Prior Segmentation of Old Arabic Manuscripts by Separator Word Spotting

Conference Paper · September 2014

DOI: 10.1109/SOCPAR.2014.7007977

2 authors, including:

Nabil Aouadi
University of Tunis El Manar
11 PUBLICATIONS **49** CITATIONS

SEE PROFILE

Prior Segmentation of Old Arabic Manuscripts by Separator Word Spotting

Nabil Aouadi
University of Tunis, LaTICE-ENSIT
5 Avenue Taha Husseïn 1008 Tunis B.P 56. Bab Menara Tunis
00216 93 63 43 26
Nabil.aouadi@utic.rnu.tn

Afef Kacem Echi
University of Tunis, LaTICE-ENSIT
5 Avenue Taha Husseïn 1008 Tunis B.P 56. Bab Menara Tunis.
00216 99 93 25 09
Afef.kacem@esstt.rnu.tn

Abstract— **Because of the low quality of old manuscripts, the complexity of Arabic script and the different writing styles, segmenting them is a challenging problem. This work aims to preprocess these manuscripts to be correctly segmented into independent words for text recognition. The idea is to spot separator words, detach them from neighboring words if necessary and use them to segment text-lines into words. To locate separator word in these document images, we proposed a word spotting method based on Generalized Hough Transform. This method is performed using convex theory points. Around a window centered on the group of votes of the separator word, it detects all connections below text-line baseline, analyses terminal letter morphology and tries to separate between touching or overlapping components. We tested the proposed system on Arabic historical manuscripts from the 19th century onwards conserved in the Tunisian National Archives. Experiments show very encouraging results.**

Keywords-component: Segmentation, Hough Generalized Transform, Word Spotting, Convex Point Theory, Skeleton, Baseline, angular variation. .

I. INTRODUCTION

A large number of books and historical Arabic manuscripts are kept in the Tunisian National Archives (TNA) and are threatened by extinction due to moisture and the acidity of the paper. Having digital versions of old and fragile documents is a good way for preserving them and makes them useful for several studies and projects. However, they are often large, unstructured and only available in image formats, which makes them difficult to access. But, it would be useful if they can be automatically indexed. Word spotting techniques appeared not only for indexations reasons but also because of incapacity of current Optical Character Recognition (OCR) systems to understand the contents of old documents. These techniques generally need a prior segmentation of text-lines into words. Extracting words from old Arabic manuscripts raises different challenges as the text-lines are very close together, the Arabic script is cursive and the writing styles are different. Crowded writing styles for example, muddle word boundaries as inter-words spaces become narrow and increase the overlap of components among adjacent words.

As illustrated in Figure 1, handled manuscripts are old, noisy and degraded. They consist of successive text-lines with different length and more or less fluctuating. Each text-line is a sequence of words which correspond to Arabic personal names. Note that Arabic names were historically based on a long naming system; a person's identity is a full chain of names separated by the separator word 'بن' (to mean son of; some say, it's a Hebrew origin, and the meaning is a short form of the word Benjamin). It seems that these manuscripts have just one writer since multiple instances of the same word are likely to look similar. The writer used line support and old Arabic script. In fact, some letters changed the shape. For example, the letter 'ي' can be flattened and oriented right to left (see the word 'علي' in Figure 1). Others letters have also changed the number and/or position of their diacritic points. For example, the letter 'ق' is written with a single diacritic point above the letter body instead of two points and the letter 'ف' is written with a single diacritic point below instead of over the letter body (see the words 'فرج' and 'فرحات' in Figure 1). Also, vertical and horizontal ligatures are easily introduced between parts of words and attachments occurred between the words of the same text-line or between those of successive text-lines. An example of connection zones is shown (see the third text-line in Figure 1).

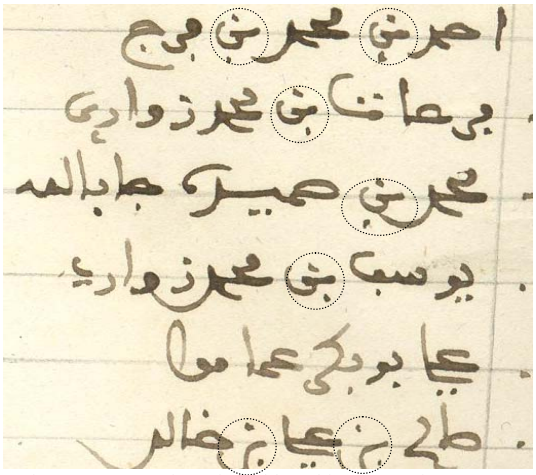

Figure 1. A portion of an old Arabic manuscript page .

Various word spotting and segmentation techniques were proposed in literature but the problem is still open. It seems necessary to address this issue from a different angle and involve mathematical concepts, including "transforms". Indeed, in a study of handwriting modeling done by [1], it is clear that the result of feature extraction could be regarded as a special form of segmentation or reduction of information. In both cases, there is a risk of losing information. We believe that the loss of information can be avoided while adopting a global approach, instead of local one, and using a mathematical transformation such as the Fourier conversion or Hough transform which are both reversible change of representation without loss of information. We are mainly interested in the Hough transform because we focused our research on the analysis and segmented writing rather than recognition or classification purposes for which other transforms are more appropriate.

The Hough Transform (HT), in its standard form, is often proved like a robust technique for the study of alignments in an image. Until then, its use is restricted to the detection of some directional primitives in Arabic [2], Latin scripts [3] or curves in pictures [4]. In its generalized form, the GHT has also been used for object Recognition [5] or for real-time object recognition [6]. The first dictionaries, proposed in [7], should include all models of PAWs. Recall that the Arabic word consists of one or more Pseudo-Arabic Words (PAWs). As noted before, our manuscripts are written by one author's hand which reduces the amount of handwriting variations that have to be compensated for.

The remainder of the paper is described as follows. In section 2, we will present a survey of techniques used for segmenting documents into words. Section 3 presents our approach to spot separator words, and using them to segment text-lines into independent words. This paper will be concluded with some obtained results and those expected in section 4 and a conclusion in section 5.

II. State of the Art

The Conventional segmentation methods can be classified into three approaches: top-down, bottom-up and hybrid approach. The top-down approach is guided by the model and recursively cuts document image into finer blocks. The bottom-up approach is a data-directed. It merges the finest entities of the document (pixels) till the complete assembly of the page. The hybrid approach combines the principle ideas of the two previous approaches. Among famous methods, based on these approaches, we can quote RLSA algorithm which is introduced by Wong [8], the RXYC algorithm, proposed by Nagy using the projection profiles [9], the method based on the analysis of blank spaces introduced by Pavlidis and Zhou [10], the Docstrum method [11], the Kise Method based on Voronoi diagrams [12], the Baird algorithm [13], etc. Several comparative studies have been led at this level [14], [15] and [16]. The segmentation methods based on texture are inspired from image processing field. They exploit a minimum of a priori knowledge since conventional methods seem to be inefficient to correctly segment ancient manuscripts. These texture based methods are used by several authors to segment old documents [17], manuscript or printed contemporary documents having complex and highly variable structure. These methods do not require binarization. Added to font recognition, they allow analyzing the document, identifying the type and differentiating the style of the text. There are also statistical approaches (Co-occurrence matrices proposed by Haralick in [18]), geometric approaches, approaches based on probabilistic models (Markov random fields, fractals) and order of frequency approaches (Fourier transforms, the Gabor transform, the wavelet transforms). In [19], the authors proceed by extracting connected components, and distances between different components are analyzed. The statistical distribution of this distance is then obtained to determine an optimal threshold for word segmentation. To determine such a threshold, Bayesian criteria of minimum classification error are employed. The projection based method is also employed for baseline detection. Thus words can be segmented by comparing this distance against a suitable threshold. In [20], the first step is pre-processing, next handwritten words are divided into smaller pieces. Then, these small pieces are segmented into candidate letters which are converted into their correspondence chain-code representation. Thereafter discrete, statistical and structural features are extracted for classification. The authors use a novel active contour based feature to increase the recognition accuracy of strongly deformed Arabic letters. They also use a decision tree to decrease the number of potential classes. A neural network is applied to compute weights for all statistical features and exploits them as an input for a k-NN classier.

III. Proposed System Overview

Our system pre-segments old Arabic manuscripts, conserved in the National Archives of Tunisia, into words using separator words such as 'بن'. The proposed system consists of three main steps: 1) separator-word spotting, 2) Segmenting touching components if necessary and 3) word segmentation. To locate separator words, we proposed a novel word spotting technique based on GHT. The system matches a given word query (here the separator word 'بن') with multiple word models, stored in a dictionary. Then, it retrieves the Hough parameters of the associate word model and uses them to spot the query word in the manuscript image as displayed in Figure 2(c). To spot the separator word 'بن' more efficiently, the system also uses Convex Point Distance (CPD), as it will be explained later, to remove confusions between the separator word 'بن' and others words or parts of words having similar morphology (case of a part of the word 'الحاج' in Figure 2(c)). When the separator word 'بن' touches or overlaps neighboring word components, the system works on separating the touching and overlapping components. To this end, the system uses a window, considering a rectangle around the voting center point and below the text-line at which a connection zone is observed. Notice that we only see below the text-line because in 'بن'

there are no ascenders. The used window is chosen large enough to include the lower part of the separator word 'بن' and its connections with the neighboring words. Then the skeleton is extracted, and the connection or junction points are detected and used to separate touching components of the separator word 'بن' from the neighboring words. For this separation, the system uses a method based on the angular variation considering the writing direction according to morphology of the Arabic terminal letter as used in [21]. More details are given in next subsections.

A. Word spotting by GHT

Our system spots Arabic words, in old manuscripts, focusing on their PAWs. For each PAW, the center of gravity O will be used as a reference point. Then, we calculate, for some contour points the triplets {b, a, d}. We group the triplets of model with the same value b and store them in an array of references. The rows are indexed by different values of b. Thus, a reference table is constructed. For more details, consult [22].

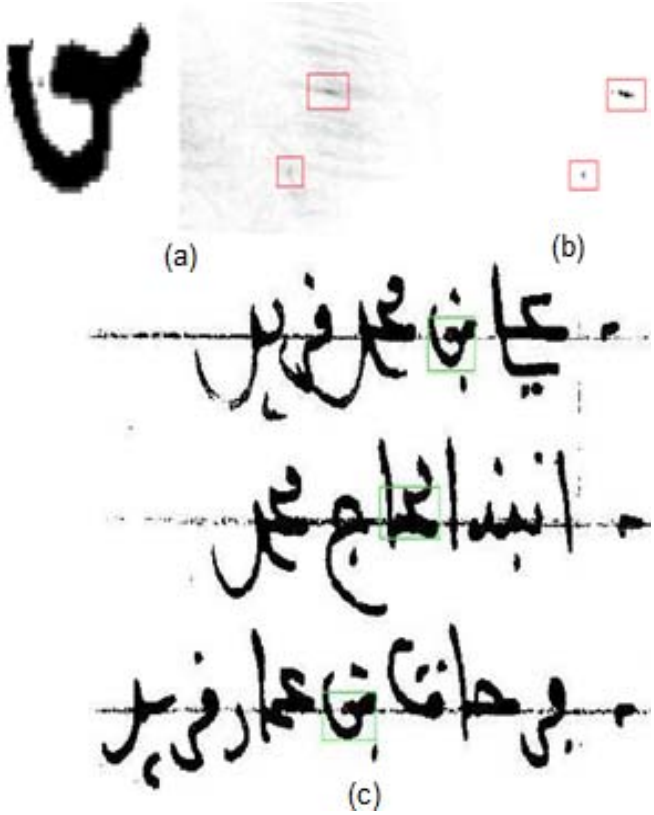

Figure 2. GHT and 'بن' spotting

We use the superposition of some sample models to construct only one model as shown in Figure 3 for the word 'بن', composed of one PAW.

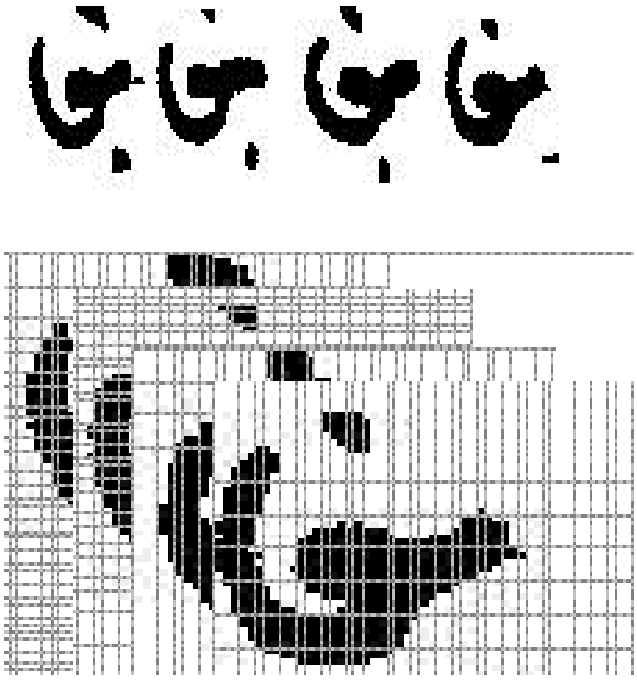

Figure 3. Superposition of several words 'بن'.

We also build an indexed and tagged dictionary using PAW reference tables. This dictionary collects information about the word queries. Generally, it includes for each word, its PAWs and for each of them we stored in the reference table its model and other parameters like Hough threshold (*s-Hough*: minimum number of points that vote for the PAW) and the distance (*d-cluster*: minimum distance between PAWs that compose the word). Notice that there is no need to use *d-cluster* to detect the separator word 'بن' since it is composed of only one PAW. Also note that threshold values are determined by empirical tests, in a training step. Now, having the 'بن' model from dictionary, the system routes the manuscript image and looks for the voting points. If the number of the voting points exceeds a fixed Hough- threshold (here s-Hough=14), a voting cluster will be formed and the query image word will be spotted. Figure 2(a) and (b) show how to find the separator word 'بن'. In red, groups of voter points are formed. Figure 2(c) shows the result of word spotting.

Note that some word parts can be confused with 'بن' like in 'الحاج' and 'علي' (see Figure 2(c)). This is due to the low value of Hough threshold (s-Hough) chosen to detect any resemblance with the separator word 'بن' and so all of its instances will be found. To solve such confusions, we used convex point's distance (CPD). In fact, these convex points seem to be useful to characterize the separator word 'بن' and distinguish it from other words computing the distance between convex points. For searching convex points, we used Sklansky algorithm [23]. Distances of convex points of the word 'بن' generally varies from 26 to

30 pixels unlike those of the words 'الحاج' and 'علي' which are around 65 pixels as shown in figure 4. Thus, by simple CPD comparison of confusing words, it has been possible to correctly identify the separator word 'بن'.

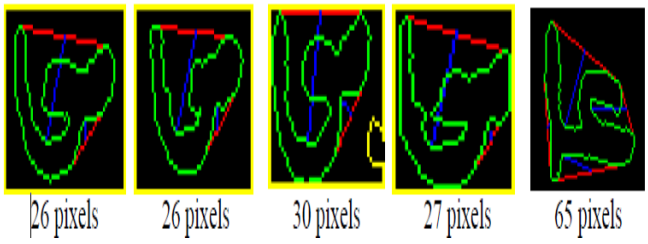

Figure 4. GHT Distinction of the separator word "بن" by CPD.

Figure 5 displays the results of 'بن' identification based on GHT and CPD. Note that for 330 occurrences of the separator word 'بن' in the used manuscripts, only 7 samples have been served for the training step. Our system has successfully spotted almost all occurrences of this separator word (97.8%) although some of them are differently written.

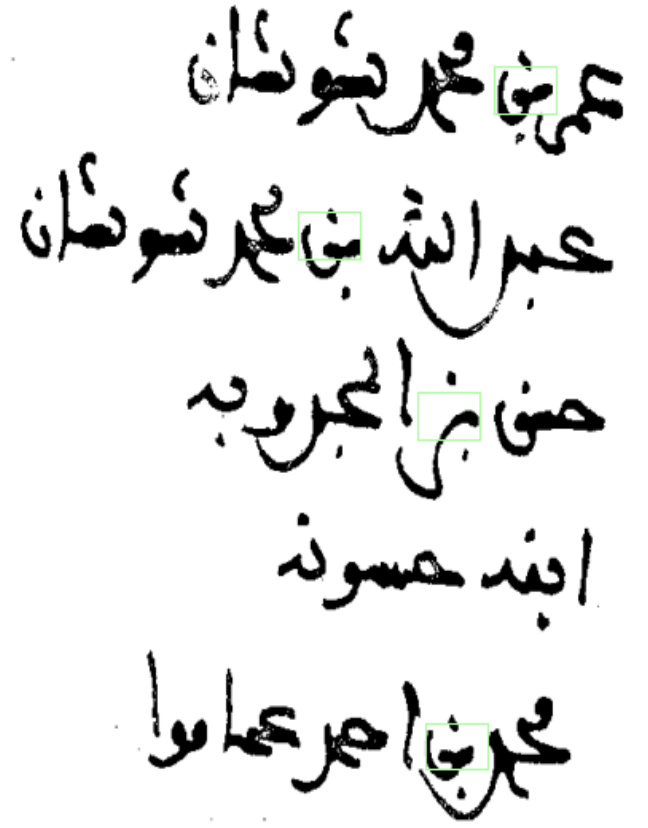

Figure 5. Some results of 'بن' spotting using GHT and CPD.

B. Segmenting touching components

Some spotted 'بن' can be attached to neighboring words or PAWs. In which case, they must be separated. For that, the baselines of the upper and the lower text-line must be extracted. Observing the handled manuscripts, we find that all connections are made below the baseline since the separator word 'بن' have a descending letter. A window covering an eventual connection zone is determined experimentally (64*64) pixels. To detect connections, the Zhang and Zuen's parallel skeleton algorithm [24] is applied on the connection zone. Because this zone is extracted from the original manuscript, it is cleaned by removing small connected components. To do this we keep only the connected component that contains the junction point Sp. To segment this connection zone (see Figure 6(a)), the junction point Sp must be detected. It is the point where all curves meet. At first, we turn the junction point to white, so we separate all the branches of the connected component. For this we apply a depth-first algorithm strategy [25]. We exploit the angular variation definition like in [21] as the angular dispersion of orientation angles along the curve. It is estimated using the formula of statistical variance where θ is the vector of angle variation of the curve and μ is the mean of θ.

$$\mathrm{var}(\theta) = \sum_{k=1}^{n} (\theta_i - \mu)^2 \qquad (1)$$

Variation of the vector angles θ of the curve is estimated using an iterative algorithm. This algorithm calculates the orientation angle θ between two consecutive pixels (in this case we choose p_i and p_{i+2}) using the following formula:

$$\theta = \left| \tan h^{-1}\left(\frac{d_{y_{i,i+2}}}{d_{x_{i,i+2}}}\right) \right| \qquad (2)$$

Then we calculate the angular variation of each branch detected (see Figure 6(b)), we sum the angular variation of each two different branches, find the minimum of the obtained values and finally label the two branches which result a minimum angular variation and the junction point with the same color. The other branches will be indicated with another color.

Connection zone is successfully segmented, as our approach succeeds to choose the pair of curves compatible with the ground truth which minimizes the sum of angular variation. Let's denote NCSS, the number of the successfully segmented connection zone, NC the total number of connection zones. The rate of well segmented connection which is defined as follows:

$$Rate\ of\ well\ segmented\ connection = \frac{NCSS}{NC} \qquad (3)$$

At this stage, there are two curves and three different types of pixels (the first type belongs to intersection, the second to the first curve and the last belongs to the second curve). Pixels of the original image have the same color of the nearest curve in the skeletonized image.

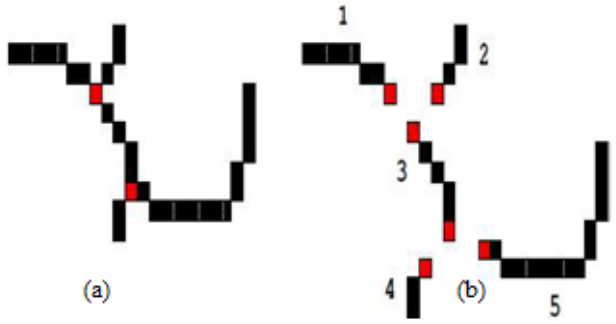

Figure 6. Touching component analysis.

C. *Word Segmentation*

Arabic words consist of a number of connected components or sub-words, and some researchers call these sub-words Pieces of Arabic Words (PAWs). In Arabic script there is no difference structure in the within word space as reported in [26], but according to our approach, once separator word 'بن' was spotted and fragmentation of connected component was correctly done, we can then segment text-line into word as reported in [27] for Arabic or in [28] for Latin script.

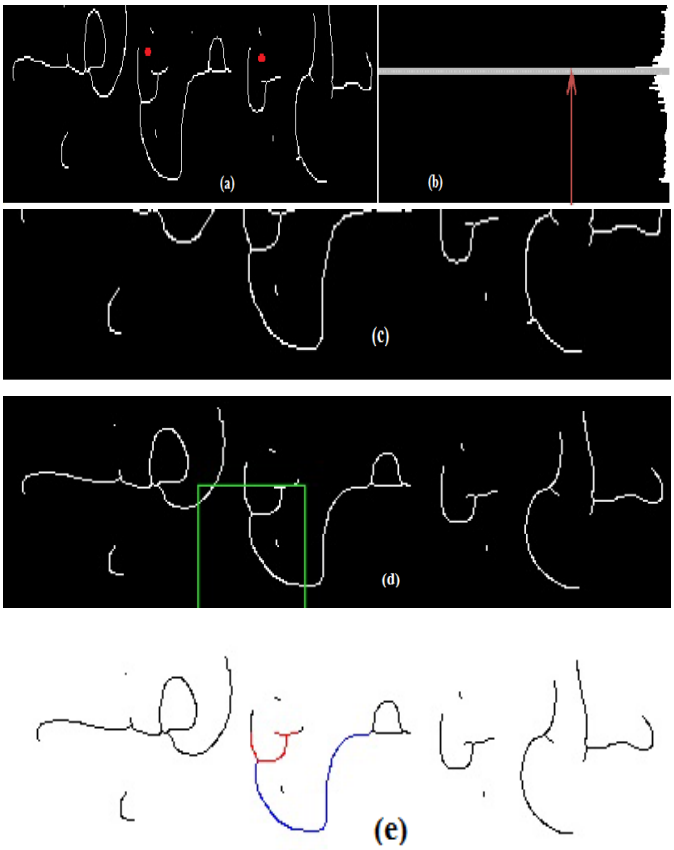

Figure 7. Touching component segmentation.

IV. Experimental Results

To evaluate the proposed prior segmentation system, we used a database of 23 manuscript images. Each image includes about 36 text-lines which are composed of 4 words on average. A total of 3312 words including 330 occurrences of the separator word 'بن' is obtained. Images are scanned with a 300 dpi resolution and binarized with Sauvaula's algorithm. Median filter and morphological operations were applied to remove noise. Notice that junction points are correctly located in 85% of cases and 75% of touching components are well segmented using the angular variation. More experiments should be carried on a larger database including Arabic personal names to assess the performance of our system. Note that our system shows failure in some cases especially when there are loops in the image which makes hard to find the main junction point or when there are diacritics and fluctuations of writing which lead to miss the baseline. But we believe that the proposed word spotting and pre-segmentation method are helpful especially when handling with these specific manuscripts. An indirect but reasonable rule is that this pre-segmentation is acceptable if it will be applied on another corpus belonging to other language, but this requires a particular separator word.

V. Conclusion

Handling with old Arabic manuscripts as we have, conventional techniques of segmentation have shown their limits. We proposed a new approach which aims to pre-segment text-lines into independent words based on the presence of separator words such as 'بن' which is commonly used to link Arabic Personal Names. Our system starts by locating all instances of the separator word 'بن' using a word spotting technique based on GHT. In case of confusions between 'بن' and others similar words or PAWs, it uses convex point distance which seems to be a helpful tool to discriminate between them. Afterwards, if separator words are connected to neighboring words, they must be segmented. So, the zone of word connection is extracted using the skeleton algorithm and baseline. Finally, to segment connection zone into two parts, so each part join its natural component, the system uses an approach based on morphology analysis of the terminal letter in old Arabic script. Note that this approach can be tuned to detect other separator word in Arabic [22], or in other languages. The proposed system performances seem very encouraging, leaving the field open to other perspectives. In fact, we believe that angular variation method can be better suited for connections between successive text-lines than between words on the same text-line. Using another method based on recognition, like in [29] might be more appropriate. In order to improve our system, we suggest looking for connections which can be occurred above the baseline. In that case, we suggest choosing the junction point which is the farthest node from the baseline. Note that it was possible to dislocate

separator word and apply only phases that come just after, but in this case, the risk is to split an entire word into two pieces or to lose some junction points which can affect recognition steps. We are aware that once a prior segmentation of the manuscript using the separator word 'بن' is correctly made, text-lines segmentation into words can be easily done.

REFERENCES

[1] Dargenton P., "Contribution à la Segmentation et à la Reconnaissance de l'Ecriture Manuscrite", Thèse de doctorat de l'institut national des sciences appliquées de Lyon, 1994.

[2] Fakir M., Sodeyaya C., "Recognition of Arabic Printed Scripts by Dynamic Programming Matching Method", IEICE trans. Inf. & syst., vol. E76-D n° 3, February 1993.

[3] Ruiz-Pinales J., Lecolinet E., "Cursive Handwriting using the Hough Transform and Neural Network", ICPR'00, p.231-234.

[4] Duda R. O., Hart P. E., "Use of the Hough Transformation to Detect Lines and Curves in Pictures", Comm. ACM, Vol. 15, p. 11 – 15, Janvier 1972.

[5] Robert S., Ivo I, Markus. M, "A Graphics Hardware Implementation of the Generalized Hough Transform for fast Image Processing and Computer Vision, vol. 27, no. 3, 1984.

[6] Ulrich M., Steger C., Baumgartner A., Ubner H., "3 Real Time Object Recognition Using a Modified Generalized Hough Transform", Eckhardt Seyfert, editor, Berlin, 2001, p. 571-578.

[7] Zarrouk S., « Reconnaissance Générale de PAW de l'arabe écrit", DEA, Ecole Nationale d'Ingénieurs de Tunis, Octobre 1998.

[8] Wong K.Y., Casey R.G., Wahl F.M., "Document Analysis System", IBM journal of Research Development, 26(6), p. 647-656, 1982.

[9] Nagy G., Seth S., "Hierarchical representation of optically scanned documents", ICPR, 1984, p. 347-349.

[10] Pavlidis T., Zhou J., " Page segmentation by white streams", ICDAR, Saint-Malo, vol.2,1991, p. 945-953.

[11] O'Gorman L., "The document spectrum for page layout analysis", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, 1993, p. 1162-1173.

[12] Kise K., Akinori S., Motoi I, "Segmentation of page images using the area Voronoi diagram", Computer Vision and Image Understanding archive,Vol.70 (3), Special issue on document image understanding and retrieval, 1998, p. 370-382. S. Scholes, Discuss. Faraday Soc. No. 50 (1970) 222.

[13] Baird H.S., Jones S.E., Fortune S.J., "Image Segmentation by shape-directed covers", ICDAR, 1991, Saint-Malo, vol.4, p.820-825.

[14] Cattoni R., Coianiz T., Messelodi S., Modena C. M., "Geometric Layout Analysis Techniques for Document Image Understanding: a Review", ITC-IRST, Via Sommarive, I-38050 Povo, Trento, Italy, January 1998.

[15] Nagy G., "Twenty Years of Document Image Analysis", in PAMI, Transactions on Pattern Analysis and machine Intelligence, Vol. 22, No 1, January 2000.

[16] Mao S., Rosenfeld A., Kanungo T., "Document Structure Analysis Algorithms: A Literature Survey", Proc. SPIE Electronic Imaging, Santa Clara, California, USA, January 2003, p.197-207.

[17] Journet N., "Analyse d'images de documents anciens : une approche texture", Thèse de doctorat, Université de La Rochelle, 2006.

[18] R.M. Haralick, K. Shanmungam, and I. Dinstein, "Textural Features of Image Classification," IEEE, vol. 3, pp. 610-621, 1973.

[19] J.H Alkhateeb, J.Jiang, J.Ren and S.S Ipson., "Component-based Segmentation of Words from Handwritten Arabic Tex", World Academy of Science, Engineering and Technology Vol:2 2008-05-28.

[20] L.Dinges, A. Al-Hamadi, M. Elzobi, Z. Al Aghbari1and H. Mustafa "Offline Automatic Segmentation based Recognition of Handwritten Arabic Words" International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 4, December, 2011.

[21] Ouwayed N., Belaïd A., "Separation of overlapping and Touching Lines within Handwritten Arabic documents". In CAIP (2009).

[22] Aouadi N. and Kacem A., "OCR-independent and segmentation-free word-spotting in handwritten Arabic Archive documents", ICEESA - 2013, IEEE xplorer, Print ISBN,978-1-4673-6302-0.

[23] Sklansky J., "Finding the Convex Hull of a Simple polygon", PRL 1 $number, p. 79-83, 1982.

[24] Zhang T. Y., Suen C. Y., "A Fast Parallel Algorithm for thinning Digital Patterns".

[25] Klette, R.,Rosenfeld A. 2004 . "Geometric Methods for Digital Picture Analysis", Print Book ISBN:9781558608610 Elsevier Store.

[26] M. Khayyat, L. Lam, C.Y. Suen Learning-based word spotting system for Arabic handwritten documents. Pattern Recognition Handwriting Recognition and other PR Applications. Volume 47, Issue 3, March 2014, Pages 1021–1030.

[27] G. R. Ball, S. N. Srihari, and H. Srinivasan. Segmentation-based and segmentation free methods for spotting handwritten arabic words. In Guy Lorette and Suvisoft, editors, Tenth International Workshop on Frontiers in Handwriting Recognition, October 2006.

[28] U.V. Marti, H. Bunke, "Text-Line Segmentation and Word Recognition in a System for General Writer Independent Handwriting Recognition", ICDAR 2001.

[29] Le Kang and D. Doermann, "Template based segmentation of touching components in handwritten text lines", ICDAR, 37, p.417–422, 2011.