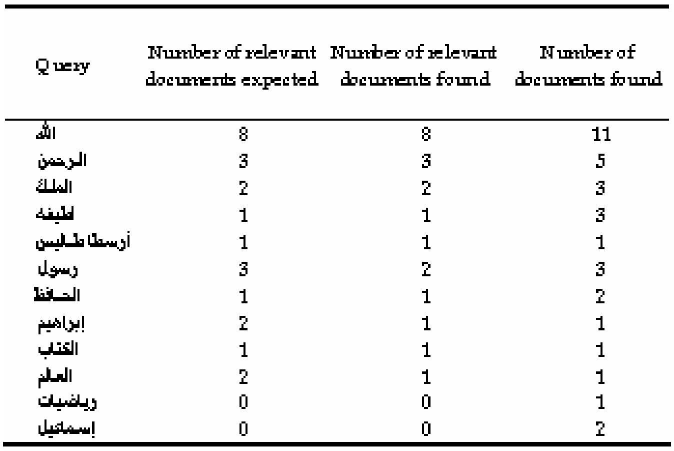


بسم الله الرحمن الرحيم وبه نستعين
الحمد لله رب العالمين وصلي الله علي سيدنا محمد وعلي
اله وصحبه اجمعين وبعد فهذه قطعه من كتاب الفراسه
وفي علم السياسه ايضا تاليف ارسطاطاليس الحكيم ارسله
منه الي الملك الاسكندر قال اعلم ايها الملك اسكندر ان
علم الفراسه علم كبير وعلي قديم الزمان استعملته الاوايل
وتفاخرت فيه بحسن الطبع وهو علم صحيح ولولا الاطاله
لاثبت العله الموجبه لصحته والذي صنف فيه واتقنه

FIG. 14- Search results detailed.

4 Preliminary experimental results

Our application was tested on 123 images of historical Arabic manuscripts from various sources covering different fields and for diverse queries. TAB.2 shows some examples.



Query	Number of relevant documents expected	Number of relevant documents found	Number of documents found
الله	8	8	11
الرحمن	3	3	5
الملك	2	2	3
لطيفه	1	1	3
أرسطا طاليس	1	1	1
رسول	3	2	3
الحافظ	1	1	2
إبراهيم	2	1	1
الكتاب	1	1	1
العالم	2	1	1
رياضيات	0	0	1
إسماعيل	0	0	2

TAB. 2 – Some examples of queries and obtained results.

The different experimentations show a recall of about 56.62% and a precision approximating 77.78%. It is not possible at this point of our investigation to speak and discuss about recall and precision. The preliminary test results are in our opinion very good in spite of the very challenging field of Arabic handwriting processing and imaged document retrieval. Our confidence goes toward Arabic sub-word processing which we consider as very promising, fast and reliable research direction rather than word processing, which is the main spotlighted track of some well-known research groups [BAL06]. In this later works, a major effort is set on word segmentation and the most recognition and retrieval errors were due to the word segmentation errors. To overcome this even open problem, we regard Arabic texts as a sequence of sub-words.

5 Conclusion

In this paper, we present and discuss an Arabic historical imaged documents retrieval system base on sub-word representation as ASCII codes. Each sub-word is assigned a sequence of codes according to its morphological features. Loops, ascenders, descenders, up and down diacritics are the characteristics we used. Each image in the database will be associated to a file code and the retrieval system will not handle the images but rather the file codes. Since these and the other more complicated features are very hard to locate and to extract, the resulting codes are not very faithful. Therefore, we proposed the use of an approximate string-matching algorithm for searching the file codes based on the Levenshtein edit distance. The test results proved the effectiveness of sub-word segmentation and representation of Arabic handwriting and the success of the retrieval method on a set of historical documents. However, some problems remain unsolved such as document slant correction and reliable feature extraction, which problems caused the performance decreasing of the coding and retrieval performances.

6 References

[BAI04] H. S. Baird, ″Difficult and Urgent Open Problems in Document Image Analysis for Libraries″, *Proc. 1st Int. Workshop on Doc. Image Analysis for Libraries (DIAL'04)*, pp.22-28, 2004.

[LEE02] G. Leedham, S. Varma, A. Patankar and V. Govindaraju "Separating Text and Background in Degraded Documents Images – A Comparison of Global Thresholding Techniques for Multi-Stage Thresholding", *Proc. 8th IWFHR2002, Niagara-on-the-Lake*, pp. 244-249, 2002.

[SPI95] 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, 1995, pp.382–389.

[SME97] Smeaton and A. Spitz, ″Using character shape coding for information retrieval″, *Proc. 4th Intern. Conf. on Doc. Anal. & Recogn.*, IEEE Computer Society Press, pp.974–978, 1997.

[CHE98] F. Chen and D. Bloomberg, "Summarization of imaged documents without OCR", *Comp. Vision & Image underst.*, vol.70, no.3, 1998.

[MAN96] R. Manmatha , C. Han and E. Risemen, ″word spotting: a new approach to indexing handwriting″, *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'96)*, pp.631-637, 1996.

[SAR07] T. Sari and M. Sellami, "State of the art of Off-line Arabic Handwriting Segmentation", *Intern. Journ. of Computer Processing of Oriental Languages*, 2007, vol. 20, no.1, pp.53-73.

[NAV01] G. Navarro, "A guided tour to approximate string matching", *ACM Computing Surveys*, 2001, vol.33, no.1, pp.31-88.

[BAL06] G.R. Ball, S.N. Srihari, H. Srinivasan, ″Segmentation-Based and Segmentation-Free Methods for Spotting Handwritten Arabic Words″, Proc. (IWFHR06), pp.20-26, 2006.