A Hybrid Approach for Line Segmentation in Handwritten Documents

Hande Adiguzel, Emre Sahin, Pınar Duygulu
Department of Computer Engineering
Bilkent University
{adiguzel,iesahin,duygulu}@cs.bilkent.edu.tr

Abstract

This paper presents an approach for text line segmentation which combines connected component based and projection based information to take advantage of aspects of both methods. The proposed system finds baselines of each connected component. Lines are detected by grouping baselines of connected components belonging to each line by projection information. Components are assigned to lines according to different distance metrics with respect to their size. This study is one of the rare studies that apply line segmentation to Ottoman documents. Further, it proposes a new method, Fourier curve fitting, to detect the peaks in a projection profile. The algorithm is demonstrated on different printed and handwritten Ottoman datasets. Results show that the method manages to segment lines both from printed and handwritten documents under different writing conditions at least with 92% accuracy.

1. Introduction

Large archives of historical documents attract many researchers from all around the world. The increasing demand to access those archives makes automatic retrieval and recognition of these documents crucial.

Ottoman archives are one of the largest collections of historical documents; they include more than 150 million documents ranging from military reports to economic and political correspondences belonging to the Ottoman era [1]. Many researchers from all around the world are interested in accessing the archived material [2]. Unfortunately, many documents are in poor condition due to age or are recorded in manuscript format.

Line segmentation is usually a crucial preprocessing step in most of the document analysis systems. Although text line segmentation is a long standing problem, it is still challenging for hand-written degraded documents. The problems of handwritten texts can be categorized into 2 parts: (i) line-based problems such as, variance of interline distances, inconsistent baseline skews and multi-oriented text lines; and (ii) character-based problems such as, broken characters due to degradation, small-sized diacritical components and variance of character size.

Even though there are many advanced methods designed for complex datasets [3-7], the studies in text line segmentation are dominated by projection profile and connected component based approaches.

Projection profile based methods are usually successful on machine printed documents [8] nevertheless; they can be extended to deal with slightly curved text lines [7]. Besides, projection profile based methods are easy to implement and fast thru the basic intuition of straightness of text lines.

Connected component based methods are appropriate for more complex documents where interline distances and baseline skews change. However, most of the connected component based methods work directly on the input image where each pixel is treated equally and a change of one pixel may result in a different result [7].

In this study, we use a hybrid approach for line segmentation which combines both connected components and projection information. Rather than obtaining the projection profile directly from the input image or straightly using connected components for line detection, baselines of connected components are extracted and passed to second phase where projection profiles are used. This process also allows some skew tolerance.

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.156

The contributions of this paper are threefold. First we address the line segmentation problem for historical Ottoman documents which are rarely studied. We intend to apply our method to considerably large historical datasets with multiple authors from various time periods. Therefore, our intuition was to use simple and fast line segmentation methods rather than non-trivial, complex ones. To achieve this, bottlenecks of projection profile based and connected component based methods are prevented through a hybrid approach. Additionally, a new approach, Fourier curve fitting is suggested for determining the peaks and valleys in projection profile analysis which is still considered as a problematic issue [7].

The rest of the paper is organized as follows. Section 2 overviews the related work on text line segmentation. In Section 3, we describe the proposed approach. Then, we explain our datasets and evaluation results. Finally, we give our conclusion and future work.

2. Related Work

Text line segmentation algorithms can be mainly categorized as projection profile based [8, 9] and connected component based [10, 11]. Projection based methods makes the assumption of text lines being parallel and straight thus, they are effective for machine printed documents.

Further, for handwritten documents where interline gaps are small or lines have considerably high skew, piece-wise projection approaches are used [9]. In these approaches documents are divided into vertical strips and vertical projection profiles of strips are combined to obtain the results.

Connected component based methods [10, 11] extracts geometrical information such as shape, orientation, position and size from connected components to group or merge them into lines. They are more appropriate for complex documents than the profile based methods. However, they are sensitive to small changes in connected component structures.

There are few studies [1, 12] that apply line segmentation on Ottoman datasets. In [1] it is assumed that baselines will have more number of black pixels than the other rows. With this intuition projection profiles of the documents are analyzed and peaks of the profile are detected according to some predefined threshold. Due to inconsistent baseline skews, multi-oriented text lines and small interline gaps observed in our dataset, directly applying projection profile method will not succeed. Further, different threshold values need to be set for different types of writing styles or writers.

Another study that demonstrate their results on Ottoman documents, constructs a Repulsive-Attractive Network for line segmentation [12]. In this network, attractive and repulsive forces are defined and baseline units' y-coordinates are iteratively changed according to these forces until local convergence is obtained. Nevertheless, the lines must have similar lengths and each baseline is detected according to previously examined one where a detection error can trigger other ones.

3. Methodology

The details of the approach are given in the following sections. First, preprocessing steps are explained. Then, baseline extraction is given and finally, the procedure of assigning each connected component to lines is described.

3.1. Binarization

Binarization is one of the important preprocessing steps of segmentation. Global binarization methods use a single threshold value to classify pixels into foreground or background classes. However, they do not always yield to satisfactory results especially on historical documents that are degraded, deformed and not in good quality due to faded ink and stained paper and may be noisy because of deterioration.

After the original documents are converted into gray scale, adaptive binarization method [13], which calculates multiple threshold values according to the local areas, is used for binarization. Then, small noises such as dots and other blobs are cleaned by removing connected components which are smaller than a predefined threshold.

3.2. Connecting Broken Characters

To connect broken characters, first Manhattan distance between adjacent foreground pixels are calculated then pixels are connected if the measured distance is smaller than a predefined threshold.

3.3. Page Segmentation

The documents in our datasets are scanned in 2-page format. Therefore, before line segmentation the documents must be segmented into pages. The horizontal projection profile of each document is calculated and then the two largest peaks of the profile are observed for segmenting the two pages. To detect the widest peaks, a Fourier curve [14] is fitted to the horizontal projection profile and then the image is cropped according to the smallest value of the profile that lie between the two peaks.

3.4. Small-Sized Connected Components

Ottoman language has some common properties with Arabic; most notably the alphabet and the writing style which relies on dots and diacritics heavily. However, these small-sized components may produce ambiguous results for line segmentation since they usually lie between the text lines. In [15, 16] it is mentioned that diacritical points can generate false separating or extra lines.

Some line segmentation studies applied on languages that include diacritical symbols [7, 16] does not filter these small connected components during line segmentation and then apply a post processing step for correcting the approximate results. On the other hand, some studies [3, 17] eliminate those small-sized components during segmentation and reconsider them to generate the final line segmentation results.

We propose a method that ignores small-sized components during line segmentation to obtain results more accurately without post processing. After we detect all connected components, the small ones are marked so that they will not be used during detection of the lines.

To find the small-sized components, each connected component's filled area is calculated and then components which have a smaller filled area than a predefined threshold are marked as small. After the lines are detected, small-sized components are reconsidered and assigned to related lines. Figure 1.b shows the document image without small-sized components. As it can be observed, the text line structure is enhanced.

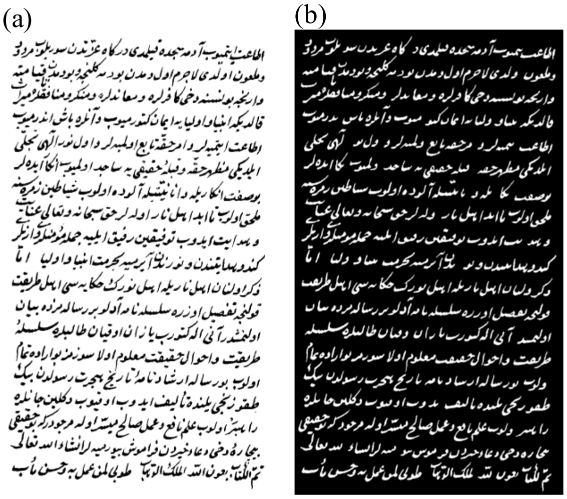

Figure 1. (a) Binarized document image, (b) image constructed without the small-sized components.

3.5. Baseline Extraction

Baseline is the fictitious line which follows and joins the lower part of the character bodies in a text line [18]. Thus, each connected component has baseline pixels that fit or come close on its baseline.

In this study, for baseline extraction first each connected component's baseline pixels are found approximately. To find those pixels, contour image of the connected component is obtained (Figure 2.b). Then, left-to-right (Figure 2.c) and right-to-left (Figure 2.d) gradient image, measuring the horizontal change in both left and right directions are calculated.
Also, the bottom-to-top (Figure 2.e) gradient image which shows the vertical change in the upward direction is obtained. Then these 3 gradient images are subtracted from the contour image which results in the group of pixels that approximately lie on the baseline (Figure 2.f).

To obtain the exact baseline pixels, first the y-coordinates' standard deviation (s) and mean (m) values are calculated. Then the pixels where |p-m| > s are considered as outliers and removed from the group. The rest of the pixels are used as baseline pixels (Figure 2.g).

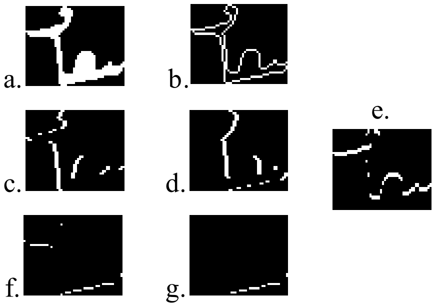

Figure 2. Procedure of extracting baseline pixels. (a) connected component, (b) contour image, (c) left-to-right, (d) right-to-left, (e) bottom-to-top gradient images, (f) approximate baseline pixels, (g) exact baseline pixels.

This procedure is applied for each connected component and a new image is reconstructed from these obtained baseline pixels which can be seen from Figure 3. Then to detect the baselines of each line, vertical projection profile of the reconstructed image consisting of baseline pixels is obtained (Figure 3). The peaks of this profile can be interpreted as lines and the valleys as interline gaps.

To detect the peaks, a Fourier curve [14] is fitted to the profile and local maxima points are found. Fourier curve can capture the repetitive pattern of lines.

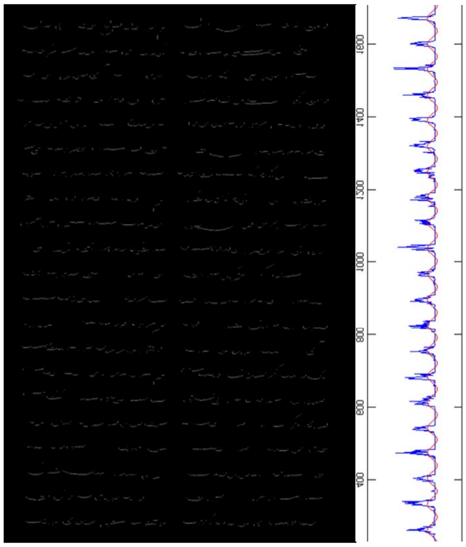

Figure 3. Right part shows the image reconstructed from the baseline pixels of connected components, left part is the vertical projection profile of the reconstructed image. A Fourier curve is fitted to the projection.

Then, for each two adjacent peaks of the curve, the smallest value in the profile that lie between these peaks, which is usually zero, is obtained as a cut point. Thus, for each gap a cut point is calculated respectively. These points can be considered as an approximate of the interline gaps and are used for separating the baseline pixels that belong to different adjacent lines (Figure 4).

After obtaining the baseline pixels that belong to each line (Figure 4.a), polynomial curves are fitted to each group of those pixels to calculate the actual baselines (Figure 4.b). Line fitting can also be used however; we preferred to use a 4th degree polynomial to tolerate some amount of curvature.

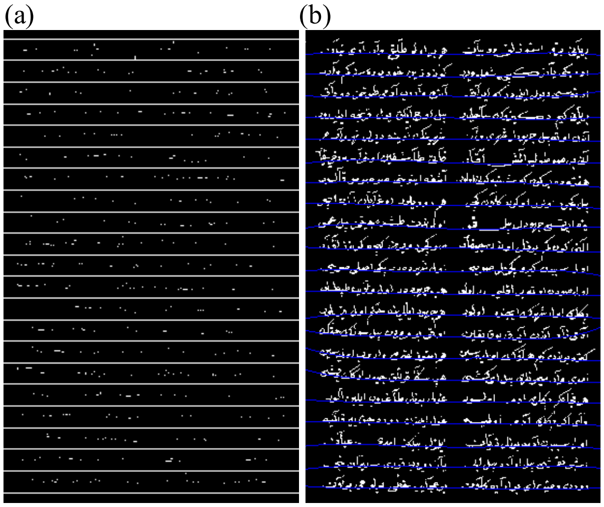

Figure 4. (a) Approximate of interline gaps, (b) computed baselines.

3.6. Assigning Connected Components to Lines

After the baseline curves are extracted, the connected components which are not marked as small are assigned to their closest curves. To find the closest curve of a component, the distance function is obtained from the curve's equation and the component's midpoint. Then, the derivative of the distance function is computed to find the closest distance between the midpoint and the curve. Finally, the component is assigned to the curve which has the minimum distance to its midpoint.

To finalize the results, removed diacritical components are assigned to lines. First each small-sized connected component's nearest neighbors in 4 directions (right, left, up, down) are found. The nearest neighboring components should not be small-sized thus, must be assigned to some line.

The 4 nearest neighbors' assigned lines are voted accordingly to their distances to the small-sized component. To, illustrate if the nearest neighboring component in some particular direction is closest to the small-sized component, its line id gets the highest vote. With this voting scheme each small-sized connected component has at most 4 different line candidates with their votes calculated according to the distances. Finally, each small-sized component is assigned to the line which has the highest vote.

4. Experiments

To evaluate our results, we constructed 3 different Ottoman datasets. First one and the second one consist of the text pages from 2 different books. First book is a printed one and includes 120 pages while the second one is a handwritten one and includes 50 pages. These 2 datasets are constructed to compare the algorithm's performance for handwritten and printed texts (Table1).

The third dataset includes 240 pages taken from 6 different books, 20 pages from each and the documents are all handwritten. This dataset is constructed to evaluate the performance under different writing styles and writers. Table 1 shows the number of lines and connected components in the ground truth and the results for each dataset.

We evaluated our results using a pixel-based matching score (MS) [19] which is calculated as follows:

$$\mathrm{MS}(r_i, g_j) = \frac{T(P(r_i) \cap P(g_j))}{T(P(r_i) \cup P(g_j))} \qquad (1)$$

where $\mathrm{MS}(r_i, g_j)$ is a real number between 0 and 1 and it is the matching score between the result zone r_i and the ground truth zone g_j. P represents the foreground pixels and T is an operator that counts the number of pixels in the zone.

Table 1. Results obtained with different printed and handwritten datasets with different writing styles and writers.

	#of Lines in GT	Detected Lines	#of CCs	Correct Detected CCs
Book1	923	921	5208	4847
Book2	880	879	4634	4315
Book3	871	869	4489	4154
Book4	795	793	4132	3884
Book5	764	763	3795	3375
Book6	836	834	4208	3824
Printed	3210	3210	16157	15510
Handwritten	1068	1041	5573	5245

The matching-scores between all the result zones and the ground-truth zones are obtained. If, the matching score is above a predefined threshold then the result zone is counted as a True positive (TP). Result zones which are not matched to any ground truth zones are False positives (FP) and the ground truth zones which are not matched to any result zones are False-negatives (FN). Precision, recall and the F1-Score is calculated as described in these studies [19].

As it can be observed from Table 2 and 3 the line segmentation results for both printed and handwritten datasets are considerably high at MS thresholds 95% and 90%. The F1-Score for the 6 different handwritten books is nearly 93% and 94% for MS threshold 95% and 90% respectively. The segmentation accuracy is nearly same for different books. Thus, it can be concluded that, the change in the writing styles or writers which means different interline spacing, character sizes and line skews does not have an impact on segmentation results.

Also, as it was expected Table 3 shows that the segmentation accuracy increases for printed texts. However, there is not much difference between them and the results of handwritten documents which means the algorithm is successful for segmenting both printed and handwritten documents.

Also, we observed that most of the errors are due to inadequate binarization, noisy components such as page numbers or ornamentation and assigning small-sized connected components to wrong lines. Binarization errors are due to dark stains that cannot be separated from the ink pixels and different ink colors used in the document. The noisy components can be detected as a separate process or removed manually before segmentation. Moreover, most of the wrong assigned small components are very hard to classify without language dependent metrics.

We have implemented our algorithm in C++. The average time taken for processing a single image with 2000*1500 pixels is 3.1 seconds on a 2.67GHz CPU and 4GB RAM.

Table 2. Results obtained on 6 different handwritten books with MS threshold of 95% and 90%.

0.95	Precision	Recall	F1 - Score
Book1	96.65%	90.67%	93.57%
Book2	97.03%	91.01%	93.92%
Book3	95.52%	90.30%	92.83%
Book4	97.10%	90.24%	93.54%
Book5	95.68%	89.37%	92.42%
Book6	95.44%	89.73%	92.50%

0.90	Precision	Recall	F1 - Score
Book1	98.93	90.85	94.72
Book2	99.32	91.20	95.09
Book3	98.75	91.33	94.90
Book4	99.26	90.85	94.87
Book5	97.92	90.57	94.10
Book6	98.58	90.69	94.47

Table 3. Results obtained on 2 different printed and handwritten datasets with MS threshold of 95% and 90%.

0.95	Precision	Recall	F1 - Score
Printed	98.64	94.56	96.56
Handwritten	97.75	92.29	94.94

0.90	Precision	Recall	F1 - Score
Printed	99.56	95.03	97.24
Handwritten	99.33	92.37	95.72

5. Conclusion and Future Work

In this study, we have presented a hybrid approach for line segmentation based on both connected components and vertical projection profile. Projection profile based methods are simple, easy to implement and can deal with a certain amount of curve. Besides, connected component based approaches are successful for more complicated documents whose interline distances vary or baseline skews are inconsistent. Thus, by extracting baseline pixels from connected components and then using the projection profile information we managed to segment the lines from both handwritten and printed documents. The effectiveness of the algorithm is demonstrated on different datasets and it is shown that the algorithm is successful for different kind of writing styles and writers.

Although the proposed method was designed for Ottoman documents, it can adapt well to other scripts since it does not use any language dependent information. Besides, other scripts like English, German, French and Greek which uses Latin alphabet do not include as much small-sized components and diacritics as Ottoman language. This feature can bring an improvement to the segmentation results since some of the errors were caused by assigning small-sized components to wrong lines.

As a future work, we are planning to apply our algorithm on much larger Ottoman datasets and to other scripts. At the same time, we believe that using more advanced techniques for preprocessing steps such as binarization, diacritics detection and noisy component removal will improve the segmentation results.

Acknowledgements

This study is partially supported by TUBITAK grant no 109E006.

References

[1] E. Ataer, P. Duygulu, “Retrieval of Ottoman documents”, Proceedings of the 8th ACM international workshop on Multimedia information retrieval, October 26-27, 2006.

[2] Ottoman Text Archive Project (OTAP), http://courses.washington.edu/otap/

[3] Saabni, R.; El-Sana, J.;, "Language-Independent Text Lines Extraction Using Seam Carving," International Conference on Document Analysis and Recognition (ICDAR), 2011.

[4] X. Du, W. Pan, T. D. Bui, “Text line segmentation in handwritten documents using Mumford–Shah model”, Pattern Recognition, Volume 42, Issue 12, December 2009.

[5] A. Alaei, U. Pal, P. Nagabhushan, “A new scheme for unconstrained handwritten text-line segmentation”, Pattern Recognition, Volume 44, Issue 4, April 2011.

[6] Bukhari, S. S., Shafait, F., and Breuel, T. M. “Script independent handwritten textlines segmentation using active contours”. In Proceedings 10th International Conference on Document Analysis and Recognition, 2009.

[7] Yi Li; Yefeng Zheng; Doermann, D.; Jaeger, S.; Yi Li; "Script-Independent Text Line Segmentation in Freestyle Handwritten Documents," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.30, no.8, pp.1313-1329, Aug. 2008.

[8] Y. Lu, “Machine printed character segmentation: an overview,” Pattern Recognition 28, 67-80, 1995.

[9] N. Tripathy and U. Pal, “Handwriting Segmentation of Unconstrained Oriya Text,” Proc. Ninth Int’l Workshop Frontiers in Handwriting Recognition, pp. 306-311, 2004.

[10] S. Jaeger, G. Zhu, D. Doermann, K. Chen, and S. Sampat, “DOCLIB: A Software Library for Document Processing,” Proc. SPIE Document Recognition and Retrieval XIII, pp. 63-71, 2006.

[11] A. Simon, J.-C. Pret, and A. Johnson, “A Fast Algorithm for Bottom-Up Document Layout Analysis,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 3, pp. 273-277, Mar. 1997.

[12] E. Öztop, A.Y. Mülayim, V. Atalay, F. Yarman-Vural, “Repulsive attractive network for baseline extraction on document images”, Signal Processing, Volume 75, Issue 1, 5 January 1999.

[13] Rosenfeld, A., Kak, A. C.: “Digital Picture Processing”, 2nd edition, Academic Press, New York, 1982.

[14] Bochner S., Chandrasekharan K., “Fourier Transforms”, Princeton Univ. Press, Princeton, 1949.

[15] Zahour, A.; Likforman-Sulem, L.; Boussalaa, W.; Taconet, B.;, "Text Line Segmentation of Historical Arabic Documents," Ninth International Conference on Document Analysis and Recognition, 2007.

[16] Zahour, B. Taconet, L. Likforman-Sulem and Wafa Boussellaa, "Overlapping and multi-touching text- line segmentation by Block Covering analysis", Pattern Analysis and Applications, Vol. 12, pp. 335-351, 2008.

[17] N. Ouwayed and A. Belad. “Multi-oriented text line extraction from handwritten Arabic documents”. In 8th IAPR Int. Workshop on Document Analysis Systems, pages 339-346, 2008.

[18] Likforman-Sulem L., Zahour A., Taconet B.: “Text line segmentation of historical documents: a survey.” Int. J. Doc. Anal. Recogn. 9(2), 123–138, 2007.

[19] Jayant Kumar, Wael Abd-Almageed, Le Kang, and David Doermann. “Handwritten Arabic text line segmentation using affinity propagation”. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems , New York, 135-142. 2010.