DOI:https://doi.org/10.15276/aait.03.2019.3
UDC 004.93

THE TECHNIQUE OF EXTRACTION TEXT AREAS ON SCANNED DOCUMENT IMAGE USING LINEAR FILTRATION

Alesya V. Ishchenko[1)]
ORCID ID: 0000-0002-7882-4718, alesya.ishchenko@gmail.com
Marina V. Polyakova[1)]
ORCID: http://orcid.org/0000-0002-1597-8867, marina_polyakova@rambler.ru
Alexandr G. Nesteryuk[1)]
ORCID: http://orcid.org/0000-0002-0806-8259, nesteryuk@opu.ua
[1)] Odessa National Polytechnic University, Avenue Shevchenko, 1. Odesa, 65044,Ukraine

ABSTRACT

The method of selection of text areas on the image of the scanned document from the background is proposed. Text areas of the image have approximately the same intensity values inside these areas. Therefore, linear filtering and threshold image transformation are used. Linear filtering allows you to smooth out the intensity values of pixels inside homogeneous areas. In the case of a threshold transformation, the threshold value is used, which makes it possible to isolate homogeneous areas of the image that make up the text fragments from the background.A study was conducted on the selection of a threshold value for highlighting homogeneous areas of text, which showed that the threshold value is better to choose among the pixel intensities at the base of the histogram peak, which corresponds to the background. It is proposed to select the threshold by the value of the second derivative for the image histogram after linear filtering. Therefore, the intensity of the local maximum of the histogram, which is closer than the other local maxima to the right end of the image intensity interval, is chosen as the threshold. For this purpose, an analysis of the histogram of the distribution of image pixel intensity values is carried out after linear filtering by rows and columns at each step. Testing of the proposed method of separating textual image areas was carried out for segmentation of textual images of scanned archival newspapers from the MediaTeam documents database at the University of Oulu (Finland).The proposed method of extracting text fragments from the background using linear filtering and threshold conversion allowed to improve the quality of selection of these areas compared to the similar method in the percentage of correct recognition of text areas by 12 %, which is important for the task of image segmentation.

Keywords: Image Segmentation; Text Areas; Scanned Document; Linear Filtering; Image Processing

***For citation*: Alesya V. Ishchenko, Marina V. Polyakova, Alexandr G. Nesteryuk. The Technique of Extraction Text Areas on Scanned Document Image Using Linear Filtration. *Applied Aspects of Information Technology.* 2019; Vol.2 No.3: 206–215. DOI: https://doi.org/10.15276/aait.03.2019.3**

INTRODUCTION

Today a huge part of the information is stored in electronic form. Automated search and retrieval of necessary information is performing using intelligent image processing systems.

In this way one of the problems that arise when processing electronic archives of scanned documents is the problem of text recognition. One of the image processing stages that precede this problem is segmentation of the scanned document. Segmentation is the most important step in the recognition of textual image regions, which consists in partitioning the image into regions that are homogeneous on some feature.

Extracting text from scanned documents images has many applications for analyzing documents, for example, searching images by keywords, searching a document by its contents, page segmentation, address location, etc. The rapid development of digital technologies led to the digitization of a large number of documents of all categories, including archival documents of state archives and commercial enterprises, libraries, universities, etc. In this way the problem of processing of a large number of scanned documents with sufficient segmentation quality arises. Therefore requirements for efficiency of their processing increase.

Therefore automating of the segmentation of scanned documents with high segmentation quality and low processing time is an important problem when creating electronic archives. To navigate through images of documents stored electronically, it is necessary to select such structural elements of the image as text and illustrations.

The **aim** of the paper is to elaborate a method for extracting text regions of a scanned document image in order to improve the quality of segmentation and the speed of segmentation performance of images of scanned documents for further processing and storage.

FORMULATION OF THE PROBLEM

The document image is usually mixed, which implies the presence of text regions and regions of illustrations.

The types of illustrations may be different depending on the specific type of document, and usually they are presented in the form of graphs, photos,

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

etc. Text regions contain characters and digits that, in turn, form words and sentences used to describe graphic elements of a document image. Extracting text regions in the image is a difficult problem due to the fact that the elements that make up the graphics, such as lines, can have different lengths, thickness and orientation; various geometric shapes, such as polygons and circles, can be painted or not painted. This can lead to an erroneous identification of these elements as large-sized text characters, for example, a header. In turn, text components may vary in font styles and sizes.

The text extraction of the image is one of the most important stages in the analysis of documents due to the fact that the text contains basic information about the document.

To increase the speed of image processing, the scanned document image first is segmented into illustration regions on a uniform background and text regions on a uniform background. This paper solves the problem of text regions extracting on an image containing text on a uniform background.

ANALYSIS OF RECENT PUBLICATIONS

There are many methods for text extracting from images, but all of them are elaborated to process certain documents. Such methods for the segmentation of scanned documents images are analyzed and compared in [1].

In [2], the document image is partitioned into homogeneous rectangular blocks of fixed size, for each of which the coefficients of the discrete Fourier transform are calculated, and then clustering is performed using the k-means method. The result of the algorithm is two binary masks: for text and for illustrations.

In [3] it is assumed that the text is horizontal. Color image of scanned document is transformed into a grayscale image. Then the image is partitioned into blocks, and in each block the coefficients of the discrete wavelet transform are calculated, then the boundaries of the text fragments are detected and non-text fragments are removed.

In [4], for the image blocks, the wavelet transform coefficients are calculated, with the help of which the network is trained on the basis of the hidden Markov model.

In [5], when segmenting images, linear classifiers were used to classify blocks, decision trees were used in [6], and neural networks were used for scanned document image segmentation in [7].

In the considered papers, on the basis of the coefficients of spectral transforms the features for the extraction of text are calculated. The use of such features makes it possible to sufficiently accurately extract text regions, but calculation of spectral features requires a significant time of processing.

In [8], pre-processing is first used to extract text regions, and then wavelet transform and run-length coding is used. After that, the image is partitioned into blocks and a map of illustrations is constructed, which is refined using the methods of optimization and image enhancement. This method is characterized with a high quality of segmentation however it has a low speed of segmentation performance due to the complexity of the calculations.

In [9], the FAST algorithm was used to extract text regions on images. First, the image is partitioned into blocks and the density is estimated in each block by counting the number of critical points. More dense blocks are related to text regions, and less dense blocks are related to regions of illustrations or noise. Then the connectivity of the blocks was tested, and they were grouped so as to separate the text region from the illustration regions. The quality of extraction of text regions by this method is quite high, but it is not effective for large fonts.

There are also methods for extracting text areas that begin processing with text characters, which are then combined into paragraphs, and columns, until whole text regions are extracted [9].

Such methods include, for example, the analysis of connected components [11-14]. The processing by such methods begins with the analysis of the smallest objects of the image (usually pixels), then they are combined into connected components. In [11], to isolate the connected components, the image of the document was half toned. The analysis took into account that the connected components correspond to the characters of the text, usually smaller, in contrast to the components of the illustrations.

In [12], the connected components were extracted on a grayscale image of a scanned document using graph theory. Further, the size of the extracted connected components was estimated, and then the threshold classifier was applied. The BESUS method [14] consists of several modules based on morphology. The text is extracted, considering the spatial relationships between pairs of text strings, which are identified based on the similarity and distribution of connected components.

In [15], the image was first half toned. Next, the connected components were determined, and then they were classified based on the analysis of the mutual location and properties of the connected components. In [16], thresholding, vertical and horizontal smoothing of the image is performed. Then the

boundaries of connected components are detected and connected components were classified. In the reviewed papers [11-16], processing in the neighborhood of each pixel is used. The proposed in [11-16] methods process complex shape regions with a high quality, but such methods take more time to process the image due to the fact that they first processes the neighborhood of each pixel, and then document fragments.

There are image segmentation algorithms based on the Voronoi diagram. This approach is based on the use of centers or extreme points of connected components, which are called critical points. The diagram is a partition of the image plane into regions. Each region contains one critical point and is a set of points of the plane for which this critical point is closer than the others critical points [17].

In [18], the Voronoi diagram was first used to solve the problem of segmentation of images that contain text. The advantage of this segmentation algorithm is high quality segmentation, but it has a low speed of segmentation performance because of the complexity of the calculations.

This paper proposes a technique for extracting text regions of a scanned document image to reduce the time of the image processing.

Research methods

At elaborating the algorithm for the extraction of text components on the scanned document image, digital image processing methods were used.

Main material

Let the image of the scanned document be represented as separate images [19], each of which contains only one class of regions such that text on a uniform background and graphics and photos on a uniform background. First, the regions of illustrations are extracted from the scanned document image using averaging filtering [20]. Text regions on a uniform background with the mentioned method were extracted with an accuracy of 99.4%. Therefore, this paper discusses a simplified model for presenting an image as text on a uniform background.

In this paper, the representation of an image $i(x, y)$ containing text on a uniform background is considered as a structural texture [21] $i(x, y)$, which describes the spatial organization of text characters $t(x, y)$ of the image [19]:

$$i(x, y) = t(x, y) * \sum_{k=1}^{L_y} \sum_{l=1}^{L_x} \delta(x - l\Delta x,\ y - k\Delta y), \quad (1)$$

where: $\delta(\cdot, \cdot)$ is the delta function;

– Δx, Δy are texture parameters that determine the distance between text characters in the column and image row, respectively;

– L_x, L_y are texture parameters that determine the number of characters in a column and image row;

– "*" is a convolution operator;

– $t(x, y)$ is the function of changing the intensity of pixels of a text symbol in spatial coordinates x, y, the value of this function varies from one symbol to another.

The regions of the image containing the title and the regions containing the main text have the same pixel intensity and differ in the size of the text characters and the distance between them. Therefore, the following inequalities are true:

$$\left|\Delta x_h - \Delta x_n\right| > x_{\min}, \ \left|\Delta y_h - \Delta y_n\right| > y_{\min}, \quad (2)$$

where: Δx_n, Δy_n are the distances between the characters of the main text in the image column and row respectively;

– Δx_h, Δy_h are the distances between the characters of the title text in the image column and row respectively;

– $x_{\min}$, $y_{\min}$ are the model parameters which depend on the size and type of character font.

The following characteristics are described in the model of representation of the image containing the text. These are the distance Δx between the characters of the text, and the distance Δy between the lines of the text which differ from each other. These characteristics are the defining parameters when processing the rows and columns of the image.

Further on the image it is necessary to extract text regions. To this end, it is proposed to use linear filtering and thresholding. In this case, linear filtering is a kind of preprocessing that allows taking into account the structural features of homogeneous regions and smoothing the image intensity values inside them.

We assume that the region are homogeneous if the pixel intensity values within this region are in the neighborhood of some parameter ε. At the thresholding the defined threshold is compared with the intensity values of the smoothed image. This allows extracting homogeneous regions of the image containing the same intensity values. In Fig. 1 a diagram of the proposed method for extracting scanned document image regions containing text fragments is shown. Consider the technique in more detail.

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

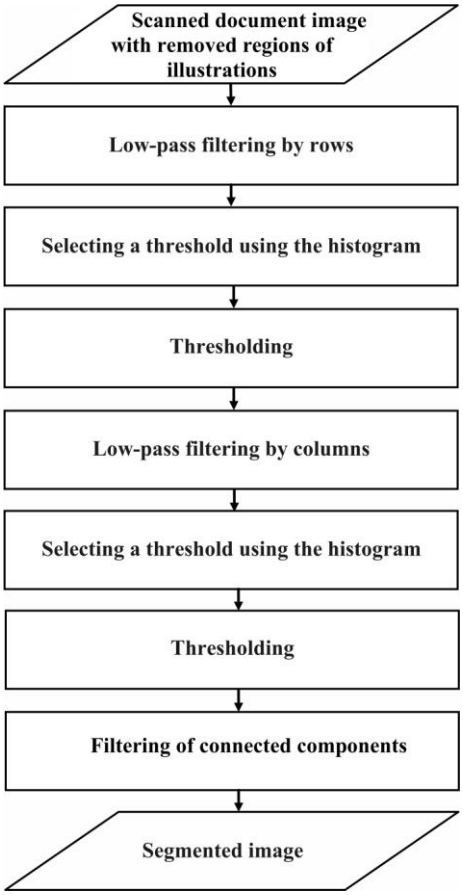


Fig. 1. **The scheme of extraction of text regions of the scanned document image**

***Source*: compiled by the author**

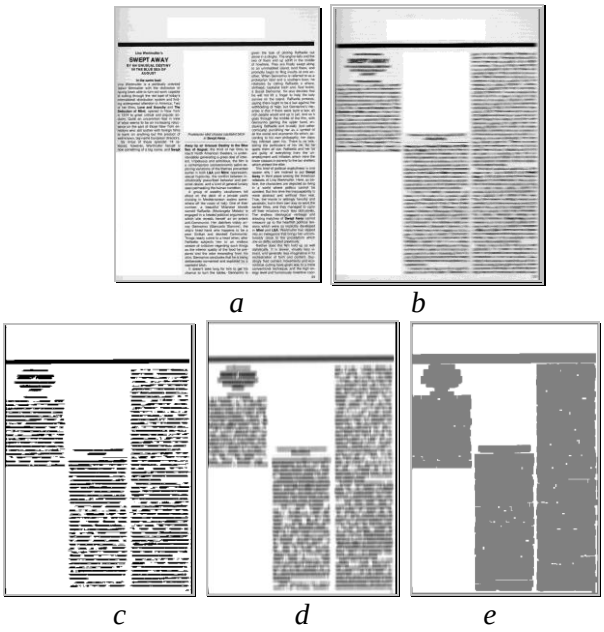


Fig. 2. **Extraction of text regions on the image of the scanned document:**

a – the original image; b – result of linear filtering by rows; c – the result of the thresholding after filtering by rows; d – result of linear filtering by columns; e – the result of the thresholding after filtering by columns

***Source*: compiled by the author**

Therefore, when processing an image of a scanned document containing text on a uniform background, first filtering is performed by rows (Fig. 2 b), and a thresholding is applied to the result (Fig. 2c). Then linear filtering is performed on the columns of the image which was obtained as a result of the thresholding at the previous step, and the thresholding is again performed (Fig. 2d,e). The linear filter mask was chosen empirically depending on the font size of the image, i.e., depending on the values of Δx_n, Δy_n, Δx_h, and Δy_h, and was a sequence of units of length from 25 to 50.

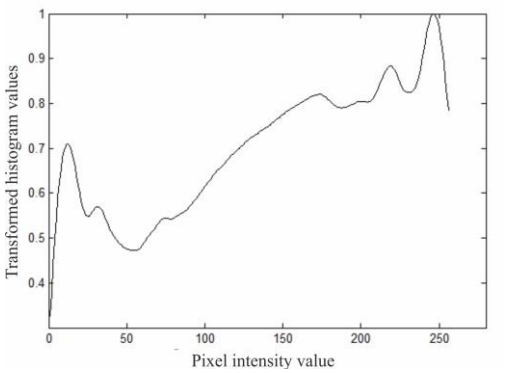
To select the threshold with which homogeneous regions of the text are extracted, several researches have been performed. The meaning of the threshold value is to split the image into regions of pixels of light intensities corresponding to the background and regions of dark intensities corresponding to the text. The text region is a set of those pixels whose intensity is below the threshold value, namely, $I(x, y) < T$, and the background is the set of other pixels whose intensity exceeds the threshold value, as follows $I(x, y) > T$.

The selecting of the threshold is an important step in the processing of the algorithm. Selecting too small a threshold value does not ensure the extraction of text regions from the background, and vice versa, a too large threshold value causes the algorithm to miss fragments of characters and they are removed from the image.

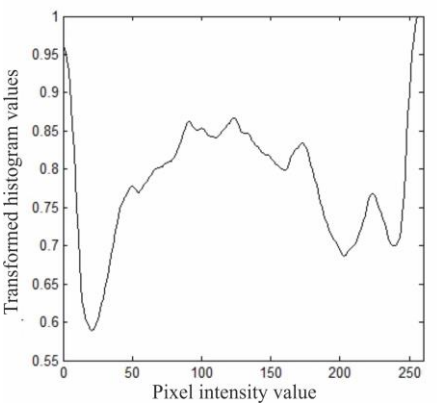
Consider the selecting of threshold in more detail. We construct a histogram of the image obtained after linear filtering by image rows, which we smooth by applying a Gaussian filter. Then a logarithmic transform is applied to the smoothed histogram using the natural logarithm function [22]. A unit is added to the values of the transformed histogram (Fig. 3a). This is necessary in order to increase the contrast of the histogram peaks. Then we construct a histogram of the image obtained after linear filtering by the columns of the image, and apply the same operations to it as described above (Fig. 3b).

Note that the obtained histogram contains a peak in the area of light intensities. This peak corresponds to the background. The histogram also contains peaks in the area of dark intensities corresponding to the textual region of the image. In the area of dark intensities there are several peaks, since after linear filtering the range of intensities that correspond to the

symbols of the text extends. They correspond to the local maxima of the image histogram.



a



b

Fig. 3. **Transformed image histogram after image linear filtering:**
a – in rows; b – in columns
***Source*: compiled by the author**

Many segmentation methods use the Otsu method to select the threshold value by the histogram [23]. The idea of the method Otsu is to determine the threshold between the two classes in such a way that the variance of the intensity distribution within the classes is minimal. At the same time, minimizing the intraclass variance is equivalent to maximizing the interclass variance.

Research has shown that if selecting text regions on a uniform background using the threshold obtained by the Otsu method then it appears undervalued. This can be explained by the fact that linear filtering leads to smoothing of homogeneous text regions, and smoothing of individual characters of the text region leads to blurred boundaries of this region. Therefore, it was concluded that the use of the threshold calculated by the Otsu method is inexpedient.

It was empirically found that to extract a text region on the image after smoothing individual characters, the threshold value is better to choose among the pixel intensities at the base of the transformed histogram peak, which corresponds to the background (Fig. 4).

Research has shown that at the base of the peak of the transformed histogram, the rate of change of the slope of the curve corresponding to this function is high. It is known that if the rate of change of the slope of the curve at some point is maximal, then the value of the second derivative of the function corresponding to this curve is also maximal at such a point.

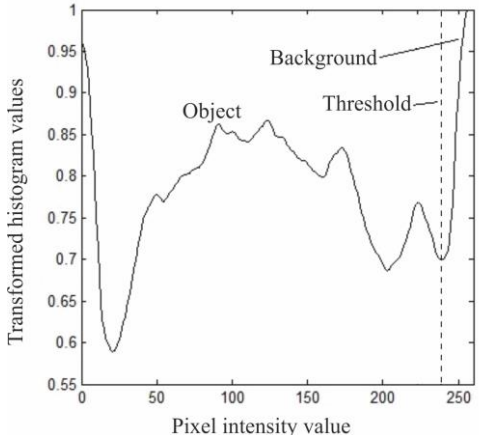


Fig. 4. **Threshold selecting for extracting text regions**
***Source*: compiled by the author**

Therefore, the histogram was twice differentiated as a function of the frequency of appearance of intensity values from these values. Then the maximum of the values of the second derivative will determine the value of the pixel intensity at the base of the peak of the histogram corresponding to the background, which we estimate.

Therefore, as the threshold, the intensity of the local maximum of the second derivative of the histogram was selected, which is closer than the other local maxima to the right end of the image intensity interval. In Fig. 4 is shown a histogram after a logarithmic transform of the contrast increasing of the histogram peaks and unit adding. This threshold was used to extract the text region of the image.

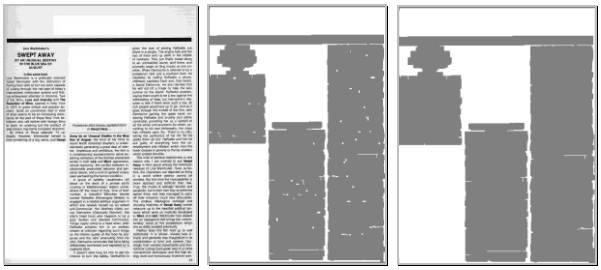
After the thresholding the resulting binary image contained small, separately located white regions corresponding to the background (Fig. 2e). Such fragments negatively influence on the quality of extraction of text regions on the image of the scanned document. Therefore, the filtering of connected components was applied to the resulting image. For this, connected components of pixels were found with a degree of connectivity equal to 8. A component of pixels is considered connected if for each pixel from this component there is a neighbor pixel from the same component [24].

In our case, with a degree of connectivity equal to 8, there are pixels whose neighbors are all pixels adjacent to the pixel being analyzed, that is, from above, below, to the left, to the right and diagonally. Let Y be a connected component from the set A contained in the image $I'(x, y)$. In each connected component Y_i, $i = 1, \ldots, K$, where K is the number of

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

connected components, there are components corresponding to the background, and their size d_i is calculated, i.e. the number of pixels belonging to this component. If the size of a specific component $d_i < T'$, where T' is a threshold, then the pixel intensities of the component Y_i are set equal to 0, i.e. this component is filled with pixels corresponding to the text.

As a result, we obtain a segmented image, where the black blocks correspond to the text fragments of the original image (Fig. 5c).



a *b* *c*

Fig. 5. **The result of the extraction of text regions on the image of the scanned document:**
a – the original image; b – result after low-pass filtering and thresholding; c – the result of the filtering of connected components
***Source*: compiled by the author**

The original scanned document image with already extracted regions of illustrations, containing text on a uniform background is shown in Fig. 5a. After linear filtering and thresholding, we obtain an image with extracted text regions, but containing small white components corresponding to the background (Fig. 5b). Filtering of connected components allows processing the neighborhoods of the pixels of the text and filter out the background pixels, thereby obtaining extracted text regions of the image of the scanned document (Fig. 5c).

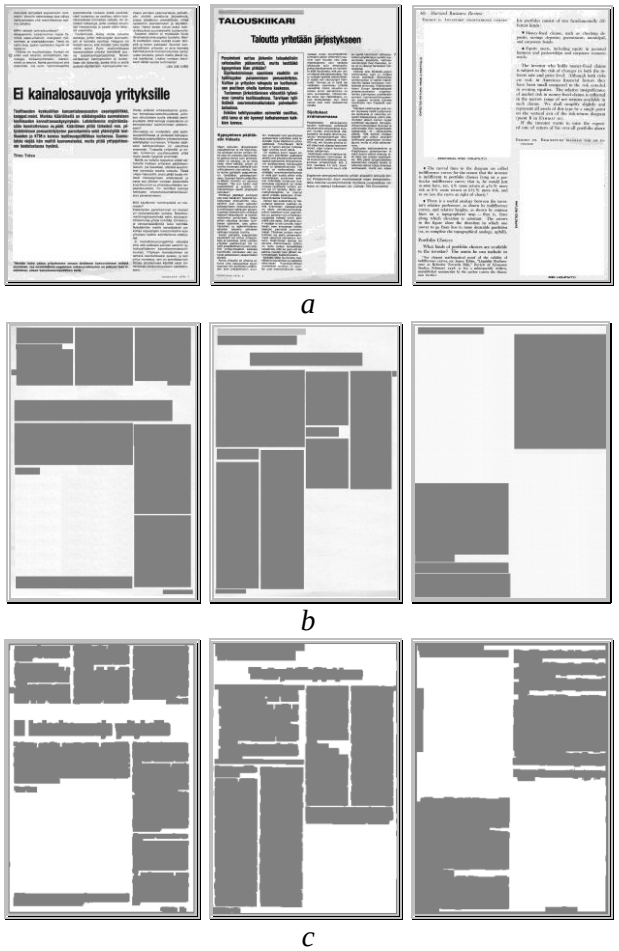
To estimate the quality of the extraction of text regions of scanned document images of using the proposed technique, images of articles and journals from the database of documents by MediaTeam Oulu [25] were used.

Documents from the database [25] contain text and/or illustrations such as headings, main text, graphics, and photographs. Related images are scanned in high quality 300 dpi, and are 3200 × 2300 pixels in size.

To extract text regions in the image, 60 test images of scanned documents were selected after processing by the segmentation method, which separates the regions of illustrations from the text and background regions [20].

Experimental research was performed using an Intel Core i5-3210 processor, 2.5 GHz CPU, 6 GB memory, and Windows 7 operating system, 64-bit. The experiment results were compared with the results of the segmentation method by Erkilinc M.S. et al. [8]. The use of this method allows obtaining high quality image segmentation, and also has a sufficiently high speed of segmentation performance. The technical characteristics and the system characteristics used for the experiment [8] correspond to the technical and software characteristics of this research.

The examples of original images with already removed regions of illustrations, on which text regions should be extracted, are shown in Fig. 6a, and the ground-truth segmentation results for these images are shown in Fig. 6b. The results of the extraction of text regions on the scanned document images using linear filtering and thresholding are shown in Fig. 6c. Text regions are marked in dark gray, background is white.



a

b

c

Fig. 6. **The results of the text regions extraction on scanned document images: a – original images; b – ground-truth segmentation results; c – the results of the text regions extraction**
***Source*: compiled by the author**

The quality of the extraction of text regions on the image using linear filtering and thresholding and the method of image segmentation of Erkilinc M.S. et al. [8] was estimated using a confusion matrix

[26]. These methods were used to segment images into 2 classes, so the confusion matrix has a dimension of 2x2. The rows of the matrix correspond to the ground-truth class labels for the pixels of images. The columns of the confusion matrix correspond to class labels obtained using the researched segmentation method.

The elements of the confusion matrix show what percentage of the image pixels corresponding to the ground-truth class are assigned to the class of pixels determined by the researched method and were calculated as follows:

– pixels of the background regions are classified as background pixels. This is the percentage of pixels assigned to the non-text regions in the segmented image and in the ground-truth segmentation image relative to the total number of pixels corresponding to non-text regions in the ground-truth segmentation image;

– pixels of a text regions are classified as text pixels. This is the percentage of pixels assigned to the text regions in the segmented image and in the ground-truth segmentation image relative to the total number of pixels corresponding to the text regions in the ground-truth segmentation image;

– pixels of the background regions are classified as pixels of the text region. This is the percentage of pixels assigned to non-text regions in the segmented image and assigned to text regions in the ground-truth segmentation image relative to the total number of pixels corresponding to the text regions in the ground-truth segmentation image;

– pixels of a text regions are classified as pixels of background. This is the percentage of pixels assigned to text regions in the segmented image and assigned to non-text regions in the ground-truth segmentation image relative to the total number of pixels corresponding to the non-text regions in the ground-truth segmentation image.

The averaging results of the segmentation quality estimation for all examined test images for the proposed method and the Erkilinc M.S. et al. segmentation method are shown in Table 1.

*Table 1.***The averaging results of the segmentation quality estimation and time of processing for the proposed method and the analyzed segmentation method**

Class label	Image segmentation using linear filtering	Erkilinc M.S. et al. segmentation method [5]
Text	90,0 %	88 %
Background	89 %	97 %
Time of processing	2,33 s	14 s

***Source*: compiled by the author**

Conclusions and future research

The results show that the proposed method of extracting text regions on an image using linear filtering and thresholding compared to a similar known method [8] is 2 % higher in the percentage of correct segmentation of text regions. Background regions are segmented by the proposed method about 7 per cent lower than this obtaining by known method [8].

The inaccuracy of segmentation using the proposed technique was that small areas of text can be erroneously defined as background due to the fact that the boundaries of text characters are very blurred after using linear filtering. The proposed method for extracting text regions reduces the average image processing time compared to the segmentation method of the authors Erkilinc M.S. and others almost 6 times.

Further research may be aimed at improving the quality of the extraction of text regions compared to the ground-truth segmentation due to the assumption that the extracted regions of the text are rectangular, as is done when extracting the illustrations on the image in [20].

REFERENCES

1. Antonacopoulos, A., Gatos, B. & Bridson, D. “ICDAR 2005 page segmentation competition”. In Proc. ICDAR. Seoul: Korea. 2005. p.75–80.

2. Sasirekha, D., & Chandra, Dr. E. “Enhanced Techniques for PDF Image Segmentation and Text Extraction”. *International Journal of Electronics and Computer Science Engineering*. 2012. p.1833–1839.

3. Gupta, N. & Banga, V. K. “Image Segmentation for Text Extraction”. *2nd International Conference on Electrical, Electronics and Civil Engineering (ICEECE'2012).* Singapore: 28-29 April 2012, pp. 182-185.

4. Kumar, S., Gupta, R., Khanna, N. [et al.] “Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model”. *IEEE Transactions on Image Processing.* Vol. 16 No. 8: 2117–2128. DOI: https://doi.org/10.1109/tip.2007.900098.

5. Wong, K. Y., Casey, R. G. & Wahl, F. M. “Document analysis system”. *IBM Journal of Research and Development.* 1982; Vol.26(6): 647–656. DOI: https://doi.org/10.1147/rd.266.0647.

ISSN 2617-4316 (Print)
ISSN 2663-7723 (Online)

6. Esposito, F., Malerba, D. & Semeraro, G. “A knowledge-based approach to the layout analysis”. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*. 1995; Vol.1: 466–471. DOI: https://doi.org/10.1109/ICDAR.1995.599037.

7. Li, L., Yu, S., Zhong, L. & Li, X. “Multilingual text detection with nonlinear neural network”. *Mathematical Problems in Engineering.* 2015; Vol.2015, 431608: p. 7. DOI: https://doi.org/10.1155/2015/431608.

8. (2011). Erkilinc, M. S., Jaber, M., Saber E. [et al.]. “Analysis and classification for complex scanned document”. SPIE Newsroom. DOI**:** https://doi.org/10.1117/2.1201107.003819**.**

9. Mathur, G. Ms. & Rikhari, S. “Text Detection in Document Images: Highlight on using FAST algorithm”. *International Journal of Advanced Engineering Research and Scienc.* 2017; Vol.4 No. 3: 275–284. DOI: https://doi.org/10.22161/ijaers.4.3.43.

10. Shafait, F., Keysers, D. & Breuel, T. M. “Performance Evaluation and Benchmarking of SixPage Segmentation Algorithms Pattern Analysis and Machine”. *Intelligence, IEEE Transactions on.* 2008; Vol. 30: 941–954. DOI: https://doi.org/10.1109/TPAMI.2007.70837.

11. Bukhari, S. S., Shafait F. & Breuel, T. “Improved document image segmentation algorithm using multiresolution morphology” [Text]. *In Proc. of the 18th Document Recognition and Retrieval Conf. Document Recognition and Retrieval XVIII – DRR 2011*. San Jose: CA, USA. DOI: https://doi.org/10.1117/12.873461.

12. Zirari, F., Ennaji, A., Nicolas, S. & Mammass, D. “A document image segmentation system using analysis of connected components”. *Proceeding of the 12th Int. Conf. ICDAR 2013 (Document Analysis and Recognition).* Washington: DC. USA. p. 753–757. DOI: https://doi.org/10.1109/icdar.2013.154.

13. Smith, R. W. “History of the Tesseract OCR engine: what worked and what didn’t”, Proceedings of SPIE. 2013. Vol. 8658, 865802. DOI: https://doi.org/10.1117/12.2010051.

14. Das A. K. & B. ChandaD. “Segmentation of text and graphics in document image: A morphological approach”. *In Proc. Inf. Conf. Computational Linguistics, Speech and Document Processing*. Calcutta: India. Dec. 1998. p. A50–A56.

15. Vil’kin, A. M., Safonov, I. V. & Egorova, M. A. “Algorithm for page segmentation”. *Digital Signal Processing and its Applications.* 2011.

16. Rege, P. P. & Chandrakar, C. A. “Text-Image Separation in Document Images Using Boundary”. Perimeter Detection. *ACEEE International Journal on Signal & Image Processing*. 2012; Vol.4 No. 1: 10–14. DOI: https://doi.org/01.ijsip.03.01.70.

17. Skvortsov, A. V. “Triangulyatsiya Delone i yeyo primeneniye”. [Delaunay triangulation and its application]. *Izd-vo Tomskogo un-ta* (in Russian). Tomsk: Russian Federation. 2002. 128 p. ISBN: 5-7511-1501-5.

18. K. Kise, A. Sato & M. Iwata. “Segmentation of page images using the area Voronoi diagram”. *Computer Vision and Image Understanding.* 1998; Vol.70 Issue 3:370–382. DOI: https://doi.org/10.1006/cviu.1998.0684.

19. Ishchenko, A., Polyakova, M., Kuvaieva, V. & Nesteryuk, A. “Elaboration of structural representation of regions of scanned document images for MRC model”. *Eastern-European Journal of Enterprise Technologies.* 2018; No.6/2 (96): 32–38. DOI: https://doi.org/10.15587/1729-4061.2018.147671.

20. Polyakova, M., Ishchenko, A. & Huliaieva, N. “Document image segmentation using averaging filtering and mathematical morphology”. *14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET).* Lviv-Slavske: Ukraine. 2018. p. 966–969. DOI: https://doi.org/10.1109/TCSET.2018.8336354.

21. Haralick, R. M. “Statistical and structural approaches to texture”. *Proceedings of the IEEE.* 1979; Vol.67: 786–804. DOI: https://doi.org/10.1109/proc.1979.11328.

22. Gonsales, R. S., Vuds, R. E. & Eddins, S. L. “Cyfrovaya obrabotka izobragenij v srede Matlab”. [Digital image processing using MATLAB]. *Publ. Tehnosfera* (in Russian). Moscow: Russian Federation. 2006. p.616

23. Otsu, N. “A threshold selection method from gray-level histograms” [Text]. *IEEE Trans. Syst. Man, Cybern.* V. SMC-9. 1979. p. 62–66. DOI: https://doi.org/10.1109/tsmc.1979.4310076.

24. Gonsales, R. & Vuds, R. “Cyfrovaja obrabotka izobragenij”. [Digital image processing]. *Publ. Technosfera* (in Russian). Moscow: Russian Federation. 2005. p.1072

25. Sauvola, J. & Kauniskangas, H. “MediaTeam Document Database II”. [Electronic resource]. A collection of document images, University of Oulu. Oulu: Finland. (CD-R). 1999.

26. “Confusion_matrix”. – Available from: http://en.wikipedia.org/wiki/Confusion – Active link: – 20.06.2019.

Conflicts of Interest: the authors declare no conflict of interest.

Received 24.04.2019
Received after revision 05.06.2019
Accepted 17.06.2019

DOI:https://doi.org/10.15276/aait.03.2019.3
УДК 004.93

МЕТОДИКА ВИДІЛЕННЯ ТЕКСТОВИХ ОБЛАСТЕЙ НА ЗОБРАЖЕННІ ВІДСКАНОВАНОГО ДОКУМЕНТА З ВИКОРИСТАННЯМ ЛІНІЙНОЇ ФІЛЬТРАЦІЇ

Олеся Володимирівна Іщенко[1)]
ORCID: http://orcid.org/0000-0002-7882-4718, alesya.ishchenko@gmail.com
Марина Вячеславівна Полякова[1)]
ORCID: http://orcid.org/0000-0002-1597-8867, marina_polyakova@rambler.ru
Олександр Геннадійович Нестерюк[1)]
ORCID: http://orcid.org/0000-0002-0806-8259, nesteryuk@opu.ua
[1)] Одеський національний політехнічний університет, пр. Шевченка, 1. Одеса, 65044, Україна

АНОТАЦІЯ

Запропоновано методику виділення текстових областей на зображенні відсканованого документа з фону. Текстові області зображення мають приблизно однакові значення інтенсивності всередині цих областей. Тому використовується лінійна фільтрація і порогове перетворення зображення. Лінійна фільтрація дозволяє згладити значення інтенсивності пікселів всередині однорідних областей. При пороговому перетворенні використовується значення порога, яке дозволяє виділити однорідні області зображення, що становлять текстові фрагменти, з фону. Проведено дослідження вибору порогового значення для виділення однорідних областей тексту, яке показало, що значення порога краще вибирати серед інтенсивностей пікселів у підставі піку гістограми, який відповідає фону. Вибір порога запропоновано здійснювати за значенням другої похідної для гістограми зображення після лінійної фільтрації. Тому в якості порога вибирається значення інтенсивності локального максимуму гістограми, який знаходиться ближче інших локальних максимумів до правого кінця інтервалу інтенсивностей зображення. Для цього проводиться аналіз гістограми розподілу значень інтенсивності пікселів зображення після лінійної фільтрації по рядках і по стовпцях на кожному кроці. Апробація запропонованої методики виділення текстових областей зображення проведена для сегментації текстових зображень відсканованих архівних газет з бази даних документів MediaTeam університету Оулу (Фінляндія). Запропонована методика виділення текстових фрагментів з фону з використанням лінійної фільтрації та поро-гового перетворення дозволила підвищити якість виділення цих областей у порівнянні з аналогічним методом за відсотком правильного розпізнавання областей тексту на 12%, що актуально для задачі сегментації зображень.

Ключові слова: сегментація зображень; текстові області; відсканований документ; лінійна фільтрація; обробка зображень

DOI:https://doi.org/10.15276/aait.03.2019.3
УДК 004.93

МЕТОДИКА ВЫДЕЛЕНИЯ ТЕКСТОВЫХ ОБЛАСТЕЙ НА ИЗОБРАЖЕНИИ ОТСКАНИРОВАННОГО ДОКУМЕНТА С ИСПОЛЬЗОВАНИЕМ ЛИНЕЙНОЙ ФИЛЬТРАЦИИ

Алеся Владимировна Ищенко[1)]
ORCID: http://orcid.org/0000-0002-7882-4718, alesya.ishchenko@gmail.com
Марина Вячеславовна Полякова[1)]
ORCID: http://orcid.org/0000-0002-1597-8867, marina_polyakova@rambler.ru
Александр Геннадиевич Нестерюк[1)]
ORCID: http://orcid.org/0000-0002-0806-8259, nesteryuk@opu.ua
[1)] Одесский национальный политехнический университет, пр. Шевченко, 1. Одесса, 65044, Украина

АННОТАЦИЯ

Предложена методика выделения текстовых областей на изображении отсканированного документа из фона. Текстовые области изображения имеют приблизительно одинаковые значения интенсивности внутри этих областей. Поэтому используется линейная фильтрация и пороговое преобразование изображения. Линейная фильтрация позволяет сгладить значения интенсивности пикселей внутри однородных областей. При пороговом преобразовании используется значение порога, которое позволяет выделить однородные области изображения, составляющие текстовые фрагменты, из фона. Проведено исследование выбора порогового значения для выделения однородных областей текста, которое показало, что значение порога лучше выбирать среди интенсивностей пикселей в основании пика гистограммы, который соответствует фону. Выбор порога предложено осуществлять по значению второй производной для гистограммы изображения после линейной фильтрации. Поэтому в качестве порога выбирается значение интенсивности локального максимума гистограммы, который находится ближе остальных локальных максимумов к правому концу интервала интенсивностей изображения. Для этого проводится анализ гистограммы распределения значений интенсивности пикселей изображения после линейной фильтрации по строкам и по столбцам на каждом шаге. Апробация предложенной методики выделения текстовых областей изображения проведена для сегментации текстовых изображений отсканированных архивных газет из базы данных документов MediaTeam университета Оулу (Финляндия). Предложенная методика выделения текстовых фрагментов из фона с использованием линейной фильтрации и порогового преобразования позволила повысить качество выделения этих областей по сравнению с аналогичным методом по проценту правильного распознавания областей текста на 12 %, что актуально для задачи сегментации изображений.

Ключевые слова: сегментация изображений; текстовые области; отсканированный документ, линейная фильтрация, обработка изображений

ABOUT THE AUTHORS

Alesya Vladimirovna Ishchenko, Senior Lecturer of Department of Applied Mathematics and Information Technologies, Odessa National Polytechnic University, 1, Shevchenko Avenue. Odessa, 65044, Ukraine
alesya.ishchenko@gmail.com. ORCID ID: 0000-0002-7882-4718
Research field: Artificial Intelligence, Image Processing, Neural Networks

Алеся Владимировна Ищенко, старший преподаватель кафедры Прикладной математики и информационных технологий института компьютерных систем. Одесский национальный политехнический университет, пр. Шевченко, 1. Одесса, 65044, Украина

.

Marina Vyacheslavovna Polyakova, Doctor of Technical Sciences, Associate Professor Department of Applied Mathematics and Information Technologies, Odessa National Polytechnic University, 1, Shevchenko Avenue. Odesa, 65044, Ukraine
marina_polyakova@rambler.ru. ORCID: http://orcid.org/0000-0002-1597-8867
Research field: Artificial Intelligence, Wavelet Analysis, the Theory of Distributions

Марина Вячеславовна Полякова, доктор технич. наук, доцент кафедры Прикладной математики и информационных технологий. Одесский национальный политехнический университет, пр. Шевченко, 1. Одесса, 65044, Украина

Alexandr Gennadievich Nesteryuk, Candidate of Technical Sciences, Associate Professor Department of Computer Systems, Odessa National Polytechnic University, 1, Shevchenko Avenue. Odesa, 65044, Ukraine
nesteryuk@opu.ua. ORCID: http://orcid.org/0000-0002-0806-8259
Research field: Discrete-Continuous Systems; Discrete-Continuous Nets, as well as their Properties; Continuous, Discrete, and Hybrid Petri Nets; Hybrid Systems; Neural Networks; Artificial Intelligence; Software Development

Александр Геннадиевич Нестерюк, кандидат технич. наук, доцент кафедры Компьютерных систем. Одесский национальный политехнический университет, пр. Шевченко, 1. Одесса, 65044, Украина