Multi-Skewed Text Line Extraction from Handwritten Documents using Seam Carving

Shubham Talbar and Pranali Pawar

July - December 2016

1 Abstract

A very important step in the handwriting recognition process is that of text line extraction: it aims at extracting individual text lines from document images. In this work, we implement a novel text line extraction algorithm for multi-skewed text line documents. The algorithm used is based on seam carving to compute separating seams between text lines.

2 Introduction

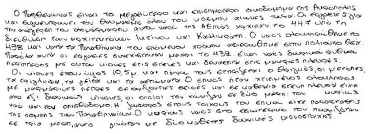
The large collections of handwritten historical manuscripts existing in libraries, museums, and private houses around the world are valuable human heritage. The rising interest in these collections and the recent effort to digitize them reveal interesting problems, which call for theoretical and applied research in Historical Document Image Analysis. These include document image binarization, writer identification, page layout analysis, keyword searching, indexing, and script recognition. These procedures are essential in helping scholars easily access and analyze digital copies of historical documents. However, the low quality of these document images, the lack of constraints on page layout, and the complexity of handwriting, pose real challenges for processing such document images automatically.

In our project we implement a language independent global method for automatic text-line extraction[2]. The algorithm used computes an energy map of the input text block image and determines the seams that pass across text lines. The crossing seam of a line, l, marks the components that make the letters and words along l.

3 Our Approach

Our proposed method consists of two stages:

1) *Medial* seam computation using a projection profile matching approach.



(a) Input Image

(b) Energy Image

2) *Separating* seam computation using a modification of the seam carving procedure.

In the following two sections we describe these two stages in detail. We use the convention that an image $I \epsilon R^{n \times m}$ to gray scale 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 at the upper left corner of the image.

A. Medial Seam Computation

Our medial seam computation method is inspired by the projection profile matching approach of [1]. We split the page vertically into r slices, each one of width $w = \lfloor m/r \rfloor$. We apply the Sobel operator to I to compute its edge image $S \epsilon R^{n \times m}$. We calculate smoothed horizontal projection profiles $P_g^c of S$ in each slice independently:

$$P_i^c = \sum_{j=k}^{k+w-1}, P^c = (P_i^c)_{i=1}^n, P_g^c = g(P^c),$$

$c = 1, ..., r, k \epsilon \{1, 1+w, ..., 1+(r-1)w\}$ (1), where g is a cubic spline smoothing filter. We denote the local maxima locations of the c-th profile by

$L_h^c; h = 1, ..., l$ and those of (c+1)-th by $L_{h'}^{c+1}; h' = 1; , , , l'$ Here, l and l' denote the, potentially different, number of maxima found at profiles c and *c+1* respectively. For each maximum location of profile c, we find the closest maximum location of profile *c+1* and for each maximum location of profile *c+1*, we find the closest maximum location of profile c:

$$match(L_h^c) = arg \min_{L_{h'}^{c+1}} |L_h^c - L_{h'}^{c+1}|, h = 1, ..., l, (2)$$

$$match(L_h^{c+1}) = arg \min_{L_h^{c+1}} |L_{h'}^{c+1} - L_h^c|, h' = 1, ..., l'. (3)$$

If the above matched locations in (2) and (3) agree, they are connected with a line. The above procedure is repeated until all slices are processed. The text line locations can now be represented in matrix form $L_{h,j}; h = 1, ..., l; j = 1, ..., m$, where each element $L_{h,j}$ contains the i-th coordinate of the h-th line, and l is the final number of lines found. The proposed method creates piece-wise linear seams that approximate the medial axis of the text lines in the manuscript page. Any two consecutive seams define a region in which the seam carving computation is constrained. This constraint enforces the separating seam to

pass between two consecutive text lines, and thus, it prevents it from assigning text parts to wrong lines.

B. Separating Seam Computation

We adapt the seam carving algorithm proposed in [3] to compute the separating seams. We include the regional constraints of the computed medial seams and modify the seam computation so that it can handle non-rectangular image regions. The energy map is the derivative image of the grayscale manuscript page:

$$E_{ij} = |\frac{I^{\sigma}_{i,j+1} - I^{\sigma}_{i,j-1}}{2}| + |\frac{I^{\sigma}_{i+1,j} - I^{\sigma}_{i-1,j}}{2}| (4)$$

where I^{σ} is the original grayscale image smoothed with a Gaussian filter of standard deviation σ. On this map, high-energy regions correspond to text components and low-energy regions correspond to parchment background. Let us denote the energy map between two text lines by $E_h = E_J$, where J is a two-dimensional grid of width m, where the j-th column contains all the intermediate i coordinates between two text line locations, that is, $\mathbf{J}_j = \{L_{h,j}, ..., L_{h+1,j}\}^T, h = 1, ..., l-1, j = 1, ..., m$. A seam that passes horizontally through an image grid E_h can be defined as

$$S_h = \{S_{h,j}\}_{j=1}^{m} = \{(y_h(j), j)\}_{j=1}^{m}, |y_h(j) - y_h(j-1)| \leq 1, y_h(j) = L_{h,j}, ..., L_{h+1,j} (5)$$

where $y_h : [1, ..., m] \rightarrow [L_{h,j}, ..., L_{h+1,j}]$.

The seam computation is done using dynamic programming in a similar way to [1]. We look for the optimal seam in the image grid E_h that minimizes the following constrained optimization problem:

$$s^*_h = arg\min_{s_h} \sum_{j=1}^{m} E_{s_{h,j}}, s.t. L_{h,j} \leq L_{h+1,j} (6).$$

The first step is to traverse the image grid Eh from left to right and to compute the cumulative minimum energy M for all possible connected seams for each pixel location $(y_h(j), j)$:

$$\mathbf{M_{y_h(j),1} = E_{y_h(j),1}, M_{y_h(j),j} = E_{y_h(j),j} + \min\left\{ M_{y_h(j)-1,j-1}, M_{y_h(j),j-1}, M_{y_h(j)+1,j-1} \right.} (7)$$

The minimum value of the last column in M will indicate the end of the minimal connected horizontal seam. In the second step we traverse the cumulative energy M backwards to find the path of the optimal seam. The above procedure is repeated for each image grid E_h, until the whole manuscript page is processed.

C. Parameter Selection

The parameters of our algorithm are the number of slices r for the medial seam computation, the smoothing parameter b of the cubic spline filter (function *csaps* in MATLAB) and the standard deviation σ of the Gaussian filter for the gradient image computation. In Table I we show the selected values for the above

parameters on the applied datasets. There is no automatic way to tune these parameters, because they depend on the type of manuscript under investigation. Different parameters were used inside the collections due to the different type of pages contained in them.

The standard deviation σ does not heavily affect the algorithm's accuracy. A positive value can be used when the manuscript images contain some amount of bleed-through noise, which can result in a more robust separating seam computation. The number of slices r depends on the image resolution and text layout. A value of r = 4 works relatively well for an average manuscript page. In the case of Aline, the value of r = 8 is used due to the higher resolution of the image and the different layout: many text lines span only part of the page width. The smoothing parameter b depends on the handwriting and script complexity. Heavy smoothing would create fewer local maxima, resulting in merged text lines. On the other hand, insufficient smoothing would create additional medial seams between text, resulting in nonrobust text lines.

4 Experimental Results

We have evaluated our algorithm on some images of ICDAR 2013 Handwriting Segmentation Contest data set including English, French, Greek and German.

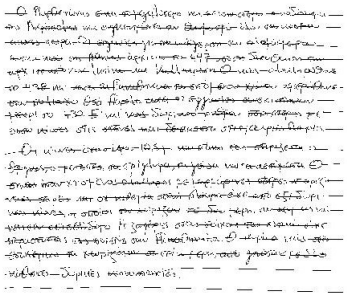


Figure 2: Seam lines on Greek document

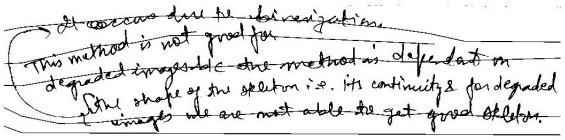


Figure 3: Seam lines on English document

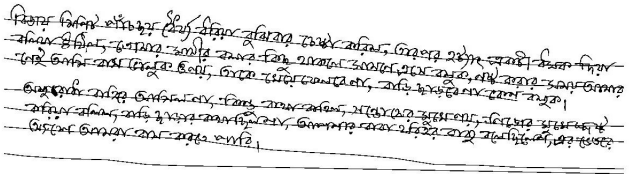


Figure 4: Seam lines on Gurumukhi document

5 Future Work

In future work, we would like to improve our energy accumulation process to reduce the computation time. Moreover, we will improve the performance of splitting large components which touch multiple text lines. [1]

References

[1] M. liwicki, e. indermuhle, h. bunke, "on-line handwritten text line detection using dynamic programming", international conference on document analysis and recognition, vol. 1, pp. 447-451, 2007.

[2] Nikolaos arvanitopoulos, sabine süsstrunk,"seam carving for text line extraction on color and grayscale historical manuscripts".

[3] S. avidan, a. shamir, "seam carving for content-aware image resizing", acm transactions on graphics, vol. 26, no. 3, pp. 10, 2007.