UDC 004.93
DOI: 10.15587/2312-8372.2019.173913

Ishchenko A.,
Zhuchkovskyi V.

ELABORATION OF THE HIERARCHICAL APPROACH TO SEGMENTATION OF SCANNED DOCUMENTS IMAGES

Об'єктом дослідження є процес розпізнавання областей зображень відсканованих документів. У роботі запропоновано ієрархічний підхід до сегментації зображень відсканованих документів. Даний підхід представляє собою зображення відсканованого документа у вигляді багаторівневої структури. На кожному рівні структури виділені зображення, що містять структурні області. Об'єкти нижнього рівня строго співвідносяться з певною областю зображення верхнього рівня: області фото та графіки співвідносяться із зображенням, що містить ілюстрації, а області тексту та фону – із зображенням, що містить і текст, і фон одночасно. Використання ієрархічного підходу дозволяє виконувати обробку окремо для кожної області зображення, а саме: спочатку на оригінальному зображенні відсканованого документа за допомогою аналізу зв'язкових компонент виділяються області ілюстрацій. Тим самим перший рівень ієрархії складають зображення, що містить ілюстрації, та зображення, що містить текст із фоном. Потім області ілюстрацій розділяються на фото та графіку за допомогою розбиття областей ілюстрацій на блоки певного розміру, а текстові області виділяються з фону за допомогою обробки в околиці кожного пікселя. Тим самим другий рівень ієрархічної структури представляють зображення, що містять однорідні області: фото, графіку, текст і фон. Ієрархічний підхід до сегментації дозволив скоротити час обробки в середньому в 80 разів. Скороченню часу обробки зображення сприяло те, що на кожному рівні і в, свою чергу, в окремій частині ієрархічної структури була можливість врахувати структурні ознаки однорідної області зображення, яке відповідає даному рівню. А також вибрати ознаки ідентифікації цих областей з високою обчислювальною ефективністю, використання яких також дозволило скоротити час обробки відсканованого документа.

Ключові слова: *ієрархічний підхід, відскановані документи, сегментація зображень, обробка зображень, однорідні області.*

Received date 27.11.2018
Accepted date 12.12.2018
Published date 28.06.2019

1. Introduction

The rapid development of digital technologies has led to the conversion into electronic form of all types of materials, including documents from archives, libraries, and enterprises to create electronic archives.

An important stage in the processing of images of scanned documents is segmentation, which consists in dividing the image into homogeneous areas that are similar in one feature or set of features.

In the literature, there are 2 main approaches to the segmentation of scanned documents images [1]: descending and ascending. Segmentation methods that use a descending approach [2–4], first determine the objects of a higher level of the page structure – text and graphic elements, and then – columns of text, paragraphs, lines, symbols of text. These methods are characterized by high speed, but low quality segmentation, since it is not always possible to process non-rectangular areas of text or headings that occupy several columns of text. Segmentation methods that use an upward approach to segmentation [5, 6] begin processing with text characters that are combined into paragraphs, paragraphs, and columns. Next, the resulting objects are classified as text areas until the entire text areas are highlighted. These methods are distinguished by high quality segmentation, since they process images with a complex shape well, but have low speed, due to the fact that they require processing of each pixel first, and then the document areas. When digitizing a large number of printed documents, the requirements for the speed of their processing increase. As a consequence, an urgent task, which is solved in this work, is increasing the speed of processing images of scanned documents by reducing their processing time, using a hierarchical approach to image segmentation. Therefore, *the object of research* is the process of recognizing the areas of images of scanned documents. The *aim of research* is using a hierarchical approach to the segmentation of images of scanned documents to reduce the processing time of images with sufficient quality segmentation.

2. Methods of research

The basis of any scanned document is its structure that is the mutual arrangement of graphic material and text. Regions that include uniform content, such as text only, graphics only, or only photos, form structural regions.

One of the main stages of image processing of scanned documents is image segmentation. Areas of text, graphics and photos are highlighted when segmentation on the images. Each of these structural areas has different properties; therefore, it is difficult to select a feature system for selecting text areas, as well as graphic and photo areas from the background.

The existing methods of segmentation of scanned documents images do not simultaneously satisfy the requirements for short processing time with sufficient segmentation quality.

The need for a hierarchical organization of the system is necessary if its implementation requires the expenditure of a large amount of time, which is unacceptable for this system. Therefore, in order to reduce the processing time of a document image in this paper, let's propose to use a hierarchical approach to the segmentation of images of scanned documents. This approach consists in the representation of an image in the form of a multilevel structure, in which there is a division of the set of its constituent objects into subsets of different levels with the property of integrity [7]. The hierarchical structure allows the processing of individual information arrays. That is, this approach allows processing for each level of image representation. According to the hierarchical approach, the image is first decomposed into areas of illustrations, including both photos and graphics, followed by their classification, as well as areas containing text and background, followed by highlighting text areas from the background. As a result of image segmentation using the proposed approach, the image of the scanned document is presented as separate areas: text, graphics, photo and background. The hierarchical approach is shown in Fig. 1.

According to the hierarchical approach to image segmentation (Fig. 1), the first level corresponds to the original image of the scanned document. As the base model for the representation of this image, the most frequently used model of mixed raster content in literature is Mixed Raster Content [8]. According to this model, the image of the document is represented as an image of a mask, as well as images of the foreground and background. Each of these images contains objects of a certain class and is independently compressed by certain encoders. The mask contains information on the relative position of the foreground and background objects in the image. In order to be able to extract information from structural regions, it is necessary to present an image that represents these structural regions in separate images.

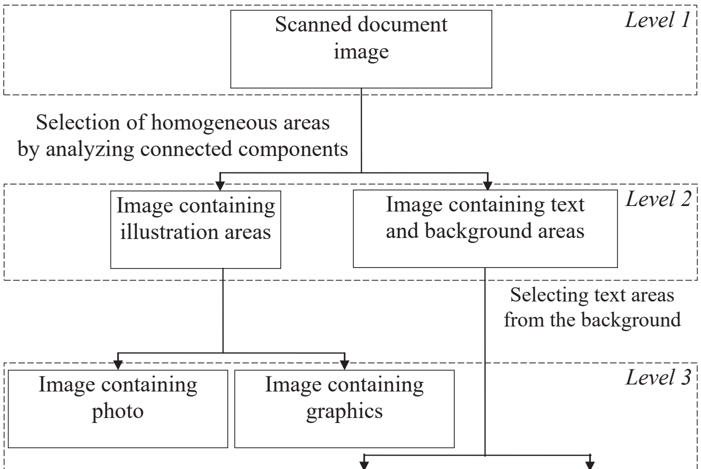
Level 1
Scanned document image
Selection of homogeneous areas by analyzing connected components
Level 2
Image containing illustration areas
Image containing text and background areas
Selecting text areas from the background
Level 3
Image containing photo
Image containing graphics
Image containing text
Image containing background

Fig. 1. The scheme of the hierarchical approach to the segmentation of images of scanned documents

Therefore, in this research, the model of work [9] is used as an image representation. This model is different in that it represents an image of a scanned document as a set of images, each of which contains one class of a uniform area – graphics, or photos, or text on a uniform background. This image model represents text areas as a structural texture with text symbols as non-derivative elements, and illustration areas as areas of constant intensity. Such a representation of the image of the scanned document allows to take into account the structural properties of homogeneous areas and choose a system of signs for their identification, which have high computational efficiency, which entails a reduction in the processing time of the image with sufficient quality segmentation.

The next step is dividing the image of the scanned document into an image containing illustrations and an image containing text areas and a background. To do this, use the method developed in [10] to select illustrations on an image of a scanned document using averaging filtering. According to this method, homogeneous areas of text and illustrations are distinguished by analyzing connected components. The use of analysis of connected components allows to distinguish the areas of illustrations from areas of text, since they are connected components that correspond to the characters of the text, differ from the connected components corresponding to the illustrations, in their form and periodicity. That is, the size of the connected components corresponding to the characters of the text acts as a sign for identifying the text. Using a hierarchical approach to segmentation allows to compute simple, computationally, features for text areas and segment the original image of the scanned document into 2 images: one of them contains illustrations on a uniform background, the other is text and background. The use of «simple» signs reduces the processing time of the image, and the use of averaging filtering when selecting areas of illustrations preserves a sufficiently high image segmentation quality.

According to the hierarchical approach (Fig. 1), the third level of the hierarchy presents images containing photos, graphics, text, and background, separately from each other.

First, the areas of illustrations are divided into homogeneous areas, which differ in their structure: photos and graphics. To do this, let's consider the part of the hierarchical structure in which the image is located, containing areas of illustrations. To this end, a method proposed in [11] was proposed for identifying graphic and photo areas using statistical and geometric features. To reduce image processing time when identifying photo and graphics areas, the illustration areas are divided into blocks of fixed size. When block processing is usually the quality of segmentation is reduced, it is therefore necessary to use such signs of identifying homogeneous areas that have high computational efficiency, that is, provide sufficient quality segmentation with a short image processing time. Therefore, it is proposed to use the aspect ratio of objects and the estimate of the expectation of the height of the intensity drop at the borders of homogeneous areas as identification signs in [11]. As the first, a feature is chosen that characterizes an object representing graphics as an image that contains linear objects [12], represented by contours of different widths and lengths. The second feature allows to distinguish the graphic areas from the photo, given that the graphic areas are characterized by fewer shades of gray than the photos, which entails an increase in the height of the intensity differences between the gradations.

Next, the text areas on the images of scanned documents are separated from the background. For this purpose, the method of selection of text areas [13] is used, which allows to take into account the distances between text characters when processing image lines and the distance between text lines when processing images in columns. Processing in the vicinity of each pixel of the image reduces processing time and maintains a sufficiently high image quality.

3. Research results and discussion

To estimate the processing time of scanned documents images, using a hierarchical approach to image segmentation, images of scanned articles and journals of the MediaTeam Oulu document database [14] were used.

Experimental studies were conducted using a computer that has the following technical and system characteristics: Intel Core i5-3210 processor, 2.5 GHz CPU, 6 GB RAM, 64-bit Windows 7 operating system. 92 test images of scanned documents were selected that contained the main text, headlines, photos and graphics. Images of the used document base are 3200×2300 pixels in size and scanned with a resolution of 300 dpi, which gives a rather high quality of the scanned document. Examples of analyzed images are presented in Fig. 2.

The results of the experiment were compared with the results of [2, 3, 5], in which the studies were performed using a 2.4 GHz dual-core processor and a 64-bit Windows 7 operating system. The technical and system characteristics of this study and the analyzed works are comparable, therefore the results of these works can be added to the comparison.

The results of the processing time to use a hierarchical approach to segmentation and known approaches are given in Table 1.

The system and hardware resources of computers that were used for experiments in [2, 3, 5] and in this study had comparable characteristics. The results presented in Table 1 allow to conclude that the use of the proposed hierarchical approach to the segmentation of images of scanned documents has significant advantages in speed.

Further studies can be aimed at solving the segmentation problem in order to increase speed by introducing parallel processing of homogeneous areas at one level of the structure.

4. Conclusions

In this article it is proposed to use a hierarchical approach in the segmentation of scanned documents images. This approach allows homogeneous regions of an image to be represented using a tree structure in which the regions of the lower level strictly correlate with a specific region of the image of the upper level. According to this approach, the areas of illustrations and areas containing text and background are objects of the same level of structure. A field of photos, graphics, text and background presented at a different level. This allows to perform processing separately for each image area, as well as significantly reduce processing time at each level.

Acknowledgments

The author expresses gratitude and deep appreciation for the valuable and constructive advice and comments when working on this research to colleagues from the Department of Applied Mathematics and Information Technologies of the Odessa National Polytechnic University (Ukraine):

– Doctor of Technical Sciences, Associate Professor M. Poliakova;

– Doctor of Technical Sciences, Prof. V. Krylov.

Fig. 2. Examples of scanned documents images of the MediaTeam Oulu database [14]

Table 1

Image processing time by the analyzed methods

Cluster center-based classification method [2]	Fisher classifier method [2] descending approach	Method [3] descending approach	Method [5] ascending approach	Hierarchical approach
158 s	139 s	286 s	155 s	2.3 s

References

1. Shafait, F., Keysers, D., Breuel, T. M. (2008). Performance Evaluation and Benchmarking of Six-Page Segmentation Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30 (6),* 941–954. doi: http://doi.org/10.1109/tpami.2007.70837
2. Kumar, S., Gupta, R., Khanna, N., Chaudhury, S., Joshi, S. D. (2007). Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model. *IEEE Transactions on Image Processing, 16 (8),* 2117–2128. doi: http://doi.org/10.1109/tip.2007.900098
3. Acharyya, M., Kundu, M. K. (2001). Multiscale Segmentation of Document Images Using M-Band Wavelets. *Lecture Notes in Computer Science,* 510–517. doi: http://doi.org/10.1007/3-540-44692-3_62
4. Cesarini, F., Gori, M., Marinai, S., Soda, G. (1999). Structured document segmentation and representation by the modified X-Y tree. *Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No. PR00318),* 563. doi: http://doi.org/10.1109/icdar.1999.791850
5. Baird, H. S., Moll, M. A., An, C., Casey, M. R. (2007). Document image content inventories. *Document Recognition and Retrieval XIV*. doi: http://doi.org/10.1117/12.705094
6. Vilkin, A., Egorova, M. (2010). Segmentatsiia otskanirovannykh dokumentov. *GrafiKon'2010*, 339–341.
7. Moiseev, N. N. (1981). *Matematicheskie zadachi sistemnogo analiza.* Moscow: Nauka, 487.
8. de Queiroz, R. L., Buckley, R. R., Xu, M. (1999). Mixed Raster Content (MRC) Model for Compound Image Compression. *Visual Communications and Image Processing*. San Jose, 3653, 1106–1117. doi: http://doi.org/10.1117/12.334618
9. Ishchenko, A., Polyakova, M., Kuvaieva, V., Nesteryuk, A. (2018). Elaboration of structural representation of regions of scanned document images for MRC model. *Eastern-European Journal of Enterprise Technologies, 6 (2 (96)),* 32–38. doi: http://doi.org/10.15587/1729-4061.2018.147671
10. Polyakova, M., Ishchenko, A., Huliaieva, N. (2018). Document image segmentation using averaging filtering and mathematical morphology. *Telecommunications and Computer Engineering (TCSET)*. Lviv-Slavske. doi: http://doi.org/10.1109/tcset.2018.8336354
11. Polyakova, M., Ishchenko, A., Volkova, N., Pavlov, O. (2018). Combined method for scanned documents images segmentation using sequential extraction of regions. *Eastern-European Journal of Enterprise Technologies, 5 (2 (95))*, 6–15. doi: http://doi.org/10.15587/1729-4061.2018.142735
12. Magnier, B., Montesinos, P., Diep, D. (2011). Ridges and Valleys Detection in Images Using Difference of Rotating Half Smoothing Filters. *Lecture Notes in Computer Science*. Ghent, 261–272. doi: http://doi.org/10.1007/978-3-642-23687-7_24
13. Gusak, D. E., Ishhenko, A. V. (2019). Vydelenie tekstovykh fragmentov na izobrazhenii otskanirovannogo dokumenta. *Suchasni informatsiini tekhnologii.* Odessa.
14. Sauvola, J., Kauniskangas, H. (1999) *MediaTeam Document Database II, a CD-ROM collection of document images*. University of Oulu.

Ishchenko Alesya, *Senior Lecturer, Department of Applied Mathematics and Information Technologies, Odessa National Polytechnic University, Ukraine, e-mail: alesya.ishchenko@gmail.com, ORCID: http://orcid.org/0000-0002-7882-4718*

Zhuchkovskyi Vladyslav, *Department of Applied Mathematics and Information Technologies, Odessa National Polytechnic University, Ukraine, e-mail: pzmaus45@gmail.com, ORCID: http://orcid.org/0000-0001-7364-784X*