INT J COMPUT COMMUN, ISSN 1841-9836
9(1):8-15, February, 2014.

Handwritten Documents Text Line Segmentation based on Information Energy

C.A. Boiangiu, M.C. Tanase, R. Ioanitescu

Costin-Anton Boiangiu*, Radu Ioanitescu
"Politehnica" University of Bucharest
Romania, 060042 Bucharest
*Corresponding author: icostin.boiangiu@cs.pub.ro

Mihai Cristian Tanase
VirtualMetrix Design
Romania, 060104 Bucharest
mihaicristian.tanase@gmail.com

Abstract: The first step in the text recognition process is represented by the text line segmentation procedures. Only after text lines are correctly identified can the process proceed to the recognition of individual characters. This paper proposes a line segmentation algorithm based on the computation of an information content level, called energy, for each pixel of the image and using it to execute the seam carving procedure. The algorithm proposes the identification of text lines which follow the text more accurately with the expected downside of the computational overhead.
Keywords: text line segmentation, text recognition, information energy, OCR.

1 Introduction

The identification of the boundaries of the lines of text represents an essential step in many algorithms like the ones for document structure extraction or text recognition. The research in this field has focused mostly on the development of such algorithms for printed documents. This limitation of the domain of application reduces the complexity of the problem as the printed documents are perfectly formed and the main problem that would need to be solved is the skew angle introduced in the process of printing or scanning, angle which is assumed to be the same for the entire document,. With such documents, the problem is reduced to the identification of the skew angle which is assumed to be constant for the entire page because the text lines are parallel with each other. Such methods are presented in [1]- [3].

However, when dealing with handwritten documents, the assumptions made by such algorithms do not hold anymore. There is no constant skew angle, the lines are not parallel and even the size and format similarity between the same characters found on different areas of the page cannot be assumed. Even worse, the separation between lines cannot be assumed as for printed documents because, in handwritten tex, the characters often overlap the line bellow as they are more compactly spaced on the vertical. Algorithms trying to address these increased complexities have also been attempted in [4] - [13].

The pixel of a given document stored as a digital image have different levels of information. A white pixel in a big white region or one that is found between two lines of text contains very little information while a character defining pixel would contain a lot of information. This paper proposes the association of an energy level for each pixel of the input image which tries to estimate the importance or the amount of information provided by that pixel. Although the concept can be applied to general images, printed or handwritten text, we will apply this concept for handwritten documents. The information level is then used by the seam carving algorithm for the segmentation of the text lines.

2 Related Work

The algorithms that try to locate the text lines in a document are divided mainly by the information they take from the input document. As the input documents are the results of a digitization process, they are acquired, generally through scanning, as grayscale images. This grayscale representation is converted into binary or black and white for algorithms which are designed to work with this type of documents only. The binary representation conversion is done using a previously defined threshold level. All pixels that have gray levels above the threshold are converted to black, the rest are converted to white. A too high threshold will results in an image containing too little information for text recognition to be possible and a too low threshold will result in too many artifacts, again making the text recognition process impossible.

Based on the observation that the body of the lines of text contains gray pixels and because the pixels that make the characters being of a darker shade of gray, some algorithms compute projection profiles representing the sum of all the pixels values in a given direction. The method works well for printed documents but fails to produce good results when applied to handwritten documents.

To address these problems, different approaches were used: identify the local skew of the handwritten text, calculate the accumulated space between characters, try to fill the space between characters or use attraction from the text pixels and repulsion from the previously detected line trying to estimate the text line boundaries closer.

3 Information Energy

The different pixels from an image carry different information content. A document can be viewed as a group of low information pixels representing the space between lines of text and respectively a group of high information pixels representing the actual lines of text. Each pixel in the energy map has a value associated with it that represents the amount of information that the given pixel stores in the image. If a high energy pixel is removed from the image, the resulting image has a significant drop in detail, whereas removing a low energy pixel results in a negligible information loss.

The information energy concept can be understood by trying to eliminate a continuous band of pixels from an image representing handwritten text. This concept is illustrated in 1.

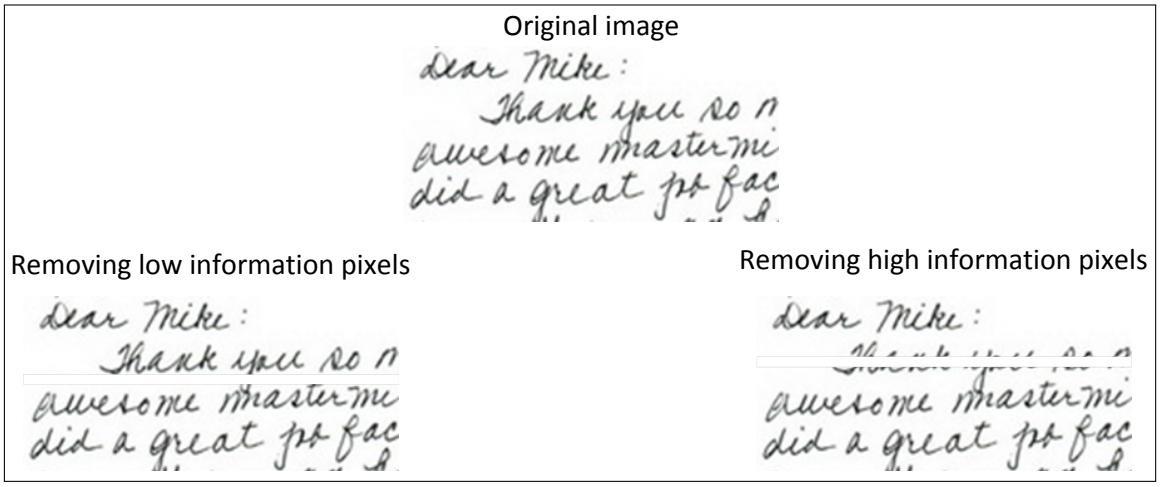

Figure 1: Information energy exemplification

A band with the same amount of pixels has been removed from a text document. It can be seen that when removing high information pixels there is not enough remaining information to detect the content, while when removing low information pixels, there is hardly any information loss and the content can be usually detected entirely.

An important observation is that for handwritten documents and in general text documents, the variation of pixel values represent information in itself. It can be seen that large continuous areas with similar pixel values constitute low energy areas, in particular spaces between lines of text, while high variation pixel values constitute high energy area. Computing the energy map of the entire document will thus provide information of the location of the high energy areas which represent the lines of text.

4 Accounting for the Text Direction

The algorithm begins with the calculation of the information energy for each pixel. Similarly to [4] and [11], for each pixel, the energy value is calculated using the next formula:

$$E(i,j) = 2 * e(i,j) + min\left(d(\frac{neighborsNumber}{2} + k) * E(i + direction, j + k)\right) \quad (1)$$

where:

- k fulfills the following condition: $-\frac{neighborsNumber}{2} < k < \frac{neighborsNumber}{2}$
- direction represents the direction of processing the energy map +1 representing left to right processing and -1 the opposite direction.

Using a neighborsNumber value of 3 and because k is a natural number since we work with discrete pixel positions, we get only the immediate neighbors. During the calculation the direction coefficient has been kept constant. As shown in images from the chapter "Test and Results" section, for documents with horizontal lines or very low skew angles we obtained good results. However, for larger skew angles the results quality decreased with wrong text line segmentation and as a result we considered accounting for the text direction during processing.

To accurately follow the text, the direction of the text lines should be taken into consideration at each pixel. For each pixel, the direction coefficient that accounts for the direction of the text line when that given pixel is reached, will be calculated.

The calculated information energy map is presented in 2 as a result from interpolating the two processing directions.

Computation algorithm

Input: image to be processed, processing window (height*width), skewing angle

For every pixel in the image to be processed:

1. For every processing window skewing angle

i. Sum up the pixels in the skewed window

ii. If the sum is less than the current minimum then

a) Update the current minimum

b) Save the window variation

2. With the minimum variation saved in step 1 ii:

i. Update the direction coefficient for the minimum variation

ii. Rescale the minimum window variation

Good results were obtained experimentally with processing window sizes that were at least the width and height of the average character in the processed text and for values of a few order

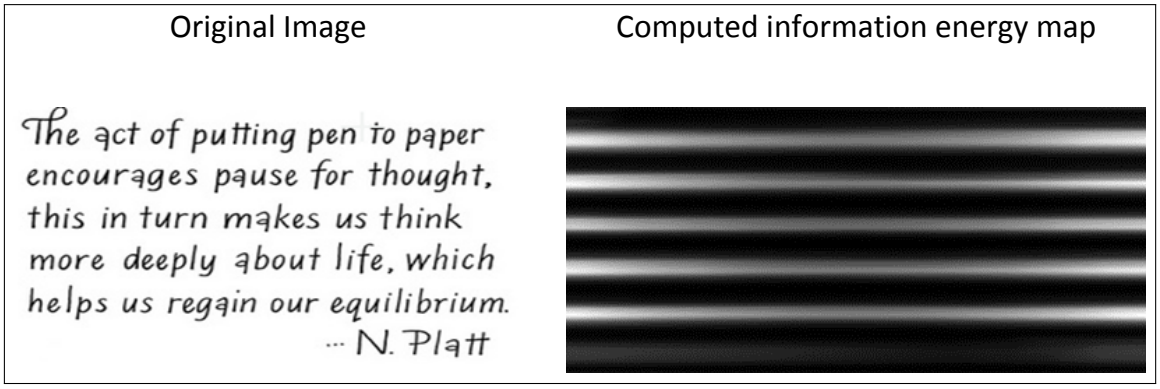

Figure 2: The information energy map of an image

of magnitude larger. If the processing windows has a smaller size then the algorithm attempts to segment the characters specific structures as separate lines of text leading to erroneous results.

Because the size of the processing text is usually roughly known and since the results are good with larger sizes for the processing window, this aspect does not represent a limitation. Similarly, the minimum and maximum variation angles can be set to 20 degrees. These variation angles represent the maximum skewing angles and combined cover a 40 degrees area that is sufficient for most documents.

The size of the processing window is fixed upfront, the calculations at each step are fixed. The total number of operations depends on the direction coefficient and height of the initial image since the algorithm is repeated for every pixel. For each pixel and for each variation of the skewing angle, the information energy level is computed.

Line identification algorithm

Input: computed information energy for each pixel in the image to be processing, neighborsNumber

Output: the number and layout of the identified lines

1. For each line (1..h)

2. For each column (1..w)

3. Select the minimum cost pixel on the right not found further away than neighborsNumber

i. If the selected pixel is not included in any line include it

ii. Else move to the next line

The algorithm looks through all the information energy levels to locate the minimal values. The values with the minimum energy level represent the blank pixel regions that separate the lines of text.

5 Tests and Results

To test the algorithm, a number of images of handwritten documents have been used. The test data files consisted of about 500 different types of documents representing old letters, library index files, patents, receipts and various printed documents. The documents showed pronounced skew angles and their layout was not trivial. The database was considered to be relevant, although the number of documents is not large, because of their variety which allowed the testing of the algorithm on a wide set of conditions.

Different methods for the computation of the information energy are used as examples to show how the algorithm depends on this type of variation and to show the general application

of the concept. The results of the algorithm are discussed in the conclusions section and future possible solutions are presented with an explanation of the associated computational costs.

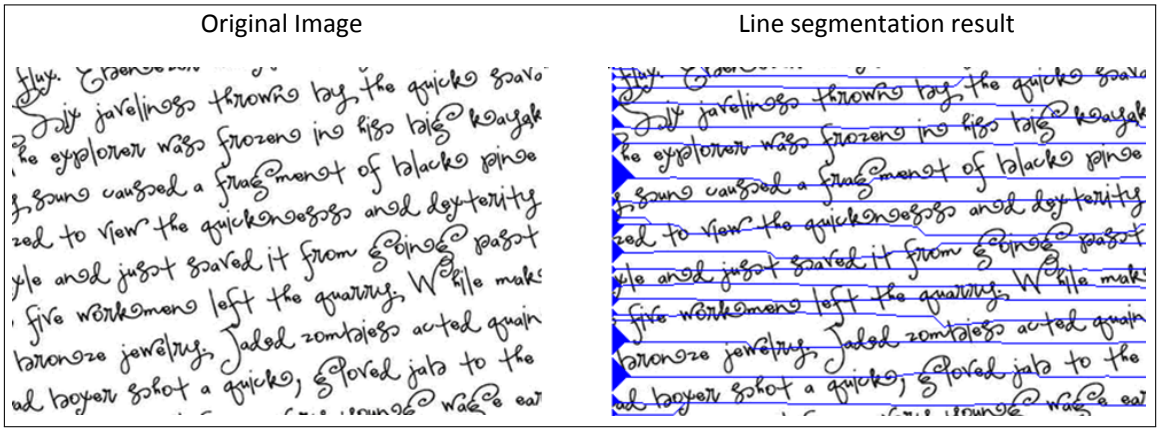

Figure 3: Gaussian first derivative energy computation Test 1

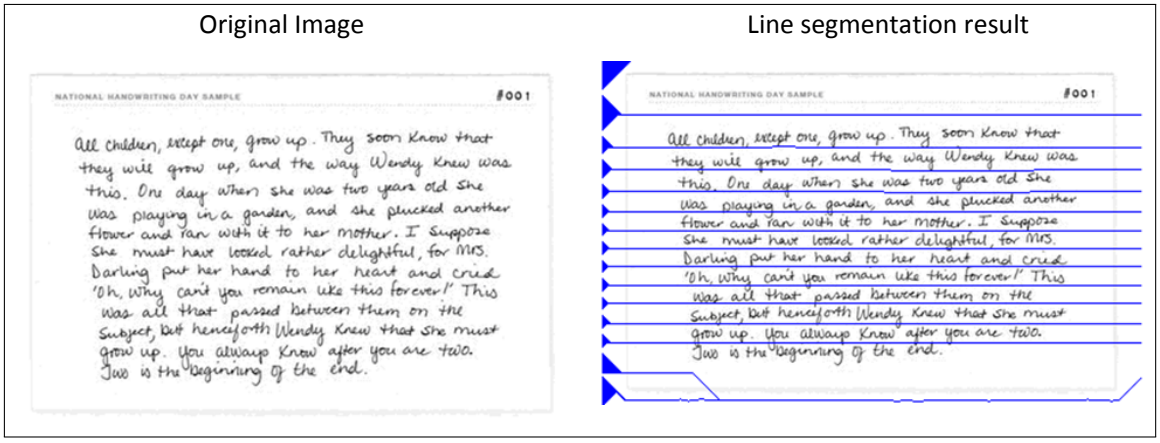

Figure 4: Gaussian first derivative energy computation Test 2

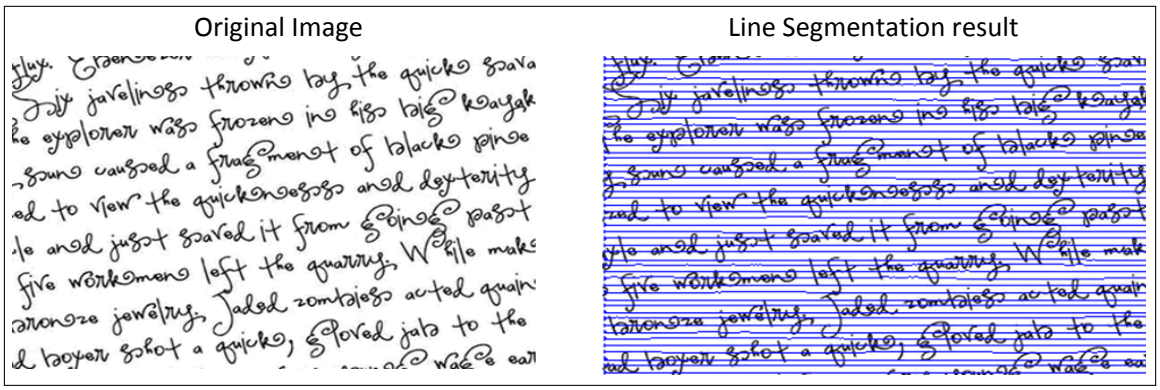

Figure 5: Magnitude of the gradient energy computation Test 1

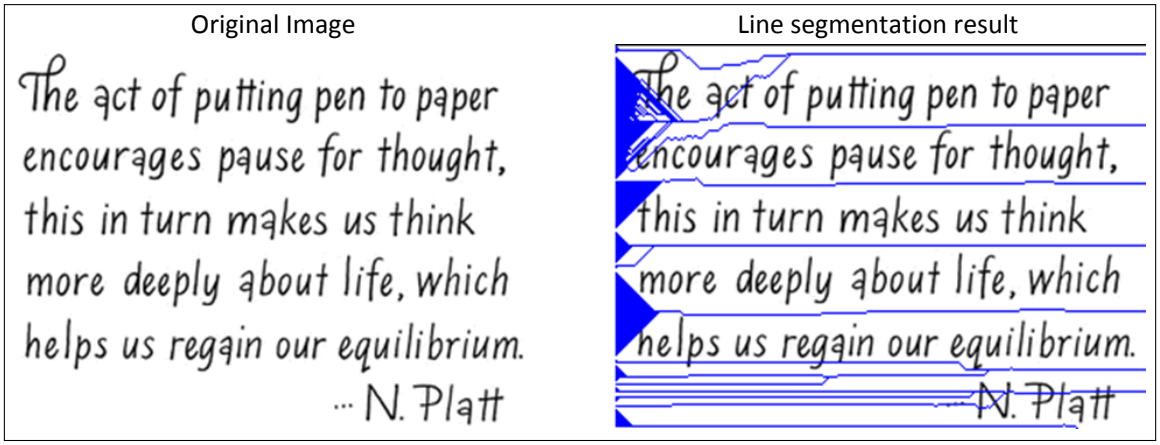

Figure 6: Magnitude of the gradient energy computation Test 2

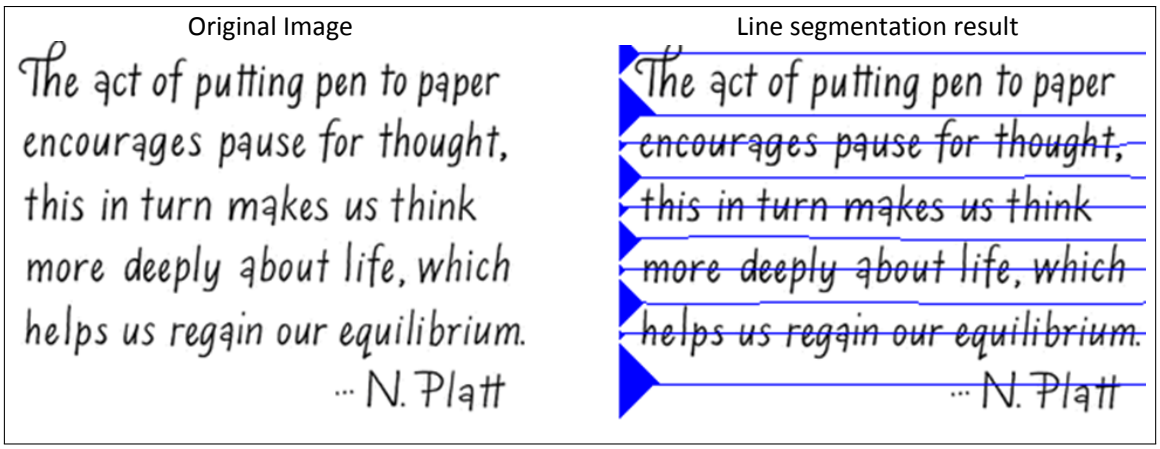

Figure 7: Inverse Distance Transform energy computation 1

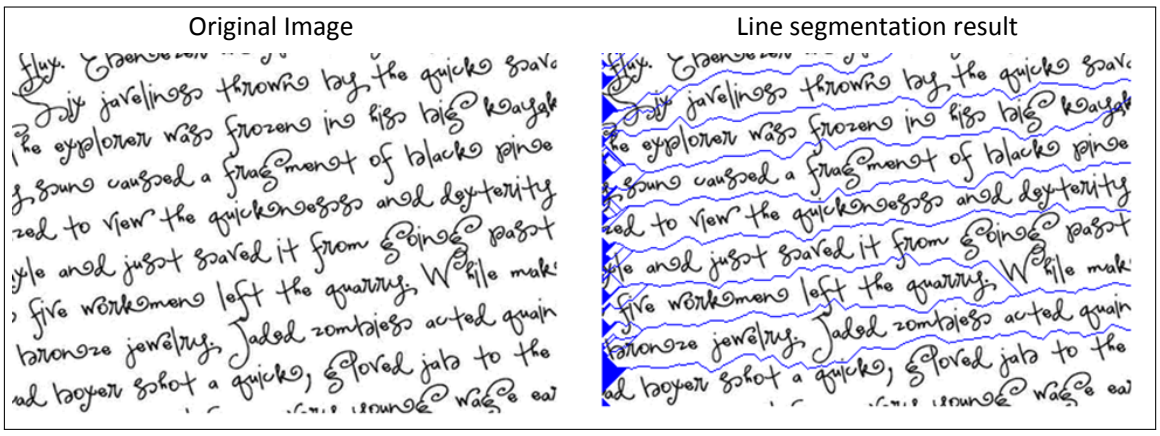

Figure 8: Inverse Distance Transform energy computation Test 2

6 Conclusions and Future Works

This paper describes a text line segmentation algorithm that shows good results even for handwritten documents. The algorithm is based on the concept of information energy which is used to estimate the text lines in the processed document.

Experimental results showed that using constant direction coefficients when calculating the information energy levels produces bad results for inputs that show high levels of skew. The algorithm addresses this problem by updating the direction coefficients with a window processing algorithm. These direction coefficients account for the local text direction and for variations of the skew angles that are not uncommon in handwritten documents. By using direction coefficients, pixels from the same line have higher probability of selection when calculating the minimum values in the information energy map.

The algorithm that computes the direction coefficient has a large computation complexity. Since the computation time depends on the size of the processing window, one solution is to use a smaller size. Experimental testing showed that at minimum, the height and width of the average character should be used for the size of this window. Alternatively the skew angle could be assumed to be smaller eliminating parts of the computation cases. Another change could be the limitation of the possible line curvature angle for the detected document lines to a lower interval. By constraining the line curvature angle. This limitation would also address cases in which a detected line would follow the space between two words and reach the previous or the next line which would represent a wrong line detection.

A possible future direction is the evaluation of the robustness of the algorithm on larger images datasets and with a variation of the document types which would allow to more extensively evaluate the accuracy of the handwritten text segmentation algorithm.

The work presented in this paper is a building block of a much bigger project: a complete, modular, fully automatic content conversion system developed for educational purposes. In the near future, with the completion of the system and the running in automatic batch processing of large image databases of all kind of skewed documents (containing handwriting or not) the algorithm will be fully evaluated in order to assess its real potential as a preprocessing phase for OCR applied on handwritten documents.

Acknowledgement

The work presented in this paper was funded by the Sectorial Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the financial agreement POSDRU/89/1.5/S/62557.

Bibliography

[1] dos Santos, R.P. et al, Text Line Segmentation Based on Morphology and Histogram Projection, *Document Analysis and Recognition (ICDAR)*, 651- 655, 2009.

[2] Saha, S. et al, A Hough Transform based Technique for Text Segmentation, *Journal of Computing*, 2(2):135-140, 2010.

[3] Arivazhagan, M. et al, A Statistical approach to line segmentation in handwritten documents, *Proceedings of SPIE*, 2007.

[4] Strand, L. et al, Minimal Cost-Path for Path-Based Distances, *Image and Signal Processing and Analysis*, 379-384, 2007.

[5] Avidan, S. et al, Seam Carving for Content-Aware Image Resizing, *ACM Siggraph*, article 10, 2007.

[6] Saabni, S. et al, Language-Independent Text Lines Extraction Using Seam Carving, *Document Analysis and Recognition (ICDAR)*, pp. 563-568, 2001.

[7] Papavassiliou, V. et al , Handwritten document image segmentation into text lines and words, *Pattern Recognition*, 43(1):369-377, 2010..

[8] Du, X. et al, Text Line Segmentation in Handwritten Documents Using Mumford-Shah Model, *Pattern Recognition*, 42(12):3136-3145, 2009.

[9] T ripathy, N.; Pal, U., Handwriting segmentation of unconstrained Oriya text, *Frontiers in Handwriting Recognition*, 306-311. 2004.

[10] Kennard, D.J., Barrett, W.A., Separating Lines of Text in Free-Form Handwritten Historical Documents, *Document Image Analysis for Libraries*, 12-23, 2006.

[11] Asi, A. et al, Text Line Segmentation for Gray Scale Historical Document Images, *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*, pp. 120-126, 2011.

[12] Bar-Yosef, I., Input sensitive thresholding for ancient Hebrew manuscript, *Pattern Recognition Letters*, 26(8):1168-1173, 2005.

[13] Bar-Yosef, I. et al, Line segmentation for degraded handwritten historical documents, *Document Analysis and Recognition*, 1161-1165, 2009.