General text line extraction approach based on locally orientation estimation

Nazih Ouwayed, Abdel Belaïd, François Auger

To cite this version:

Nazih Ouwayed, Abdel Belaïd, François Auger. General text line extraction approach based on locally orientation estimation. Document Recognition and Retrieval XVII - DRR 2010, 17th Document Recognition and Retrieval Conference, Jan 2010, San Jose, CA, United States. pp.1-10. hal-00488359

HAL Id: hal-00488359

https://hal.archives-ouvertes.fr/hal-00488359

Submitted on 1 Jun 2010

General Text Line Extraction Approach based on Locally Orientation Estimation

Nazih Ouwayed[a], Abdel Belaïd[a] and François Auger[b]

[a]LORIA-University Nancy 2, Campus Scientifique B.P. 239, 54506 Vandoeuvre-Ls-Nancy, France.

[b] University Nantes, IREENA, 44600 Saint-Nazaire, France.

ABSTRACT

This paper presents a novel approach for the multi-oriented text line extraction from historical handwritten Arabic documents. Because of the multi-orientation of lines and their dispersion in the page, we use an image paving allowing us to progressively and locally determine the lines. The paving is initialized with a small window and then its size is corrected by extension until enough lines and connected components were found. We use the Snake for line extraction. Once the paving is established, the orientation is determined using the Wigner-Ville distribution on the histogram projection profile. This local orientation is then enlarged to limit the orientation in the neighborhood. Afterwards, the text lines are extracted locally in each zone basing on the follow-up of the baselines and the proximity of connected components. Finally, the connected components that overlap and touch in adjacent lines are separated. The morphology analysis of the terminal letters of Arabic words is here considered. The proposed approach has been experimented on 100 documents reaching an accuracy of about 98.6%.

Keywords: Handwritten Arabic documents, text line segmentation, skew angle estimation, Snake, Wigner-Ville distribution, overlapping and touching lines.

1. INTRODUCTION

Ancient handwritten Arabic documents are for the most of them multi-oriented. In fact, in the meantime, the documents were required to be completed. Thus, annotations have been added. For problems of space, these additions have been accomplished in the margins creating multi-oriented text boxes with sinuous lines (see Figure 1).

In this context, classical approaches for text line segmentation cannot succeed. Already, all methods, global in nature and working at the page level, like the projection techniques,[1] the nearest neighbor clustering technique,[2] Hough transform,[3] etc. must be discarded. Besides, the consideration of Arabic document needs some specific precaution. In fact, the local orientation of isolated characters or PAWs (Part of Arabic Words) does not obligatory correspond to the main orientation of the global word. Some ligatures are vertical or oblique while the word is horizontal. This is why we oriented our research towards local techniques.

This paper is organized as follows: In Section 2, our approach for text lines extraction using a multiple steps (paving, oriented zone detection, text line extraction and separation of connected lines) is detailed. Experiments results are showed in Section 3 and last section concludes the paper.

2. SYSTEM OVERVIEW

The system operates following several steps. First, an automatic paving is operated dividing the page image in windows of the same size. Then, the orientation is estimated in each window using the energy distribution on the local projection histogram profile. Afterwards, the oriented zones (a zone is a set of neighbor windows with the same orientation) are detected. Finally, the lines are extracted in each zone using a method based on the follow-up of connected components. The connected adjacent lines are separated in a post-processing step using statistical information about terminal letters of Arabic words.

Further author information: (Send correspondence to A.A.A.)

A.A.A.: E-mail: nazih.ouwayed@loria.fr, abelaid@loria.fr, Telephone: +33 3 83 59 20 61

B.B.A.: E-mail: francois.auger@univ-nantes.fr

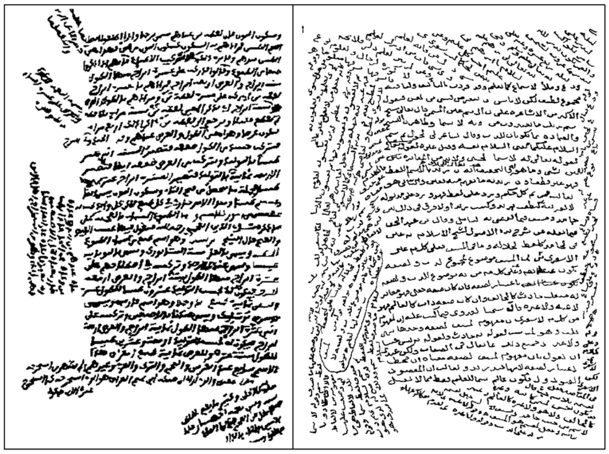

Figure 1. Examples of multi-oriented ancient Arabic documents.

2.1 Paving

In this step, the document image is partitioned in small windows of $(w \times h)$ size (see Figure 2.c). This size is automatically generated, based on the idea that a window must approximately contain 3 lines in each window, to be able to produce a histogram profile representative of the orientation. This follows several steps. First, an initial window of arbitrarily size (15 × 15 pixels) is placed in the middle of the image, supposed representing a horizontal alignment (see Figure 2.a). Then, the Snake approach is applied to determine the lines (see the explanation below). The window width is enlarged until the snake will give at least 3 lines. Once the lines found, the average line height $\bar{h}$ is performed as well as the average gap $\bar{g}$ between them. The gap distance is calculated using the convex hull-based metric[4] (see Figure 2.b). The final window size is equal to $(w \times h)$ where $w = h = 3 \times \bar{h} + 2 \times \bar{g}$.

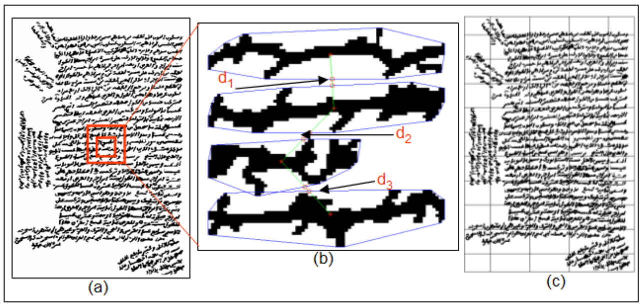

Figure 2. Automatic paving algorithm (d_1,d_2,d_3 are the gap distances).

Active Contour Model (Snake)

The active contour models is defined as an energy minimizing spline.[5] From an initial contour, a deformation of this contour is operated to make the most closest possible of the initial shape. The energy functional minimized is a weighted combination of internal and external forces. The internal forces stems from the shape of the snake, while the external forces emanates from the image.

The snake is parametrically defined as $v(s) = (x(s), y(s))$ where $x(s), y(s)$ are x, y co-ordinates along the contour and s is from [0,1]. The energy functional to be minimized is:

$$\begin{aligned} E^*_{snake} &= \int_0^1 E_{snake}(v(s))ds \\ &= \int_0^1 [E_{int}(v(s)) + E_{image}(v(s)) + E_{con}(v(s))]ds \end{aligned} \qquad (1)$$

where E_{int} is the internal spline energy that serves to regulate its shape, E_{image} is the external spline energy which attracts the snake to the contours sought in the image, and E_{con} which expresses some additional constraints that may be imposed by the user to the snake he wants to get.

The traditional external energy is located on the external contours. This forces to initialize the snake near the target contour. Moreover, the gradient values are opposite from the two sides of the same contour, which prevents the snake entering the concavities. For that reason, Xu et al.[6] developed a new kind of external energy that permits the snake to start far from the object, and forces it into boundary concavities. This energy is named gradient vector flow, or GVF. We used the GVF on the major axis of the connected components within each line of the window (see Figure 3).

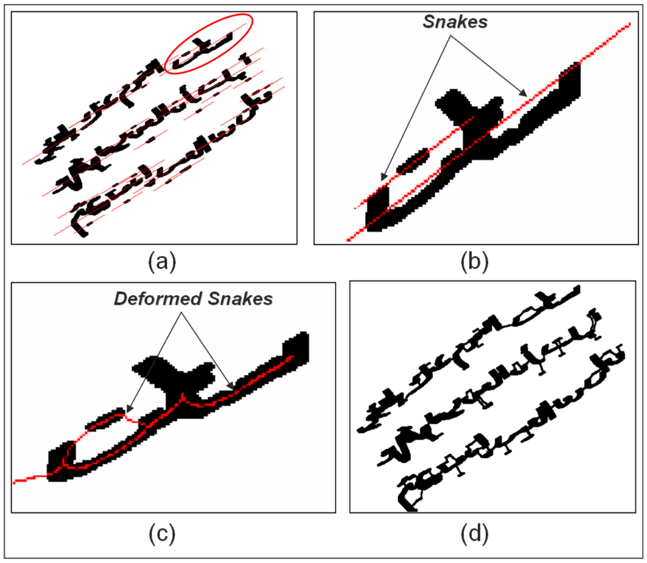

Figure 3. Application of the Snake for line detection, (a) Major axis drawn for each connected component of the lines. The ellipse encapsulates the initial connected component of (b), (c) shows the deformed snake of (b), (d) gives the final result showing the connected components gathered in each line.

2.2 Oriented zone detection

This step consists to extract first the local orientation in the windows and to extend them carefully to the neighborhood. These two steps runs as follows.

2.2.1 Orientation estimation in the windows

Usually, we perform the projection profile histogram to determine the local orientation in the window. However, the presence of noise in raw document images can create local maxima which disturb the projection histogram analysis. Hence, to avoid this drawback, we propose to use Cohen's class distributions[7] on the projection histogram profiles obtained using different projection angles. These distributions reduce the importance of these local maxima and are fitted to the non-stationary nature of these histogram profiles.

We focused on the Wigner-Ville distribution (WVD), whose properties can be more responsive to the presence of peaks that the other distributions of Cohen's class (see equation 2).

$$W_x(t,f) = \int_{-\infty}^{+\infty} x(t+\tau/2)x^*(t-\tau/2)e^{-j2\pi f\tau}d\tau \tag{2}$$

To estimate the orientation angle, we use the analytic signal $x_a(t)$ of the centered projection profile $x(t)$ of the document or of the selected area. The analytic signal is the signal $x(t)$ without its negative frequencies.

The profile $x(t)$ is determined by projecting each document with a chosen orientation. To calculate all possible profiles, we turn the image around its center of gravity (which gives us a point deduced from the image content and not from its size and framing) and we choose the horizontal axis as an arbitrary reference for the zero degree angle. Then, we compute a time-frequency representation for each projection profile.

The angle corresponding to the profile with the highest maximum value of its time-frequency representation is chosen as the estimated angle of the document. Figure 4 shows the highest values of the Wigner-Ville distribution (WVD) of the histogram profiles of the document shown on the the top-left level of the figure.

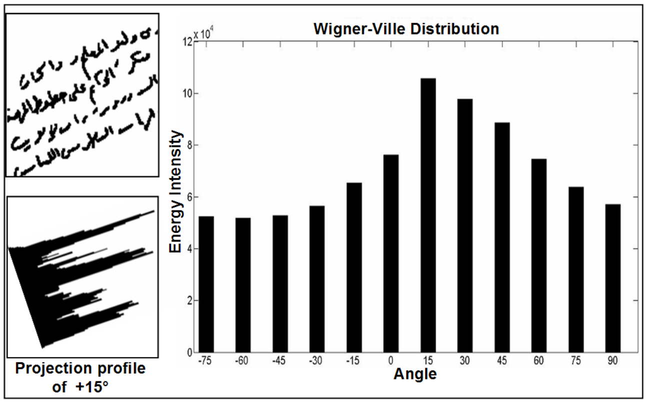

Figure 4. Energy intensity of the text window in the top-left of the Figure (true angle is +15 °).

2.2.2 Window correction and extension

Two cases may happen after the orientation calculation:

The first case occurs when the orientation is satisfactory. Hence, we try to extend the window area to its neighbors. For this, we perform the orientation of the window reassembled with each one of its neighbors. If the orientation of the total is equal to the orientation of one of them, we really merge them. In the opposite, we keep the first configuration (see figure 6.a).

The second case is related to non orientation satisfactory. In this case, we try to split the window in smaller ones allowing us to reach better precise orientation.

2.3 Text line extraction

The text line follow-up starts in the first window on the right side of the page. The algorithm starts by looking for the new maxima (see Figure 5.a). Each peak represents the starting point P_s of the baseline bl_j. The ending point P_e of the baseline is calculated using the P_s, the orientation, the width and the height of each window (see figure 5.b). The baseline bl_j is calculated basing on the two points (P_s, P_e) and the orientation of the window. The connected components that belong to a baseline are looked for construct the text line (see Figure 5.c).

A step of text line correction follows the text line detection to assign the non-detected components and the diacritical symbols to the appropriate text line (see Figure 5.c and d). A distance method is used to address this problem. First, the distance between the centroid of non-detected component or diacritical symbol C_i and the text line is calculated. C_i is assigned to the text line l_j if $d_{ci,lj} < d_{ci,lj+1}$ else to $l_j + 1$. For each zone (see figure 6.a), the text lines are clustered to form the zone text lines. Then the relations between the text lines zones are studied to form the document text lines (see figure 6.b). For this, we see if a connected component is up to two text lines between two zones. If it is the case, we merge the two text lines in one text line

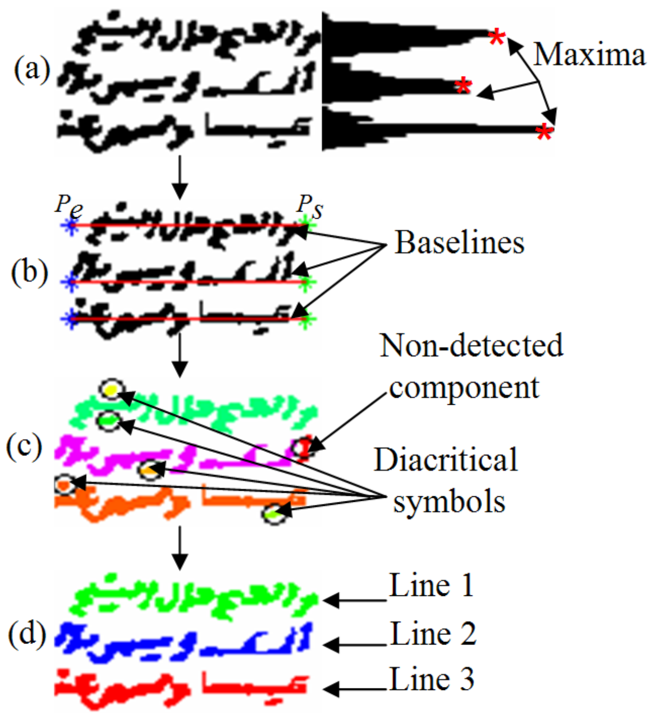

Figure 5. Text line detection steps for a window, (a) maxima detection, (b) baselines estimation, (c) assignment of each connected component and diacritical symbol to its appropriate line, (d) extracted lines.

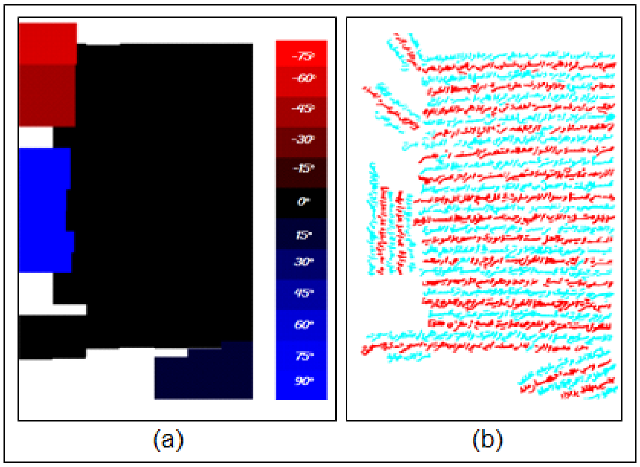

Figure 6. (a) Multi-skewed zones detection, (b) text lines extraction of document in Figure 2.a.

2.4 Separation of connected lines

In Arabic, the lines can be connected essentially at the level of terminal letters in the PAWs (Part of Arabic Word). Hence, we propose a method considering the morphology of these letters. To face this connection problem, the analysis will be focused on the connection zones (see figure 7).

The zones are determined considering a rectangle around the intersection point S_p of the two connected components which size is fixed manually. The starting ligature point B_p is the highest point in the zone close to the baseline. The descender direction is determined according to B_p relative to S_p. According to these characteristics, our idea consists to follow the skeleton pixels within the zone using the starting point B_p and the right descender direction. The follow-up will then cross the intersection point S_p and continues in the right direction that we have to determine. Figure 8 shows the final result of the separation algorithm (for more details see[8]).

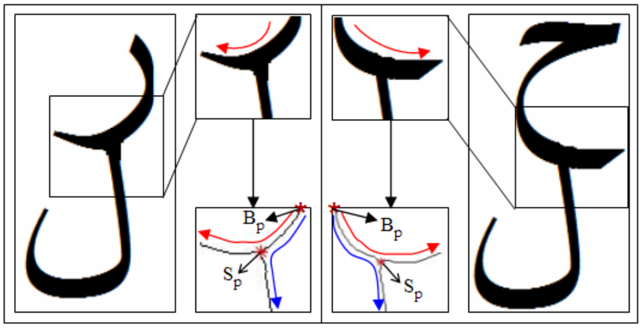

Figure 7. Writing direction according to the terminal letters morphology (true direction indicated by red arrow and the false by blue).

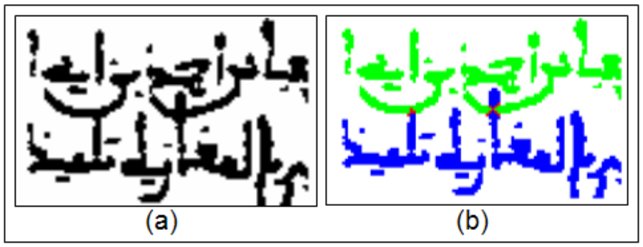

Figure 8. (b) Overlapping and touching connected components separation result of the window in (a).

3. EXPERIMENTAL RESULTS

To study the effectiveness of our approach, we experimented it on 100 handwritten Arabic documents that contain 2500 text lines. They are manuscripts belonging to the Tunisian National Library,[9] National Library of Medicine in the USA[10] and the National Library and Archives of Egypt.[11] The tests were prepared after a manual indexing of documents by zones and text lines using a truth files. The orientation experimented varies between −75 ° and +90 °. The tests were carried out on a PC equipped with a microprocessor Pentium M 1.4 GHz and 1 GB of memory under Windows XP. The application has been developed with MATLAB R2007b. In the multi-skew (zones) detection level, we have achieved a level of accuracy of 97%, this rate increases to 99% if we don't take care with the small zones that are not detected. The 1% error rate is due to the paving and the false inclination. In the text line segmentation level, the accuracy reaches 98.6%. The 0.5% of text lines not detected is due to the multi-skew detection algorithm. The 0.9% error rate is due to the presence of diacritical symbols in the beginning of the lines that create false maxima. Figure 9 illustrates the usability of our algorithm on a sample of 3 documents chosen arbitrarily from the 100 documents processed. To identify the lines, each consecutive pair of lines is presented by two different colors.

Figure	Document size	Resolution (dpi)	$w \times h$ of paving (pixels)			Execution time (s)			Zone number			
			a	b	c	a	b	c	True	Detected		
										a	b	c
a	390×632	96	56×35	75×75	50×50	57	34	37	5	5		
b	572×800	72	52×53	75×75	50×50	77	59	84				
c	750×820	120	125×75	90×90	50×50	52	47	141				
d	362×500	72	91×36	120×120	50×50	39	26	57				

Table 1. Results of the multi-skew estimation for the documents of the Figure 13.

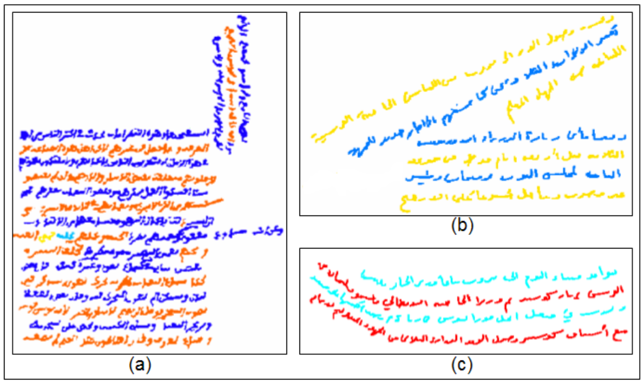

Figure 9. Some results of the multi-oriented text line extraction approach.

4. CONCLUSION

An original approach is proposed in this article, which aims to extract the text lines from the multi-oriented handwritten Arabic documents. First, the multi-skewed zones are detected using an automatic paving and the Wigner-Ville Distribution. Then, the text lines are extracted based on the orientation of each zone and the baselines. Finally, the connected adjacent lines are separated using an approach based on the morphology analysis of the terminal letters of Arabic words. The 98.6% of extraction rate show the efficacy and the performance of our approach. In the next step, we will try to generalize this approach to latin, urdu, persian scripts and to documents containing multiples scripts.

REFERENCES

[1] Zahour, A., Likforman-Sulem, L., Boussellaa, W., and Taconet, B., "Text line segmentation of historical arabic documents," in [*9th Int. Conf. on Document Analysis and Recognition*], 138–142 (2007).

[2] Likforman-Sulem, L. and Faure, C., "Extracting lines on handwritten documents by perceptual grouping, in advances in handwiting and drawing: multidisciplinary approach," *C. Faure, P. Keuss, G. Lorette, A. Winter (Eds)* , 21–38 (1994).

[3] Pu, Y. and Shi, Z., "A natural learning algorithm based on hough transform for text lines extraction in handwritten document," 637–646 (1998).

[4] Mahadevan, U. and Nagabushnam, R. C., "Gap metrics for word separation in handwritten lines," in [*ICDAR '95: Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1)*], 124 (1995).

[5] Kass, M., Witkin, A., and Terzopoulos, D., "Snakes: Active contour models," *Proc. 1st ICCV* , 259–268 (June 1987).

[6] Xu, C. and Prince, J. L., "Gradient vector flow: A new external force for snakes," *Proc. IEEE Conf. on Comp. Vis. Patt. Recog. (CVPR)* , 66–71 (June 1997).

[7] Cohen, L., "Generalized phase-space distribution functions," *J. Math. Phys.* **7**(5), 781–786 (1966).

[8] Ouwayed, N. and Belaïd, A., "Separation of overlapping and touching lines within handwritten arabic documents," in [*the 13th International Conference on Computer Analysis of Images and Patterns (CAIP'2009)*], to appear in september (2009).

[9] http://www.bibliotheque.nat.tn/.

[10] http://www.nlm.nih.gov/hmd/arabic/welcome.html.

[11] http://portal.unesco.org/ci/photos/showgallery.php/cat/559.