Journal of Theoretical and Applied Information Technology
10th December 2014. Vol.70 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FEATURE EXTRACTION IN SEGMENTED WORDS FOR SEMI-AUTOMATIC TRANSCRIPTION OF HANDWRITTEN ARABIC DOCUMENTS

NOUREDDINE EL MAKHFI AND RACHID BENSLIMANE
Information Processing and Transmission Lab,
Higher School of Technology,
Sidi Mohamed Ben Abdellah University,
Fez, Morocco
E-mail: n.elmakhfi@gmail.com

ABSTRACT

Scanning is a widely used solution for the preservation of ancient manuscripts. However, this solution gives masses of document images which content is not easily exploitable. In this work, we propose a new method that reduces considerably the manual transcription. The aim is to explore the content of digitized manuscripts. The proposed method is based on two main phases: the first one consists to segment the digitized manuscripts on words, while the second searches in a database of image words and their equivalent in text mode the corresponding text words of each image word if it exists. This search phase is based on a matching operation between features extracted words and those stored in the database. To characterize each word, we use SIFT (Scale Invariant Feature Transform) algorithm as a feature extractor for interest words points. The results of the comparison provide a set of words in text mode which helps transcription of Arabic manuscripts

Keywords: *Semi-Automatic Transcription; SIFT; Interest Points; XML; TEI; Word Segmentation; Image / Text; Arabic Manuscripts.*

1. INTRODUCTION

Transcription of old Arabic manuscripts is an essential step for indexing and diffusing the contents of these manuscripts. The cursive nature of the Arabic script has a disability for software OCR (Optical Character Recognition). A transcript with the text processing software or the html coding is not a best solution because the content is not respecting the exchange format. The parsers and analyzers have difficulties of research in these files. The complex structure of Arabic manuscripts can be compared to a hierarchical model. The strong dependence between the description of the data structure and how they are stored on the physical media, offers rigorous structures and paths, while maintaining relative simplicity of implementation.

The search for information in the images of Arabic and Latin manuscripts is a critical issue, despite abundant research in this field. However, we note that most research that exist have focused on segmentation and recognition of Arabic script printed. While the Arabic handwritten text present problems in recognition. These problems manifest themselves in overlapping letters. This makes access to content of the manuscripts images by text recognition impossible.

We can mention some projects that have worked on the transcription of Latin manuscripts. These projects are focused on the conversion of the structured content of digital documents to a content-oriented presentation. The structured content respects the exchange format. The integration of style-sheets allows passage of an exchange format to presentation format. The limitation of this method lies in its manual aspect that requires considerable effort and the problem of independence of the transcribed text with its location on the original images.

BAMBI (Better Access to Manuscripts and Browsing of Images) [1][2]: hypermedia support system for the study of ancient manuscripts in facilitating the work of researchers from the story text. The authors of BAMBI have worked on another project: ARMARIUS « A Living Online Archive for Ancient Manuscripts» [3]. This project is a model of collaborative digital library for annotating and transcribing ancient manuscripts online. « User Trace-Based Recommendation

System for a Digital Archive » [4] is an extension of the previous project. It comprises a system which tracks the important users' actions. It saves these actions as traces composed of hierarchical events. These events are considered as a reusable case.

DEBORA (Digital accEss to Books Of RenAissance) [5]: is a European project that developed tools for remote access to collections of documents from the 16th century. The technique of semi-automatic transcription [6] in this project is based on the segmentation of document pages in characters. Then it applies a clustering method on characters typed or printed.

EAMMS (Electronic Access to Medieval Manuscripts) [7], it is a North American project, which defined the recommendation for encoding and electronic storage of descriptions of medieval and renaissance manuscripts.

DIAMM (Digital Image Archive of Medieval Music) [8], it is a project involving the archives of digital images of music notes to the medieval period. The DIAMM project uses the annotations and the transcription tools of text to create the collections of images.

MASTER (Manuscript Access through Standards for Electronic Records) [9]. It is a generic system flexible enough to allow its application in various fields of cataloguing manuscripts. The technology chosen is based on international standards SGML (Standard General Mark-up Language) and XML (eXtensible Markup Language). However, the MASTER project does not describe the transcription of a manuscript, but the characteristics of cataloguing.

In the context of previous work, we have developed a web application using metadata and annotations [10]. However, we found that the metadata and annotations do not offer a good way to reproduce the entire content of Arabic manuscripts. In this way, we are interested in the transcription of Arabic manuscripts so as to help them access the content.

In this paper, we propose a new method for semi-automatic transcription which is based on identifying features of words segmented Arabic manuscripts. We are based on the Scale Invariant Feature Transform (SIFT) [11] [12] algorithm to extract the interest points of characterizing the segmented words. We use a consistent structure to encoding TEI (Text Encoding Initiative) [13] used to describe and transcribe these manuscripts. This structure offers a dual privilege, it serves to show the contents of manuscripts in image mode or in text mode, which main benefit is to help to access in the Arabic manuscript contents.

2. PRINCIPLE OF THE PROPOSED METHOD

Scanned documents stored as images can be structured according to an XML model. The exchange of information contained in these images requires a transcript for a transformation of image information in the form of textual data. These data can be structured according to the recommendations of TEI [13] that ensure compatibility with existing tools for electronic dissemination. The following block diagram summarizes the steps of the proposed method:

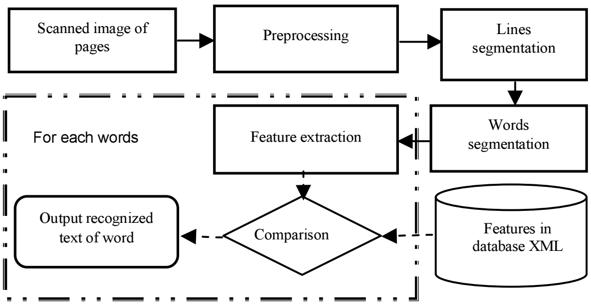
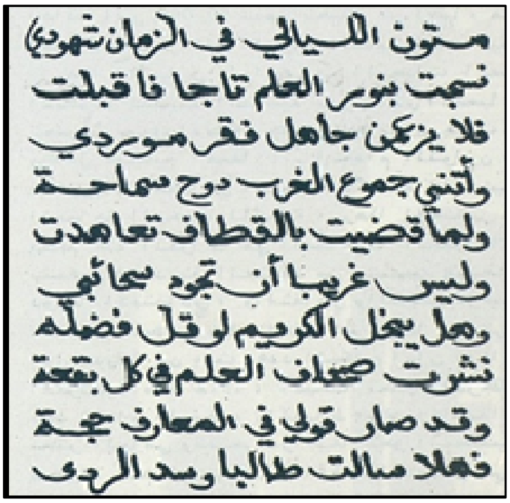


Figure 1: Block diagram of the proposed method for the transcription

2.1 Preprocessing

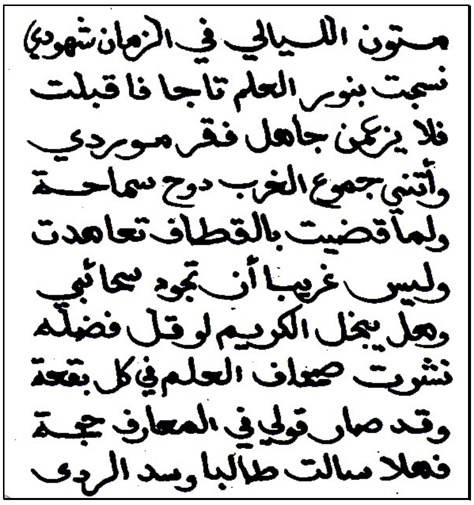
This process involves the conversion of the input image to grayscale and the grayscale image should be binarized. This binarization is done through the application of a global thresholding method of the raw image obtained after acquisition. We opted for the method of global thresholding for various processed images that have a bimodal histogram expressing both classes of pixels of the image background (paper texture) and pixel content (diagrams, characters and all the spots ink). In this work, we used the global optimal thresholding method of Otsu [14].

The binary image produced by binary segmentation often contains a noise whose removal is effected by the application of closing binary morphological filtering [15].



متون الليالي في الزمان شهودي
نسجت بنور العلم تاجا فاقبلت
فلا يزعمن جاهل فقر موردي
وأتتني جموع الغرب دوح سماحة
ولما قضيت بالقطاف تعاهدت
وليس غريبا أن تجود سحائبي
وهل يجهل الكريم لوقل فضله
نشرت صحاف العلم في كل بقعة
وقد صار قولي في المعارف حجة
فهلا سالت طالبا وسد الردى

Figure 2: Original image from journal A-jawahir [16]



متون الليالي في الزمان شهودي
نسجت بنور العلم تاجا فاقبلت
فلا يزعمن جاهل فقر موردي
وأتتني جموع الغرب دوح سماحة
ولما قضيت بالقطاف تعاهدت
وليس غريبا أن تجود سحائبي
وهل يجهل الكريم لوقل فضله
نشرت صحاف العلم في كل بقعة
وقد صار قولي في المعارف حجة
فهلا سالت طالبا وسد الردى

Figure 3: Result of the binarization and the morphological filtering (closing)

2.2 Text line segmentation

The method used for segmentation is the vertical projection lines of binary images. This method is effective if the lines of the text do not overlap. Moreover, it has the advantage of being very fast. In the case of images of printed text, the lines are generally spaced enough for not creating overlap, making segmentation of lines rather delicate. In this case, it is common to use the method of the vertical projection. This method proceeds by first calculating the profile of the vertical projection of the filtered image.

Figure 4 illustrates the vertical projection profile calculated for the image of Figure 3. It provides various features on the main lines on the background of the binary image and the row heights. The separation zones between the lines corresponding to the detected vertical projection profile of the trays. The width of the tray defines the separating strip between two lines. The minima correspond to the observed baseline.

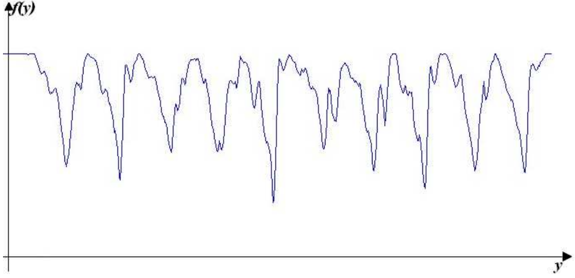
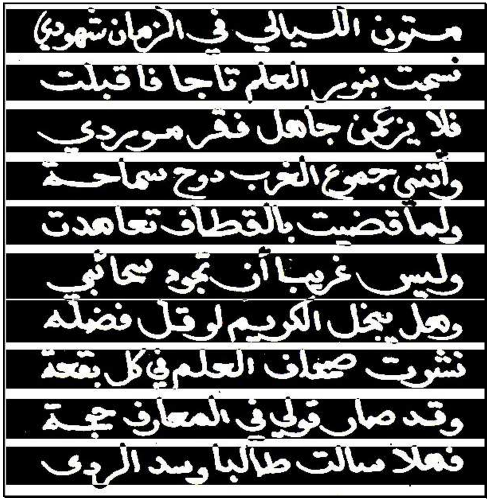


Figure 4: Vertical projection profile

Figure 5 illustrates the result of the method used for segmentation lines.



متون الليالي في الزمان شهودي
نسجت بنور العلم تاجا فاقبلت
فلا يزعمن جاهل فقر موردي
وأتتني جموع الغرب دوح سماحة
ولما قضيت بالقطاف تعاهدت
وليس غريبا أن تجود سحائبي
وهل يجهل الكريم لوقل فضله
نشرت صحاف العلم في كل بقعة
وقد صار قولي في المعارف حجة
فهلا سالت طالبا وسد الردى

Figure 5: Results of text-line segmentation

2.3 Words segmentation

The principle of the proposed word segmentation method is based on the selection of words, pseudo-words and characters in a first step and a morphological dilation of these pseudo-words in a second step.

In the first step, we proposed the segmentation of characters and pseudo-words. For this purpose, we apply the labels of the binary image to extract the connected components in each line. The following figure shows the result of the segmentation of pseudo-words and the isolated letters.

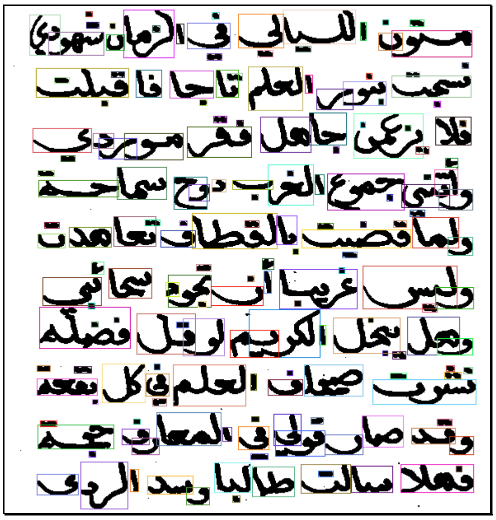


Figure 6: Results of the segmentation of pseudo-words and isolated characters

In the second step, our goal is to extract all relevant information in each word to make it easier for transcription. To this achieve, we apply a morphological dilation of the binary image to the right to allow the merger of isolated characters and pseudo-words. The horizontal projection at each line provides words. The following figure shows an example of segmentation of words segmentation of words.

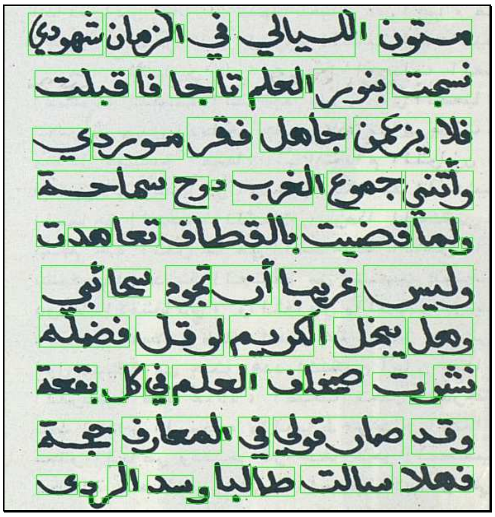


Figure 7: Results of words segmentation

2.4 Feature extraction

The principle of extraction of the words features is based on SIFT [12] algorithm. This algorithm consists of detecting the interest points in several invariant transformations: rotation, scale, illumination and the minor viewpoint changes. The SIFT algorithm is based on finding the extremes in scale space. The Gaussian scale space of an image I (x, y) is defined by the convolution of the intensity of the image with a Gaussian filter:

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y,\sigma) \quad (1)$$

Where G (x, y, σ) is the Gaussian filter is defined by:

$$G((x,y)',\sigma) = \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}) \quad (2)$$

I(x, y, σ) is the function image.
Lowe [12] simplified the calculation by introducing DOG (Difference of Gaussian)

$$DoG(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma)$$

The DOG which will be used to detect local extremes

$$DoG(x,y,\sigma) = G(x,y,k\sigma) - G(x,y,\sigma) \quad (3)$$

$$DoG(x,y,\sigma) \approx (k-1)\sigma^2\nabla^2 G \quad (4)$$

In practice, each octave of scale space (doubling of σ) is divided at intervals (s intervals). The subtraction images offer adjacent difference images of Gaussian (DOG). The highlight of SIFT is the simplification of Gaussian convolutions with the difference of Gaussian DOG. So the DOG function is fast to compute (with subtraction images) providing an approximation of the Laplacian. The maximum are then pixels which have an intensity maximum or minimum relative to its immediate neighbourhood in the image (8 neighbours), as well as those in the scale space (9 neighbours in the previous level and 9 neighbours in the following scale). The minima and local maxima are detected. Note that the detector of Lowe is invariant at scaling. However, the descriptor associated with it is quite robust and gives excellent results in the case of rotation and change of scales. The following table describes the steps followed to extract interest points by the SIFT algorithm:

TABLE 1. STEPS OF INTEREST POINTS DETECTION (SIFT)

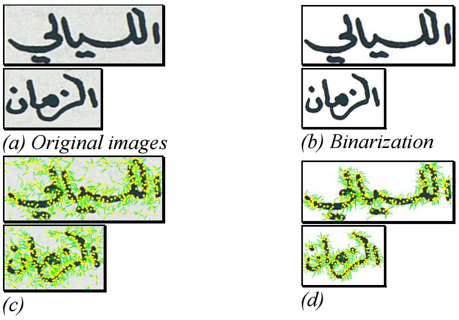
Steps		Results
1	Convert the input image to grayscale	Image size [M, N]
2	Scale-space extrema detection	Points detected at different scales

3	Location of interest points	Points at original scale
4	Assign dominant orientation	Orientation
5	Interest point descriptors	Points, vectors of 64 components

2.5 Results of detection and comparison of words:

2.5.1 Interest points detection*:*

Interest points determined by the SIFT method are robust to changes in invariant point scales, and change in brightness, and are immune to noise. We found that these points follow the form of the writing so as to represent the variations of handwriting. In our case, we have exploited these interest points for the recognition of forms of writing Arabic uniform structure. Therefore, our proposed method is based on the detection of interest points from images segmented words. We treated the extraction results of two interest points segmented words (الليالي nights, الزمان time). The following figure shows the concentration of interest points on writing to characterize its shape. In (c), the number of points is greater than that of (d). This explains why the texture effect can cause some unwanted points. We eliminated these points with the binarization process as shown by the result (d).



(a) Original images *(b) Binarization*

(c) *(d)*

Figure 8: Examples of interest points extracted from the SIFT method

2.5.2 Comparison of interest points*:*

The comparison between two interest points can be achieved by several methods. The choice of such a method can be realized depending on the cost of treatment. The two interest points are characterized by their properties: (x, y, scale, orientation and descriptor [64 and 128]). The comparison can be done therefore in two steps:

- ✓ The first comparison step between two points is performed by examining the dominant orientations of interest points.
- ✓ The second step of comparing is to compute the distance between two interest points of descriptors vectors.

The methods most commonly used are based on the comparison and correlation calculating distances between vectors. Calculating the distance between two vectors u and v of the descriptors of interest points can be done with the Euclidean distance or Mahalanobis distance. In general case, the comparison of interest point SIFT can be done directly on a query image and a target image as shown in the following figures.

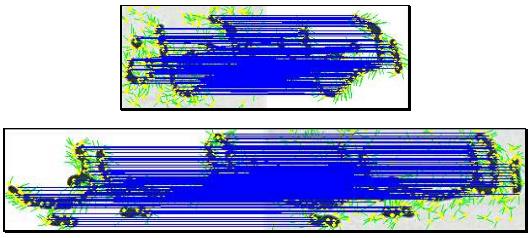
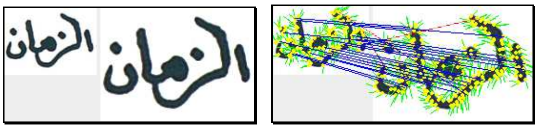


Figure 9: Comparison of interest points between two words

Invariance to changes of scale causes a reduction in the number of interest points. This reduction is dependent on the resolution of the image. The number of interest points increases if the image resolution is high. We note that the comparison result (Fig.10b) preserves the form of writing even if the number of matched points is less than the overall number. In (b), we obtained three false points that are detected on the letters “L ل”, “Z ز” and “N ن”.



(3) *(b)*

Figure 10: Result of scale-invariant interest point detection between two words

The direct method of image queries with the comparison target image cannot be applied for the case of a transcription system. A backup of the interest points in a database is essential. It is the use of interest points recorded with their corresponding words in text mode whose main interest is to speed semi-automatic transcription.

2.6 Feature detection in data base XML

Search words similar to the query image are done by comparing features of an XML model. To realise this, we proposed a model of the XML database that stores the words in text mode and interest points. The root element of the XML model is called "Library". It contains one or more "Document". This last element is any document (manuscript, book …). Some metadata is needed to identify the document title, the reference (the id attribute) and language (title, language). The "featuresType" element specifies the type of detector and descriptor of interest points. We developed the XSD schema with the software [17] as shown in the figure:

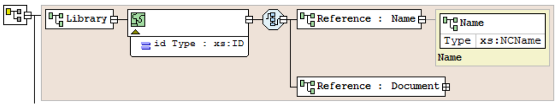


Figure 11: Schema diagram for 'Library' root element

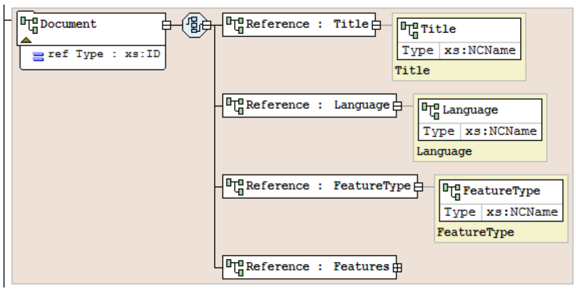


Figure 12: Schema diagram for database

The element "Features" as shown in the following diagram contains one or more "Feature". These are the elements that have saved the features of handwritten words. The text equivalent for each word image is saved in the "text" element. The "Feature" element therefore characterizes a word and it contains one or more interest points. Each word is identified by a unique attribute "idf" in the "Feature" element. Interest points associated with the manuscript are recorded in the "InterestPoint" element. The "InterestPoint" element includes the following:

- The coordinates (x, y) in the image;
- Scale;
- Dominant orientation of the interest point: ori;
- Size of the descriptor: dl (64 or 128);
- Descriptor.

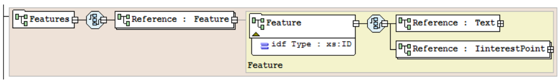


Figure 13: Schema diagram for 'Feature' element

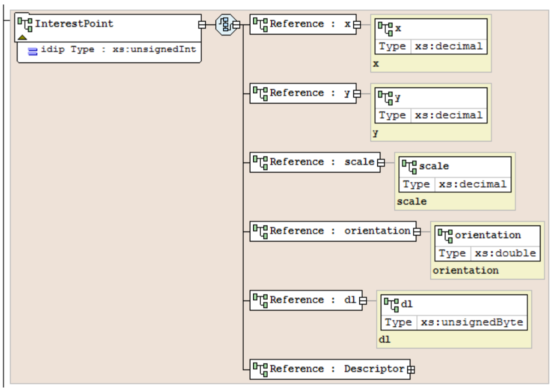


Figure 14: Schema diagram for 'pInteret' element

The element "Descriptor" contains a single element "d". The latter occupies the vectors characterizing the point of interest. Each vector can have a dimension of 64. However, the dimension 128 gives high accuracy, but at the price congestion database. The dimension of the vectors is listed in the "dl" element. The "n" attribute allows organizing the feature vectors of the point of interest. It serves as an identifier.

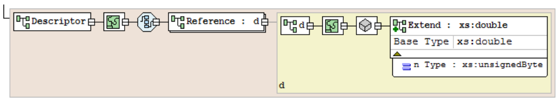


Figure 15: Schema diagram for 'Descriptor' element

2.7 Transcription of words

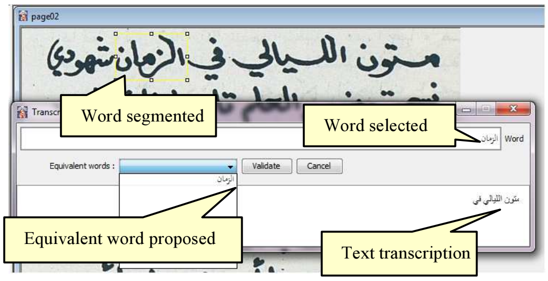
Transcription of words is a way to transform the information contained in the documents as digital images identifiable information by the machine. Transcription of words segmented into Arabic manuscripts images requires an XML encoding conforms to the recommendations as TEI "Text Encoding Initiative". It is a TEI encoding that respects the W3C XML standard [18], which aims to standardize the coding of these documents and to facilitate their exploitation vis-à-vis their trade, their exploration and dissemination online or offline.

In this section, we will use the metadata according to an XML schema. This scheme provides a mapping between the physical structure and logical structure of manuscripts respecting the TEI encoding. Our goal is to ensure the transcription, indexing and exploration of Arabic manuscripts on the basis of XML data. We can classify these metadata into two classes:

In the first class, we use metadata about the document entity, they provide essential information about the document as a whole, and we can encode the documents manuscripts clearly: title, author, date of publication, keywords, etc.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In the second class, we treat the metadata contained in the body of the document as a transcript of the content. The transcription of each word in a singular form of word annotation segmented. The concept of semi-automatic transcription allows giving the equivalent of each word picture segmented form text. The following figure represents the interface of transcription that we have developed.



Word: الزمان

Equivalent words : الزمان

متون الليالي في

Figure 16: Transcription interface

3. RESULTS AND DISCUSSION

Search by metadata is not exhaustive. Several information included in images of manuscripts are inaccessible. The annotation technique also does not allow full access to the content of the images of manuscripts. This technique consists in associating with each page (digital image) a group of words that defines the textual content of the image. The indexing and retrieval of image content is based mainly on annotations that relates to images in XML files. Transcription is a better solution since it allows a complete transformation of the image content in the form of structured textual content. For this purpose, there are several studies that refer to transcription by providing a means of converting structured content to a content-oriented presentation. However, this technique suffers from its manual appearance. The information is difficult to identify on the images of manuscripts. For this reason, we opted for the semi-automatic transcription of Arabic manuscripts. We facilitated the task of transcription by automatically locating the words to transcribe images. We solved the problem of the independence of the image content and structured text content. Direct access to the information contained in images provides a means for research and the establishment of Arabic manuscripts.

The result of the proposed method depends on the segmentation of words. In the case of overlapping of lines or words of a passage by manual transcription is required.

4. CONCLUSION

In this article, we have developed a model of semi -automatic transcription of Arabic manuscripts. We began by segmenting words into pages of manuscripts. To do this, we extracted features of segmented words aimed to help the semi -automatic transcription of image content at an exchange format. We experienced an exchange format based on XML encoding. We can see that the transcription TEI is a solution that provides opportunities like the description and analysis of the content of manuscripts, cataloguing support, structuring information and the presentation of information contained in these manuscripts. Indeed, it provides a uniform description of these manuscripts for indexing, improve trade and facilitate the sharing of textual information on computer networks.

For the perspectives of future work, we think of automate transcription incorporating algorithms segmentation and character recognition.

REFERENCES

[1] Stéphane Nicolas, Thierry Paquet, Laurent Heutte, 2003. “Digitizing Cultural Heritage Manuscripts”: the Bovary Project. in ACM Symposium on Document Engineering, ACM Doc Eng 2003, Grenoble, France, pp. 55-57

[2] 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.

[3] ARMARIUS- A Living Online Archive for Ancient Manuscripts. REIM Doumat, E. Egyed-Zsigmond, J.M. Pinon. Dans 11ème Colloque International sur le Document Electronique (CIDE11), Rouen, France. pp. 44-56. 2008.

[4] Reim Doumat, Elöd Egyed-Zsigmond, Jean-Marie Pinon. “User Trace-Based Recommendation System for a Digital Archive”. In Proceedings of ICCBR 2010. pp.360-374, Italie. pp. 360-374. Lecture Notes in Computer Science 6176. Springer . ISBN 978-3-642-14273-4.

[5] DEBORA: projet européen n°. LB 5608 A. Coordinateur R. Bouché, juin 2000.179 pages.

[6] Le Bourgeois F., Emptoz H., « DEBORA: Digital AccEss to BOoks of the RenAissance », IJDAR, vol. 9, p. 193-221, April, 2007.

[7] http://www.hmml.org/eamms/index.html

[8] http://www.diamm.ac.uk

[9] BURNARD, Lou. ; ROBINSON, PETER. Vers un standars européen de description des manuscrits: le projet Master. Document numérique. 1999, vol 3, no 1-2, p.151-169.

[10] O. El Bannay, R.Benslimane, N. El Makhfi and N. Rais "Searching in Arab Manuscripts Using Metadata and Annotation" European Journal of Scientific Research ISSN 1450-216X Vol.28 No.1 (2009), pp.155-164 EuroJournals.

[11] David. G. Lowe, "Distinctive ImageFeatures from Scale-Invariant Keypoint". January 5, 2004.

[12] David G. Lowe, Object Recognition from Local Scale-Invariant Features. Proc. of the International Conference on Computer Vision, Corfu, 1999.

[13] TEI Consortium, "TEI P5: Guidelines for Electronic Text Encoding and Interchange", edited by Lou Burnard and Syd Bauman 1.9.1. Last updated on March 5th 2011.

[14] N. Otsu, «A Threshold Selection Method from Gray-Level Histograms», IEEE transactions on Systems, Man and Cybernetics, 9(1), p. 62-66, 1979.

[15] J.Serra , Image Analysis and Mathematical Morphology, Academic Press, 1982

[16] Poetry by Mohamed Al-Idrissi during the cultural week organized at Al-Quaraouiyine school in early March 1985, Fez, Journal of Al-jawahir (1985).

[17] http://www.w3c.org/xml

[18] Stylus Studio XML Editor, http://www.stylusstudio.com