Запропоновано комбінований метод сегментації зображень відсканованих документів, в якому, на відміну від відомих, проводиться попереднє відокремлення області графічних і фотозображень від текстових областей і фону. При цьому проводиться аналіз зв'язкових компонент, які є різними для графічних зображень, фотозображень та текстових областей. Для класифікації виділених областей, на області фото і графіки використовується блоковий метод. Встановлено, що такий спосіб розбиття областей на блоки менше впливає на якість сегментування в порівнянні з застосуванням блочного методу безпосередньо до вихідного зображення. Для відділення більш складних за формою текстових областей від фону застосовано обробка околиці кожного пікселя.

Для виділення на зображеннях відсканованих документів границь ілюстрацій використовувався метод Блумберга. Для поділу на фото і графіку запропоновано розбиття ілюстрації на блоки пікселів. Кожному блоку пікселів відповідає вектор з двох ознак: середнього значення величини локального градієнта і середнього значення функції, що локалізує на зображеннях відсканованих документів лінійні об'єкти (графіка і символи тексту). Отримані вектора ознак класифікувалися машиною опорних векторів.

При виділенні текстових фрагментів використовувалися низькочастотна фільтрація і порогове перетворення.

Практичне відпрацювання комбінованого методу проведено для сегментації тестових зображень відсканованих статей газет з бази даних документів MediaTeam університету Оулу (Фінляндія). Встановлено, що комбінований метод характеризується підвищеною швидкодією сегментації зображень при високій якості обробки

Ключові слова: сегментація зображень, відсканований документ, блочний метод, графічне зображення, фотозображеня, текстовий фрагмент, зв'язкова компонента, метод Блумберга

UDC 004.93
DOI: 10.15587/1729-4061.2018.142735

COMBINED METHOD FOR SCANNED DOCUMENTS IMAGES SEGMENTATION USING SEQUENTIAL EXTRACTION OF REGIONS

M. Polyakova
Doctor of Technical Sciences,
Associate Professor*
E-mail: marina_polyakova@rambler.ru
A. Ishchenko
Senior Lecturer*
E-mail: alesya.ishchenko@gmail.com
N. Volkova
Senior Lecturer*
E-mail: volkovanp30@gmail.com
O. Pavlov
Senior Lecturer*
E-mail: pavlov.ol.al@gmail.com
*Department of Applied Mathematics and Information Technologies
Odessa National Polytechnic University
Shevchenka ave., 1, Odessa, Ukraine, 65044

1. Introduction

Intelligent systems for processing the images of scanned documents are applied in printing and e-commerce, in order to automate document input at enterprises, as well as in the technologies of electronic document management and the Internet search. The most important stage of processing the images of scanned documents is the stage of their segmentation [1, 2]. It is known [1, 2] that the segmentation of images of scanned documents is used to solve the following tasks:

– to store documents in a digital form and to send them via computer network;

– to search for and to store text parts of documents in large databases;

– to optically recognize characters;

– to save scanned documents in a PDF format.

Segmentation quality affects the quality of processing the entire image of the scanned document [1, 2]. Insufficient quality of segmentation could lead to [1, 2]:

– inaccurate determining the boundaries of text region, which might be the cause of incorrect character recognition or the blurred boundaries of text characters as a result of the application of an encoder designed for illustrations (photographs and graphics);

– lower quality (blur, loss of contrast) of an illustration in the fragment of a non-text region;

– inaccurate determining the boundaries of a text region.

Thus, at present, strict requirements are put forward to the quality of segmentation when processing the images of scanned documents. In addition, a large part of tasks on processing the images of scanned documents requires conducting a segmentation at high performance speed [1, 2].

Known methods for the segmentation of images of scanned documents [2–12] are not capable to simultaneously satisfy modern requirements to the quality of segmentation and to the high-speed performance of segmentation of images of scanned documents. Block processing methods [3, 4] are characterized by high-speed performance, but the result

of segmentation depends on the partition of the processed image into blocks. At the same time, pixel processing [5–9] requires a significant time cost although it makes it possible to obtain high quality of segmentation. For methods that involve analysis of connected components [2, 10–12], the processing time and the quality of segmentation are determined based on the result of the identification of connected components.

Thus, there is currently an exacerbated contradiction between modern requirements to the quality and high-speed segmentation of images of scanned documents and the capabilities of contemporary segmentation methods.

2. Literature review and problem statement

Methods for the segmentation of images of scanned documents are categorized in the scientific literature into block processing methods, pixel processing methods, and methods for processing connected components [3–12].

The methods from the first group split the image of the scanned document into blocks of pixels and then classify the obtained blocks into predefined classes. Paper [3] used, as the features of image blocks, the intensity histogram values. Four neural networks were built accordingly to assign the vectors of these features to one of the 4 classes: text, photographic image, graphics, and background. Study [4] compared the two proposed method. The first method employed *k*-means clustering of the vectors of features constructed on the basis of the discrete cosine transform coefficients in the image blocks. In the second method, the vectors of features, computed based on the histogram of image blocks intensities, were treated with a threshold processing. The first method makes it possible to obtain the higher quality of segmentation, however, it requires more processing time compared to the second method. In general, block processing reduces the time for segmenting an image of the scanned document. However, the result of segmentation in this case depends heavily on partitioning the processed image into blocks.

The methods of the second group classify pixels of the scanned document into the predefined classes with respect to vectors of features computed in the neighborhood of each pixel. Such a processing implies a significant time cost although quality of the obtained segmentation is typically higher than that when applying methods from the first group. Thus, authors in [5] first determined the image edges using a two-fold differentiation. They then conducted the correlation processing of each line in the image, specifically, the convolution of each line of the image with a reference signal and comparison of the convolution result with a threshold. Paper [6] proposed a method for the segmentation of images of natural scenes and color images from documents into text and non-text regions. The original image was processed at different scales and with a different orientation using the M-band frame wavelet packages. Vectors of features were constructed by calculating the local energy of decomposition coefficients in the neighborhood of each pixel at each level of the representation. The derived vectors of features were clustered using the fuzzy c-means method. In [7], in order to segment images of the scanned documents, a feature vector was computed in the neighborhood of each pixel. The vector includes the intensity of the neighborhood central pixel, the mean and standard deviation of the intensity for the neighborhood. The vectors of features were classified using the improved method of fuzzy c-means that takes into consideration not only the proximity of vectors in the space of features, but also the spatial proximity of the respective pixels. Paper [8], in order to improve the speed of segmentation performance of method [7], applied an ant colony optimization algorithm.

Authors in [9], in order to segment the images of scanned documents into text, photographs and separators in the form of lines, sequentially applied the pixel and block processing using 5 modules. The first module involved the preprocessing of images of scanned documents, which included scaling, image improvement, and a transition from the RGB color space to the CIE La*b* space. In the second module, the authors extracted text fragments using the wavelet transform and run-length encoding. In the third module, they extracted illustrations (photographs and graphics) by splitting the image into blocks and performing the pre-segmentation using projections onto the vertical axis of the blocks' lines. Next, the result of preliminary segmentation was refined by using the criterion of the maximum *a posteriori* probability that was applied to the parameters of the Markov random field in the neighborhood of each pixel in a block. The fourth module, in order to extract separators, employed methods for edge detection, edge linking, the approximation of lines via straight line segments, and the Hough transform. In the fifth module, the results of extracting the text, illustrations, and separators in the form of individual maps for each class of segments, were combined using the method of *k*-means, thereby forming the result of segmenting the original image. Method [9] is distinguished by the high quality of segmentation, but low speed.

The quality of segmentation and the time of processing the images of scanned documents for methods from the third group depend on the number of connected components, separated by using the binarization. In order to extract the photographs and graphics, the image in paper [10] was first binarized to determine the connected components. It was then taken into consideration that the characters in a text are typically smaller than the components that correspond to the objects in photographs and graphics and, by using the mathematical morphology, the authors erased from the image those connected components that corresponded to the characters in the text. The connected components corresponding to graphics and photographs decreased in size as a result of morphological processing, which is why they had to be enlarged to select the regions of graphics and photographs. In [11], the extraction of connected components in the images of scanned documents was performed not for a binary image, but for halftone one, using a graph theory. The authors then computed, for the obtained connected components, their size as a feature, and then applied a threshold classifier. In [12], the segmentation of images of scanned documents was carried out by classifying the connected components of these images into the text and non-text components. The connected components of an image were scaled so that each of them is represented by a 40x40 matrix composed of zeros and unities. Next, based on the elements of these matrices, and the four shape features, the authors determined vectors of features of the connected components whose classification was performed using a multilayer perceptron. Method [12] is distinguished by the high quality of segmentation, but low-speed performance as a result of large dimensionality of feature vectors.

An analysis of [3–12] that examined the methods for segmenting the images of scanned documents has revealed that the main characteristics of these methods are the quality of segmentation and the time of processing. The values

for these characteristics depend on the size of blocks of the image's pixels. If the blocks of an image's pixels are large, the quality of processing is typically low at a high- speed performance. When choosing small blocks of an image's pixels, the situation is reversed.

Thus, the analysis of known methods for segmenting the images of scanned documents [3–12] that we performed has revealed the following drawbacks of the known methods:

– the complexity of training neural networks, the dependence of weights on the training set of images;

– the instability of methods against images' artifacts of various kinds;

– the high quality of a segmentation at a significant time cost;

– fast image segmentation at low quality.

A number of such tasks as saving the scanned documents or searching for documents in large databases require methods for segmenting the images of scanned documents with high-speed performance at high-quality segmentation. To improve the performance speed when segmenting the images of scanned documents, it is appropriate to process images within blocks of pixels whose size is chosen depending on size of the connected components in the segments of a particular class (text, photographs, and graphics). Thus, when extracting illustrations (photographs and graphics), one could apply a block segmentation method, and when extracting text fragments – processing in the neighborhood of each pixel. The quality of determining the boundaries of illustrations prior to applying a block method can be improved by running an analysis of the connected components.

3. The aim and objectives of the study

The aim of this work is to elaborate a combined method for segmenting the images of scanned documents with a sequential application of the analysis of connected components, block processing, and the processing of a neighborhood of each pixel, which would improve the performance speed of segmentation while ensuring the required quality of processing.

To accomplish the aim, the following tasks have been set:

– to elaborate stages in the combined method for segmenting the images of scanned documents in order to sequentially extract the regions of photographs, graphics, and text;

– to experimentally research the combined method using the images of scanned documents from a test database.

4. Elaboration of stages in the combined method for segmenting the images of scanned documents

We propose a combined method for segmenting the images of scanned documents (Fig. 1), according to which we first extract the regions of graphical images and photographs from the text regions and background within an image. In this case, we run an analysis of connected components because regions of graphical images and photographs differ from text regions in shape and periodicity of the connected components. Since the boundaries of regions with graphical images and photographs are defined, the classification of the selected regions into the regions of photographs and graphics employs a block method. In this case, the technique to split regions into blocks affects the classification quality less compared to applying a block method directly to the original image. And, finally, to extract text regions from the background, it is proposed to apply the processing of each pixel's neighborhood because text regions are typically characterized by a complex shape.

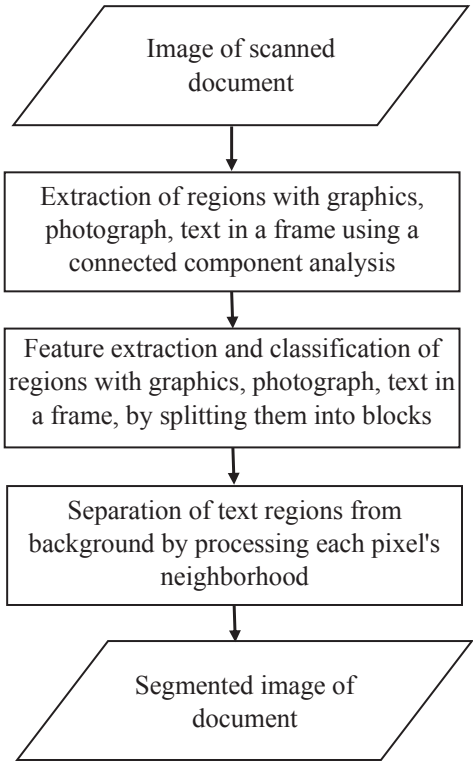


Fig. 1. The proposed combined method for segmenting the images of scanned documents

To extract text in a frame, the halftone photographic and graphical images in the images of scanned documents, we applied a Bloomberg method with filling the holes [10] based on mathematical morphology (Fig. 2). The method, which is based on the analysis of connected components, removes those connected components in the image that correspond to the characters of the text, leaving the non-text connected components. The choice of method [10] is predetermined by its high performance as a result of using mathematical morphology and a relatively high quality compared to other methods for analyzing connected components [11, 12].

As a result of the first stage in the proposed method of segmentation we extract, by analyzing the connected components on the image of the scanned document, the non-overlapping regions $I_k(x, y)$, k=1,..., K, which correspond to graphics, photographs, and text in a frame

Then the image of scanned document $I(x, y)$, x=1,..., n; y=1,..., m, where n is the number of lines, m is the number of columns, is represented in the following form

$$I(x,y)=\sum_{k=1}^{K} I(x,y)\chi_k(x,y)+I(x,y)\chi(x,y). \quad (1)$$

Here $\chi_k(x, y)$, k=1,..., K; – the characteristic function of the k-th selected region of graphics, photographs, or text in a frame, K is the number of such regions; $\chi(x, y)$ is the characteristic function of the remaining regions of text and a background, and

$$\sum_{k=1}^{K}\chi_k(x,y)+\chi(x,y)=E,$$

where E is the matrix, composed of unities, of the same size as the image matrix $I(x, y)$. Then $I_k(x, y)=I(x, y)\ \chi_k(x, y)$, $k=1,...,K$.

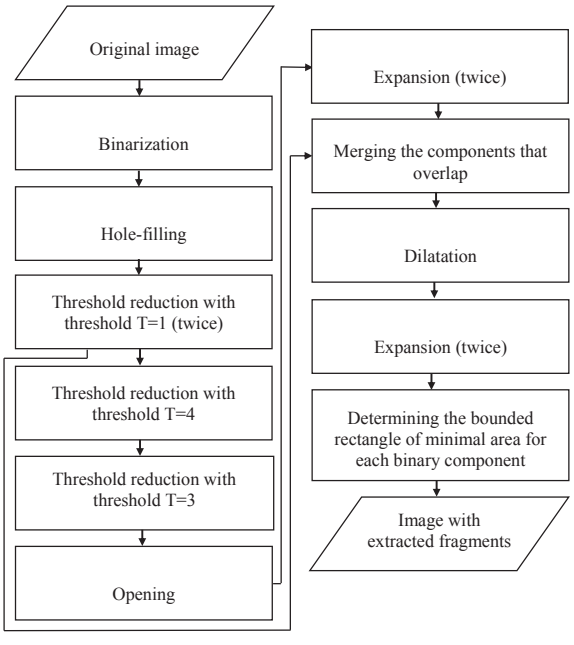


Fig. 2. Schematic of the Bloomberg method involving the filling of holes for extracting the fragments of an image of the scanned document

The boundaries of regions $I_k\ (x,\ y)$, $k=1,...,\ K$ that correspond to graphics, photographs, and text in a frame were identified at the first stage of the elaborated method of segmentation. At the second stage, with respect to the representation (1) of the scanned document image, we performed the feature extraction and classification of these regions. To this end, regions $I_k(x,\ y)$, $k=1,...,\ K$; $x=1,...,\ n_k$; $y=1,...,\ m_K$; of the scanned document image were split into blocks the size of $N\times N$, which are denoted $I_k^{ij}(x,\ y)$, $i=1,...,\ [n_k/N]$; $j=1,...,\ [m_k/N]$. Then each block $I_k^{ij}(x, y)$ was assigned a vector of two features $f(i,j)=(f_1(i,j),\ f_2(i,j))$. Compared to processing the neighborhood of each pixel, partitioning a region into blocks reduces the processing time and the amount of memory required to store the feature vectors.

To select the first feature $f_1(i,\ j)$, note that the regions of graphics and text are typically characterized by a much smaller number of grayscale than photographs. That makes it possible to distinguish the regions of graphics and text from photographs. Reducing the number of levels of gray predetermines an increase in the intensity difference between gradations. In order to estimate the intensity difference between gradations in the image, it is advisable to use the gradient magnitude. In the present work, to evaluate the gradient magnitude, we convoluted the region of graphics, photographs, or text in a frame $I_k(x, y)$, $k=1,...,\ K$; with masks G_1, G_2 from the Prewitt filter [13]:

$$G_{k1}(x, y)=I_k(x, y)*G_1,\ G_{k2}(x, y)=I_k(x, y)*G_2.$$

The derived image matrices $G_{k1}(x,\ y)$, $G_{k2}(x,\ y)$ were transformed according to the following formula:

$$G_k(x, y)=\sqrt{{G_{k1}}^2(x, y)+{G_{k2}}^2(x, y)}.$$

To enhance contrast, the images of regions $G_k(x, y)$ were treated with a histogram equalization [13]. During further extraction of the feature of these regions, such a transform considerably improved partitioning the object classes.

The gradient magnitude estimation $G_k(x,\ y)$ for each region $I_k(x, y)$ with graphics, photographs, or text in a frame was split into blocks the size of $N\times N$, and we then calculated in each block the mean value for the $G_k(x, y)$ – feature $f_1(i, j)$.

The choice of the second feature $f_2(i,\ j)$ to identify the regions of graphics, photographs, or text in a frame, is due to that graphics is an image containing mostly linear objects represented by contours of different width and length. Text characters could also be considered as linear objects, while a photographic image is mostly composed of areal objects represented by regions of uniform intensity. Linear objects are characterized in the scientific literature by ridges that is lines consisting of points that belong to an object, at which a local maximum is achieved in the direction of normal relative to this line. It is assumed in this case that a linear object is lighter than the background. If a linear object is darker than the background, it is characterized by valleys that are lines consisting of points that belong to an object, at which a local minimum is achieved in the direction of normal relative to this line. Therefore, in order to distinguish regions with graphics and text from photographs, we shall search for the ridges and valleys in these regions. Next, if the examined region of an image contains ridges or valleys, this region would be identified as graphics or text in a frame. Otherwise, the examined region of the image would be characterized as a photograph.

To identify graphics, text in a frame and photographs on the images of scanned documents, it is proposed to detect ridges and valleys through the analysis of eigenvalues of the Hessian matrix at each point in the examined region of the image [14]. This approach is characterized by high-speed performance, although by the low quality of detection in comparison with other groups of methods.

To find the partial derivatives $I_{xx}(x, y)$, $I_{xy}(x, y)$, $I_{yx}(x, y)$, $I_{yy}(x, y)$, for each region $I_k(x, y)$ with graphics, photographs or text in a frame we applied a filter with coefficients {–1, 1}. Next, according to the method from paper [14], we computed the Hessian matrix $H(x, y)$ for each pixel (x, y) within region $I_k(x, y)$, using formula

$$H(x,y)=\begin{pmatrix} I_{xx}(x,y) & I_{xy}(x,y) \\ I_{yx}(x,y) & I_{yy}(x,y) \end{pmatrix} \quad (2)$$

and determined the eigenvalues of this matrix $\lambda_h(x,\ y)$ and $\lambda_l(x,\ y)$, where $|\lambda_h(x,\ y)|\geq|\lambda_l(x,\ y)|$. It was assumed that the graphics and text are represented by dark lines against a lighter background; they then could be separated using the following conditions for the eigenvalues of the Hessian matrix at each point of the image: $\lambda_h(x, y)>>\lambda_l(x, y)$, $\lambda_l(x, y)\approx 0$, $\lambda_h(x,\ y)>0$.

The first two conditions make it possible to detect linear objects on the image. The last condition means that such objects are highlighted by dark color against a light background.

Next, at the image's points in which $\lambda_h(x,\ y)>0$ is satisfied, we compute two functions: $S(x, y)$ and $R_b(x, y)$. Function

$$R_b(x,y)=\frac{|\lambda_l(x,y)|}{|\lambda_h(x,y)|}$$

allows us to distinguish a linear object in the image from an area object. Since $|\lambda_h(x,y)|>>|\lambda l(x,y)|$, then $R_b(x, y)$ acquires values from segment [–1, 1], and for pixels in the linear objects the $R_b(x, y)$ values are close to zero. Function $S(x, y)$ characterizes the noise immunity of the image's representation using the eigenvalues of the Hessian matrix and is computed as the Frobenius norm of this matrix:

$$S(x,y)=\|H(x,y)\|_F=\sqrt{|\lambda_l(x,y)|+|\lambda_h(x,y)|}. \quad (3)$$

Values of function $S(x, y)$ are small in the image regions of homogeneous intensity. At the image points for which condition $\lambda_h(x, y)>0$ is not satisfied, functions $S(x, y)$ and $R_b(x, y)$ are considered equal to zero.

Based on functions $S(x, y)$ and $R_b(x, y)$, we determine function $G_{c,b}(x, y)$, x=1,..., N; y=1,..., M; which localizes in the images of scanned documents the linear objects that make up the graphics and text characters.

$$G_{c,b}(x,y)=\begin{cases}e^{-R_b^2(x,y)/2b^2}(1-e^{-S^2(x,y)/2c^2}), & \text{if } |\lambda_h(x,y)|>0,\\ 0, & \text{otherwise;}\end{cases} \quad (4)$$

where c, b are parameters. In [14], b was recorded equal to 0.5; c was considered equal to half the value, maximal for $\|H(x,y)\|_F$, x=1,..., N, y=1,..., M.

Values of function $G_{c,b}(x, y)$) for each region $I_k(x, y)$ with graphics, photographs or text in a frame were split into blocks the size of $N\times N$, and then in each block with indices i, j we calculated the mean intensity value for this function – feature $f_2(i, j)$.

The derived values of features for each region $I_k(x, y)$ with graphics, photographs or text in a frame were normalized using the following formula:

$$g_l(i,j)=\frac{f_l(i,j)}{f_{l\max}},\quad l=1,...,2, \quad (5)$$

where $g_l(i, j)$ is the normalized value for feature l in the block of image with indices i, j; $f_l(i, j)$ is the original non-normalized value for feature l in the same block; $f_{l\max}$ is the maximum value for feature l in a set of all blocks of regions with graphics, photographs, or text in a frame for all segmented images. Such a technique of normalization, in contrast to the standard normalization, does not introduce distortions to the geometry of feature space, resulting in overlapping of classes [15].

The derived vectors of features of blocks of pixels in the regions of photographs, graphics, and text in a frame on the images of scanned documents were classified by a polynomial support vector machine by optimizing the function of a mean root square error [16]. To construct a training set, we selected a set of images of scanned documents, for which we knew the result of ground-truth segmentation. At each such image we extracted regions corresponding to graphics, photographs, or text in a frame. These regions were split into blocks the size of $N\times N$, and then in each block we computed an feature vector $f^{train}(i,j)=(f_1{}^{train}(i,j), f_2{}^{train}(i,j))$, which was normalized using formula (5).

The training set consisted of input vectors representing the normalized feature vectors $g^{train}(i,j)=(g_1{}^{train}(i, j), g_2{}^{train}(i, j))$ of blocks of pixels in regions $I_k(x, y)$ of images and the target values. A target value $q(i, j)$ for each block i, j in the region of the ground-truth segmented image was determined as the most frequently occurring value for pixels' labels in a given block.

When forming a training set, we took into consideration that the input vectors and the target values depend on 4 indices, specifically: No. i, j of blocks horizontally and vertically in the image region $I_k(x, y)$, No. k of the image region, and the number of the image itself. Such a system of indices is appropriate when labeling of images, however, it complicates training a support vector machine and subsequent classification of feature vectors of the test set by a trained machine. Therefore, we constructed a mutually equal representation, which assigns to each set of 4 indices of feature vector of the image block a serial number p of this vector in the training set. Then we included to the training set the input vectors $g_p{}^{train}=(g_p1^{train}, g_p2^{train})$ and their corresponding target values q_p, p=1,..., P, where P is the number of vectors in a training set.

Fig. 3 shows examples of the training set in the space of two normalized and, accordingly, dimensionless features g_1, g_2. The feature g_1 is the normalized mean value of the local gradient for the image block, g_2 is the normalized mean value of the function that selects ridges and valleys for the block of image pixels. Markers «**p**», «o», «x» denote examples of the training set with target values "photograph", "text in a frame", "graphics", respectively.

It follows from Fig. 3 that the examples of the training set are separated nonlinearly, with class 3 objects (graphics) scattered over a large area of feature space. It is therefore preferable to separate, first, the class 1 objects (photographs) from the rest, then the class 2 objects (text in a frame) from the class 3 objects.

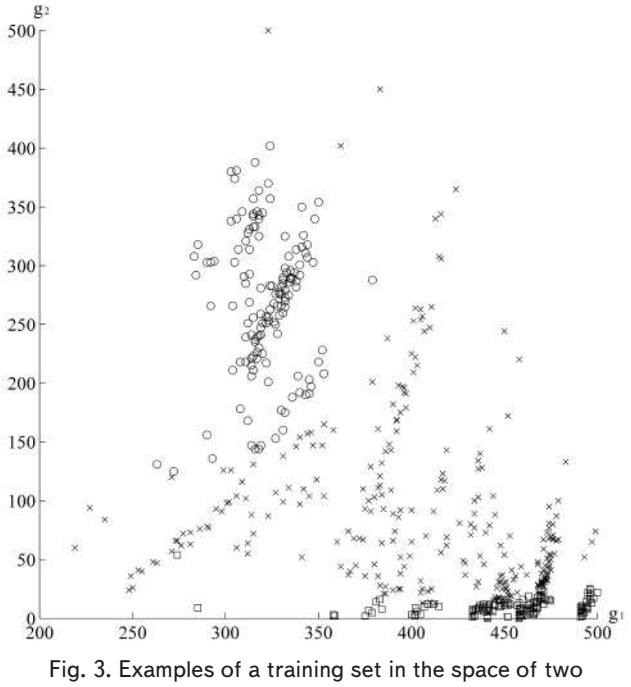


Fig. 3. Examples of a training set in the space of two normalized features

A support vector machine is designed for classification data into 2 classes. Because the problem under consideration requires that data should be classified into 3 classes: photographs, text in a frame, graphics, we built 2 support vector machines. The first support vector machine was trained to

separate the data class 1 objects (photographs) from other objects. Once a given support vector machine selected the data class 1 objects, they were removed from the training set. The remaining objects were used to train the second support vector machine to separate the data class 2 objects (text in a frame) from the remaining objects (graphics).

To highlight the text fragments in the images of scanned documents, it is proposed to use:

– a low-frequency filtering that smooths the image intensity values within homogeneous regions;

– a thresholding that selects homogeneous regions in the original image by comparison with the threshold value for the intensity of the smoothed image.

When implementing a low-frequency filtering, image processing is executed either in rows or in columns, or in a two-dimensional neighborhood of each pixel in the image. The second implementation of the low-frequency filtering is appropriate when the spectral representation of rows in the image is similar to the spectral representation of columns; in this case, the method is easier to configure. The spacing between text characters that defines the method parameters when processing rows in the image differs from the distance between text lines that define the method parameters when processing images based on columns. Therefore, it is more expedient to apply the first implementation for text extraction from the image of the scanned document.

Therefore, for the image of a text against a homogeneous background, the first to apply is the row-based filtering, whose result undergoes the thresholding. Next, the low-frequency filtering is performed for the columns of the image derived from the thresholding. Then again, the thresholding is performed. The low-frequency filter mask represents a sequence of 15 unities.

To choose the threshold following the low-frequency filtering of the image based on rows, we constructed an image histogram. The resulting histogram was smoothed by the Gaussian filter, and then the smoothed histogram was treated with logarithmic transformation to enhance the contrast of the peaks. The transformed histogram contained a peak in the area of light intensities corresponding to the background, as well as peaks in the area of dark intensities, corresponding to the text fragment. There are several peaks in the region of dark intensities, because, upon smoothing, the range of intensities corresponding to the text characters expanded.

Determining a threshold based on a histogram is performed in many segmentation methods using the Otsu method [17]. This method was developed to separate objects into 2 classes based on a single feature using a threshold. When extracting a text fragment from a homogeneous background, the threshold derived by the Otsu method turned out to be too low. This is explained by the fact that following the smoothing of individual characters in a text fragment, the fragment boundaries were blurred.

Preliminary study has shown that in order to detect the boundaries of a text fragment after smoothing the individual characters, the threshold is better to choose under the histogram peak corresponding to the background. After smoothing the individual characters, a text fragment in the histogram was matched by several weakly expressed peaks in the area of dark intensities. It is therefore proposed to differentiate the histogram twice as frequency function from the values of intensities. Then the base of the histogram's peak corresponding to the background matches the maximum of values for the second derivative. Therefore, we chose, as a threshold for extracting a text fragment, the argument (intensity) of the meaningful local maximum of the histogram that is the closest among the rest of local maxima to the right end of the interval of the image intensities.

5. Experimental research of the combined method using a test database

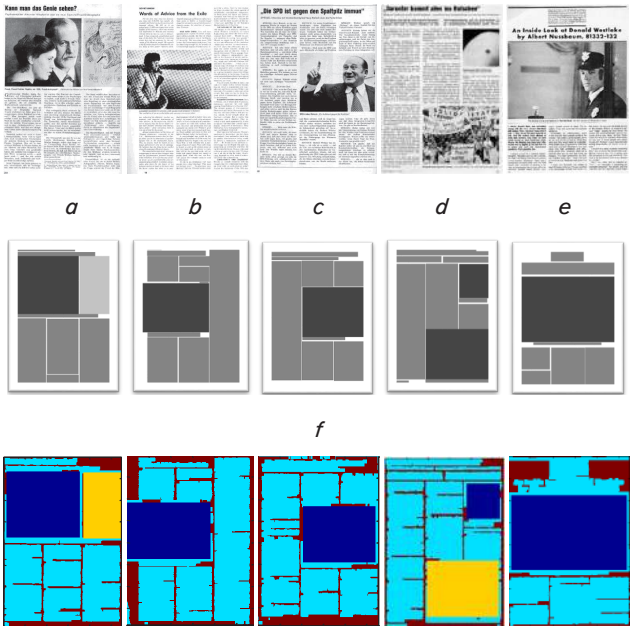
By conducting an experiment, we compared the quality of segmenting the images of scanned documents for the proposed combined method and methods elaborated in [7, 9].

We estimated segmentation quality for 100 test images of the scanned newspaper articles from the document database MediaTeam at Oulu University (Finland) [18]. These images of the size of 3,300×2,600 pixels were scanned at a resolution of 300 dpi. Such resolution makes it possible to obtain high-quality images, however, the size of these images imply a large amount of computation. To reduce the amount of computation, the image was processed by a smoothing filter with a mask of 11×11 units, and then it was scaled by a factor of 0.25 vertically and horizontally using decimation [9]. Next, the reduced image was treated with gamma correction at coefficient γ=2.2.

The images that we used contained headers, text, charts, pictures, photographs, as well as separators in the form of lines and/or frames. For each test image, by using the information included in the database, the corresponding ground-truth segmentation is formed. Fig. 4, 5 show the original images of scanned documents, the ground-truth segmentation for each such image, and the result of segmentation using the proposed combined method. Fig. 6, 7 demonstrate the original images of scanned documents and the result of segmentation applying methods from papers [7] and [9], respectively.

The quality of segmentation when using the proposed method and methods from papers [7, 9] in comparison to the ground-truth segmentation was assessed by constructing confusion matrices. A confusion matrix is a square matrix whose rows correspond to the labels of classes for ground-truth image pixels. Columns in a confusion matrix match the labels of classes for image pixels, obtained with the help of the examined method. The element of the confusion matrix that is at the intersection of the i-th row and j-th column shows the percentage of the image pixels from a class with the i-th label assigned to the class with the j-th label. The sum of the elements in each row in the confusion matrix is 100 %.

The proposed method and methods from papers [7, 9] were applied in order to extract on the images of scanned documents the segments of 4 classes: photographs, text, graphics, and background. That is why the size of the confusion matrix is 4×4 elements (Tables 1–3). Tables 1, 2 give the confusion matrices for the proposed method with different processing blocks; Table 3 – for a method from paper [9], which used the same test image database. The represented confusion matrices were derived by averaging the results of the segmentation quality assessment for all examined test images.

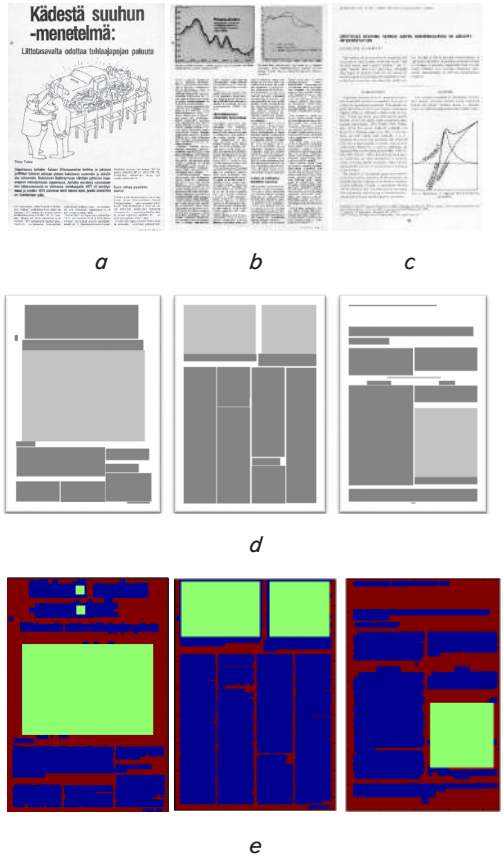


a *b* *c* *d* *e*

f

g

Fig. 4. Results of segmenting the images of scanned documents using the proposed method: *a*–*e* – original image; *f* – ground-truth segmentation; *g* – result of segmentation with the proposed method (blue color denotes text, red – background, deep blue – photographs, yellow – graphics)

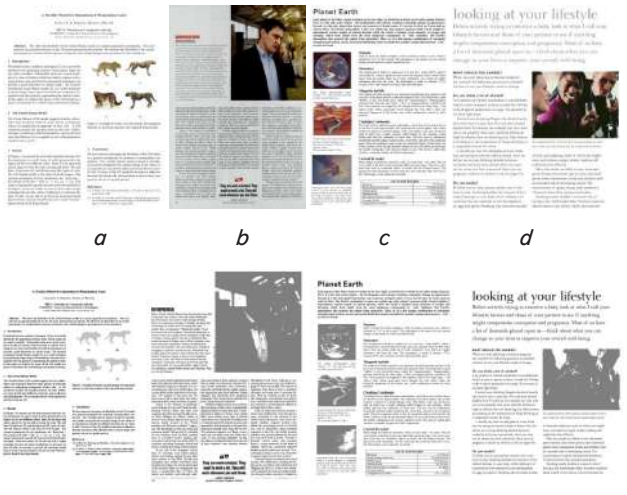


a *b* *c*

d

e

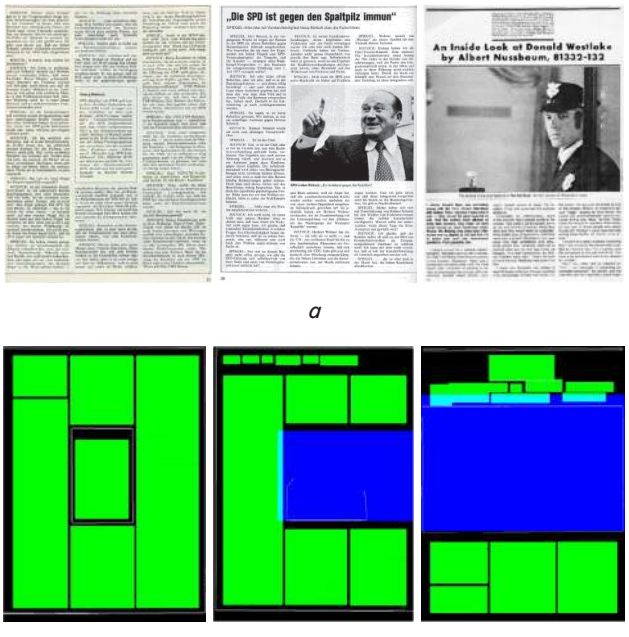
Fig. 5. Results of segmenting the images of scanned documents using the proposed method: *a*–*c* – original image; *d* – ground-truth segmentation; *e* – result of segmentation with the proposed method (deep blue color denotes text, red – background, cyan – graphics)



a *b* *c* *d*

e

Fig. 6. Results of segmenting the images of scanned documents using a method from paper [7]: *a*–*d* – original image; *e* – result of segmentation (grey color denotes photographs)



a

b

Fig. 7. Results of segmenting the images of scanned documents using a method from paper [9]: *a* – original image; *b* – result of segmentation (green color denotes text, blue – separator, deep blue – photographs)

Table 1

Confusion matrix for the proposed method, processing the 32×32-pixel blocks

Results of ground-truth segmentation, %	Results of image segmentation by the proposed method, %			
	Photograph	Text	Graphics	Background
Photograph	91.05	6.19	0	2.76
Text	0	89.95	0.08	9.97
Graphics	0	5.05	84.30	10.65
Background	0	11.12	0	88.88

Table 2

Confusion matrix for the proposed method, processing the 48×48-pixel blocks

Results of ground-truth segmentation, %	Results of image segmentation by the proposed method, %			
	Photograph	Text	Graphics	Background
Photograph	90.04	4.24	3.53	2.19
Text	0	89.97	0	10.03
Graphics	0	5.05	84.30	10.65
Background	0	11.13	0	88.87

Table 3

Confusion matrix for a method from paper [9]

Results of ground-truth segmentation, %	Results of image segmentation by a method from paper [9], %		
	Photograph	Text	Background
Photograph	96	0	4
Text	3	88	9
Background	2	1	97

The segmentation accuracy values for other methods, given in Table 4, are taken from papers [7–9].

Table 4

Comparative assessment of segmentation accuracy when using the proposed method and methods known from the scientific literature

Reference, year of publication	Segmentation accuracy estimation, %
[10], 2011	89.54
[12], 2010	95.01
[9], 2012	93.67
[7], 2016	95.21
[8], 2016	96.95
Combined method, processing the 32×32-pixel blocks	88.54
Combined method, processing the 48×48-pixel blocks	88.47

The speed of performance of the proposed method of segmentation, compared to methods from papers [7–9], was assessed based on the processing time. Table 5 gives values for the time of processing the images from the test database when using the proposed method. The authors of this paper employed the Intel Core i5-7400 processor, 3 GHz CPU, 16 GB memory, Windows 10 operating system, 64 bits. In paper [9], the study was performed using the dual-core processor, 2.4 GHz CPU, the average processing time for the image the size of 3,000×2,000 pixels was 15 seconds. Table 6 gives values for the time of processing the images from a test database using methods from papers [7, 8]. Papers [7, 8] exploited the Intel Core i7 processor, 2.7 GHz CPU, 16 GB memory, Windows 7 operating system, 64 bits. Those papers used a test database, which is different from the one used in this paper. The average time to process images in [7] was 107.28 seconds; in [8] – 95.8279 seconds. The results regarding the processing time when using the proposed method and known methods are given in Tables 5, 6.

Table 5

Image processing time when using the combined method, 32x32-pixel blocks

Original image	Image size	File size, MB	Processing time, s
P00539.jpg (Fig. 4, *a*)	3,130×2,195	1,16	2.520586
P00616.jpg (Fig. 4, *b*)	3,101×2,309	1.58	2.016973
P00531.jpg (Fig. 4, *c*)	3,134×2,220	1.07	1.866177
P00533.jpg (Fig. 4, *d*)	3,112×2,207	1.12	2.141444
P00545.jpg (Fig. 4, *e*)	3,062×2,195	1.21	2.322482
P00838.jpg (Fig. 5, *a*)	3,396×2,334	1.00	2.566932
P00834.jpg (Fig. 5, *b*)	3,388×2,338	1.71	2.053916
P00811.jpg (Fig. 5, *c*)	2,711×1,789	0.61	1.123459

Table 6

Image processing time when using methods from papers [7, 8]

Original image	Image size	File size, MB	Processing time [7], s	Processing time [8], s
24.png (Fig. 6, *d*)	2,161×2,776	2.18	149.078	138.907
1.jpg (Fig. 6, *c*)	2,479×3,508	1.46	179.772	161.591
2.jpg (Fig. 6, *b*)	2,303×3,136	1.95	175.154	166.587
A24.bmp (Fig. 6, *a*)	739×1,123	2.54	96.184	87.911

Table 7 gives the average processing time of an image at each stage of the proposed method and the average time of image segmentation when using the proposed method (last column).

Table 7

Average image processing time when using the proposed method, in stages

Pre-processing, s	Analysis of connected components	Block processing, s	Pixel neighborhoods processing, s	Total processing time, s
0.498410	0.559466	1.156980	0.139885	2.332494

Although the hardware and software resources of computers employed in the experiments in papers [7–9] and those applied in this paper differ, the proposed method has a significant advantage in performance speed.

6. Discussion of results of research the speed of performance and quality of the proposed method for image segmentation

The results from Tables 1, 2, 4 demonstrate that the quality characteristics of segmenting the images of scanned documents when using the proposed method almost do not change at a different size of the processed block. We compared the confusion matrices derived when performing image segmentation using the proposed method (Tables 1, 2)

and applying a method from paper [9] (Table 3) exploiting the same database of test images. The proposed method, in terms of the percentage of correct recognition of photograph regions, is inferior to a method from paper [9] by 5–6 %; for text regions, it is better by 1 %; for background regions, it is worse by 8 %, compared to [9]. When compared with the image segmentation methods for scanned documents known from the scientific literature, the proposed method is inferior in terms of segmentation accuracy by 8 % (Table 4). At the same time, the following errors in segmentation occurred when the proposed method was used:

– when a graphics region includes multiple pictures, they could be extracted as separate segments;

– if a graphics region includes the separately located open lines, the fragments of the graphics that correspond to these lines are labeled as a background;

– if a photograph region includes a small object against the background of uniform intensity, this region can be extracted as graphics;

– fragments of a text region with large font-size could be detected as graphics.

However, when the proposed method of segmentation is used, the image processing of scanned documents is performed sequentially. That reduces, in comparison with known segmentation methods, the percentage of the errors, contained in the confusion matrix beyond the main diagonal (Tables 1–3).

Further research could address resolving the following tasks:

– extraction of the connected components in order to determine the boundaries of illustrations: a more sophisticated technique to choose the image binarization threshold could improve the quality of determining the boundaries of the illustrations, provided the performance speed of the combined segmentation method is not compromised;

– the approximation of the extracted text fragments by rectangular regions would also improve the quality of the resulting segmentation when compared with a ground-truth segmentation.

7. Conclusions

1. We have proposed a combined method for segmenting the images of scanned documents. A given method differs from the segmentation methods known from the literary sources in that it sequentially detects, first, the boundaries of illustrations using an analysis of the connected components, followed by the separation of illustrations into photographs and graphics applying the block segmentation method. Next, the text fragments are extracted employing the processing in the neighborhood of each pixel.

When detecting the boundaries of illustrations, the image was first binarized, next, the closed contours were filled, and then the Bloomberg method, known from the literature, was applied. To separate into photographs and graphics, the illustrations were split into blocks of pixels. Each block of pixels was assigned with a vector of 2 features: the mean value of the local gradient magnitude, and the mean value of the function that localizes on the images of scanned documents the linear objects that make up the graphics and text characters. The derived feature vectors were classified using a polynomial support vector machine.

When extracting the text fragments, we used a low-frequency filtering, enabling to smooth the image intensity values within homogeneous regions, as well as the thresholding, by comparing the intensity values for the smoothed image with the threshold. These two transformations were applied sequentially, first in rows, then in columns, and allowed us to extract the homogeneous regions of interest in the original image.

The proposed method has made it possible to improve the performance of segmenting the images of scanned documents at high quality processing.

2. The results of experiments have shown that the proposed method is characterized by a significant advantage in performance speed compared to the methods for segmenting the images of scanned documents, known from the scientific literature, although it is inferior by 8 % in terms of the segmentation accuracy. Therefore, the proposed method for the segmentation of images of scanned documents could be recommended to apply for tasks that require enhanced performance speed. The further research will focus on improving the quality of detecting the boundaries of illustrations using an analysis of the connected components.

Acknowledgements

Authors express their gratitude and deep appreciation to Dr. of Eng., Professor of Department of Applied Mathematics and Information Technologies at Odessa National Polytechnic University (Ukraine), Mr. V.N. Krylov for valuable and constructive advice and comments in the preparation of this paper.

References

1. Haneda E., Bouman C. A. Text Segmentation for MRC Document Compression // IEEE Transactions on Image Processing. 2011. Vol. 20, Issue 6. P. 1611–1626. doi: https://doi.org/10.1109/tip.2010.2101611
2. Polyakova M., Ishchenko A., Huliaieva N. Document image segmentation using averaging filtering and mathematical morphology // 2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET). 2018. doi: https://doi.org/10.1109/tcset.2018.8336354
3. Muralikrishna V., Sai Ram M. S. Image segmentation of scanned document using back-propagation artificial neural network based technique // International Journal of Computers and Communications. 2012. Vol. 6, Issue 14. P. 183–190.
4. Sasirekha D., Chandra E. Enhanced techniques for PDF image segmentation and text extraction // International Journal of Electronics and Computer Science Engineering. 2012. Vol. 10, Issue 9. P. 1833–1838.
5. Obnaruzhenie i lokalizaciya tekstovyh oblastey na polutonovyh cifrovyh izobrazheniyah / Korennoy A. V., Yudakov D. S., Dedov S. V., Strazhnik V. P. // Vestnik VGU. Sistemnyy analiz i informacionnye tekhnologii. 2015. Issue 4. P. 65–72.

6. Kundu M. K., Dhar S., Banerjee M. A new approach for segmentation of image and text in natural and commercial color documents // 2012 International Conference on Communications, Devices and Intelligent Systems (CODIS). 2012. doi: https://doi.org/10.1109/codis.2012.6422142
7. Abdullah H. S., Jassim A. H. Improved fuzzy c-means for document image segmentation // British Journal of Science. 2016. Vol. 14, Issue 2. P. 1–15.
8. Abdullah H. S., Jasim A. H. Improved Ant Colony Optimization for Document Image Segmentation // International Journal of Computer Science and Information Security (IJCSIS). 2016. Vol. 14, Issue 11. P. 775–785.
9. Text, photo, and line extraction in scanned documents / Erkilinc M. S., Jaber M., Saber E., Bauer P., Depalov D. // Journal of Electronic Imaging. 2012. Vol. 21, Issue 3. P. 033006. doi: https://doi.org/10.1117/1.jei.21.3.033006
10. Bukhari S. S., Shafait F., Breuel T. M. Improved document image segmentation algorithm using multiresolution morphology // Document Recognition and Retrieval XVIII. 2011. doi: https://doi.org/10.1117/12.873461
11. A Document Image Segmentation System Using Analysis of Connected Components / Zirari F., Ennaji A., Nicolas S., Mammass D. // 2013 12th International Conference on Document Analysis and Recognition. 2013. doi: https://doi.org/10.1109/icdar.2013.154
12. Document image segmentation using discriminative learning over connected components / Bukhari S. S., Al Azawi M. I. A., Shafait F., Breuel T. M. // Proceedings of the 8th IAPR International Workshop on Document Analysis Systems – DAS '10. doi: https://doi.org/10.1145/1815330.1815354
13. Gonsales R., Vuds R. Cifrovaya obrabotka izobrazheniy. Moscow: Tekhnosfera, 2005. 1072 p.
14. Multiscale vessel enhancement filtering / Frangi A. F., Niessen W. J., Vincken K. L., Viergever M. A. // Lecture Notes in Computer Science. 1998. P. 130–137. doi: https://doi.org/10.1007/bfb0056195
15. Mandel' I. D. Klasterniy analiz. Moscow: Finansy i statistika, 1988. 176 p.
16. Chu W., Keerthi S. S., Ong C. J. A general formulation for support vector machines // Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02. 2002. doi: https://doi.org/10.1109/iconip.2002.1201949
17. Otsu N. A Threshold Selection Method from Gray-Level Histograms // IEEE Transactions on Systems, Man, and Cybernetics. 1979. Vol. 9, Issue 1. P. 62–66. doi: https://doi.org/10.1109/tsmc.1979.4310076
18. Sauvola J., Kauniskangas H. MediaTeam Document Database II: a collection of document images. University of Oulu. Finland, 1999.