Serdica J. Computing **6** (2012), 101–114

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

APPLICATION OF WAVELET DECOMPOSITION TO DOCUMENT LINE SEGMENTATION

Lasko Laskov

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

1. Introduction. During the past several 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-source software products exist, designed to process document images which contain texts in various languages. On the other hand, the standard available software is not applicable to many problems which emerge from different scientific fields. Such problems are image processing of historical documents, handwriting recognition, recognition of mathematical formulae, etc.

One example is the recognition of neume writing in historical documents [7], [4]. The problem of computer processing of this ancient notation emerges

ACM Computing Classification System (1998): I.7, I.7.5.
Key words: document image processing, document segmentation, wavelets.

from and has various applications in the fields of history, musicology, theology. The efforts to design a software system able to process manuscripts which contain neume notation lead to various nonstandard algorithms for document image binarization, segmentation and symbol recognition [10].

Another example is the recognition of handwritten digits in astronomical logbooks containing metadata of astronomical photographical plates. This problem emerges from the need of creating a 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 [19], [20].

Another specific example is the collection of Bulgarian folklore texts, which are a subject of research in the project described in [17]. The texts are in handwritten form, in old notebooks dating from the 60s and 70s of the 20$^{\text{th}}$ century.

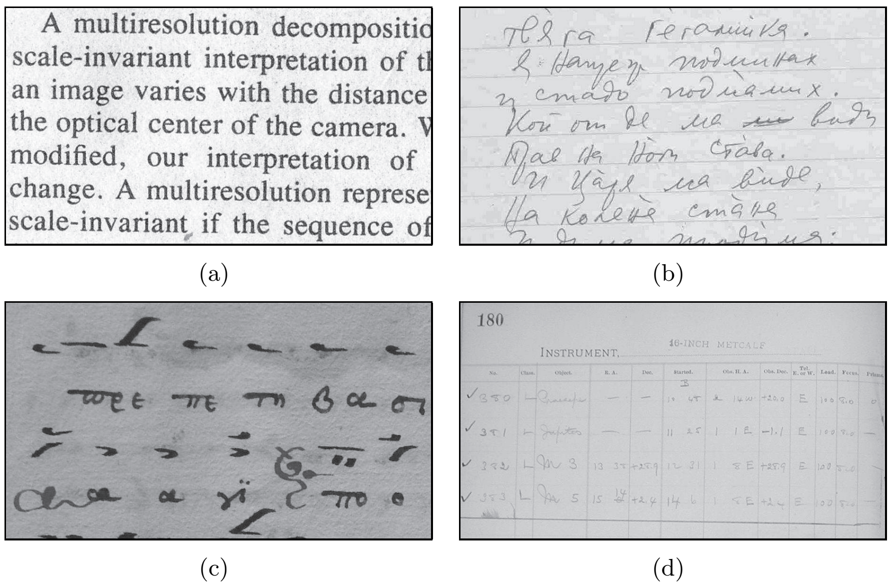


Fig. 1. Fragments from different document images used in the experiments: (a) standard printed document; (b) notebook with Bulgarian folklore text in handwriting; (c) medieval manuscript containing neume notation; (d) astronomical logbook

These examples and many others are very different in nature (see Figure 1) but their computer processing consists of certain common stages like:

- document image binarization—the process of separating object pixels from background pixels;

- document segmentation—the process of analysis of document structure and lines and extraction of symbols;
- classifier design and learning and recognition of symbols.

In the literature there are several methods for document line segmentation, based on different approaches like rescaling and convolution techniques, statistical techniques, Hough transform, etc. (for more details see Section 2). The method presented in this paper is based on another popular approach—wavelet transform [13].

Wavelet transform has numerous applications in many problems in the fields of signal processing and pattern recognition. In document image analysis it is usually adopted for extraction of textual areas (see for example [1]), or symbol feature space definition and classifier design (see [3]). On the other hand, the effectiveness of wavelet transform in tasks like signal compression and noise removal leads to the idea of developing a one-dimensional clustering technique that can be applied in the task of segmentation of document lines.

The above two observations give us the confidence to claim that the method presented in this paper is a novel approach, even though it is based on such a popular technique as wavelet transform. To verify its robustness and effectiveness, the developed algorithm is tested on a number of completely different sets of input images (printed texts, notebooks containing handwritten text, medieval manuscripts, astronomical logbooks).

This paper is organized as follows. In the next section a concise description of the problem is given and some existing approaches are mentioned. Section 3 discusses briefly a method for document line extraction, which is used both as a starting point and for comparison with the proposed approach. In Section 4 the proposed method is presented, followed by a detailed analysis of the experimental results in Section 5, and finally conclusions and some directions for future work are given.

2. Problem description. Structural analysis of a document image is the process of document structure extraction where paragraphs, text lines, separate words and symbols are discovered. Document structural analysis can be as complex as a document itself can be: it can be composed of textual parts, images, tables and other graphical structures. In particular, text line segmentation is the stage in which the text lines are located in the document image. It is an important stage, since the consequent word and symbol extraction is highly dependent on its accuracy. Text line segmentation can be a complex task by itself, especially in non-standard handwritten texts.

A number of methods for document line segmentation exist and they are based on very different approaches. For example, in [18] a method is proposed which is based on document image rescaling and convolution techniques. In [2] text lines are modeled as Gaussian densities and a statistical approach is adopted to extract the document lines. A popular approach is based on the (ρ, θ) version of the Hough transform [6], as the method described in [12]. Maybe the broadest class of methods found in the literature is based on the so-called *projective profile* of the image [16] because of the intuitive nature and applicability to a broad variety of text documents.

The method described here adopts the projective profile as well. Also, it is assumed that a single paragraph is analyzed and that the document lines are relatively straight and horizontal. This allows the application of the horizontal projective profile for analysis of the document's vertical structure. If this condition is not fulfilled, the Hough transform can be adopted to discover the angle to which the text lines are rotated.

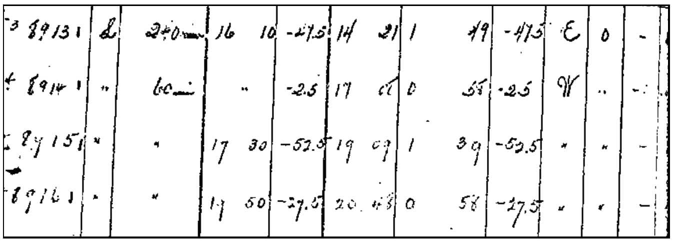
The other assumption is that the document images are binarized in advance, for example with one of the techniques [15], [9], [8], and the images examined are composed of black (object) and white (background) pixels. It has to be noted that the images can contain various types of noise due to specific 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 spacing between the lines, despite of the low quality of the images and noise.

3. Related work. An algorithm which segments text lines, based on the projective profile of the image, is given in [11], where the specific problem of segmentation of neume notation in ancient manuscripts is examined. The idea of the method which is proposed here is based on some observations on this approach. 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 which correspond to noise and parts of symbols crossing the line gaps, leaving only those that correspond to the document lines.

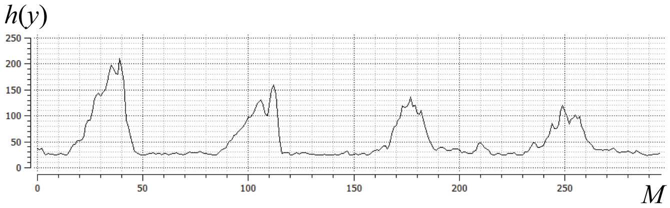
Given a binary image $I_{M\times N}(x, y)$ with M rows and N columns the horizontal projective profile is defined (see also Figure 2):

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

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



(a)



(b)

Fig. 2. (a) A binary image containing a fragment of an astrological logbook; (b) its horizontal projective profile $h(y)$

defined:

$$\Phi_m(y) = \begin{cases} 1, & |y| \le m/2 \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

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

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

If we denote the number of local minima of $\tilde{h}(y)$ with n, then $S(m) : m \to n$ 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 a nearly constant value and $\tilde{h}(y)$ will contain only the minima and maxima of interest, and the document lines are discovered.

Nevertheless this method gives relatively good results, especially in the case of standard printed texts and the special type of manuscripts that are its aim. It has the following drawbacks:

- m is not known in advance, which leads to an 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 with respect to image height, and it can recognize more than one line as a single line.

The above disadvantages are the motivation for the following approach.

4. The proposed approach. As in the case of the method described above, the approach which is proposed in this paper is based on an analysis of the horizontal projective profile h of the input image. Again, the objective of the algorithm is to detect the local extrema of h that correspond to the document lines, by separating the extrema of interest from noise local minima and maxima. The algorithm is based on the approximation $A^d_{2^j}h$ of h at a given resolution 2^j, $j \in \mathbb{Z}$, $j > 0$, performed by discrete wavelet decomposition.

The theoretical background that motivates the steps of the proposed algorithm relies on two basic and well-known features of the wavelet transform. The first one (see [13]) is that the approximated function $A^d_{2^j}h$ is the approximation at the resolution which mostly preserves the main characteristics of h. In other words, if g is another approximation of h at resolution 2^j, then

$$\|g(y) - h(y)\| \geq \|A^d_{2^j}h(y) - h(y)\|. \tag{4}$$

The other property that is in the basis of the described algorithm is the time-scale view of the input signal performed by the wavelet decomposition. The time information is not lost as in the case of Fourier transform (see [14]), which makes possible to locate the extrema of interest in the projective profile h even if the document lines are not uniformly distributed.

The first stage of the algorithm is the projective profile accumulation, as given in (1). The discrete approximation of h at resolution 2^j is then calculated [13], which is expressed by the convolution:

$$A^d_{2^j}h = \left\{ (h(u) \circ \phi_{2^j}(-u))\,(2^{-j}n) \right\}_{n\in\mathbb{Z}}, \tag{5}$$

where $\phi(x)$ is a scaling function, 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 (5) is calculated using the pyramidal algorithm proposed by Mallat where at each step the signal is filtered with a low-pass filter $\tilde{H}$ and after that is downsampled with factor 2.

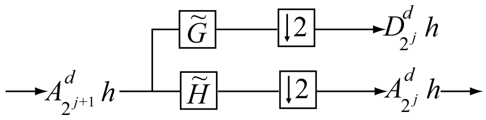


Fig. 3. One step of the pyramidal algorithm proposed by Mallat [13]. $A^d_{2^j}h$ is the approximation of $h(y)$ at resolution 2^j, $D^d_{2^j}h$ is the signal detail at resolution 2^j. $\boxed{\tilde{H}}$ and $\boxed{\tilde{G}}$ denote convolutions with the quadrature mirror filters $\tilde{H}$ and $\tilde{G}$ respectively, and $\boxed{\downarrow 2}$ denotes downsampling with factor 2

The connection between the scaling function $\phi(x)$ and the impulse response $\tilde{\chi}(n)$ of the filter $\tilde{H}$ is $\tilde{\chi}(n) = \langle \phi_{2^{-1}}(u), \phi(n-u) \rangle$, where $\langle \cdot \rangle$ denotes the inner product of two functions.

Before the pyramidal algorithm is applied, we check whether the projective profile $h(y)$ has a power of 2 number of samples. If not, zero padding is performed on the discrete signal $h(y)$ to fulfill this requirement. Then the signal is filtered with a low-pass filter $\tilde{H}$ and after that it is downsampled with factor 2. The result is the profile approximation at resolution 2^{-1}. At each step the filtering and downsampling is repeated in cascade manner with the approximated signal at resolution 2^{j+1} to obtain the approximated signal at resolution 2^j (see Figure 3).

Our experiments have shown that for the purpose of document line segmentation, the decomposition at resolution of 2^{-3} or 2^{-4} is enough. We have used quadrature mirror filters $\tilde{H}$ and $\tilde{G}$ defined by the orthogonal Daubechies coefficients [5]. Note that the filter $\tilde{G}$ and the signal detail $D^d_{2^j}h$ in Figure 3 are given only for completeness, since we are interested only in the signal approximation.

After calculation of $A^d_{2^j}h$ 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 adjacent discrete samples of $\bar{h}(y)$:

- $A^d_{2^j}h(y_k)$ is a local minimum if $A^d_{2^j}h(y_k) < A^d_{2^j}h(y_{k-1})$ and $A^d_{2^j}h(y_k) < A^d_{2^j}h(y_{k-1})$;
- $A^d_{2^j}h(y_k)$ is a local maximum if $A^d_{2^j}h(y_k) > A^d_{2^j}h(y_{k-1})$ and $A^d_{2^j}h(y_k) > A^d_{2^j}h(y_{k-1})$.

The discovered minima and maxima are projected back into the original profile $h(y)$ and they correspond to the discovered document lines.

Even though the implementation of the wavelet decomposition is brought to filtering of the signal and downsampling, this approach has major advantages over a simple filtering noise-removal procedure. First of all, as mentioned in the beginning of this section, it is ensured that $A^d_{2^j}h$ preserves the most significant characteristics of h. In contrast, if a mean filter is used instead, some of the important extrema of h may be lost, if the filtering procedure exceeds a certain number of iterations and that number of iterations is not known in advance.

The second advantage of the proposed approach is that after applying the filter $\tilde{H}$, downsampling is performed. At each stage of the pyramidal algorithm this halves the number of samples of $A^d_{2^j}h$, which makes it possible to apply the simple extremum detection procedure described above. This procedure has the important advantage that no *a priori* information on the processed discrete function is required.

5. Experimental results and error measures. The proposed method is tested and compared to the approach presented in Section 3, using four data sets of document image pages that have totally different characteristics: (i) printed document, containing standard text and mathematical expressions; (ii) notebook pages, containing handwritten text; (iii) medieval manuscripts, containing neume notation; (iv) astronomical logbooks. These four data sets are chosen because they represent four different non-standard problems in the field of document image processing that are not completely solved, or are not solved at all: (i) off-line recognition of mathematical formulae, (ii) off-line recognition of handwriting (in Bulgarian in this particular case), (iii) recognition of ancient manuscripts, and (iv) recognition of handwritten digits (in the context of non-standard documents).

To evaluate the algorithms' correctness, the standard error measures of precision, recall and F-measure are adopted. The output of a document line extraction algorithm is classified in four disjoint classes:

- *true positive*, t_p—number of lines correctly discovered;
- *true negative*, t_n—number of line spacings correctly discovered;
- *false positive*, f_p—number of false lines discovered;
- *false negative*, f_n—number of lines marked as line spacings.

Afterwards precision, recall and F-measure are defined:

$$P_r = \frac{t_p}{t_p + f_p}, \tag{6}$$

$$R = \frac{t_p}{t_p + f_n}, \tag{7}$$

$$F = \frac{2}{\frac{1}{P_r} + \frac{1}{R}}. \tag{8}$$

P_r measures how many of the discovered text lines are actually lines, R measures how many of the existing lines are discovered and the F-measure combines P_r and R as their harmonic mean.

The experiments that are presented here are conducted on 80 document images, 20 of each of the four different types of documents, containing a total of 1869 text lines. For each of the text images both floating mean and wavelet decomposition segmentation algorithms are applied and for each of them the three error measures are calculated. The statistics of P_r, R, and F for both methods are given in Table 1.

Table 1. Mean μ and standard deviation σ of the error measures for both algorithms

	Floating mean			**Wavelet decomposition**		
Error measures	P_r	R	F	P_r	R	F
μ	0.8388	0.6910	0.7141	0.8576	0.9836	0.9041
σ	0.3187	0.3851	0.3694	0.1725	0.0393	0.1286

If we compare the mean of P_r, R, and F for both algorithms, it is obvious that the proposed approach gives much better results (it has $\mu(F) = 0.9041$ compared to $\mu(F) = 0.7141$ for the floating mean method). The comparison between the standard deviations of error measures shows that the method based on wavelet decomposition is much more stable, which is also evident in the F-measure graph given in Figure 4.

In order to examine in more details the experimental results for the four types of document images separately, four tables are given at the end of this paper. The image representatives are chosen randomly among the whole data set and for each the four disjoint classes t_p, t_n, f_p, and f_n are given, as well as the three error measures.

Table 2 shows that floating mean algorithms perform relatively well on

Table 2. Results of experiments with printed document data snippet

	Floating mean							**Wavelet decomposition**						
ID	t_p	t_n	f_p	f_n	P_r	R	F	t_p	t_n	f_p	f_n	P_r	R	F
1	22	23	0	2	1	0.9166	0.9565	24	25	2	0	0.9231	1	0.96
2	38	39	0	2	1	0.95	0.9743	40	41	2	0	0.9524	1	0.9756
3	3	4	2	33	0.6	0.0833	0.1463	36	37	5	0	0.878	1	0.9351
4	20	21	0	2	1	0.9091	0.9524	22	23	2	0	0.9167	1	0.9565
5	40	41	0	3	1	0.9302	0.9638	43	44	2	0	0.9555	1	0.9773
6	25	26	0	2	1	0.9259	0.9615	27	28	0	0	1	1	1
7	40	41	0	2	1	0.9524	0.9757	42	43	1	0	0.9767	1	0.9882

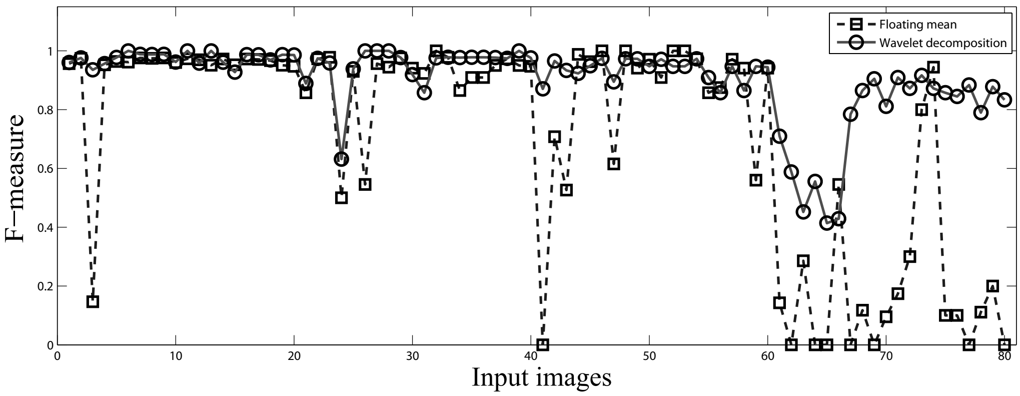


Fig. 4. F-measure graph of floating mean and wavelet decomposition algorithms for each of the 80 data set representatives. If the F-measure is close to 1, the recognition of the lines can be considered successful and if it is close or equal to zero, then the method actually failed

printed documents. The algorithm tends to miss about 2 lines per document and discovers no false lines, and for that reason the method has $Pr = 1$ in nearly all cases. An exception is the image ID3 where the floating mean method actually failed. On the other hand, wavelet decomposition method tends to find false (empty) lines, but it does not miss a single line, and that is why this method has $R = 1$.

The experiments with handwritten notebooks show similar results (Table 3) with a notable exception—image ID46, where the floating mean method missed 10 lines while wavelet decomposition did not miss a line.

The images used to form the third data set are much more complex than

Table 3. Results of experiments with handwritten notebook data snippet

	Floating mean							**Wavelet decomposition**						
ID	t_p	t_n	f_p	f_n	P_r	R	F	t_p	t_n	f_p	f_n	P_r	R	F
21	12	13	0	4	1	0.75	0.8571	16	17	4	0	0.8	1	0.8889
22	18	19	0	1	1	0.9474	0.973	19	20	1	0	0.95	1	0.9743
23	22	23	0	1	1	0.9565	0.9778	23	24	2	0	0.92	1	0.9583
24	2	3	0	4	1	0.3333	0.5	6	7	7	0	0.4615	1	0.6316
25	20	21	0	3	1	0.8696	0.9302	23	24	3	0	0.8846	1	0.9388
26	6	7	0	10	1	0.375	0.5454	16	17	0	0	1	1	1
27	22	23	1	1	0.9565	0.9565	0.9565	23	24	0	0	1	1	1

Table 4. Results of experiments with medieval manuscripts data snippet

	Floating mean							Wavelet decomposition						
ID	t_p	t_n	f_p	f_n	P_r	R	F	t_p	t_n	f_p	f_n	P_r	R	F
41	0	0	41	42	0	0	*fail*	37	38	7	4	0.8409	0.9024	0.8706
42	23	24	0	19	1	0.5476	0.7077	42	43	3	0	0.9333	1	0.9655
43	15	16	0	27	1	0.3571	0.5263	42	43	6	0	0.875	1	0.9333
44	41	42	0	1	1	0.9762	0.9879	42	43	7	0	0.8571	1	0.923
45	26	27	0	2	1	0.9286	0.963	28	29	3	0	0.9032	1	0.9492
46	19	20	0	0	1	1	1	19	20	1	0	0.95	1	0.9743
47	8	9	0	10	1	0.4444	0.6154	17	18	3	1	0.85	0.9444	0.8947

the printed documents and handwritten notebooks. For the experiments, shown in Table 4, pages from three medieval manuscripts were used. The floating mean method shows good results in three of the representatives which have close characteristics to the handwritten notebooks. In the other cases the floating mean method practically failed. On the contrary, the wavelet decomposition approach shows both good precision and good recall for all images in the data set. The worst result is in the case of image ID41, where wavelet decomposition missed 4 lines and floating mean found no lines at all.

The fourth data set consists of the most complex images used for the experiments. Astronomical logbooks have a complex, table-like structure of handwritten digits and printed table elements. The quality of the images is very low and as a result the binary images contain a lot of noise. Besides the data in table form, these images often contain subsidiary handwritten notes, which additionally impedes the line segmentation.

In astronomical logbooks we shall consider a document line a line of the table of digits. The images are not processed in advance, not counting the image

Table 5. Results of experiments with astronomical logbook data snippet

	Floating mean							Wavelet decomposition						
ID	t_p	t_n	f_p	f_n	P_r	R	F	t_p	t_n	f_p	f_n	P_r	R	F
61	1	2	2	10	0.3333	0.0909	0.1428	11	12	9	0	0.55	1	0.7097
62	0	0	3	10	0	0	*fail*	10	11	14	0	0.4167	1	0.5882
63	2	3	5	5	0.2857	0.2857	0.2857	7	8	17	0	0.2917	1	0.4516
64	0	0	3	10	0	0	*fail*	10	11	16	0	0.3846	1	0.5555
65	0	0	2	6	0	0	*fail*	6	7	17	0	0.2609	1	0.4138
66	3	4	2	3	0.6	0.5	0.5454	6	7	16	0	0.2727	1	0.4286
67	0	0	3	20	0	0	*fail*	20	21	11	0	0.6452	1	0.7843

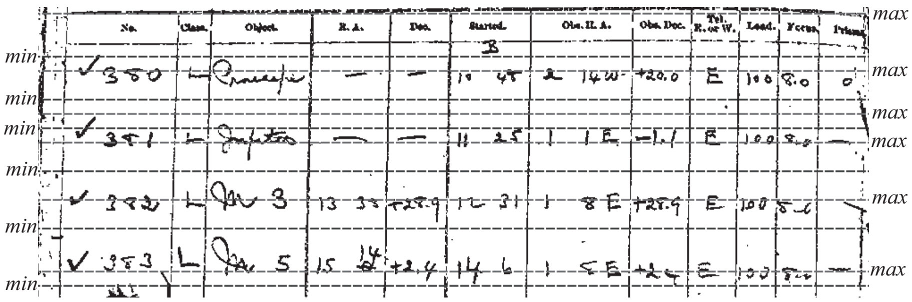


Fig. 5. Fragment of the logbook with local extrema detected by the wavelet decomposition approach. The discovered local maxima and minima show that all the text lines are segmented, in spite of the noisy image and non-standard structure of the document

binarization, which means that no preliminarily allocation of the textual parts of the images is performed. The results of the experiments in Table 5 show that the wavelet decomposition approach finds all the document lines (see also Figure 5). It also detects many of false lines, but they can be easily rejected in a next step, using vertical projection profile or some other technique. In this data set the floating mean approach shows poor results, except in the case of image ID66.

The experiments with the four data sets prove that document segmentation based on wavelet decomposition is an efficient and stable method, since it nearly always has a recall $R = 1$ or close. On the other hand, it has the tendency to detect false (empty) lines. The reason for the false line detection is in the rudimental procedure which finds local extrema in the approximated signal. This procedure was chosen because it is very sensitive and actually proves the stability and efficiency of the approach. Also, it has to be noted that the proposed method is much more efficient from a computational point of view compared to the floating mean approach, since no iterative procedure with an unpredicted number of iterations is involved. For example, to process the image ID45 floating mean took 15 sec., while wavelet decomposition finished in less then a second.

6. Conclusion. In this paper a document line segmentation method based on wavelet decomposition of the horizontal projective profile of the image was presented. The method takes advantage of the fact that at a given level of decomposition the approximated signal preserves the main characteristics of the initial signal, while those which correspond to noise and small details are suppressed. The method is tested on four completely different data sets and proved to be stable and efficient.

As future work, the presented approach will be extended towards segmentation of and symbol extraction in non-standard documents. Another direction of future research is the application of wavelet decomposition to more general one-dimensional clustering problems.

REFERENCES

[1] Acharyya M., M. K. Kundu. Document image segmentation using wavelet scale-space features. *IEEE Transactions on Circuits and Systems for Video Technology*, **12** (2002), No 12, 1117–1127.

[2] Arivazhagan M., H. Srinivasan, S. Srihari. A statistical approach to line segmentation in handwritten documents. In: Proceedings of the SPIE, Document Recognition and Retrieval XIV SPIE, Vol. **6500**, 2007, 65000T–65000T–11.

[3] Chen G. Y., T. D. Bui, A. Krzyzak. Contour-based handwritten numeral recognition using multiwavelets and neural networks. *Pattern Recognition*, **36** (2003), No 7, 1597–1604.

[4] Dalitz C., G. K. Michalakis, C. Pranzas. Optical recognition of psaltic byzantine chant notation. *International Journal on Document Analysis and Recognition*, **11** (2008), No 3, 143–158.

[5] Daubechies I. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, **41** (1988), 909–996.

[6] Duda R. O., P. E. Hart. Use of the Hough transformation to detect lines and curves in pictures. *Commun. ACM*, **15** (1972), 11–15.

[7] Gezerlis V. G., S. Theodoridis. Optical character recognition of the orthodox hellenic byzantine music notation. *Pattern Recognition*, **35** (2002), No 4, 895–914.

[8] Sauvola M. P. J. Adaptive document image binarization. Pattern Recognition, **33** (2000), No 2, 225–236.

[9] Kittler J., J. Illingworth. Minimum error thresholding. *Pattern Recognition*, **19** (1986), No 1, 41–47.

[10] Laskov L. Processing of byzantine neume notation in ancient historical manuscripts. *Serdica Journal of Computing*, **5**(2011), No 2, 183–198.

[11] Laskov L., D. Dimov. Segmentation of ancient neumatic musical notation. In: Proceedings of the Int.Conf. Automatics & Informatics'08, 2008, 21–24.

[12] Louloudis G., B. Gatos, I. Pratikakis, C. Halatsis. Text line and word segmentation of handwritten documents. *Pattern Recognition*, **42** (2009), No 12, 3169–3183.

[13] Mallat S. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11** (1989), 674–693.

[14] Mallat S. A wavelet tour of signal processing (2. ed.), Academic Press, 1999.

[15] Otsu N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, **9**(1979), No 1, 62–66.

[16] Papavassiliou V., T. Stafylakis, V. Katsouros, G. Carayannis. Handwritten document image segmentation into text lines and words. *Pattern Recognition*, **43** (2010), No 1, 369–377.

[17] Peycheva L., N. Kirov. Bulgarian folk songs in a digital library. In: Proc. of the Int. Conf. Digital Preservation and Presentation of Cultural and Scientific Heritage (Ed. Radoslav Pavlov), Veliko Tarnovo, Bulgaria, 2011, 60–68.

[18] Shi Z., S. Setlur, V. Govindaraju. Text extraction from gray scale historical document images using adaptive local connectivity map. In: Proceedings of the Eighth International Conference on Document Analysis and Recognition, Washington, DC, USA, 2005, IEEE Computer Society, 794–798.

[19] Tsvetkov M. Wide-field plate database: a decade of development. Virtual Observatory: Plate Content Digitization, Archive Mining and Image Sequence Processing, iAstro workshop, Sofia, Bulgaria, 2006, 10–41.

[20] Tsvetkov M., W. Osborn. Wide-field plate database—one major step with the pari North American initiative. In: Proc. of the VI Serbian-Bulgarian Astronomical Conference (Eds.: M. S. Dimitrijevic, M. Tsvetkov, L. Popovic, V. Golev), Belgrade, Serbia, 2009, Rudjer Boskovic, Vol. **9**, 113–142.

Lasko Laskov
New Bulgarian University
Informatics Department
21, Montevideo Str.
1618, Sofia, Bulgaria
e-mail: llaskov@nbu.bg

Received March 2, 2012
Final Accepted June 28, 2012