

 Open access • Proceedings Article • DOI:10.1145/2037342.2037362

Text line segmentation for gray scale historical document images — [Source link](#)

Abdelkadir Asi, Raid Saabni, Jihad El-Sana

Institutions: Ben-Gurion University of the Negev, Tel Aviv University

Published on: 16 Sep 2011

Topics: Seam carving, Line (text file) and Distance transform

Related papers:

- [Text line segmentation of historical documents: a survey](#)
- [Line Segmentation for Degraded Handwritten Historical Documents](#)
- [ICDAR2009 handwriting segmentation contest](#)
- [Script-Independent Text Line Segmentation in Freestyle Handwritten Documents](#)
- [Language-Independent Text Lines Extraction Using Seam Carving](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/text-line-segmentation-for-gray-scale-historical-document-14sz19yjhx>

Text Line Segmentation for Gray Scale Historical Document Images

Abedelkadir Asi¹ Raid Saabni^{2,3} Jihad El-Sana¹

¹Ben-Gurion University of the Negev, Beer Sheva, Israel

²Tel-Aviv University, Tel-Aviv, Israel

³TRDC, Kafr Qarea, Israel

{abed,saabni,el-sana}@cs.bgu.ac.il

ABSTRACT

In this paper we present a new approach for text line segmentation that works directly on gray-scale document images. Our algorithm constructs distance transform directly on the gray-scale images, which is used to compute two types of seams: *medial seams* and *separating seams*. A medial seam is a chain of pixels that crosses the text area of a text line and a separating seam is a path that passes between two consecutive rows. The medial seam determines a text line and the separating seams define the upper and lower boundaries of the text line. The medial and separating seams propagate according to energy maps, which are defined based on the constructed distance transform. We have performed various experimental results on different datasets and received encouraging results.

Keywords

Seam Carving, Line Extraction, Multilingual, Signed Distance Transform, Dynamic programming, Handwriting

1. INTRODUCTION

Historical handwritten documents are valuable cultural heritage, as they provide insights into both tangible and intangible cultural aspects from the past. The need to preserve these documents demands global emerging efforts to analyze and manipulate them by utilizing techniques from various science fields. Handwritten historical documents pose real challenges for automatic processing, such as image binarization, writer identification, page segmentation, and keyword searching and indexing. A considerable number of algorithms address these tasks; some provide acceptable results and already integrated into working systems.

Document image segmentation into text lines is a major prerequisite procedure for various document image analysis tasks, such as word spotting, key-word searching, and text alignment [1, 2, 3, 4, 5, 6]. Extracting text lines from handwritten document images poses different challenges than those

in machine-printed documents [7]. These challenges can be related roughly to two main factors: writing style and image quality. Writing styles differs among writers and give rise to various text-line segmentation difficulties. Baseline fluctuation due to pen movement, and as a result the baseline may be straight, sequence of straight segments, or curved. Variability in skew among different text lines is a real challenge that complicates the extraction process. Crowded writing styles muddle text line boundaries as interlines spaces become narrow and increase the overlap of components' bounding boxes among adjacent text lines. The presence of touching components from two adjacent text lines poses an obstacle for both text line segmentation approaches; those that search for separating lines and those that search for aligned physical units. Punctuation and diacritic symbols, which are located in-between text lines, complicate the deciphering process of the physical structure of handwritten text lines. Historical document images are usually of low quality due to aging, frequent handling, and storage conditions. They often include various types of noise, such as holes, spots, broken strokes, which may entangle the extraction process and produce binarization errors once binarization is applied.

Several text-line extraction algorithms for handwritten documents have been proposed (see Section 2). Saabni and El-Sana [8] presented an interesting line extraction algorithm, which is based on the seam carving technique introduced by Avidan and Shamir [9]. Their algorithm is limited to binary images and as a result it inherits the limitations of image binarization, which introduce noise and various artifacts. Preprocessing and post-processing steps were introduced to determine text line boundaries, cope with touching/overlapping components, and collect additional strokes. In this paper, we built upon their work and developed a robust line extraction algorithm that works directly on gray-scale images and overcomes the limitations mentioned above.

Our algorithm constructs distance transform directly on the gray-scale images and computes *medial seams* and *separating seams*, which determine the text lines in a document image (see Figure 1). The medial seam determines the middle of the text row and the separating seams, which are generated with respect to a medial seam, define the upper and lower boundaries of the text line. The medial and separating seams propagate according to different energy maps, which are defined based on the constructed distance transform. The inability to determine the boundaries of text lines forces Saabni and El-Sana [8] to recompute the energy map

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HIP '11, September 16 - September 17 2011, Beijing, China
Copyright 2011 ACM 978-1-4503-0916-5/11/09...\$10.00.

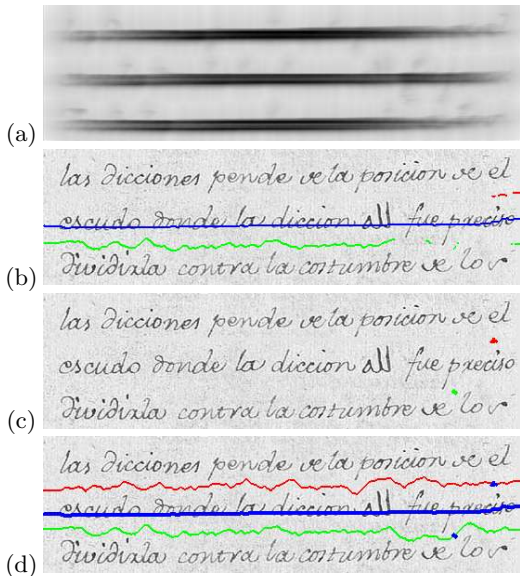


Figure 1: Algorithm flow: (a) the seam-map, (b) medial seam (blue) and seam fragments (green and red), (c) seam seeds (green and red), and (d) medial seam and separating seams.

for the entire page image after the extraction of each text line. The separating seams determine the text line boundaries, define the region to be updated, and overcome the limitation of recomputing the energy map.

In the rest of this paper we briefly review related work, describe our approach in detail, and report experimental results. Finally we conclude and discuss directions for future work.

2. RELATED WORK

Determining the text lines of a document image is a basic procedure for various document processing applications and have received tremendous attention over the last several decades.

Image smearing was among the earliest approaches used to determine text lines; Wong *et al.* [10] applied image smearing to binarized printed document images and Bar-Yosef *et al.* [11] used it for historical documents. *Projection profiles* along a predetermined direction is used in top-down approaches to estimate the paths separating consecutive text lines [12, 13, 14, 11, 15, 16]. *Adaptive local projection profiles* is employed to handle multi-skew in document images [11, 17]. *Hough transform* is used to compute the direction to apply projection profile; and to generate good text line hypotheses [18, 19]. *Fuzzy run length matrices* and *adaptive local connectivity maps* are applied directly to the gray-scale document images [20, 4, 21]. Tracking minima points to follow the white-most and black-most pixels along horizontal paths are used to estimate the boundaries and baselines of text lines [22]. *Seam carving approach* is used to find the seams, which resemble the baseline of text row, using a signed-distance-transform based energy map [8]. Various *grouping techniques*, such as heuristic rules, learning algorithms, nearest neighbor, and search trees are applied to

segment text-line in document images [23, 24, 25, 26, 27, 28, 29]. These approaches are applied to binary images, as they often require the isolation of basic elements, such as strokes and connected components. *Level-set techniques* are applied to extract text lines and handle multiple orientations, touching, and overlapping characters [5, 30]. A *painting technique* is employed to smear the foreground of the document image and enhance the separability between the foreground and background to simplify the detection of text-lines [31]. Dynamic programming is used to determine text lines through computing minimum cost paths between consecutive text lines [32].

3. OUR APPROACH

Humans tend to perceive text line patterns by tracking the concentration of ink along lines without actually reading the written text. Spaces between text lines play a major role in perceiving the layout of text lines despite the existence of touching/overlapping components, which are usually sparse and rarely disrupt the human text line recognition. These observations motivated the development of our approach, which find the medial seams that determine the text line and the two seams that separate it from its previous and next text lines.

We build upon the approach proposed by Saabni and El-Sana [8], which applies distance transform to binary images to generate energy maps; and then computes the minimal energy seams, which crosses the components along text lines. Connected components labeling in binary document images is a crucial step in their algorithm. Our approach works directly on the gray-scale images and it does not require binarizing the image neither labeling the components. It extracts a stripe that resembles the text line directly from the gray-scale image, which eliminates the need for post-processing to determine the components of the text line that do not intersect the computed seam. In addition, separating seams naturally cope with overlapping components during seam propagation. Saabni and El-Sana [8] recompute the energy map for the entire page after the extraction of each text line, because of the inability to bound the influence of the extracted text line. In contrast, our approach locally updates the energy map to cancel the influence of the extracted line's region, and thus saves recomputing the energy map.

Next we discuss in detail the two main steps of the line extraction procedure: constructing the energy maps and computing seams.

3.1 Energy Map

The search for a chain of pixels that either passes across a text line (i.e., the medial seam) or lies as far as possible from text lines (i.e., separating seam) calls for an energy function that provide a sufficient distance measure. The distance transform was adopted to generate the energy map, where local minimum points determine medial seams and maximum points define the separating seams.

To determine a seam that passes along the medial axis of a text line and crosses its components, we use an energy map based on a gray-level distance transform, introduced by Levi and Montanari [33]. Gray-level distance transform

is defined as modified geodesic distances; i.e., the distance between pixels p and q is the minimum of the lengths of the paths $[p = p_0, p_1, \dots, p_n = q]$. The length of the path, $l(p)$, is defined by Equation 1, where f is a distance and $d(p_i, p_{i+1})$ corresponds to the slope between two consecutive pixels.

$$l(p) = \sum_{i=0}^{n-1} f(d(p_i, p_{i+1}), I(p_i), I(p_{i+1})) \quad (1)$$

Toivanen [34] developed the distance transform on curved space (DTCOS) and the weighted distance transform on curved space, where the distance metric is defined as difference between the gray values of the pixels along the minimal path; i.e., $d(p_i, p_{i+1}) = |I(p_i) - I(p_{i+1})| + 1$.

The Gray-level Distance Transform (GDT) assigns values to pixels according to their distance from the nearest minimal points (reference points). In contrast to binary document images, paths between components pixels and background pixels in gray-scale images are curved and not straight lines, as the straight lines between two pixels can be blocked by obstacles consisting of higher or lower gray-values.

The distance transform of noisy document images may include small fluctuation that influence seam generation. To overcome this limitation, we apply Gaussian filter to smooth the image before generating the distance transform.

3.2 Seam Generation

Seams are computed using dynamic programming which relies on generating an energy map that encodes the minimal cost of the valid paths. We refer to this energy map as the *seam-map*, which is computed similar to [8] with slight modification to generate salient line structure. We replace the equal weights for the horizontal and diagonal distances by different weights that reflects the actual distance on the image (see Equation 2, where $w_0 = 1$ and $w_1 = w_{-1} = \frac{1}{\sqrt{2}}$). We found that this modification generates accurate energy maps and produce robust seams. The algorithm determines the minimal cost path by starting with the minimal cost on the last column (right column) and traversing the seam-map backward – from right to left.

$$map[i, j] = 2GDT(i, j) + \min_{l=-1}^1 (w_l * map[i+l, j-1]) \quad (2)$$

3.2.1 Medial Seam

We noticed that computing the seam-map in one pass as shown above generates a seam-map which rows are more salient on the right side than on the left side, as shown in Figure 3(b). The weak seam-map on the left side allows the medial seam to jump to adjacent lines as shown in Figure 2(a). To improve the saliency of the rows and retain the seam on the medial of text lines, we construct the seam map using two passes – from left-to-right and from right-to-left – and then bi-linearly interpolate the resulting two seam maps into the final seam-map, as shown in Figure 3(d). In this scheme, the generated seam-map is well-defined along text lines and faithfully resembles their structure (see Figure 3),

which forces the medial seam to propagate along the text lines, as shown in Figure 2(b).

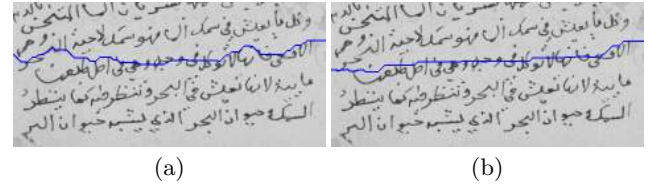


Figure 2: Medial seam generation using (a) one pass, (b) two passes.

3.2.2 Separating Seams

The separating seams define the upper and lower boundaries of text lines; i.e., determine the text strip, which is necessary to assign in-between component to the right text lines and accurately determine the pixels that need to be updated in the seam-map to avoid recomputing the seam-map after each line extraction. Separating seams of a text line are generated with respect to the medial seam of their text line.

Seam seeds are defined with respect to a medial seam as the global maxima (on the Gray-scale Distance Transform) along the vertical segment connecting two consecutive medial seams. Separating seams are grown from seam seeds toward the two sides of a page image (left and right). However, the need to determine the separating seams of a medial seam before determining its neighboring medial seams complicates computing the seam seeds.

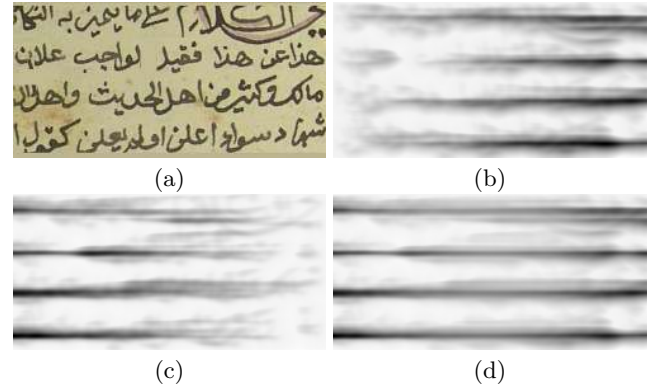


Figure 3: Seam-map generation: (a) original image, (b) left-to-right, (c) right-to-left seams, and (d) the interpolated final seam-map.

To generate the seam seeds for a medial seam, s_m , from each pixel, p_x on the medial seam we search for the maximum points on its lower and upper sides along the column including p_x . The absence of the adjacent medial seams makes it hard to distinguish between local and global maximum points. To resolve such dilemma, for each suspected maximum point, p_{max} , we proceed searching until we reach the first minimum point, p_{min} . If p_{min} is connected through a minima path (valley) to the starting seam, s_m , then p_{max} is a local maximum, which belongs to the processes seam, and there is a need to continue searching for another maximum point, otherwise p_{max} is considered a global maximum (an in-between-lines point). To verify that p_{min} does not belong

to the seam s_m , we extend p_{min} along minimal points with respect to the GDT map and test whether the extended path reaches the processed medial seam, s_m , or not. If the path returns to the s_m , then p_{min} belongs to the processed seam s_m , otherwise it is not. Since local minima cannot be completely avoided, a noncontinuous seams that consists of separated short seams, which are denoted *seam fragments*, are generated (see Figure 1). Erroneous fragments are filtered out based one the *length* of the fragment and its *distance* from the medial seam.

Short fragments are not reliable, as they may indicate local minima with respect to the segment between the two medial seams. Therefore, seam fragments are sorted in ascending order according to their length and the fragments in the lower fraction are filtered out (in current implementation we ignore the lower half). For each of the remaining fragment we associate a *certainty value*, $certainty(f, s)$, which aims to resemble the probability of a seam fragment to coincide with the corresponding separating seam. The certainty value is the sum of the distances of fragment pixels from the medial seam, as shown in Equation 3, where f_s and f_e are the first and the last column of the fragment f , and ms is its corresponding medial seam. We sort the remaining fragments according to their certainty value in a descending order and ignore the lower fraction (in our current implementation we ignore the lower half).

$$certainty(f, ms) = \sum_{i=f_s}^{f_e} |ms(i) - f(i)| \quad (3)$$

The fragments with the highest certainty values form the *seed candidate* set. We extend each seed candidate to the left and right sides of the page image, by propagating the seam along the maximal points of the GDT map. We then prioritize the candidates based on the number of fragments each extended seam passes through. The candidate with the maximal priority is taken as the seam seed. Note that we select two seam seeds one below the medial seam and one above it.

3.2.3 Seam Propagation

Growing a seam seed into a separating seam should maintain an appropriate distance from the corresponding medial seam. The separating seam is guided by the GDT map, which is computed based on the topography (gray levels) of the image. The fork of ridges leads to the existence of separating seams with small differences in their weights, where the maximal-weight seam may not be the sought seam (see Figure 5). To overcome this limitation we incorporate the distance from the medial seam into the propagation scheme of the separating seam by integrating a *spring model* within the seam prorogation scheme. The applied force of the spring model is used as a weight in the propagation scheme; i.e., $F = k(|d_r - d|)$, where d_r and d are the rest distance and the distance from the medial seam, respectively, and k is the spring constant. The rest distance is the average distance between the medial seam and the currently computed separating seam. This scheme pushes the separating seam away from the medial seam, when it is too close and attracts the seam toward the medial seam when it moves away from

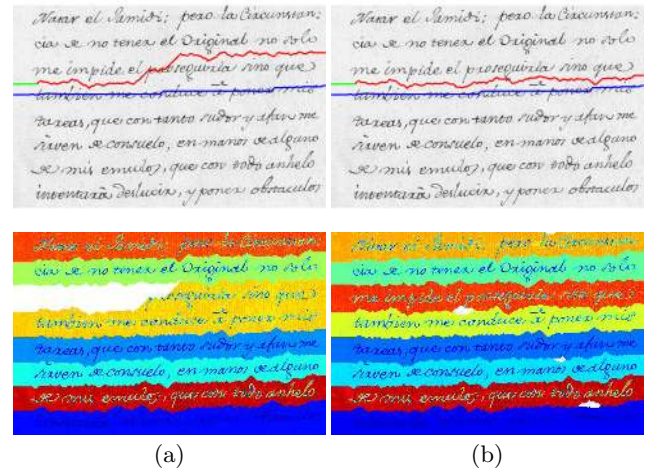


Figure 4: Spring model: (a) the medial seam (blue) and a possible separating seam (red), (b) the resulting separating seam after applying spring force.

the seam (see Figure 4). The spring factor k was determined experimentally, and we have found out that we need small values of k , usually $1/d_r$.



Figure 5: A document image and its distance transform, where two fork examples are marked with red rectangles.

4. EXPERIMENTAL RESULTS

Several evaluation methods for line extraction algorithms have been proposed in the literature. Some, evaluate the results manually, while others use predefined line areas to count misclassified pixels. Since the proposed approach works directly on gray-scale images and it is not possible to rely on connected components for automatic evaluation, the experimental results were evaluated manually. The correctness of an extracted text line is evaluated based on its three seams: the medial and the two separating seams. The medial seam is expected to go through the middle of the same text line – the ink (dark) area. The fraction of a medial seam that goes between text lines or jump to another text line is defined as faulty. Similarly, a separating seam is expected to go between the lines and faulty fractions go through (or touch) text area or jump to another in-between area. We measure the correctness, $correctness(s)$, of a seam s as the ratio of the correct fractions; i.e., the ratio of the sum of pixels in the correct fractions to the line width in pixels. We measure the correctness of an extracted text line, l , using Equation 4, where $medial(l)$, $upper(l)$, and $lower(l)$ are the medial, upper, and lower seams of the text line l . Equation 4 return a value in the range $[0.0, \dots, 1.0]$, 1 for perfectly extracted lines and 0 for completely wrong extraction.

$$\begin{aligned}
correctness(l) &= \frac{correctness(medial(l))}{2} \\
&+ \frac{correctness(upper(l))}{4} \\
&+ \frac{correctness(lower(l))}{4}
\end{aligned} \quad (4)$$

Equation 4 nicely measure the correctness of a text line, but underestimates the intersection of a separating seam with a descender or ascender, as it usually occupy a small number of pixels. Therefore, we count the number of such intersection for each line and measure the percentage of such intersection separately, as shown in Table 1.

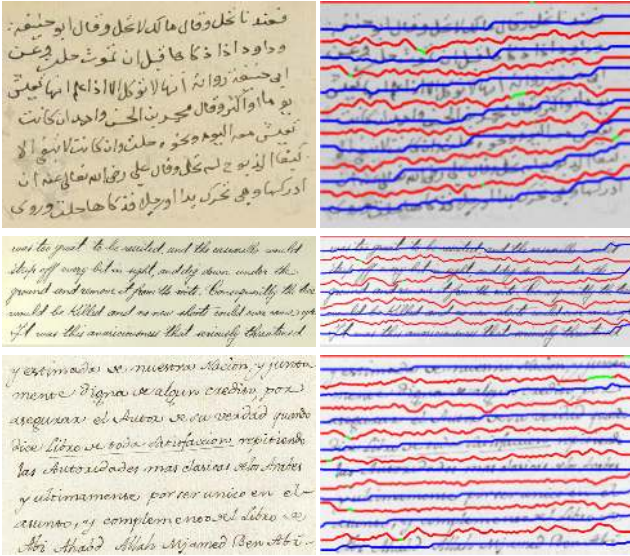


Figure 6: Random samples from the tested document images: Arabic, English and Spanish.

The absence of publicly availability database for evaluating line extraction algorithm on gray-scale images drove the development of our own dataset, which consist of various historical manuscripts in different languages. We have evaluated our system using 97 Arabic pages (900 lines) from *Juma Al-Majid Center for Culture and Heritage* [35], 70 pages (1050 lines) from *Wadod Center for Manuscripts* [36], 40 pages (420 lines) from a 19th-century master thesis collection in the American University of Beirut(AUB) [37] and 10 pages (150 lines) from Thomas Jefferson manuscripts located at the *Congress Library*. Our dataset includes Arabic, English, and Spanish handwritten document images. The images have been selected to have multi-skew, touching/ overlapping components and both regular and irregular spacing between lines.

Table 1 shows the average performance of our algorithm using various datasets of different qualities. Figure 6 presents samples from the tested datasets. As can be seen, it performs well independent of the used script and manages to generate very good results for languages that include delayed strokes, dots, and diacritics.

DataSet	Correctness(%)				Stroke (%) Crossing
	Medial	Upper	Lower	Line	
Wadod	99	97	97	98	9
Al-Majid	98	96	97	97	2
AUB	96	95	94	95	9
Congress L.	95	93	94	94.2	11

Table 1: The performance of our algorithm on various dataset written in different languages.

Our approach enables the separating seams to split touching components along the path passing between the lines and separate them, not necessarily on the optimal position. This procedure may split fractions of bracket-shape ascenders or descenders that besiege the propagating seam and force it to pass through, as shown in Figure 7. Nevertheless, it is easy to fix this in a post-processing procedure that examines the cases where the separating seam passes through local minima. Propagating along local minima path usually reveals whether the crossed shape was a touching component, ascender, or descender.



Figure 7: The last word on the second line(right-to-left), descender besieges the propagating seam and force it to pass through.

In languages that include many dots and diacritics, such as Arabic, hand-writers may not respect the closeness rule and place dots or diacritics closer to the above or below text line. Our algorithm may fail to detect such a case, as it requires recognizing the written text to determine for which text line those dots or diacritics belong. However, it is noticeable that even for a human it is not easy to assign the misclassified diacritics to the adequate text line without reading the text.

Zhixin *et al.* [4] presented an interesting approach for text line segmentation on gray scale images. They adaptively binarize the local connectivity map to focus on the line locations and superimposed the binarized ALCM on a binary version of the original document image to collect components that touch the line patterns. Their approach still requires adaptive binarization for component extraction and labeling, whereas our approach is binarization-free and does not include any component labeling step. We also used documents from the Congress Library to test our approach, but since their documents were randomly selected for testing, it is not easy to provide accurate comparison.

5. CONCLUSIONS AND FUTURE WORK

We presented a language independent approach for text line segmentation for gray-scale images. Our approach constructs an energy map directly on the gray-scale document image; and then computes a medial seam and separating seams for

each text line. A medial seam passes along the presumed text lines and the separating seams define the upper and lower boundaries of the text line.

Our approach avoids applying image binarization, which introduces noise and various artifacts. It also does not extract connected components and does not need to deal with text fragmentation. Instead, it directly computes the distance transform on the gray-scale images. Determining the boundary of text line enables updating the seam-map locally, and hence saves recomputing the seam-map and distance transform for the entire image after each text line extraction.

We see the scope on future work on extending this approach to determine the page layout and the component of a text line directly on gray-scale document images.

6. ACKNOWLEDGMENT

This research was supported in part by the Israel Science Foundation grant no. 1266/09, DFG-Trilateral Grant no. 8716, the Lynn and William Frankel Center for Computer Sciences at Ben-Gurion University, Israel. We would like to thank the reviewers for their insightful comments which led to several improvements in the presentation of this paper.

7. REFERENCES

- [1] T. Rath and R. Manmatha, "Word image matching using dynamic time warping," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, 18-20 June 2003, pp. II-521-II-527vol.2.
- [2] T. M. Rath, R. Manmatha, and V. Lavrenko, "A search engine for historical manuscript images," *Annual ACM Conference on Research and Development in Information Retrieval*, pp. 369-376, 2004.
- [3] E. M. Kornfield, R. Manmatha, and J. Allan, "Text alignment with handwritten documents," *Document Image Analysis for Libraries, International Workshop on*, vol. 0, p. 195, 2004.
- [4] S. Zhixin, S. Srirangaraj, and G. Venu, "Text extraction from gray scale historical document images using adaptive local connectivity map," in *ICDAR '05: Proceedings of the Eighth International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 794-798.
- [5] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, "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.
- [6] R. Saabni and J. El-Sana, "Keyword searching for arabic handwriting," in *International conference of frontiers in Handwriting recognition ICFHR, Montreal, Canada.*, August 2008, pp. 271-276.
- [7] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *International Journal on Document Analysis and Recognition*, vol. 9, pp. 123-138, 2007, 10.1007/s10032-006-0023-z. [Online]. Available: <http://dx.doi.org/10.1007/s10032-006-0023-z>
- [8] R. Saabni and J. El-Sana, "Language-independent text lines extraction using seam carving," in *International Conference on Document Analysis and Recognition, (to appear)*, 2011.
- [9] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graph.*, vol. 26, no. 3, p. 10, 2007.
- [10] K. Y. Wong, R. G. Casey, and F. M. Wahl, "Document analysis system," *IBM Journal of Research and Development*, vol. 26, no. 6, pp. 647-656, 1982.
- [11] I. B. Yosef, N. Hagbi, K. Kedem, and I. Dinstein, "Line segmentation for degraded handwritten historical documents," in *ICDAR, 2009*, pp. 1161-1165.
- [12] T. Pavlidis and J. Zhou, "Page segmentation by white streams," in *1st Int. Conf. Document Analysis and Recognition. (ICDAR) Int. Assoc. Pattern Recognition*, 1991, pp. 945-953.
- [13] S. Nagy and S. Stoddard, "Document analysis with expert system," in *Proceedings of Pattern Recognition conference in practice II*, 1985.
- [14] E. Bruzzone and M. Coffetti, "An algorithm for extracting cursive text lines," in *ICDAR 99: Proceedings of the Fifth International Conference on Document Analysis and Recognition. IEEE Computer Society*, 1999, p. 749.
- [15] J. He and A. C. Downton, "User-assisted archive document image analysis for digital library construction," in *ICDAR '03: Proceedings of the Seventh International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 2003, p. 498.
- [16] A. Zahour, B. Taconet, P. Mercy, and S. Ramdane, "Arabic hand-written text-line extraction," in *ICDAR, 2001*, pp. 281-285.
- [17] F. LeBourgeois, "Robust multifont ocr system from gray level images," in *ICDAR '97: Proceedings of the 4th International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 1997, pp. 1-5.
- [18] S. Vladimir, G. Georgi, and S. Vassil, "Handwritten document image segmentation and analysis," *Pattern Recogn. Lett.*, vol. 14, no. 1, pp. 71-78, 1993.
- [19] L. Likforman-Sulem, A. Hanimyan, and C. Faure, "A hough based algorithm for extracting text lines in handwritten documents," *Document Analysis and Recognition, International Conference on*, vol. 2, p. 774, 1995.
- [20] S. Zhixin and G. Venu, "Line separation for complex document images using fuzzy runlength," in *DIAL '04: Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)*. Washington, DC, USA: IEEE Computer Society, 2004, p. 306.
- [21] Z. Shi, S. Setlur, and V. Govindaraju, "A steerable directional local profile technique for extraction of handwritten arabic text lines," *Document Analysis and Recognition, International Conference on*, vol. 0, pp. 176-180, 2009.
- [22] A. Nicolaou and B. Gatos, "Handwritten text line segmentation by shredding text into its lines," in *ICDAR '09: Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer

Society, 2009, pp. 626–630.

- [23] L. Gorman, “The document spectrum for pagelay-out analysis,” *IEEE Trans. Pattern Analysis and Machine Intelligence.*, vol. 15, no. 11, pp. 1162–1173, 1993.
- [24] L. forman Sulem and C.Faure, “Extracting text lines in handwritten documents by perceptual grouping,” in *Advances in handwriting and drawing: a multidisciplinary approach .Winter Eds, Europa, Paris*, p. 117–135, 1994.
- [25] I. S. I. Abuhaiba, S. Datta, and M. J. J. Holt, “Line extraction and stroke ordering of text pages,” in *ICDAR*, 1995, p. 390.
- [26] A. Simon, J.-C. Pret, and A. P. Johnson, “A fast algorithm for bottom-up document layout analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 3, pp. 273–277, 1997.
- [27] Y.Pu and Z.Shi, “A natural learning algorithm based on hough transform for text lines extraction in handwritten documents,” in *Proceedings sixth International Workshop on Frontiers of Handwriting Recognition*, 1998, p. 637–646.
- [28] K. Koichi, S. Akinori, and I. Motoi, “Segmentation of page images using the area voronoi diagram,” *Comput. Vis. Image Underst.*, vol. 70, no. 3, pp. 370–382, 1998.
- [29] S. Nicolas, T. Paquet, and L. Heutte, “Text line segmentation in handwritten document using a production system,” in *Proceedings of the Ninth International Workshop on Frontiers in Handwriting Recognition*, 2004, p. 245–250.
- [30] S. S. Bukhari, F. Shafait, and T. M. Breuel, “Script-independent handwritten textlines segmentation using active contours,” in *ICDAR*, 2009, pp. 446–450.
- [31] A. Alaei, U. Pal, and P. Nagabhushan, “A new scheme for unconstrained handwritten text-line segmentation,” *Pattern Recognition*, vol. 44, no. 4, pp. 917–928, 2011.
- [32] M. Liwicki, E. Indermuhle, and H. Bunke, “On-line handwritten text line detection using dynamic programming,” in *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 01*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 447–451. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1304595.1304766>
- [33] G. Levi and U. Montanari, “A grey-weighted skeleton.” *Information Control*, vol. 17, pp. 62–91, 1970.
- [34] P. Toivanen, “New geodesic distance transforms for gray scale images.” *Pattern Recognition Letters*, vol. 17, pp. 437–450, 1996.
- [35] “Juma Al-Majid Center for Culture and Heritage,” <http://www.almajidcenter.org>, online; accessed June, 2011.
- [36] “Wadod center for manuscripts,” <http://wadod.com>, online; accessed June, 2011.
- [37] “Master thesis in pharmacy, American University of Beirut,” <http://ddc.aub.edu.lb>, online; accessed June, 2011.