Binarization-free Text Line Extraction for Historical Manuscripts

Nikolaos Arvanitopoulos and Sabine Süsstrunk
School of Computer and Communication Sciences, EPFL, Switzerland

1 Introduction

Nowadays, large collections of old historical manuscripts, which contain valuable information about our cultural heritage, exist in libraries around the world. Recently, there has been much interest in their digitization for preservation reasons, since many of the available manuscripts' quality has deteriorated from exposure to the environment. Digitization though is only the first step to make the information contained in manuscripts accessible to researchers and to the interested public. What we create after digitization is only a "digital image" of the page and further processing steps need to be applied during the *handwriting recognition process*, so that the manuscript's content is transformed into a form that is interpretable by a computer.

One important step in the handwriting recognition process is that of *text line extraction*, which aims at extracting individual text lines from the manuscript page. In this paper, we propose a binarization-free text line extraction method using seam carving [1]. The main idea is to compute an energy map of the input text blocks and determine minimum energy paths that pass through them. The energy map is constructed in a way so that gaps between text lines have low energy values. Therefore, a minimum energy path will pass only through these regions and will successfully separate two text lines.

Our algorithm has the following two advantages:

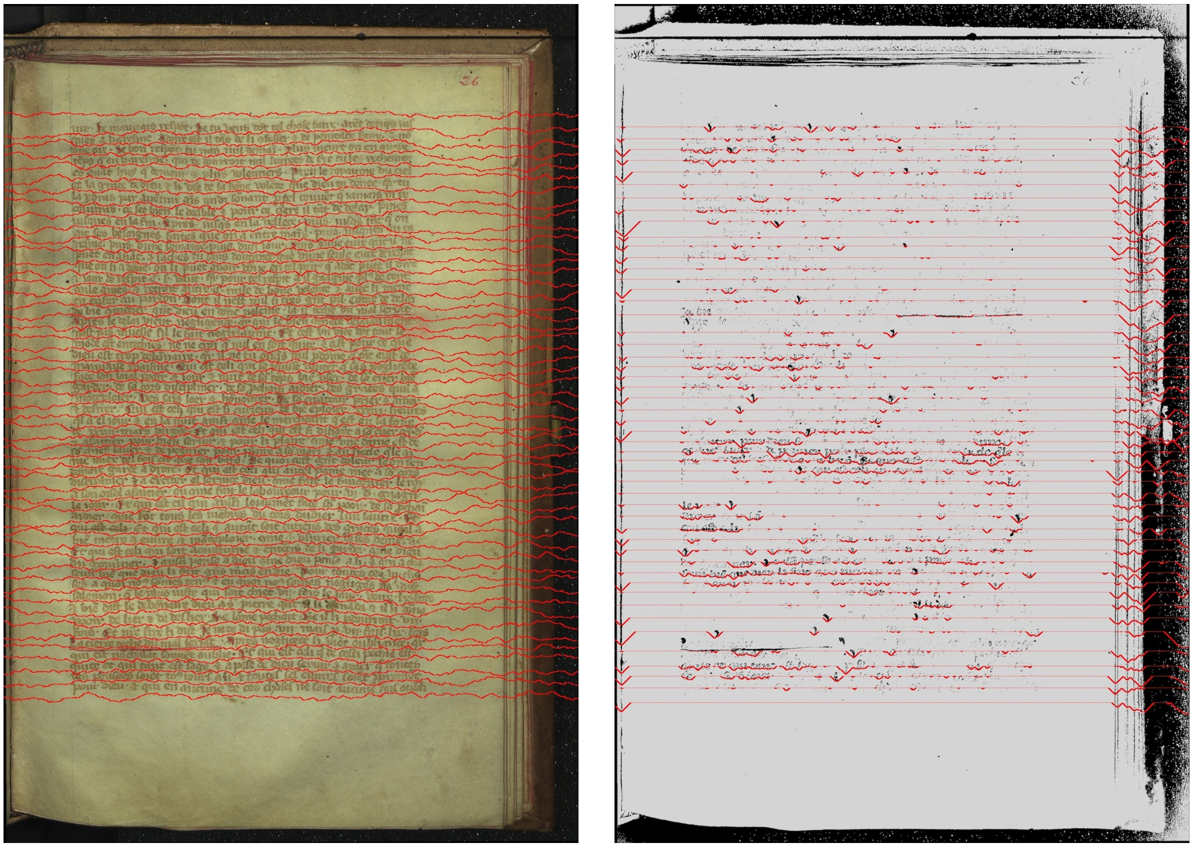
1. We make direct use of the original image representation of the manuscript page without any need for prior binarization, which can introduce information loss. This loss can produce unreliable results for the text line extraction algorithm (see Figure 1).

2. Our algorithm is general and can be applied to diverse manuscripts of different time periods and handwritings. Results in Figures 3 and 4 show the applicability of our algorithm to diverse historical manuscripts.

2 Related Work

We briefly summarize the research that has already been done for text line extraction.

Most of the state-of-the-art approaches operate on a binary image of the historical manuscript. One method based on dynamic programming computes the paths with minimum cost between two consecutive lines [11] and has been extensively used in automatic transcription and ground truth creation of historical documents [8, 7]. The work of [6] is based on horizontal projection profiles of black pixel changes. An additional post-processing step is applied, which follows the contour of the ink obtaining curve-linear line separators. Another similar approach is proposed in [13] where the output of the horizontal projections is post-processed based on properties of the computed connected components. The works of [12, 10] are based on the Hough transform, which is able to detect straight lines in images. Smearing methods, such as the ones in [17, 16, 14], try to fill-in the white pixel gaps with black pixels if their distance is less than a threshold. That way, homogeneous blocks of the document page are grouped together. Other approaches use multi-oriented filters and active contours for text line extraction [4, 5].



(a) Gautier de Metz, *Le mireour du monde.* Copy from François Buffereau, secretary of Antoine de Gingins (1475-1500).

(b) Gautier de Metz, *Le mireour du monde.* Copy from François Buffereau, secretary of Antoine de Gingins (1475-1500). Binarized version.

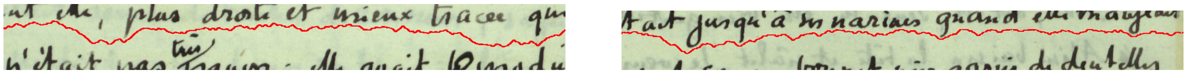
Figure 1: Left: seams generated using the original scanned image of the manuscript as input. Right: seams generated using a binary version of the original manuscript scan as input. The information loss in the binary version is so extensive that the generated seams do not clearly separate text lines of the original manuscript scan.

A notable exception of an algorithm, which does not depend on binarization, is the work of [9], where the text lines are found using extracted features from interest points of the original manuscript image. A very recent work [3] uses a framework similar to that in [2] adapted this time to the text line extraction problem.

Our method is closely related to the one in [15], where the authors use seam carving to generate seams that pass through connected components of a binary image. Unclassified components, which do not belong to any text line are assigned in a post-processing step according to their position and geometric characteristics. The main difference in our approach is that we do not need to binarize the input, which can lead to information loss. Additionally, we are able to generate robust text contours even for manuscripts of deteriorated quality (see Figures 1, 4). The text contours can always be overlaid on the original manuscript scan, even if they have been generated using as input a binarized version of the original scan. However, the technique of [15], which assigns text components to lines is not able to extract lines from the original manuscript, since the binarization process is not reversible. Binary text components contain only a subset of the information available in the original manuscript image.

3 Our Approach

Our proposed algorithm is inspired by *Seam Carving*, a computer vision algorithm used for image resizing [1]. We build upon this idea and propose a seam carving algorithm, which operates on the original color image and extracts lines in a sequential way. First, an energy map is calculated and the minimum energy path is computed based on dynamic programming. From the peaks of the horizontal projection profile of the derivative image we can find horizontal line positions. In each such region between two consecutive lines, we apply our seam carving algorithm sequentially until the whole manuscript image has been processed. In Figure 2 we show some examples of seams between two such lines.



(a) C.F. Ramuz, *Aline* (1905), p. 4

(b) C.F. Ramuz, *Aline* (1905), p. 4

Figure 2: Examples of image blocks and their computed seams.

In the following we use the convention that an image $\mathbf{I} \in \mathbb{R}^{n \times m}$ converted to grayscale has n rows and m columns. The notation $I(i,j)$ denotes the image value at the i-th row and j-th column. The coordinate system has its origin in the upper left corner of the image.

3.1 Energy Map

We modify the energy function proposed in [1] so that it can be effectively used for generating text line separation contours. First, we compute an edge image as

$$E(i,j) = \left| \frac{I(i,j+1) - I(i,j-1)}{2} \right| + \left| \frac{I(i+1,j) - I(i-1,j)}{2} \right| . \tag{1}$$

Let us denote the energy map between two text lines by $E_b = E(J,:) \in R^{l \times m}$, where $J = \{s, \ldots, e\}$ is the set of i coordinates between the start and the end of the energy block map and $l = e - s + 1$. This energy block is weighted by the following linear function, which penalizes the larger i coordinates more:

$$w(i) = \frac{1}{n-1} i + \frac{n-2}{n-1}, \quad i = 1, \ldots, l, \quad w(i) \in [1,2], \tag{2}$$

and the final energy map for this block is

$$E_{b,f}(i,j) = E_b(i,j) * w(i), \quad i = 1, \ldots, l, \quad j = 1, \ldots, m. \tag{3}$$

The idea behind this weighting is the observation that we want our seam to be closer to the upper line than the lower one. This will correct for situations where the author has written words in the gap between lines, which always belong to the lower line (see Figure 2a).

3.2 Seam Computation

A seam that passes horizontally through an image block can be defined as

$$\mathbf{s}_b^h = \{\mathbf{s}_{b,j}^h\}_{j=1}^m = \{(y(j), j)\}_{j=1}^m, \; \forall j, |y(j) - y(j-1)| \leq 1, \; y(j) = 1, \ldots, l. \tag{4}$$

The seam computation is done using dynamic programming. We look for the optimal seam in the image block that minimizes

$$\mathbf{s}_b^* = \arg\min_{\mathbf{s}_b} \sum_{j=1}^m E_{b,f}(\mathbf{s}_{b,j}^h). \tag{5}$$

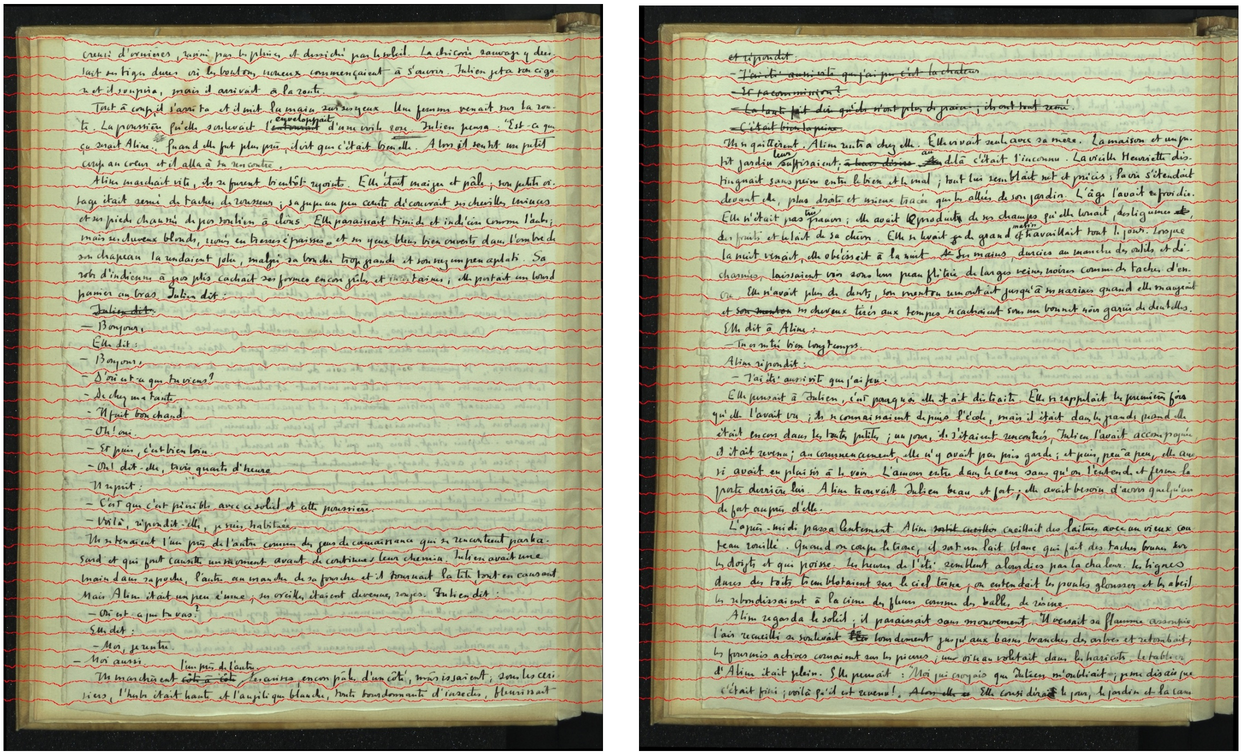
The first step is to traverse the image block and compute the cumulative minimum energy M_b for all possible connected seams for each pixel position:

$$M_b(i,j) = E_{b,f}(i,j) + \min(M_b(i-1,j-1), M_b(i,j-1), M_b(i+1,,j-1)). \tag{6}$$

The minimum value of the last column in M_b will indicate the end of the minimal connected horizontal seam. Therefore, in the second step we traverse the cumulative energy M_b backwards to find the path of the optimal seam.

4 Experimental Results

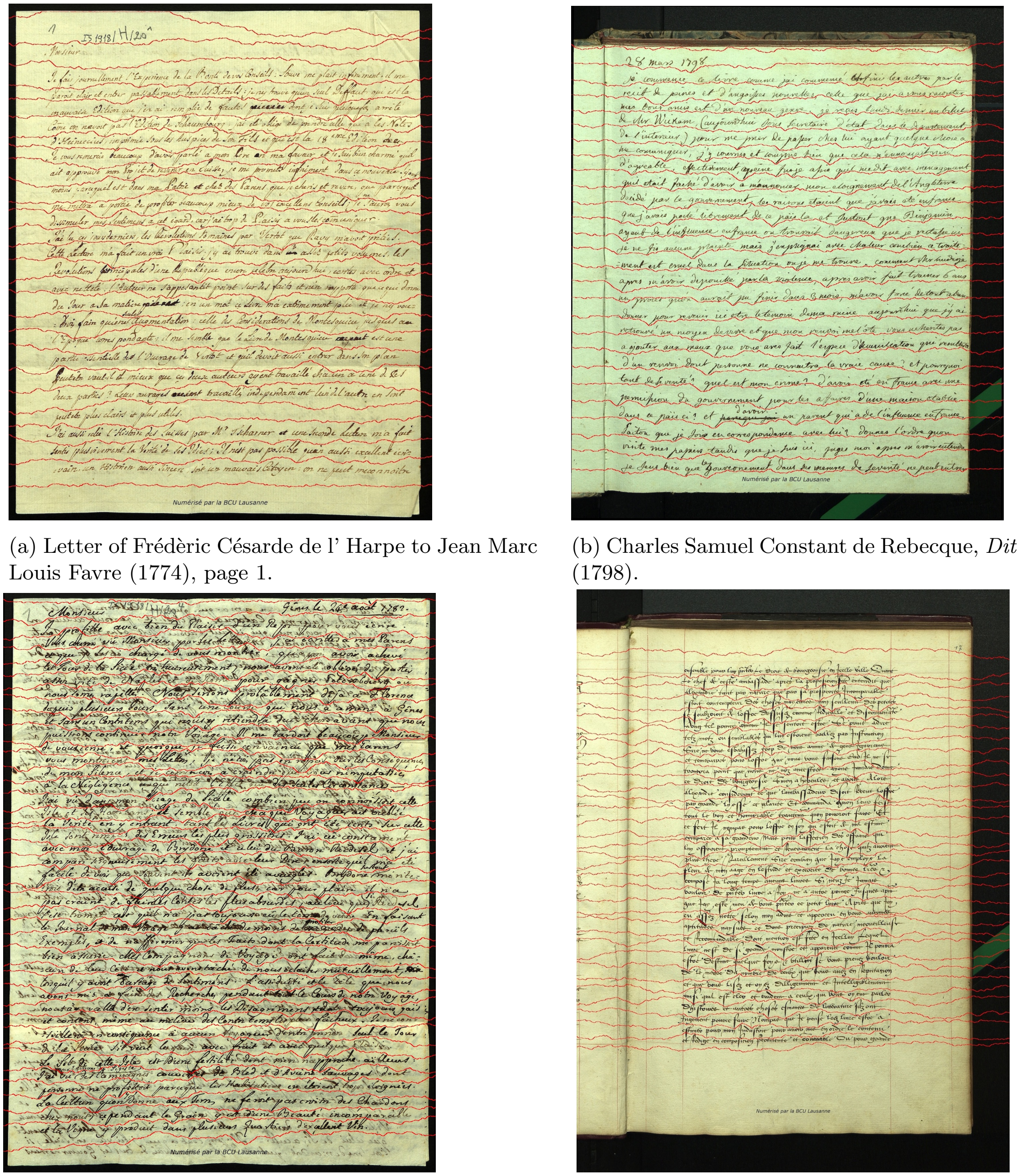
We apply our algorithm to original manuscript pages of the work *Aline* by the important Swiss-French writer Charles-Ferdinand Ramuz. Some examples of manuscript pages overlaid with the text line extraction seams are shown in Figure 3. We observe that our algorithm creates seams that pass through parchment regions, successfully segmenting the text lines. Even when the writer corrects a line or a word and writes above, the seam is able to avoid cutting the text and assigns the word to the line below it. In order to illustrate the ability of the algorithm to generalize to diverse manuscripts, we provide in Figure 4 results on manuscripts of the 16-th and 18-th century. We observe that our algorithm can be applied to manuscripts of very different quality and handwriting styles.



(a) C.F. Ramuz, *Aline* (1905), p. 2.

(b) C.F. Ramuz, *Aline* (1905), p. 4.

Figure 3: Seam carving results on two pages of *Aline* (1905), C.F. Ramuz.



(a) Letter of Frédèric Césarde de l' Harpe to Jean Marc Louis Favre (1774), page 1.

(b) Charles Samuel Constant de Rebecque, *Dit le chinois* (1798).

(c) Letter of Jean Marc Louis Favre to Frédèric Césarde la Harpe (1782), page 1.

(d) Guillaume Budé, *Commentaire et mémorial au roi François 1er* (1522).

Figure 4: Seam carving results on manuscripts of the 16-th and 18-th century respectively. Even in the lower left manuscript with extreme bleed-through, our algorithm is able to produce a robust result.

5 Conclusion

We propose a novel text line extraction algorithm for color scans of historical manuscripts based on seam carving. We show that we can obtain state-of-the-art results on these color images without any prior binarization. The next step after the text line extraction process is the application of a learning algorithm for handwritten word recognition in each extracted text line.

References

[1] Shai Avidan and Ariel Shamir. Seam Carving for Content-Aware Image Resizing. *ACM Transactions on Graphics*, 26(3):10, 2007.

[2] M. Baechler and R. Ingold. Multi Resolution Layout Analysis of Medieval Manuscripts Using Dynamic MLP. In *International Conference on Document Analysis and Recognition*, pages 1185–1189, 2011.

[3] M. Baechler, M. Liwicki, and R. Ingold. Text Line Extraction using DMLP Classifiers for Historical Manuscripts. In *International Conference on Document Analysis and Recognition*, 2013.

[4] S.S. Bukhari, F. Shafait, and T.M. Breuel. Script-Independent Handwritten Textlines Segmentation Using Active Contours. In *International Conference on Document Analysis and Recognition*, pages 446–450, 2009.

[5] S.S. Bukhari, F. Shafait, and T.M. Breuel. Text-Line Extraction Using a Convolution of Isotropic Gaussian Filter with a Set of Line Filters. In *International Conference on Document Analysis and Recognition*, pages 579–583, 2011.

[6] 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 *International Conference on Document Analysis and Recognition*, pages 357–361, 2007.

[7] A. Fischer, E. Indermühle, H. Bunke, G. Viehhauser, and M. Stolz. Ground truth creation for handwriting recognition in historical documents. In *IAPR International Workshop on Document Analysis Systems*, pages 3–10, 2010.

[8] A. Fischer, M. Wuthrich, M. Liwicki, V. Frinken, H. Bunke, G. Viehhauser, and M. Stolz. Automatic Transcription of Handwritten Medieval Documents. In *International Conference on Virtual Systems and Multimedia*, pages 137–142, 2009.

[9] A. Garz, A. Fischer, R. Sablatnig, and H. Bunke. Binarization-Free Text Line Segmentation for Historical Documents Based on Interest Point Clustering. In *IAPR International Workshop on Document Analysis Systems*, pages 95–99, 2012.

[10] L. Likforman-Sulem, A. Hanimyan, and C. Faure. A Hough based algorithm for extracting text lines in handwritten documents. In *International Conference on Document Analysis and Recognition*, volume 2, pages 774–777, 1995.

[11] M. Liwicki, E. Indermuhle, and H. Bunke. On-line handwritten text line detection using dynamic programming. In *International Conference on Document Analysis and Recognition*, volume 1, pages 447–451, 2007.

[12] G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis. Text line detection in handwritten documents. *Pattern Recognition*, 41(12):3758–3772, dec 2008.

[13] U.V. Marti and H. Bunke. Using a Statistical Language Model to Improve the Performance of an HMM-based Cursive Handwriting Recognition System. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(01):65–90, 2001.

[14] 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*, 28(4):590–604, April 2010.

[15] R. Saabni and J. El-Sana. Language-Independent Text Lines Extraction Using Seam Carving. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 563–568, 2011.

[16] Z. Shi and V. Govindaraju. Line Separation for Complex Document Images Using Guzzy Runlength. In *International Workshop on Document Image Analysis for Libraries*, pages 306–312, 2004.

[17] K.Y. Wong and F.M. Wahl. Document analysis system. *IBM Journal of Research and Development*, 26:647–656, 1982.