


## DOCUMENT LINE SEGMENTATION BASED ON WAVELET TRANSFORM

Lasko LASKOV

**Abstract:** In this paper a novel approach to document lines segmentation is presented. The algorithm is based on wavelet transform of the horizontal projective profile of the document image. The projective profile is examined as a one-dimensional signal which is decomposed using the pyramidal wavelet algorithm up to a precise level, where local minima and maxima are discovered. These local extrema, projected into the input signal, correspond to the blank spaces between document lines and to the pivots of the lines. The method is tested to a broad set of printed and handwritten documents and has proved to be stable and efficient.

**Keywords:** *document image processing, document segmentation, wavelets.*

**ACM Classification Keywords:** *I.7 Document and text processing.*



## Introduction

During the past few decades optical character recognition (OCR) has been considered a solved problem in the case of standard texts written in contemporary alphabets. A number of commercial and open sources software exist, designed to process document images which contain texts in various languages. On the other hand, the standard software available is not applicable to a number of problems which emerge from different scientific fields. Such problems are historical documents image processing, handwriting recognition, mathematical formulae recognition, etc.

A concrete example is the neume writing recognition in historical documents. The problem of computer processing of this ancient notation emerge from the scientific fields of history, musicology, theology and leads to various nonstandard algorithms for document image binarization, segmentation and symbol recognition.

Another example is the recognition of handwritten digits in astronomical logbooks containing metadata of astronomical photographic plates. This problem emerges from the need of creation of digital database of astronomical plates where the stage which slows down the process of digitalization is actually the process of metadata extraction from the logbooks.

These examples and many others are much different in nature but they are consisted of certain common stages like:

- document image binarization – the process of object pixels separation from the background pixels;
- document segmentation – the process of document structure analysis and lines, and symbols extraction;
- classifier design and learning, and symbols recognition.

In this paper, a novel approach for document lines segmentation is presented which is applicable in various nonstandard OCR applications. The algorithm is based on wavelet transform and is tested on a number of completely different sets of input images (standard printed texts, medieval manuscripts, astronomical logbooks, notebooks containing handwritten text) to verify its robustness and effectiveness.

This paper is organized as follows: in the next section a brief description of the problem and related work are given; next the algorithm description is presented, followed by experimental results, and finally conclusions and some directions for future work are discussed.

## **Problem Description and Related Work**

The text line segmentation is the stage of an OCR system in which the text lines are discovered and extracted from the document image. It is an important step since the consequent symbol extraction is highly dependent on it. A number of methods for document line segmentation exist and many of them are based on projective techniques [Papavassiliou, 2010]. Also, the Hough transform [Duda, 1972] in its  $(\rho, \theta)$  version is often adopted as a method for document line skew detection based on a voting scheme.

In this paper it is assumed that a single paragraph is analyzed and that the document lines are relatively straight and horizontal. This allows the application of horizontal projective profile for document vertical structure analysis. If this condition is not fulfilled, the Hough transform can be adopted to discover the angle in which the text lines are rotated.

The other assumption is that the document images are binarized in advanced for example with one of the techniques [Otsu, 1997], [Kittler, 1986], [Sauvola, 2000], or other, and the images being examined are composed of black (object) and white (background) pixels. It has to be noted that the images can contain various types of noise due to the

specific characteristics of the nonstandard documents like age, paper or parchment degradation and document image acquisition. Then the goal of the line segmentation algorithm is to discover the pivot of each textual line and the space between the lines, despite the low quality of the images and noise.

One possible solution is given in [Laskov, 2008] where the problem of segmentation of neume notation in ancient manuscripts has been examined. The method is based on the horizontal projective profile of the document image and a floating mean filtering to remove the “false” local minima and maxima leaving only those which correspond to the document lines.

Given a binary image  $I_{M \times N}(x, y)$  with  $M$  rows and  $N$  columns the horizontal projective profile is defined:

$$h(y) = \sum_{i=0}^{N-1} I(i, y), y = 0, 1, \dots, M - 1. \quad (1)$$

The floating mean filter is the one-dimensional discrete integrating filter  $\Phi_m(y)$ , defined:

$$\Phi_m(y) = \begin{cases} 1, & |y| \leq m/2 \\ 0, & |y| > m/2 \end{cases} \quad (2)$$

where  $m$  is the domain width. Then the filtered projective profile is expressed by the convolution:

$$\tilde{h}(y) = (\Phi_m \circ h)(y) = \frac{1}{m} \sum_{i=-m/2}^{m/2} h(y+i), y = 0, 1, \dots, M - 1. \quad (3)$$

Having denoted the number of local minima of  $\tilde{h}(y)$  with  $n$ , then  $S(m): m \rightarrow n$  is a function which gives the correspondence between

the filter domain width and the number of local extrema of the filtered profile. Since the extrema which correspond to the text lines dominate over the noise extrema, it can be expected that when  $\tilde{h}(y)$  is smooth enough,  $S(m)$  will have nearly constant value and  $\tilde{h}(y)$  will contain only the minima and maxima being of interest, and the document lines are discovered.

Nevertheless this method gives relatively good results, especially in the case of the special type of manuscripts being its aim, it has the following drawbacks:

- $m$  is not known in advanced which leads to a iterative search of the correct filter domain width;
- the number of iterations is highly dependent on the image size and average text line width;
- as a result of the filtering process, the method is not stable in the case of thin text lines and it can recognize more than one line as a single line.

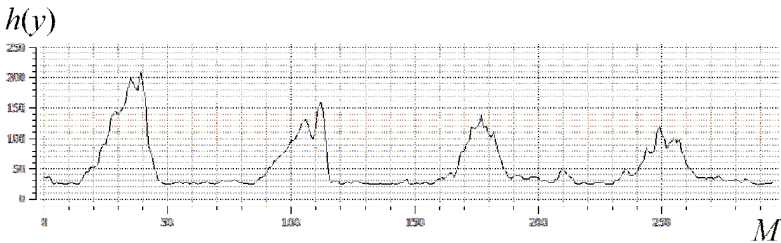
The above disadvantages are the motivation for the following approach.

## The Proposed Method

The method which is proposed here is based on the horizontal projective profile of the input image (1) (see also Fig.1(a)) and on its wavelet decomposition. The intuition behind this approach is that the wavelet decomposition at a certain level represents a compressed version of the input signal and contains its major characteristics. Also, the approximation part of the wavelet decomposition is a denoised projective profile and it can be expected that the local minima and maxima which correspond to noise in the image are removed.

89131	L	240	16	10	-27.5	14	21	1	49	-47.5	E	0	-
89141	"	60	"	"	-2.5	17	0	0	55	-2.5	W	"	-
89151	"	"	17	30	-52.5	19	09	1	39	-52.5	"	"	-
89161	"	"	17	50	-27.5	20	48	0	58	-27.5	"	"	-

(a)



(b)

**Figure 1** (a) Binary image containing a fragment of an astrological logbook; (b) and its horizontal projective profile  $h(y)$ .

After the accumulation of the projective profile  $h(y)$  its discrete approximation at resolution  $2^j$  is calculated [Mallat, 1989], which is given by the convolution:

$$A_{2^j}^d h = \left\{ (h(u) \circ \phi_{2^j}(-u)) (2^{-j}n) \right\}_{n \in \mathbb{Z}}, \quad (4)$$

where  $\phi(x)$  is a scaling function or a low-pass filter, and  $\left\{ \sqrt{2^{-j}} \phi_{2^j}(x - 2^{-j}n) \right\}_{n \in \mathbb{Z}}$  is an orthonormal family of functions. In practice (4) is calculated using the pyramidal algorithm proposed by Mallat where on each step the signal is filtered with a low-pass filter and

after that is downsampled at rate  $2^j$ . In the experiments with the method,  $\phi(x)$  is defined by the orthogonal Daubechies coefficients [Daubechies, 1992] and the resolution of decomposition is  $2^{-3}$ .

After calculation of  $A_{2^j}^d h$  at the needed resolution, the local minima and maxima which correspond to the document lines are found by applying a simple procedure for each triple of neighboring discrete samples of  $h(y)$ :

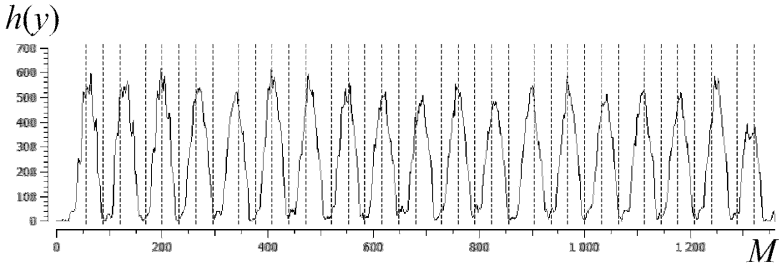
- $h(y_k)$  is a local minimum if  $h(y_k) < h(y_{k-1})$  and  $h(y_k) < h(y_{k+1})$ ;
- $h(y_k)$  is a local maximum if  $h(y_k) > h(y_{k-1})$  and  $h(y_k) > h(y_{k+1})$ .

## Experimental Results

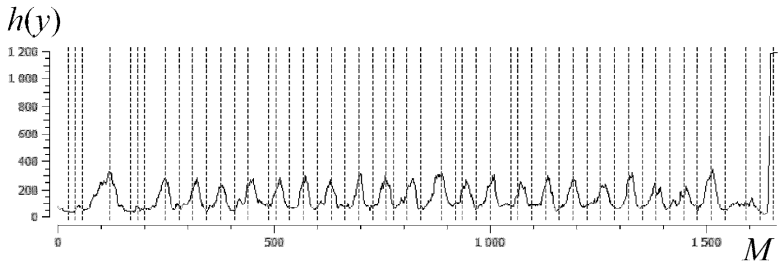
On Fig. 2 three examples of horizontal projective profiles are given with the local extrema discovered denoted with dashed lines. The images which were used for this illustration are representatives of three completely different types of documents.

Fig. 2(a) shows the profile of a standard printed document. The size of the image is 15558 x 1362 pix. and all the 19 lines contained are discovered by the algorithm.

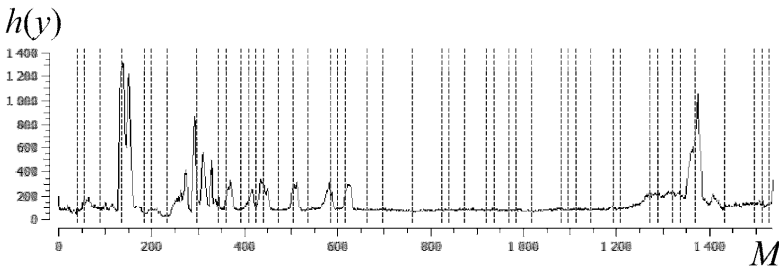
On Fig. 2(b) the projective profile of a notebook page containing handwritten text is given. The image size is 1224 x 1664 pix. and it contains 22 text lines and one line with the page number. The algorithm discovers all the text lines, the line with the page number and one false line due to the noise in the image. The false line can be easily discarded in the further processing using vertical projective profile.



(a)



(b)



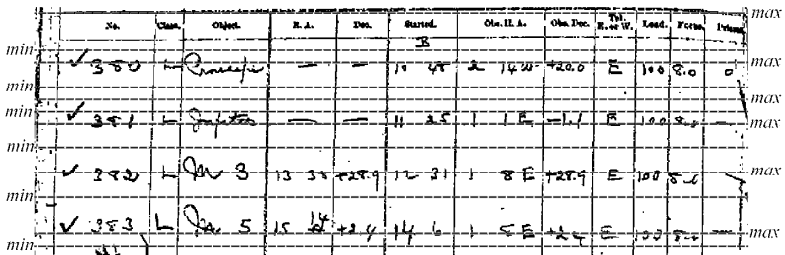
(c)

**Figure 2** Horizontal projective profiles and local extrema, represented by dashed lines, discovered by the proposed method: (a) standard printed text; (b) notebook page with handwritten text; (c) two concatenated pages from an astronomical logbook.

Fig. 2(c) contains the profile of two concatenated pages of an astronomical logbook. The structure of these documents is a table-like



and they contain straight lines which affect the projection profile if no preprocessing is performed. The image size is 2048 x 1536 but it contains only four text lines. Despite that the algorithm discovers many non-textual lines in this case, it segments all the textual lines, contained in this document (see Fig. 3). The lines which are not of interest can be rejected using vertical projection profiles or flood-fill algorithm.



**Figure 2** Fragment of the logbook whose profile is given on Fig. 2(c). The discovered local maxima and minima show that all the textual lines are segmented, in spite of the noisy image and non-standard structure of the document.

## Conclusion

In this paper a method for document lines segmentation is proposed which is based on wavelet decomposition of the horizontal projective profile of the image. The method is implemented as a part of document image processing system which is designed to process nonstandard document images. It is tested on a broad variety of different input data and it proved to be reliable even on a noisy, degraded images. The method proved to be stable with respect to the variety of image resolution and relative document image width.

As future work, a specific method for handwritten digits segmentation will be designed to be applied in astrophysical logbooks processing.

## Bibliography

- [Daubechies, 1992] I. Daubechies. Ten Lectures on Wavelets CBMS-NSF Regional Conf. Series in Appl. Math. 61, Society for Industrial and Applied Mathematics, 1992.
- [Duda, 1972] Richard O. Duda and Peter E. Hart. Use of the Hough transformation to detect lines and curves in pictures. Commun. ACM, 15, 1 (January 1972), p. 11-15.
- [Kittler, 1986] J. Kittler and J. Illingworth. Minimum error thresholding. Pattern Recognition, 1986. 19(1): p. 41–47.
- [Laskov, 2008] Laskov L., and D. Dimov, “Segmentation of Ancient Neumatic Musical Notation”, In: Proceedings of the Int.Conf. Automatics & Informatics’08, 01-04.10.08, Sofia, 2008. p. II.21–24.
- [Mallat, 1989] S. G. Mallat, A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. IEEE Trans. Pattern Anal. Mach. Intell., 1989, 11(7): p. 674-693.
- [Otsu, 1997] N. Otsu. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics, 1979. 9(1):62–66.
- [Papavassiliou, 2010] V. Papavassiliou, T. Stafylakis., V. Katsouros, G. Carayannis, Handwritten document image segmentation into text lines and words. Pattern Recognition, 2010. 43(1): p. 369-377.
- [Sauvola, 2000] J. Sauvola and M. Pietikainen. Adaptive document image binarization. Pattern Recognition, 2000. 33(2):225 – 236.

## Authors' Information



**Lasko LASKOV, PhD, New Bulgarian University, Informatics Department, llaksov@nbu.bg.**

**Major Fields of Scientific Research:** *Image processing, pattern recognition, machine learning.*