Binarization-Free Text Line Segmentation for Historical Documents Based on Interest Point Clustering

Angelika Garz
Computer Vision Lab
Vienna Univ. of Technology
1040 Vienna, Austria
garz@caa.tuwien.ac.at

Andreas Fischer
Inst. of Computer Science
and Applied Mathematics
3012 Bern, Switzerland
afischer@iam.unibe.ch

Robert Sablatnig
Computer Vision Lab
Vienna Univ. of Technology
1040 Vienna, Austria
sab@caa.tuwien.ac.at

Horst Bunke
Inst. of Computer Science
and Applied Mathematics
3012 Bern, Switzerland
bunke@iam.unibe.ch

***Abstract*—Segmenting page images into text lines is a crucial pre-processing step for automated reading of historical documents. Challenging issues in this open research field are given e.g. by paper or parchment background noise, ink bleed-through, artifacts due to aging, stains, and touching text lines. In this paper, we present a novel binarization-free line segmentation method that is robust to noise and copes with overlapping and touching text lines. First, interest points representing parts of characters are extracted from gray-scale images. Next, word clusters are identified in high-density regions and touching components such as ascenders and descenders are separated using seam carving. Finally, text lines are generated by concatenating neighboring word clusters, where neighborhood is defined by the prevailing orientation of the words in the document. An experimental evaluation on the Latin manuscript images of the Saint Gall database shows promising results for real-world applications in terms of both accuracy and efficiency.**

***Keywords*-historical documents, manuscripts, ancient documents, handwritten, text line segmentation, binarization-free**

I. Introduction

Automatic segmentation of historical document page images is an open research field; algorithms are required to be robust with respect to background artifacts such as clutter, stains and noise, as well as artifacts due to aging, and touching or interfering lines [1]. Text line segmentation, in particular, is typically needed for handwriting recognition in historical documents. Handwritten documents do not have strict layout rules and thus line segmentation methods need to be invariant to layout inconsistencies, irregularities in script and writing style, skew, and fluctuating text lines [1]. Furthermore, robustness to low contrast and rippled pages is required [2], [3].

Likforman-Sulem et al. [1] provide a detailed survey about segmentation of text lines with respect to historical documents. Well-known methods for text line segmentation in binary images include smearing [4], [5] and Hough-transform [6], [7]. Another commonly used approach is based on Projection Profiles (PP) [2], [8]–[11] for both binary and gray-scale images. Various authors [8], [9] adapted the global PP such that skewed text blocks, converging or merging text lines are segmented correctly.

Recent approaches [12]–[14] introduce seam carving [15] known from image retargeting for text line segmentation. Nicolaou and Gatos [14] use so-called local minima tracers which follow the line spacing in order to shred the document page in lines. Originally proposed for on-line documents [16], Indermühle et al. [12] use Dynamic Programming (DP) in order to find a path with the minimum cost between two lines in historical manuscripts. Asi et al. [13] apply their approach directly on gray-scale images, where a distance transform is computed from a Gaussian-blurred image, and the separating seams are established using DP.

In this paper, we introduce an efficient binarization-free method for line segmentation applicable to historical manuscripts. Thus, binarization-caused errors frequent in historical document images are not inherited. Furthermore, the proposed method follows a bottom-up approach by grouping parts-of-character interest points into text line regions. Hence, it is not necessary to extract text block regions beforehand, which can also be prone to errors for special page layouts encountered in historical documents. Touching components such as ascenders and descenders are locally split by means of seam carving.

The experimental evaluation of the proposed approach is carried out on the *Saint Gall database*[1] [17], which contains 60 pages of a Latin manuscript originating from the 9^{th} century written in Carolingian script by a single writer with ink on parchment. Two sample pages are illustrated in Figure 1. Besides the main text body, the pages contain colored initial letters and annotations located in the outer margin, which were added to the manuscript later. The text line spacing is relatively large compared to the word height and a regular page layout is present for all pages. However, there are line interconnections caused by ascenders and descenders. Stains, holes, and ink bleed-through pose

This work has been supported by Austrian Science Fund (Grant P23133) and Swiss National Science Foundation (Project CRSI22_125220).

[1] Available at `http://www.iam.unibe.ch/fki/databases`

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.23

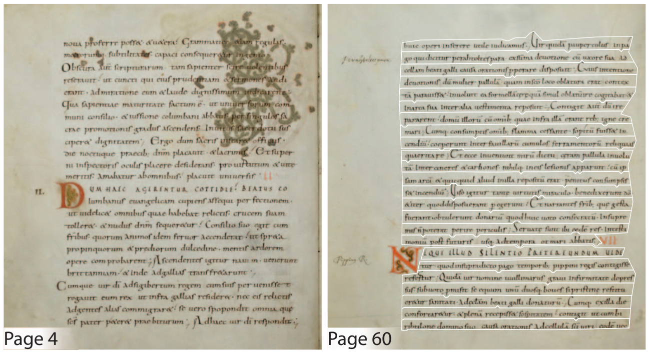

Figure 1. Two example pages of the *Saint Gall database*. Page 60 is overlaid with contours determining the lines in the ground truth.

additional challenges for line segmentation.

The remainder of this paper is structured as follows. First, the proposed method is explained; experimental results are depicted in Section III, followed by conclusions drawn in Section IV.

II. Methodology

The method proposed in this paper first transfers the gray-scale document image from the pixel domain into a domain of interest points, and thus obtains a sparse model of the image on which all consecutive steps are applied. In order to identify words, spatial clustering robust to noise is employed. Adjoined words of consecutive text lines connected by touching ascenders and descenders are identified and a path separating the lines is established using seam carving. Finally, text lines are generated connecting adjacent word clusters in the direction of the prevailing orientation of the text in the page. In the following, the successive steps of the proposed method are explained.

A. Transformation into Interest Point Domain

The first step is the extraction of interest points representing structures in the image being dissimilar to their adjacent neighborhood, e.g. in terms of intensity or color. An interest point has a defined location in the image and a definite spatial extend, which is denoted as scale.

The Carolingian handwriting in the dataset regarded has ascenders and descenders as well as capital letters having long strokes with little structure. Thus, interest points extracted by means of Difference of Gaussian (DoG) [18] are chosen according to prior studies [19]. The DoG detects interest points at locations of local minima and maxima exploiting a scale space. These local extremes represent character parts such as junctions, circles, arcs or stroke endings. The DoG is sensitive to edges, which is regarded as an unwanted property in the field of object recognition since localization along the edge is poor. However, an exact localization along edges not crucial for the given task, as interest points located on edges are needed.

Layout analysis prior to the segmentation of text lines is not a compulsive requirement due the very nature of the interest points extracted by means of DoG, which allows the selection of thresholds in both, the scale and the sensitivity in terms of pixel intensity. Thus, exclusively interest points representing parts of the text can be extracted. Considering manuscripts where a selection based on these thresholds is impossible, interest points could be classified with an approach such as the one proposed in [19], to be applied prior to line segmentation.

Note that after extraction of interest points, line segmentation can be realized very efficiently in the sparse interest point domain.

B. Identification of Word Clusters

Since interest points are mainly detected on and between characters and only few interest points are generated for background areas between text lines, words can be identified in high-density regions.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [20] is applied to the interest points' coordinates in order to group adjoining characters into word clusters. Unlike other clustering algorithms, the number of cluster centers is not determined before clustering.

The clustering algorithm requires two parameters – the minimum number of adjacent data points needed to build a cluster (we require 3 points) and the neighborhood radius, i.e. the density in the neighborhood has to exceed a certain threshold [20]. The neighborhood radius is manually estimated; it is defined to be half the height of a lower-case letter without ascenders and descenders (see Section III).

Thus, noise introduced by interest points describing accents, background clutter and large interest points representing two consecutive text lines are not taken into account owing to the neighborhood constraints.

For each word cluster, a minimum area rectangle is calculated such that one edge of the rectangle is aligned with an edge of the convex hull (see Figure 2 a).

The prevailing text orientation is determined based on the distribution of interest points of each word cluster. It is given by the median of the first principle component direction of each word cluster. Based on the orientation-normalized word clusters, the median word height used for further steps is automatically determined, where word height is defined as before.

C. Identification and Separation of Touching Components

Touching components such as ascenders and descenders as well as capital letters lead to word clusters enclosing adjacent lines. These word clusters are separated applying seam carving [15] to the interest points' coordinates. Seam carving was originally developed for image retargeting. First each pixel of the image is valued by an energy function, which is to be chosen according to the objective of the respective task. The seam is then generated by propagating a path of minimum cost through the image.

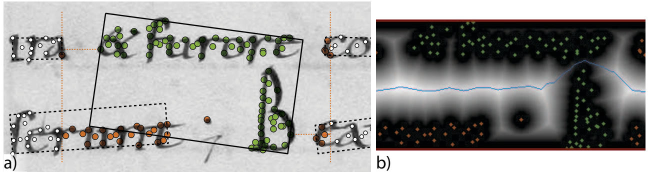

Figure 2. a) Image patch overlaid with representations of interest points (white, green and orange markers) and minimum area rectangles of the word clusters. Green markers indicate interest points belonging to the word cluster to separate, orange markers represent interest points of adjacent word clusters additionally selected to provide context. The dotted orange lines illustrate the distance in which interest points of adjacent word clusters are regarded. Further interest points of adjacent clusters not taken into account are presented as white dots.
b) Orientation-normalized distance transform overlaid with the interest points, the boundary condition (red) and the seam found (blue).

The word clusters to separate are identified based on their height. The corresponding height threshold is mainly dependent on the automatically determined median word height, but needs some fine-tuning for a given manuscript collection in practice (see Section III).

For the calculation of the seam interest points of adjacent word clusters in a local neighborhood are additionally selected in order to embed the word cluster into a line context (see Figure 2 a). Interest points within a distance of the median word height to either the left-most and right-most interest point of the word cluster to separate are considered (see Figure 2 a, dotted orange lines). Then, the interest points are normalized by the prevailing text orientation of the page. A boundary condition is introduced which incorporates the expectation that the seam is not to be propagated in the border regions of the word cluster. The energy function is calculated as distance transform of the interest points' coordinates and the boundary condition using Euclidean distance resulting in an orientation-normalized energy map (see Figure 2 b).

DP is employed in order to find the seam of maximum energy (farthest distance to the interest points and the boundary condition). For each pixel its energy and the energy of its three precedent neighbors to the left are accumulated, leading to the final energy map for DP. The highest values in the right-most column of the image represent the seams with the maximum energy. The seam of maximum energy is then propagated following the maximum pixel values in the inverse direction.

The boundary condition introduced prevents the seam from propagating into a local maximum at the border of the word cluster, which potentially happens if there is no adjacent word cluster in one of the text lines. Including interest points of adjacent word clusters provides line context and ensures calculating an optimal separating seam. Finally, the word cluster is split according to the seam.

D. Generation of Text Lines

The last step is generating text lines by concatenating word clusters; hereby, the concept of doubly linked lists is adopted. Each word cluster is connected to its nearest neighbors to the left and right in the prevailing text orientation. A fictive line is drawn from the center of mass of the word cluster and the adjacent word clusters selected are those first hit by the line in either direction. A word cluster is hit if at least one corner of its rectangle is on the opposite site of the line, which allows for a fast check. Figure 3 illustrates an example of line generation. Then, chains of connected word clusters are built and the chains with the the maximum number of word clusters are identified as text lines.

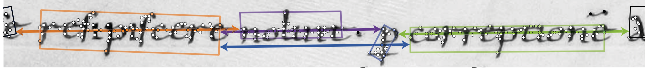

Figure 3. Concatenating word clusters. Interest points are indicated by white dots, word clusters are surrounded by rectangles. The fictive lines originating in the centers of mass of each word cluster are indicated by arrows in the respective color. The blue word cluster (*p*) is below the mean line position, however, it is hit by its neighboring clusters (violet, green), and hits the orange and the green cluster to its left and right respectively.

Handwritten documents contain curved and fluctuating lines; however, we assume the orientation not to change abruptly from one word to the next. Thus, the probability of the adjacent word cluster being hit by locally applying the prevailing text orientation as search direction is high.

As a last step, the interest points are spatially weighted with a two-dimensional Gaussian distribution with a standard deviation according to their scales (see [19]) in order to generate contours of the text lines. This leads to a probability map for each text line; these are voted against each other with the higher probability determining the text line a pixel belongs to, resulting in text line regions (see Figure 4).

III. Experimental Evaluation

The evaluation is performed on all 60 pages of the *Saint Gall database*. In a first step, a randomly selected page was used to fine-tune the system parameters. Most importantly, the threshold on the scale and sensitivity of the DoG interest points that determines the parts-of-character interest points needs to be optimized (see Section II-A). Other parameters include the neighborhood radius for DBSCAN (see Section II-B) and the threshold used for separating touching components (see Section II-C). Although reasonable defaults can be set for these parameters with respect to the automatically determined median word height, some fine-tuning improves the results in practice.

Figure 4 gives a sample result overlaid with a typical segmentation result. The text lines are represented by their contours randomly colored for easier distinction.

Figure 5 shows an image patch of the manuscript with an illumination bleeding through from the other side of the page resulting in background clutter. Nevertheless, the word clusters are found in presence of noise and lines are segmented correctly.

For evaluating the performance of the proposed system, the *pixel-level hit rate* and the *line accuracy measure* pre-

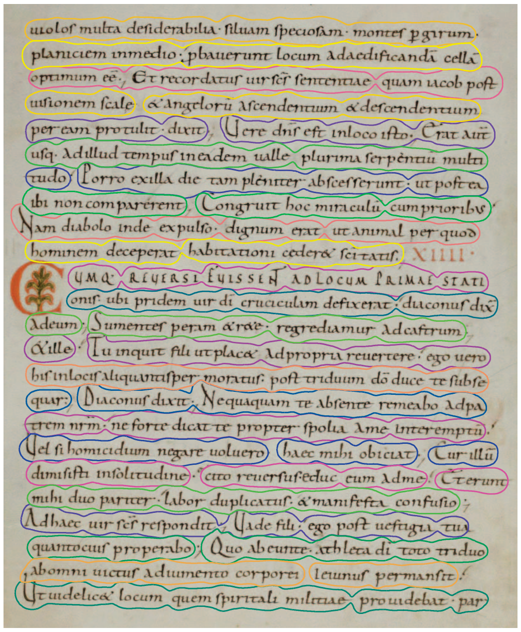

Figure 4. Page 21 of the manuscript overlaid with the contours of the detected text lines.

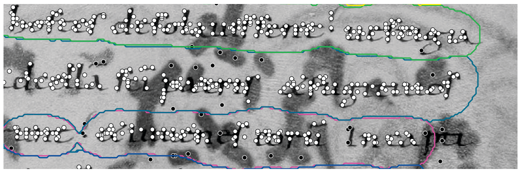

Figure 5. Line segmentation in presence of background clutter.

sented by Li et al. [21] are used. An $M \times N$ matrix P is constructed where M and N are the number of ground-truthed lines and lines detected by the algorithm respectively. The element $P_{i,j}$ $(i = 1 \ldots M, j = 1 \ldots N)$ represents the number of pixels shared between the i^{th} ground-truth line and the j^{th} detected line. For an assignment S between ground-truthed lines and detected lines, the goodness measure $G(S)$ is the total number of shared pixels $P_{i,j}$. The Hungarian algorithm [22] is applied to efficiently search for the best assignment S_o having the maximum goodness. In contrast to [21], we account for noise pixels such that we calculate the hit rate H as

$$H = \frac{G(S_o)}{|GT \cup R|} \tag{1}$$

with GT being all foreground pixels in the ground truth including those not found in the result; R represents the set of foreground pixels of the result including noise pixels (pixels additionally found by the algorithm, not being in the ground truth). Thus, segmentation errors such as splitting, merging and missing are penalized. We achieve a hit rate of 0.9865, which expresses the amount of ground truth pixels that are retrieved with an optimal assignment.

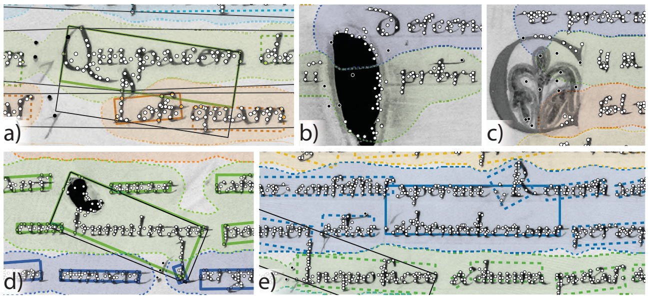

Figure 6. Image patches overlaid with markers representing interest points, minimum area rectangles color-coded according to the line they belong to, and colored contours indicating the text lines.
a) A word cluster (indicated by a black rectangle partly overlaid by the green one) split imperfectly into two word clusters (orange and green).
b,c) Interest points found on borders of holes and in initials.
d) Two lines merging due to a word cluster without a local minimum in the vertical distribution between two lines (The letter *I* of the green line is split in two parts).
e) Two lines merging due to a word cluster having a vertical extend smaller than the threshold to split (see Section II-C).

The line accuracy measure evaluates the performance on text line level. A line in the ground truth is detected if it fulfills two requirements: a) $\frac{G_{i,j}(S_0)}{|GT_i|} > 0.9$ and b) $\frac{G_{i,j}(S_0)}{|R_j|} > 0.9$. Thus, missing parts of the text line affect measure a), while noise penalizes measure b). A line accuracy of 0.9797 is achieved on the *Saint Gall database*, which has a total of 1,431 text lines.

A. *Typical Failures and Causes*

Part of the error can be traced back to imperfect separation of word clusters containing touching components. Since seam carving is done based on the interest points found, the seam might cut an ascender or descender if the distance between two adjacent interest points on the component is larger than the distance between interest points of the two components to separate (see Figure 6 a).

As pointed out in Section II-A, interest points are selected based on their scale and pixel contrast; no prior layout analysis is applied. Hence, interest points are found in initials and on the border of holes if their contrast to the background is high enough; resulting in noise in the detected lines (see Figure 6 b,c).

In one case, the split of a word cluster was not carried out because the distribution of the interest points did not show a local minimum due to the large amount of structure between the consecutive lines. Thus, the two consecutive text lines are merged (see Figure 6 d). In one case, a word cluster embracing two consecutive lines is not split since its height perpendicular to the prevailing text orientation is below the threshold to be considered for a split (see Figure 6 e).

B. *Efficiency*

Having obtained a sparse model of the image by detecting stable interest points, operations are either carried out on approximately 9,800 interest points (dependent on the actual image) or the minimum area rectangles of the word clusters instead of 16.6 million pixels. Thus, the operations can be efficiently implemented.

Generating the energy function for seam carving is expensive due to the dependencies of the pixels on each other. First identifying the word clusters which need to be split, we efficiently apply the seam carving approach only in a local area, thus further improving the efficiency of the method.

IV. Conclusion

We presented a binarization-free line segmentation method suitable for historical handwritten documents, which relies on interest points representing letters. The algorithm is carried out in the sparse interest point domain and thus, can be efficiently implemented. First, word clusters are identified in high-density regions using spatial clustering. Then word clusters containing touching or overlapping components are identified and separated by means of seam carving. Finally, adjacent word clusters in the direction of the prevailing text orientation are joined to lines.

Employing local features and spatial clustering as pivotal concept, the approach is independent of layout analysis, i.e. text block extraction. The method is not specifically adapted to the Carolingian script and thus, can be applied to other scripts with the constraint that a set of system parameters needs to be fine-tuned for a sample page of a new manuscript collection. The reported line segmentation accuracy for the Latin manuscripts of the *Saint Gall database* is promising for real-world handwriting recognition applications.

Future work includes applying the approach to manuscripts that have less restricted layouts and are harder to binarize. The proposed approach should further be tested on synthetic data in order to evaluate its robustness to fluctuation, curvature and convergence of text lines. A comparison with other state-of-the-art methods for line segmentation, possibly on pre-segmented text block areas since most of these methods are based on layout analysis prior to segmentation, is to be done.

References

[1] L. Likforman-Sulem, A. Zahour, and B. Taconet, “Text Line Segmentation of Historical Documents: A Survey,” *IJDAR*, vol. 9, no. 2, pp. 123–138, 2007.

[2] M. Bulacu, R. van Koert, L. Schomaker, and T. van der Zant, “Layout Analysis of Handwritten Historical Documents for Searching the Archive of the Cabinet of the Dutch Queen,” in *Proc. ICDAR*, vol. 1, 2007, pp. 357–361.

[3] M. Baechler, J.-L. Bloechle, and R. Ingold, “Semi-Automatic Annotation Tool for Medieval Manuscripts,” in *Proc. ICFHR*, 2010, pp. 182–187.

[4] Z. Shi and V. Govindaraju, “Line Separation for Complex Document Images using Fuzzy Runlength,” in *Proc. DIAL*, 2004, pp. 306 – 312.

[5] N. Nikolaou, M. Makridis, B. Gatos, N. Stamatopoulos, and N. Papamarkos, “Segmentation of Historical Machine-Printed Documents using Adaptive Run Length Smoothing and Skeleton Segmentation Paths,” *Image and Vision Computing*, vol. 28, no. 4, pp. 590 – 604, 2010.

[6] L. Likforman-Sulem, A. Hanimyan, and C. Faure, “A Hough Based Algorithm for Extracting Text Lines in Handwritten Documents,” in *Proc. ICDAR*, vol. 2, 1995, p. 774.

[7] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, “Text Line Detection in Handwritten Documents,” *Pattern Recognition*, vol. 41, no. 12, pp. 3758 – 3772, 2008.

[8] A. Zahour, B. Taconet, P. Mercy, and S. Ramdane, “Arabic Hand-Written Text-Line Extraction,” in *Proc. ICDAR*, 2001, pp. 281 – 285.

[9] I. Bar-Yosef, N. Hagbi, K. Kedem, and I. Dinstein, “Line Segmentation for Degraded Handwritten Historical Documents,” in *Proc. ICDAR*, 2009, pp. 1161 –1165.

[10] V. Shapiro, G. Gluchev, and V. Sgurev, “Handwritten Document Image Segmentation and Analysis,” *PRL*, vol. 14, pp. 71–78, 1993.

[11] A. Antonacopoulos and D. Karatzas, “Document Image Analysis for World War II Personal Records,” in *Proc. DIAL*, 2004, pp. 336 – 341.

[12] E. Indermühle, M. Liwicki, and H. Bunke, “Combining Alignment Results for Historical Handwritten Document Analysis,” in *Proc. ICDAR*, 2009, pp. 1186–1190.

[13] A. Asi, R. Saabni, and J. El-Sana, “Text Line Segmentation for Gray Scale Historical Document Images,” in *Proc. Workshop HIP*, 2011, pp. 120–125.

[14] A. Nicolaou and B. Gatos, “Handwritten Text Line Segmentation by Shredding Text into its Lines,” in *Proc. ICDAR*, 2009, pp. 626 –630.

[15] S. Avidan and A. Shamir, “Seam Carving for Content-Aware Image Resizing,” in *ACM Trans. Graph.*, vol. 26, no. 3, 2007, p. 10.

[16] M. Liwicki, E. Indermühle, and H. B. Bunke, “On-Line Handwritten Text Line Detection Using Dynamic Programming,” in *Proc. ICDAR*, vol. 1, 2007, pp. 447 –451.

[17] A. Fischer, V. Frinken, A. Fornés, and H. Bunke, “Transcription Alignment of Latin Manuscripts Using Hidden Markov Models,” in *Proc. Workshop HIP*, 2011, pp. 29–36.

[18] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[19] A. Garz, R. Sablatnig, and M. Diem, “Layout Analysis for Historic Manuscripts Using SIFT Features,” in *Proc. ICDAR*, 2011, pp. 508–512.

[20] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” in *Proc. KDD*, 1996, pp. 226–231.

[21] Y. Li, Y. Zheng, D. Doermann, S. Jaeger, and Y. Li, “Script-Independent Text Line Segmentation in Freestyle Handwritten Documents,” *PAMI*, vol. 30, no. 8, pp. 1313 –1329, 2008.

[22] G. Liu and R. M. Haralick, “Optimal Matching Problem in Detection and Recognition Performance Evaluation,” *Pattern Recognition*, vol. 35, no. 10, pp. 2125 – 2139, 2002.