International Journal of Soft Computing 14 (1): 28-32, 2019
ISSN: 1816-9503

Effective Method for Segmentation of Arabic Manuscript Documents

[1]Aicha Mint Aboubekrine, [1]Kamal Eddine EL KADIRI, [1]Youness Tabii, [2]Mohamed Lamine Diakite, [3]Lamarti Sefian Mohammed, [4]Nagi ould Taleb
[1]LIROSA Laboratory, Faculty of science, Abdelmalek Essaadi University, Morocco
[2]URDNI Research Unit, Faculty of Science and Technology, University of Nouakchott Alaasriya Nouakchott, Mauritania
[3]LASAD Laboratory, Abdelmalek Essaadi University, Morocco
[4]Computer Science and Systems Engineering Laboratory, Abdelmalek Essaadi University, Morocco

Abstract: In this study, we are interested in Arabic handwritten documents. We propose a method of segmentation of Arabic manuscript documents first in lines, then in words and finally, proceed to their annotation. This research poses a number of problems because of the difficulties related to the processing of Arabic manuscripts. After binarizing the images of the Arabic handwritten documents, we compute the vertical projection profile of the text image to determine the lines of text, then the horizontal projection profile for the segmentation of lines into words and finally, the detection of under words and real words. The technique we offer allows us to access the content of the handwritten documents in a very efficient and fast way.

Key words: Arabic handwritten documents, vertical projection profile, horizontal projection profile, binary image, Arabic manuscript, segmentation

INTRODUCTION

Historical handwritten documents are a valuable cultural heritage because they provide a better understanding of the tangible and intangible cultural aspects of the past. The need to preserve and manipulate these documents requires global emerging efforts and the use of various techniques from different scientific fields, like the Optical Character Recognition (OCR) as a very complex operation which constitutes the subject of researches in the digital humanities including teams of researchers in letters, human, social and computer sciences. Except that, the cursive nature of Arabic writing has a handicap for the Optical Character Recognition (OCR) algorithm (Likforman-Sulem *et al*., 2007). The processing of Arabic documents faces many problems because of the type of writing, continuity and discontinuity of writing. All these facts make the Arabic script more difficult than other scripts like Latin. A large number of documents are stored in their original format (paper) or as scanned images and remain to be exploited. Converting these images and extracting their contents in text format will allow users to automatically or semi-automatically retrieve information from text queries instead of searching manually in scanned images. In this research, we are interested in the Arabic manuscript documents and the difficulties presented by their treatments and particularly the search for information contained in the images of these manuscripts.

We note that most of the existing research has focused on the segmentation and recognition of printed Arabic characters; knowing that, the Arabic manuscript text presents more complex problems of recognition. Segmentation is an image-processing step where the lines, words and characters of the text can be determined. Text line segmentation can be less complex for handwritten documents that contain distinct spaces between lines and more complex for documents where lines of text overlap, stroke, curvilinear, space variation between lines of text and slanted texts (Al-Dmour and Fraij, 2014; Soualah and Hassoun, 2011). In this study, we propose a method to extract the lines of an Arabic handwritten text without any constraint for the writer. After detecting the lines and words based on the projection, we proceed with the extraction of the partial words of each line and then characterize the real words that have an opposite meaning.

The semi-cursive nature of the Arabic script that is characterized by its calligraphy and the presence of ascending and descending text and the overlap between the words and the diacritical points above or below the characters lead us to introduce the different stages of the approach:

Corresponding Author: Mint Aboubekrine Aiche, URDNI Research Unit, Faculty of Science and Technology, University of Nouakchott Al-Aasriya, Nouakchott, Mauritania

- The first step is the binarization of images of handwritten documents. Then we remove the noise
- The second is a method of segmenting images into lines
- The third is the extraction of words and the differentiation between real words and subwords (a part of words)

Literature review: The progresses of digital digitization and electronic storage have led to the digitization of historical documents for the preservation and analysis of cultural heritage. This development simplifies access to historical manuscripts and accelerates the search for different copies of a manuscript. Nevertheless, Arabic manuscript documents differ from other Latin languages. Comparing their copies word by word, determining and analyzing the difference between them are time consuming for the researcher (Soualah and Hassoun, 2011). While many methods have been proposed to extract textual content from handwritten documents (Arivazhagan *et al*., 2007; Ouwayed *et al*., 2010; Papavassiliou *et al*., 2010; Manmatha and Rothfeder, 2005; Souhar *et al*., 2017; Younes and Abdellah, 2015), some of them deal with different specific languages such as Chinese, Latin and Arab scripts. Since, the Arabic language differs from other languages (Makhfi and Benslimane, 2014; Saabni *et al*., 2014) because of its characteristics and the type of its writing, there must be additional study on the latter more than on other Latin languages. The segmentation of the lines is considered as the first step, to have the access to the content of the handwritten documents. Methods have been proposed and applied for the extraction of Arabic manuscripts and many of them have yielded effective results.

Earlier research attempting to segment text lines of Arabic manuscripts (Manmatha and Rothfeder, 2005; Souhar *et al*., 2017; Younes and Abdellah, 2015; Makhfi and Benslimane, 2014) combined with the methods proposed and applied for extraction yields effective results with high proximity. After the extraction of the lines, we can detect the line words. The most difficult stage during the processing of Arabic manuscripts and study in this area remains restricted. Indeed, for Arabic writing, the problem of segmenting words into full-page can be difficult compared to other Latin languages (Makhfi and Benslimane, 2014). In this study, we distinguish the projection method and we will present its different sections.

MATERIALS AND METHODS

Proposed approach: The segmentation of Arabic manuscripts into lines of text and words is an important step in making recognition systems more efficient and accurate. A line of a text can be explained as a set of aligned words. We focus then on detecting the parts of Arabic words that make up this line. From this point of view, the appropriate approach will be the one based on the analysis of the arrangement of connected components. In this research, we proposed the approach that concerns the processing of a handwritten document that will be traversed by the necessary steps to reach our goal which is the extraction of words and their annotation. Before extracting the lines, we go through the following steps.

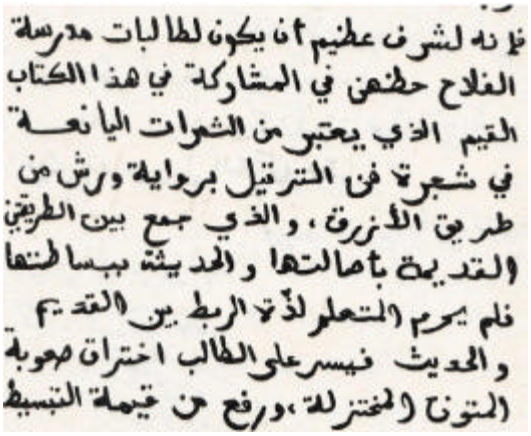

Fig. 1: The original image of an Arabic manuscript

Pretreatment: Which concerns the preparation of the document, the binarization of the images and the suppression of noise. This process involves converting the input image to grayscale. The latter must be binarized. This binarization is done by applying a thresholding method of the raw image obtained after acquisition. In this research, we used Otsu's optimal thresholding method. The binary image produced by binary segmentation often contains a noise whose elimination is performed by the application of a binary morphological filtering (Fig. 1).

Segmentation of lines: This part concerns the detection of lines of text (Fig. 2) while segmentation is one of the most important steps in any handwriting recognition system (Souhar *et al*., 2017; Younes and Abdellah, 2015). The study of segmenting lines of text from an Arabic manuscript that we propose is a method that allows us to access the contents of the manuscript and this is using the object we are looking for. In this step, we used the vertical projection explained below.

Vertical projection: Segmentation from histograms, it remains the most used approach in recognition systems

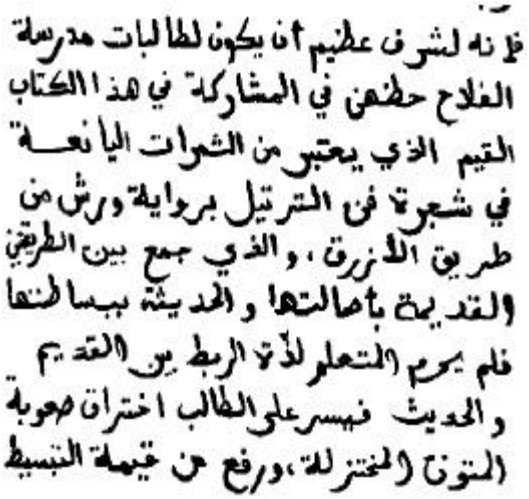

Fig. 2: The result of the binarization of the original image

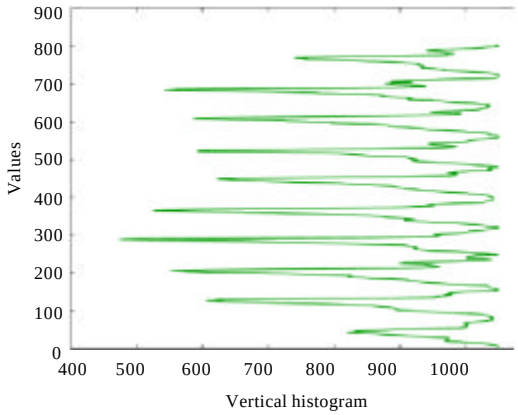

Fig. 3: The result of the vertical histogram

because of its high speed of retrieval of results (Makhfi and Benslimane, 2014) but it is sensitive when it comes to the problem of overlapping between adjacent lines. Therefore, we based ourselves on the images which have spaced lines. This method is effective if the lines of the text do not overlap. In addition, it has the advantage of being very fast. In the case of printed text images, the lines are generally spaced apart, so as to avoid overlapping which makes the segmentation of lines rather delicate. In this case, it is common to use the vertical projection method. 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 (Fig. 3). Segmentation steps in lines using the vertical projection profile (Fig. 5). Segmented line binary image (Fig. 3) projection profile of the mage (4) peak detection profile and line segmentation.

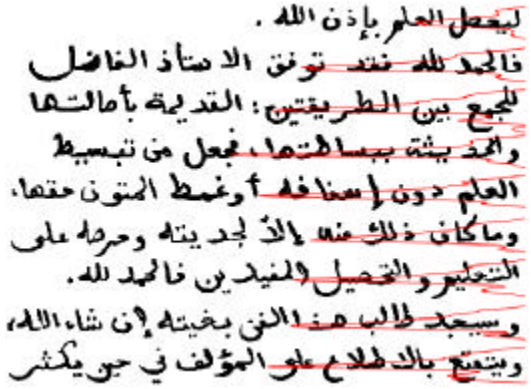

Fig. 4: The result of the segmentation from calculation of the vertical histogram

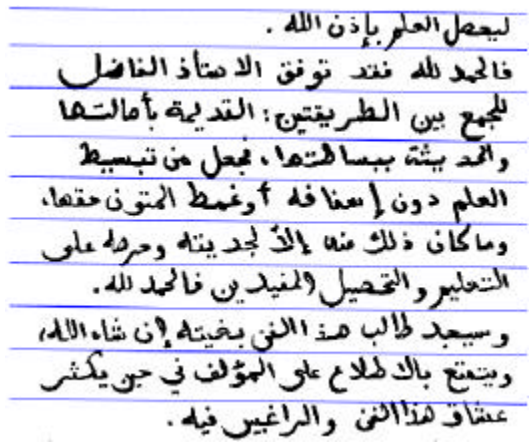

Fig. 5: The result of the segmentation in lines

Segmentation of words

Horizontal projection: Segmentation from horizontal histograms for the extraction of words is the effective approach to detect text words. This approach is used in recognition systems because of the writing of Latin languages. However, it is sensitive when it concerns the type of Arabic writing. Its treatment is mainly due to the semi-cursive nature of the Arabic script, the overlap between pieces of Arabic: (words) and also, the diacritical points located above or below the characters. An approach based on the projection profile (Al-Dmour and Fraij, 2014; Makhfi and Benslimane, 2014), involving space spacing and the average line width of the recorder is used to detect word boundaries. The principle of the proposed word segmentation method is based on the selection of words, pseudo-words and characters in a first step and calculates the morphological space of these pseudo-words in a second step.

Fig. 6: The result of the extraction of pseudo-words

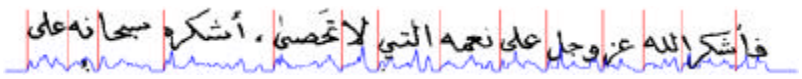

Fig. 7: The result of the extraction of words

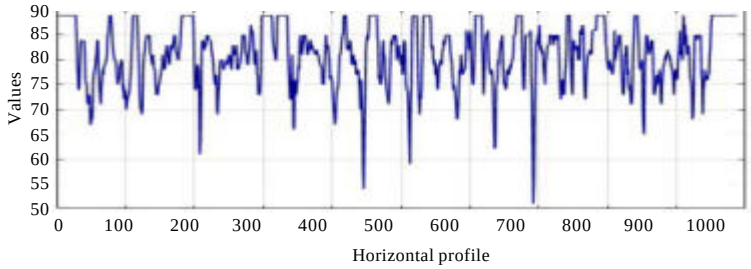

Fig. 8: Result of the peaks of the histogram which presents the spaces between the words that have a meaning

At first, we proposed the segmentation of isolated characters and pseudo-words. To do this, we apply the tags of the binary image to extract the connected components in each line. Figure 6 shows the result of the segmentation of pseudo-words and isolated letters. In the second step, our goal is to extract all the relevant information in each word to filter the real words. For this, we apply a morphological dilation of the binary image to the right to allow the fusion of isolated characters and pseudo-words. The horizontal projection at each line provides words. Figure 7 and 8 shows an example of word segmentation.

RESULTS AND DISCUSSION

Experimental evaluation: The approach is applied to a set of images of Arabic manuscript documents. At first, manuscripts were selected at spaced lines. Then, we transferred the grayscale document image to binary using Otsu binarization method. Afterwards, we applied the segmentation of text in lines and words adapted to Arabic manuscripts, based on the peaks of the histogram which from their turns, present the spaces between the words (Fig. 8).

Arabic words are different from words in other languages, since, a word can be composed of one or more pieces (Fig. 9). For this, the extraction becomes very difficult. During this extraction, we confronted cases that

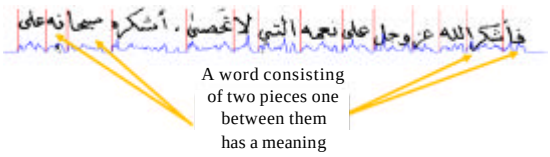

Fig. 9: Resultat of the false words has one between the mata meaning

we must take into consideration. If one piece gives meaning we can think of it as a real word, otherwise we will have to fission it with one or more other pieces to have a meaningful word. This step must be semi-automatic to obtain a collection of useful words (Fig. 9). For our research, we assume that every piece that conveys a meaning is a real word. This segmentation have done semi automatically, as some precise adjustments improve the results in practice.

CONCLUSION

In this study, we presented an approach of segmentation of Arabic manuscript documents. During our research, we noted the existence of numerous techniques for extracting lines of text from handwritten documents. The techniques we have chosen give favorable results only, if the document is flawless that is without any overlap or defect such as

connecting it between separate lines, etc. These results are encouraging, important, characterized by higher proximity and have fewer errors in the end segmentation. As a future perspective, we aim at a classification of frequent words in the manuscript and we will focus on a semi-automatic annotation based ontology.

REFERENCES

Al-Dmour, A. and F. Fraij, 2014. Segmenting Arabic handwritten documents into text lines and words. Intl. J. Advancements Comput. Technol., 6: 109-119.

Arivazhagan, M., H. Srinivasan and S. Srihari, 2007. A Statistical Approach to Handwritten Line Segmentation. In: Document Recognition and Retrieval XIV: 30 January-1 February 2007, San Jose, California, USA., Lin, X. and B.A. Yanikoglu (Eds.). SPIE, San Jose, California, ISBN:9780819466136, pp: 6500T-1-6500T-11.

Likforman-Sulem, L., A. Zahour and B. Taconet, 2007. Text line segmentation of historical documents: A survey. Intl. J. Doc. Anal. Recognit., 9: 123-138.

Makhfi, N.E. and R. Benslimane, 2014. Feature extraction in segmented words for semi-automatic transcription of handwritten Arabic documents. J. Theor. Appl. Inf. Technol., 70: 68-75.

Manmatha, R. and J.L. Rothfeder, 2005. A scale space approach for automatically segmenting words from historical handwritten documents. IEEE. Trans. Pattern Anal. Mach. Intell., 27: 1212-1225.

Ouwayed, N., A. Belaid and F. Auger, 2010. General text line extraction approach based on locally orientation estimation. Proceedings of the 17th Conference on Document Recognition and Retrieval, January 18, 2010, San Jose, California, USA., pp: 1-10.

Papavassiliou, V., T. Stafylakis, V. Katsouros and G. Carayannis, 2010. Handwritten document image segmentation into text lines and words. Pattern Recognit., 43: 369-377.

Saabni, R., A. Asi and J. El-Sana, 2014. Text line extraction for historical document images. Pattern Recognit. Lett., 35: 23-33.

Soualah, M.O. and M. Hassoun, 2011. Which metadata for ancient Arabic manuscripts cataloguing?. Proceedings of the DCMI International Conference on Dublin Core and Metadata Applications, September 21-23, 2011, Hague, Netherlands, pp: 137-146.

Souhar, A., Y. Boulid, E. Ameur and M.M. Ouagague, 2017. Segmentation of Arabic handwritten documents into text lines using watershed transform. Intl. J. Interact. Multimedia Artif. Intell., 4: 96-102.

Younes, M. and Y. Abdellah, 2015. Segmentation of Arabic handwritten text to lines. Procedia Comput. Sci., 73: 115-121.