Separation of Overlapping and Touching Lines within Handwritten Arabic Documents

Nazih Ouwayed, Abdel Belaïd

▶ **To cite this version:**

Nazih Ouwayed, Abdel Belaïd. Separation of Overlapping and Touching Lines within Handwritten Arabic Documents. The 13th International Conference on Computer Analysis of Images and Patterns - CAIP 2009, Sep 2009, Munster, Germany. pp.237-244, 10.1007/978-3-642-03767-2_29 . inria-00435250

HAL Id: inria-00435250

https://inria.hal.science/inria-00435250

Submitted on 24 Nov 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Separation of Overlapping and Touching Lines within Handwritten Arabic Documents

Nazih Ouwayed and Abdel Belaïd

LORIA, University Nancy 2, Vandoeuvre-Lès-Nancy, France
{nazih.ouwayed,abelaid}@loria.fr
http://www.loria.fr/equipes/read/

Abstract. In this paper, we propose an approach for the separation of overlapping and touching lines within handwritten Arabic documents. Our approach is based on the morphology analysis of the terminal letters of Arabic words. Starting from 4 categories of possible endings, we use the angular variance to follow the connection and separate the endings. The proposed separation scheme has been evaluated on 100 documents contains 640 overlapping and touching occurrences reaching an accuracy of about 96.88%.

Key words: Handwriting line segmentation, Arabic documents, Overlapping and Touching lines, Calligraph morphology.

1 Introduction

The text line and word extraction from a handwritten document, is seen as a labored task. The difficulty rises from the characteristics of the handwritten documents especially when they are ancient. These documents present irregular spacing between the text lines. The lines can overlap or touch when their ascenders and descenders regions belong to each other (see figure 1). Furthermore, the lines can be skewed which constitutes new orientations.

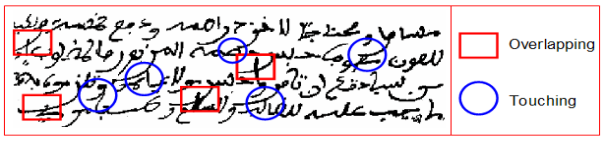


Fig. 1. Extract of a document representing some overlapping and touching lines.

In the literature, several methods have been proposed dealing with skewed lines [5–9]. Few methods have been proposed for the separation of connected words in the adjacent lines. From them, an Independent Component Analysis (ICA [4]) segmentation algorithm is proposed by Chen et al. in [1]. The ICA

converts the original connected words into a blind source matrix and calculates the weighted value matrix before the values are re-evaluated using a fast model. The readjusted weighted value matrix is applied to the blind source matrix in order to separate the word components. Louloudis et al. propose in [2] a skeleton-based algorithm. In [3], Takru and Leedham propose a technique that employs structural knowledge on handwriting styles where overlap is frequently observed. All above approaches are applicable to Latin scripts which are not easy adaptable to Arabic because of specific morphology.

In this paper, we propose a novel method considering the morphology of the terminal letters in the PAWs. The rest of the paper is organized as follows: overlapping and touching types in the Arabic document are listed in section 2. The Arabic script morphology is discussed in section 3. In section 4, our separation approach is detailed. Experiments results are showed in section 5 and last section concludes the paper.

2 Overlapping/Touching Types

The Arabic alphabet is composed of 28 letters. Among of them, 21 letters have an ascender (alif ا, Ta ط, DHa ظ, kaf ك, lam ل, heh ه), right descender (ra ر, zïn ز, sïn س, shïn ش, Sad ص, Dad ض, qaf ق, mïm م, nün ن, waw و, yeh ي) or left descender (jïm ج, ḥa ح, kha خ, ‘ain ع, ghain غ) causing the connecting lines (left or right indicates the descender starting, see figure 2.a). The connection in Arabic documents can happen in two cases: when the interlines spacing is small (see figure 2.b) or when we use a calligraph with big jambs (descenders) as Diwani (see figure 2.c).

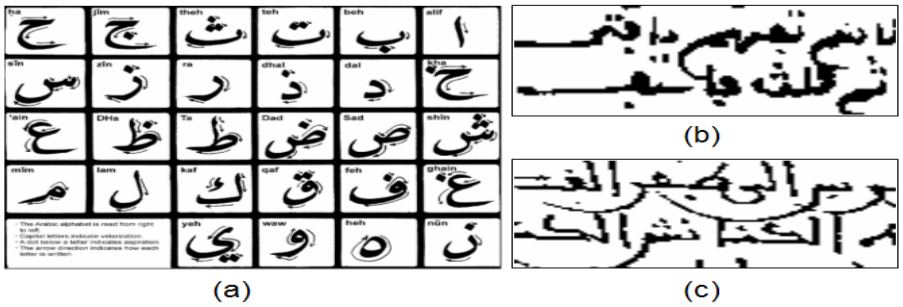


Fig. 2. (a) The Arabic alphabet chart (arrow indicates the writing direction, source: [14]), (b) Connection is due to the small interlines spacing, (c) Overlapping/touching due to the script type (Diwani).

Statistically, we have found 4 overlapping/touching types in handwritten Arabic documents. In the first type (see table 1.a), a right descender with a loop

overlaps/touches a vertical ascender. In the second type (see table 1.b), a left descender with a loop overlaps/touches a vertical ascender. In the third type (see table 1.c), a right descender overlaps/touches the curving top of lower letter. In the fourth type (see table 1.d), a left descender overlaps/touches the curving top of lower letter. In each type, the top letters stretches out to overlap/touch the bottom letters.

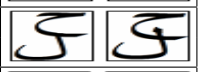
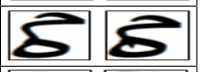
Types	Terminal Letters	Overlapping/ Touching Zones	Samples
a	Top: ر ,ز ,س ,ش ,ص ,ض ,ن ,ق ,و ,ي		
	Bottom: ا ,ط ,ظ ,ك ,ل		
b	Top: ج ,ح ,خ ,ع ,غ		
	Bottom: ا ,ط ,ظ ,ك ,ل		
c	Top: ر ,ز ,م ,و		
	Bottom: ص ,ض ,ه		
d	Top: ج ,ح ,خ ,ع ,غ		
	Bottom: ه		

Table 1. The 4 overlapping/touching types in the handwritten Arabic documents.

3 Arabic Morphology Analysis

In all cases of connection, we notice the presence of a descender connecting a lower terminal letter (see table 1, column overlapping/touching zones). These descenders are clustered in two categories : (a,c) when the descender comes from right and (b,d) when the descender comes from left. To face this connection problem, the analysis will be focused on the connection zones (see figure 3).

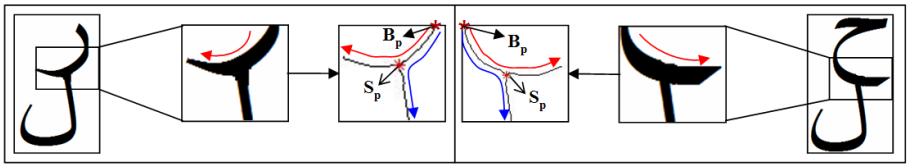


Fig. 3. Overlapping/touching connected components zones and descenders direction (rectangles represent the connection zones, right direction indicated by red arrow and the false by blue).

The zones are determined considering a rectangle around the intersection point S_p of the two connected components which size is fixed manually (see section 4.2). 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 (see section 4.2).

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.

3.1 Right Follow-up Direction

The determination of the right direction follow-up requires the study of curves in the skeleton image (i.e. each zone in the figure 3 has two curves).

In the literature, there are two main categories of curve detection methods [11]. In the first category, the Generalized Hough Transform (GHT) is used [12]. This technique is not efficient in our case because the present curves have few points and GHT needs much points for correct detection. In the second category, the chains (some of connected skeleton pixels) of points, or segments yielded by the polygonal approximation of such chains are used. The basic idea is to compute an estimation of the curvature for these chains [13]. This technique is insufficient because it does not study the continuity and the morphology of the curve.

The proposed method is based on the skeleton pixels follow-up and the angular variance. The follow-up starts from B_p and continues to the intersection point S_p. At this point S_p, the follow-up continues in multiple directions (see figure 4.a). The follow-up continues in each direction to extract the possible curves (C_{1+2}, C_{1+3} and C_{1+4}). The next step is to find the curve that represents the descender terminal letter. By experience, we found that the Arabic terminal letters have a minimum angular variance. This is explained by the fact that the terminal Arabic letters have the same orientation angle along the descender curve.

3.1.1 Angular Variance: The angular variance represents the dispersion of the orientation angles along the curve. It is estimated using the statistical variance formula:

$$Var(\Theta) = \sum_{i=1}^{n} (\theta_i - \mu)^2 \tag{1}$$

where Θ is the angles variation vector of the curve and μ is the average of Θ.

The angles variation vector Θ of the curve is estimated using an iterative algorithm that calculates the angle θ_i between two successive pixels p_i and p_{i+2} using the formula below (see figure 4.b):

$$\theta_i = \left| \mathrm{Arctan}\left(\frac{\mathrm{dy_{i,i+2}}}{\mathrm{dx_{i,i+2}}} \right) \right| \tag{2}$$

Because of the symmetric branches, the angle value must be always positive. For example, in figure 4, the angular variances are: $Var(C_{1+2}) = 703.19$, $Var(C_{1+3}) = 299$, $Var(C_{1+4}) = 572.37$. In this example, the minimum angular variance $Var(C_{1+3})$ is given by the right follow-up direction.

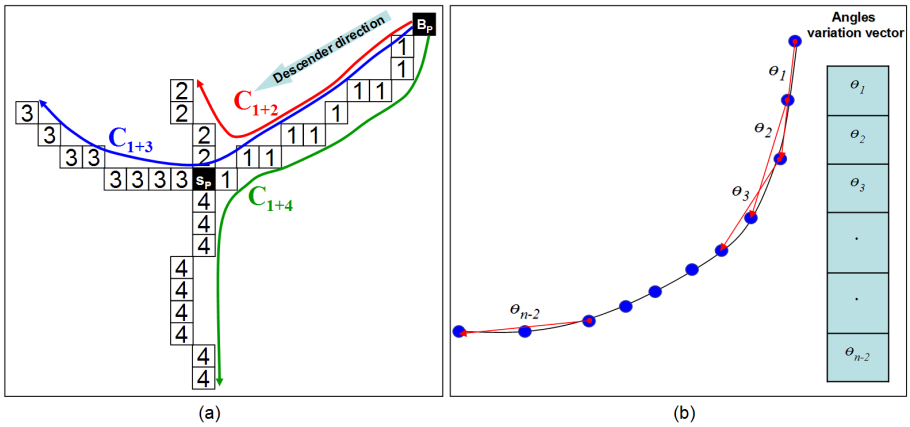


Fig. 4. (a) Example of Arabic overlapping connected components ("ra ر" overlaps "alif ا"), (b) Angles variation vector estimation algorithm.

4 Proposed Method

The method involves four steps as follows:

4.1 Step 1: Overlapping and Touching Connected Components Detection

The present paper is a continuation of our work published in [9]. The lines are extracted in [9] from the handwritten Arabic documents. Some adjacent lines can be connected in one or more connected components. This components belonging to two adjacent lines are considered as connected (see figure 5.a).

4.2 Step 2: Curve Detection

To detect the curves, the skeleton is first extracted using a thinning algorithm descried in [10]. Then, the intersection points of each connected component are detected (see figure 5.b). An intersection point is a pixel that has at least three neighbor pixels. As in Arabic script, the overlapping or touching may occur at just one intersection point S_p near the minima axis (valley between two connected lines in the projection histogram of the document, see figure 5.c). For this, S_p is the nearest point of the minima axis (see figure 5.d). Once the S_p is located, we look for the connected components zone. The center of this zone is S_p and its width (resp. height) is equal to $w_{ccx}/4$ (resp. $h_{ccx}/4$) where w_{ccx} (resp. h_{ccx}) is the width (resp. height) of the overlapping or touching connected component (4 is a determined experimentally). Since this zone is extracted from the initial document, it is cleaned by removing the small connected components.

To do it, the connected components are labeled and we keep only the connected component containing S_p (see figure 5.e). After the pre-processing, the skeleton of the zone is extracted and an skeleton follow-up algorithm is applied using the descender follow-up starting point B_p (the pixel that has the minimum y_i in the zone) and the direction follow-up (right to left if $x(B_p) > x(S_p)$ and left to right if $x(B_p) < x(S_p)$, see figure 5.f).

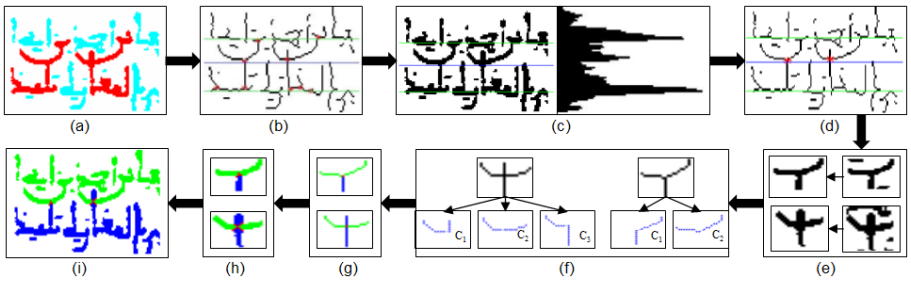


Fig. 5. Separation of overlapping and touching connected components approach steps.

4.3 Step 3: Curve Angular Variance Estimation

The angular variance of each curve is estimated using the algorithm detailed in the section 3.1.1. In the figure 5.f, the first touching components have $V(C_1) = 538.2099$ and $V(C_2) = 754.2284$. The C_1 having the minimum angular variance is the descender. The second overlapping components have $V(C_1) = 1160.8$, $V(C_2) = 438.4$ and $V(C_3) = 1208$. The C_2 having the minimum angular minimum variance is the descender.

4.4 Step 4: Pixels Assignment

In the figure 5.g, there are two curves and three different pixels types (intersection point "by red", first connected component "by green" and second connected component "by blue"). This step consists in assigning each black pixel in the initial image (see figure 5.e) to its appropriate curve. To do it, the image is scanned pixel by pixel and the 48-connected pixels of each image pixel p_i are regarded in the skeleton image. The closest branches pixel value is assigned to the initial image pixel (see figure 5.h). Finally, the pixels assigning is done at the initial document in order to obtain the final result (see figure 5.i).

5 Experimental Results and Discussion

The approach was applied to 100 handwritten Arabic documents belonging to the Tunisian National Library, National Library of Medicine in the USA, National

Library and Archives of Egypt that contain 640 overlapping and touching connected components. The tests were prepared after a manual indexing step of the overlapping and touching connected components of each document. Then, these components are clustered in 4 types (see section 2): 253 occurrences of type (a), 194 occurrences of (b), 117 occurrences of (c), 76 occurrences of (d) have been detected. The Table 2 describes the results for each type. The weighted mean of these results is equal to 96.88%. The 3.12% rate error is due to the intersection point detection algorithm because in some cases the overlapping/touching do not hold near the minima, and to the angular variance criterion because in some cases the minimum angular variance can occur for a false direction. Figure 6 illustrates the effectiveness of the algorithm on a sample of 12 representative connected components chosen from 640 occurrences.

Overlapping/ touching types	**Occurrences**	**Connections missed**	**Separations failed**	**Correctly separations rate %**
a	253	2	3	98.02%
b	194	4	2	96.90%
c	117	3	1	95.73%
d	76	1	2	94.75%

Table 2. Results of the separation of overlapping and touching connected components approach.

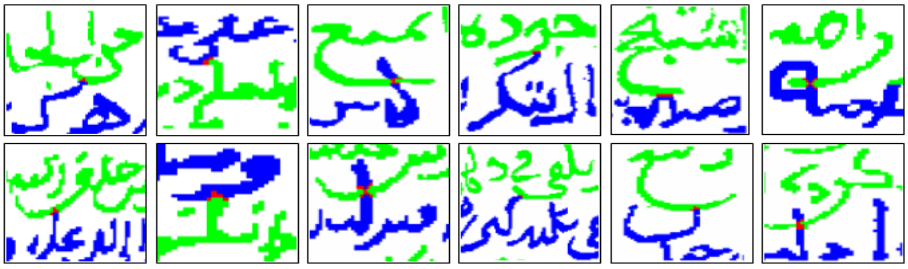


Fig. 6. Samples of our results.

6 Conclusion and Future Trends

An original method of separation overlapping and touching connected components in the adjacent text lines from the handwritten Arabic documents has been proposed in this paper. The proposed method is based on the Arabic calligraph

where overlapping and touching is most frequently observed. The approach is armed by statistical informations about Arabic writing structures. Experiments showed the efficiency and the performance of our approach. The future step of this work is related to the segmentation of the lines into single words.

References

1. Chen, Y., Leedham, G. : Independent Component Analysis Segmentation Algorithm. In: 8th International Conference on Document Analysis and Recognition, pp. 680–684, 2005
2. Louloudis, G., Gatos, B., Halatsis, C. : Line And Word Segmentation of Handwritten Documents. In: 11th International Conference on Frontiers in Handwriting Recognition, pp. 599–603, Canada 2008
3. Takru, K., Leedham, G. : Separation of touching and overlapping words in adjacent lines of handwritten text. In: International Workshop on Frontiers in Handwriting Recognition, pp. 496–501, 2002
4. Hyvarinen, A. : Survey on Independent Component Analysis. Helsinki University of Technology, Finland, 1999
5. Lüthy, F., Varga, T., Bunke H. : Using hidden Markov models as a tool for handwritten text line segmentation. In: 9th Int. Conf. on Document Analysis and Recognition, pp. 8–12, 2007
6. Zahour, A., Likforman-Sulem, L., Boussellaa, W., Taconet, B. : Text Line Segmentation of Historical Arabic Documents. In: Proceedings of the Ninth International Conference on Document Analysis and Recognition, pp. 138–142, Brazil 2007
7. Bukhari, S. S., Shafait, F., Breuel, T. M. : Segmentation of Curled Text Lines using Active Contours. In: Proceedings of Eight IAPR Workshop on Document Analysis Systems, pp. 270–277, 2008
8. Shi, Z., Govindaraju, V. : Line Separation for Complex Document Images Using Fuzzy Run length. In: Proc. Of the Int. Workshop on Document Image Analysis for Libraries, Palo, Alto, CA, 2004
9. Ouwayed, N., Belaïd, A. : Multi-oriented Text Line Extraction from Handwritten Arabic Documents. In: The Eighth IAPR International Workshop on Document Analysis Systems (DAS 2008), pp. 339–346, Japan 2008
10. Lam, L., Lee, S.-W., Suen, C.Y. : Thinning Methodologies-A Comprehensive Survey. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(9), pp. 869–885, September 1992
11. Dori, D., Liu, W. : Stepwise recovery of arc segmentation in complex line environments. In: International Journal on Document Analysis and Recognition, 1(1), pp. 62-71, February 1998
12. Ballard, D. H. : Generalizing the Hough Transform to detect arbitrary shapes. In: Pattern Recognition, 13(2), pp. 111–122, 1981
13. Rosin, P. L., West, G. A. :Segmentation of Edges into Lines and Arcs. In: Image and Vision Computing, 7(2), pp. 109-114, May 1989
14. Stanford University, USA, http://www.stanford.edu/