International Journal of Advanced Smart Convergence Vol.10 No.1 1-11 (2021)
http://dx.doi.org/10.7236/IJASC.2021.10.1.1

IJASC 21-1-1

Object detection in financial reporting documents for subsequent recognition

Petr Sokerin[1], Alla Volkova[2], Kirill Kushnarev[1]

[1]*Research laboratory «Monetary policy research and financial market analysis», Plekhanov Russian University of Economics, Russia*
[2]*KPMG Taxes and Consulting, LLC, Russia*
E-mail sokerinpo@mail.ru

Abstract

Document page segmentation is an important step in building a quality optical character recognition module. The study examined already existing work on the topic of page segmentation and focused on the development of a segmentation model that has greater functional significance for application in an organization, as well as broad capabilities for managing the quality of the model. The main problems of document segmentation were highlighted, which include a complex background of intersecting objects. As classes for detection, not only classic text, table and figure were selected, but also additional types, such as signature, logo and table without borders (or with partially missing borders). This made it possible to pose a non-trivial task of detecting non-standard document elements. The authors compared existing neural network architectures for object detection based on published research data. The most suitable architecture was RetinaNet. To ensure the possibility of quality control of the model, a method based on neural network modeling using the RetinaNet architecture is proposed. During the study, several models were built, the quality of which was assessed on the test sample using the Mean average Precision metric. The best result among the constructed algorithms was shown by a model that includes four neural networks: the focus of the first neural network on detecting tables and tables without borders, the second - seals and signatures, the third - pictures and logos, and the fourth - text. As a result of the analysis, it was revealed that the approach based on four neural networks showed the best results in accordance with the objectives of the study on the test sample in the context of most classes of detection. The method proposed in the article can be used to recognize other objects. A promising direction in which the analysis can be continued is the segmentation of tables; the areas of the table that differ in function will act as classes: heading, cell with a name, cell with data, empty cell.

Keywords: *segmentation, neural networks, RetinaNet, document analysis, machine learning.*

1. Introduction

The mission of segmentation is very important for working with documents. Segmentation makes it easier to present the scanned image for further work with it. Basically, object detection models in a scanned document are used to automate various document management processes in organizations of any field of activity. The aim of our research was to develop an adaptive method of page segmentation. The key to building a high-quality document processing tool is how we will segment documents, that is, divide documents into single-

Manuscript Received: December 10, 2020 / Revised: December. 15, 2020 / Accepted: December. 19, 2020
Corresponding Author: sokerinpo@mail.ru
Tel: +7-913-223-05-70
Research laboratory «Monetary policy research and financial market analysis», Plekhanov Russian University of Economics, Russia; KPMG Taxes and Consulting, LLC, Russia

native areas. To date, there are a number of scientific papers that offer their solutions on this topic. The main disadvantage of most of them is the low accuracy of recognition of non-standard documents, for example, with a problematic background or with a combination of non-text and text groups of pixels. The classic approach involves selecting only three types of elements for segmentation (text, figure, and table).

After investigating the shortcomings of the work of our predecessors, we collected a fairly large initial sample (more than 2000 documents taken from open sources) and segmented the documents into a larger number of homogeneous parts: text, a table with defined borders, a table without borders, a logo, a signature, a rectangular print, a round print, and a drawing.

In our article, we conduct a comparative analysis of object detection methods in the problem of document segmentation based on neural network modeling. The choice of using neural network modeling is due to the stability of these algorithms to noise, the speed of the computational process, and the relative simplicity of implementations. We chose the RetinaNet architecture, as this type of neural network architecture currently shows better results in comparison with its analogues in terms of speed and accuracy of recognition, and also allows us to solve the problem of class imbalance. The quality assessment of the constructed detector models will be carried out using the Mean average Precision metric.

2. Literature overview

The first scientific research on the segmentation of documents into blocks began in the late 70s of the twentieth century. By the beginning of the 21st century, research on the segmentation problem is receiving a new development [Cattonu et al., 1998; Agne S., Rogger M., Rohrscheneider J., 1999; Wood S., Marks J., Pearlman J., 1980; Randriamasy S., Vincent L., 1994]. George Nagy, a member of the scientific community of the Institute of Electrical and Electronics Engineers (IEEE), describes the evolution of methods for processing text documents over the last two decades of the last century [Nagy G., 2000].

In the mid-90s, Frank Hones and Jorgen Lichter, scientists at the German Center for the Study of Artificial Intelligence, presented an algorithm [Hones F., Lichter J., 1994] to separate text elements from non-text elements in a document and to group text elements as characters, words, strings and blocks. The algorithm is completely independent of the text orientation and can handle text with different fonts. It is based on a method that generates text strings by unfolding the anchors of the document image strings. An object anchor in a document is a feature that indicates where the floating object is located (that is, placed on a layer on top of the text) in relation to the text in the document. The method itself works well for documents that strictly fit the requirements on which the algorithm is trained. However, for non-standard documents, the algorithm works with less accuracy. The problems that arise include the inability to recognize symbols, in the following cases: merges non-text and text pixel groups; the presence of noise in the background (a lot of small non-text objects, usually created by a colored background, are collected as characters, provided that they are located inside the area of the string).

To date, the segmentation methodology is usually based on either algorithms based on the use of a ready-made library of OpenCV computer vision algorithms from Intel [Brahmbhatt S., 2013; Laganiere R., 2014; Taheri S. et al., 2018; Bradski G., Kaehlr A., 2008], or independently written algorithms using classical machine learning methods – for example, neural network modeling or the support vector method [Bertelli L. et al., 2011; Wei H. et al., 2013].

In 2017, Dafang Hee, Scott Cohen, and Brian Price, together with their research team, presented an algorithm [He D. et al., 2017] that solves the problem of segmenting a page into areas with pictures and areas with text, which determines the location of each table in the document. The algorithm is fed scanned documents, which are further processed by training a multitasking convolutional neural network to predict the class label for each

pixel and predict the boundaries of the instance (table or figure) for document elements. This work is one of the first solutions to solve the problem of segmenting figures and tables and determining the location of tables in a document.

Also, convolutional neural networks were used by the authors of scientific studies. [Can Y., Kabaday M., 2020; Feng Z. et al., 2017; Perner P., Imiya A., 2005; Jain A., Zhong Y., 1996] The disadvantages of the method used include difficulties with recognizing graphic images and tables without borders (or with their partial absence).

In 2005 A. Kryzhak, D. Dong, C. W. Suen, and D. Ponson proposed segmenting the document using an algorithm based on the support vector machine method [Dong et al., 2005]. In their paper, they formulated a method where the image of a word can be represented as two sequences of feature vectors in two independent channels. Using the support vector method, we solve the problem of classifying these points into two channels: local peaks in the upper outer contour and local minima in the lower outer contour. For the work of the local classifier uses the feature vectors.

In 1998, A. Antonakopos and R. Richings in their article [Antonacopoulos A., Ritch-ings R., 1994] were among the first to propose a new method of document page segmentation at that time. The main essence of the algorithm was to analyze the white space of the background surrounding the print area on the page. The white background space is filled with" tiles", and the outline of each area is identified by these white" tiles " that surround it. This method solves the problem of segmenting the image of pages with a strong skew without correcting the skew. White "tiles" in the image can also be used in subsequent document analysis processes, such as the classification of image areas. The advantages of the method proposed by the authors include the ability to recognize lines written vertically (for example, in Japanese), the absence of complex computational processes.

In 2017, scientists from Algeria – Insaf Setitra and his team presented their approach to segmentation [Setitra I. et al., 2017]. This work describes a method based on tracking work to detect lines (borders) in images of hand-written documents. Each cluster of connected pixels, which can be a word or part of a word, is viewed as an object moving in a left - to-right direction across the document. The trajectory is determined based on the regular movement of the cluster by searching for the optimal match among other clusters with the minimum angular deviation relative to its current trajectory. The approach is resistant to text distortions, since the implementation of the method preserves the history of the location of related components along their trajectories. But the algorithm does not work in cases where common connected components intersect.

To sum up, the studies we have studied mainly consider the classical approach to digitization using neural networks and the OpenCV library. Many approaches to improve the quality of recognition pre-apply segmentation of the page into objects. Usually, in a segmentation task, the text, images, and tables classes are allocated for subsequent recognition of text and tables. It can also be concluded that the main problems in recognition are the complex background of documents and intersecting objects.

3. Methodology

3.1 Description of RetinaNet architecture

When choosing the optimal neural network architecture for solving the problem of detecting objects on a document sheet, various neural network architectures were considered. The most attractive architecture currently is the RetinaNet architecture. The RetinaNet architecture achieves higher accuracy results in less or comparable time compared to neural networks of other architectures. Moreover, RetinaNet 50 can reach the AP level of 32.5 in 73 milliseconds, while other architectures, such as SDD513, R-FCN, DSSD321, require more time to reach a lower AP level (Figure 1) [Lin T. Y. et al., 2017].

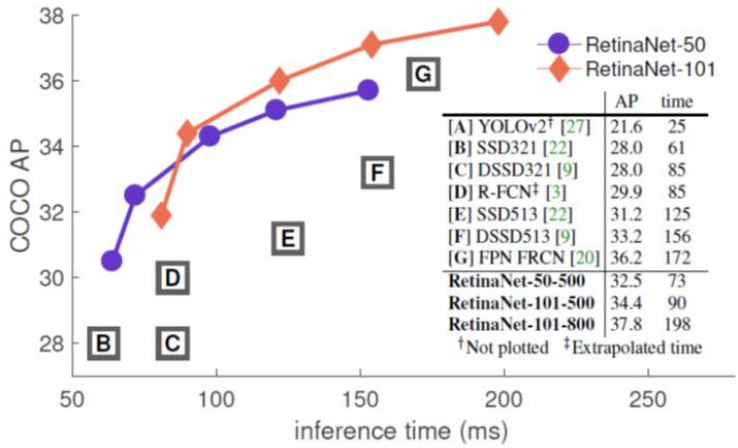

Figure 1. Comparison of RetinaNet architecture with other neural network architectures for object detection [Lin T. Y. et al., 2017]

Also, this architecture, unlike others, contains a larger number of blocks, and also solves the problem of class imbalance, which is very important for detecting different objects on a document sheet. This effect is obtained by using the loss function focal loss:

$$FL(p_t) = -(1-p_t)^{\gamma}\log(p_t) \quad (1)$$

where p_t is the probability of belonging to the class, γ is the hyperparameter of the model

Thus, the RetinaNet model is optimal for solving the problem of detecting objects on a sheet of paper. The architecture of this convolutional neural network consists of 4 parts that solve various problems [Can Y., Kabaday, 2020].

1) Backbone. This is the main part of the model, whose task is to extract features from the image. This part includes a convolutional neural network that transforms the image into data for further processing. This part of the model can be represented by various neural networks, such as ResNet, EfficientNet, VGG, etc.

2) Feature Pyramid Net (FPN) is the second part of the model. It is a convolutional neural network that has a pyramid architecture. Its task is to combine the properties of feature maps. It consists of 3 parts, which include the ascending and descending paths, as well as lateral connections. The ascending network is represented by several layers, which are successively reduced in size. It is worth noting that the upper layers have a higher semantic value, however, lower resolution and vice versa. This network has a vulnerability in the form of loss of important information when reducing the size of layers. The descending network is also represented by a pyramid that recovers from the upper layers the lower layers, which have a larger size. The lateral connections are designed to prevent the attenuation of signals when passing through the layers. The downstream network is then processed by other parts of the RetinaNet model [Feng Z. et al., 2017].

3) Classification Subnet – the third part of the model, which, using the output of the FPN, solves the problem of object classification. This neural network predicts the probability of an object belonging to a certain class.

It is represented by 4 convolutional layers. 256 feature maps are generated in each layer. On the fifth layer, the network is represented by the number of feature maps equal to the number of anchor frames multiplied by the number of classes. The last sixth layer returns a vector whose length is equal to the number of classes, in which the ranked probabilities of the object belonging to the class are located.

4) Regression Subnet – the fourth part of the model, the task of which is to extract information about the coordinates of objects in the image from the FPN. This network solves the regression problem. Its structure is similar to that of a classification network, except that layer 5 consists of the number of maps equal to 4 times the number of anchor frames, and the last layer returns the offset of the target frame relative to the anchor frame (Figure 2) [Perner P., Imiya A., 2005]. It is identical to the classification subnet except that it terminates in 4A linear outputs per spatial location. Subnet 3 and subnet 4 are fully connected, as shown in Figure 2.

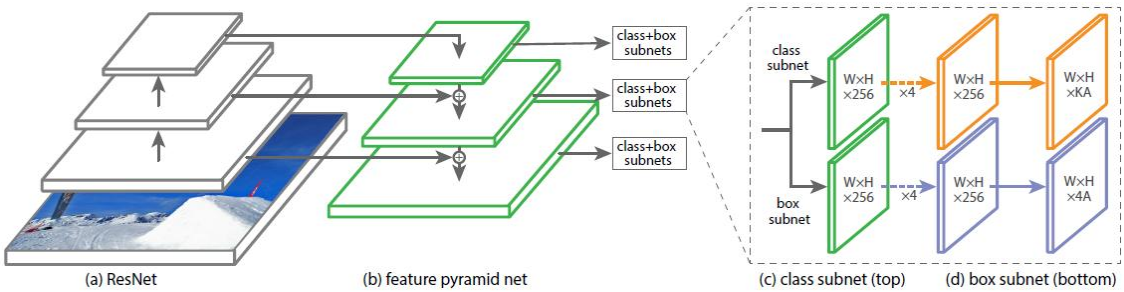

Figure 2. Architecture of the RetinaNet neural network model [Zeng N., 2018]

3.2 Description of metrics for quality assessment

To determine the quality of object detection, the mean aver-age precision (map) indicator was used. It is one of the most widely used indicators in the problem of image detection. map varies in the range from 0 to 1, and the better the detection problem is solved, the indicator is closer to 1. This indicator gives the average accuracy (precision) for the completeness values (recall) [Dong J. et al., 2005]. To determine these parameters, the IoU (Intersection over union) indicator is calculated – the ratio of the intersection areas of the predicted and original area in which the object is located to their union:

$$IoU = \frac{intersection\ square}{union\ square} \quad (2)$$

Intersection square is the common area covered by both bounding boxes or the area where one box overlaps the other box and union square is the total area covered by both the bounding boxes.

If this indicator is below a certain threshold, then the object is assigned the value False, if it is higher, then True. Next, the cumulative values of presicion and recall are calculated:

$$\text{precision} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{recall} = \frac{TP}{TP+FN} \quad (4)$$

where TP – True Positive, FP – False Positive, FN – False Negative.

The obtained data is divided along the recall axis into intervals of increasing precision in a way that within each interval there is a monotonic increase in the accuracy indicator. The maximum precision value is defined for each interval (the maximum accuracy value refers to the interpolated accuracy, that is, the highest accuracy value for a certain level of recall):

$$p' = max_{\tilde{r} \geq r} p(\tilde{r}) \quad (5)$$

where $\tilde{r}$ – is the value of the completeness function, r – s the last point of change of the increasing trend of completeness.

Next, 11 points are taken along the recall axis with the point values 0, 0.1, 0.2 ... 1.0. At these points, the indicator p' is calculated, and AP is located:

$$AP = \frac{1}{11} * (p'(0) + p'(0.1) + \ldots + p'(1.0)) \quad (6)$$

The AP metric calculates the average precision for recall in the range from 0 to 1. Averaging the AP indicator for different objects, we get the mAP indicator [Antonacopoulos A., Ritchings R., 1994]. We calculate the AP for each class and average them.

3.3 Description of the subtleties of the markup

To solve the problem of segmentation, a dataset was collected and marked up, including more than 2 thousand documents (including those with a non-standard noisy background), containing more than 14 thousand objects. In order to detect objects on the document sheet, several object classes were allocated for recognition:

1. Text. This class includes text printed by computer, located outside the tables in the form of paragraphs, as well as captions, headings, columns, etc..

2. Table. This class includes data in tabular form with clearly defined boundaries. Since numerical data is usually stored in tabular form, which is necessary for further analysis and calculations, table recognition is one of the key tasks in digitizing documentation. This is especially true for financial statements, for which tables are the main source of information.

3. Table without borders. Modern financial statements are presented in a wide variety in terms of design and appearance of documents. Often, the presented tables have no external borders and some internal borders. Moreover, it is becoming increasingly popular to separate different columns and columns in tables not by using borders, but by filling in a given color. However, such tables are extremely poorly recognized by models designed to recognize classical tables with boundaries. Also, they are often divided into several tables when recognizing a table without borders. For this reason, it was decided to separate tables without borders into a separate class during recognition.

4. Images. Many modern solutions designed for digitizing documents refer graphic images to one class of "images". However, documents may contain both simple illustrations and various diagrams, diagrams, company logos, seals, and captions, which can also be classified as images. At the same time, different graphic images within the document may be of different value for subsequent analysis of the objects found on the sheet. Therefore, the authors decided to distinguish 4 classes of graphic images: pictures, logos, seals and signatures. The images class includes graphic images, diagrams, and diagrams. This class is less valuable for subsequent analysis than other classes.

5. Logos. Unlike images, a logo is an image used by the company for identification. Often, on financial reporting documents, logo types can be displayed on each page at the top or bottom. A logo is usually much smaller in size than a picture, graph, or diagram. This image can be used to identify the company whose reports are being viewed. It should also be noted that often in large companies, external auditors help to compile reports, and some pages contain logos of the company conducting the audit. Automatic recognition of the logo may allow further solving the problem of automatic identification of the company about which

the reporting is made, as well as the company-auditor. The difference between logos and images can be distinguished in their reproducibility on financial reporting documents and relative stability in their placement on a sheet of paper. Logos are located either at the bottom of the page or at the top, are not wrapped in text and are not signed, for example, as images.

6. Seals. This class includes both seals and stamps of various shapes (round, square, rectangular, and others). One of the features of financial statements is that they are invalid without printing and writing. At the same time, many other legal documents also do not have legal force without a seal. Automatic detection of printing or its absence can allow businesses to avoid a large number of errors related to legal documents.

7. Signatures. This class includes signatures, transcripts, and other objects that are written with human writing materials, rather than printed on a computer. As in the case of seals, many legal documents are not legally valid without a signature, so when recognizing documents, it is also important to detect signatures.

3.4 Description of the learning process and functions

To recognize an object on a sheet, we decided to use the RetinaNet neural network model. The default RetinaNet parameters were used, as described in Lin T. Y. et al. (Lin T. Y. et al., 2017). The standard RetinaNet model determines the parameters number of hidden layers, number of hidden units, learning rate, activation functions, optimizer, number of epochs.

The model parameters are as follows:

- number of hidden layers - the standard version has 3-5 layers;
- learning rate – the initial learning rate is 0.01;
- activation functions - sigmoid activation function;
- optimizer – Adam optimizer;
- number of epochs - epochs 50.

At the first stage of the training process, a single neural network was used to recognize all seven classes (a model with 1 neural network). However, there was a problem with this method of recognition: the model had difficulty identifying classes with a text in it, since text on the sheets with a small number of tables and a large amount of text was often perceived as a background. To solve this problem, used a model with two neural networks: the first neural network recognized six classes of an object without the text, and then the text was defined separately.

In addition to this solution, to improve the quality of object detection, we used a model with 4 neural networks: 3 separate neural networks were used to detect different, most similar classes of objects: the first model recognized tables and tables without borders, the second model-seals and signatures, the third model – pictures and logos, and the fourth network, as in the model with 2 neural networks, identified text.

As part of the study, a separate neural network was used for each detection group. The objects were divided into 4 detection groups: images and logos, tables and tables without borders, seals and signatures.

To train neural networks, a dataset consisting of 14,246 objects containing pages of documentation, mainly financial statements, was used. Data augmentation was performed to improve the recognition quality. The essence of augmentation is to increase the dataset at the expense of existing objects by slightly changing objects and adding modified objects to the dataset [Setitra I. et al., 2017]. Image augmentation methods include:

1. Rotate the image
2. Image Shift
3. Change the brightness, contrast, and clarity of the image.

However, not all augmentation methods are suitable for the task of recognizing financial statements and documentation. Usually, for working with digitized documents, they are pre-processed in order to align the document scan and change the brightness, contrast of the image, etc. Therefore, rotations for tables and other classes cannot be used. It is also worth noting that changing the clarity of the image is a factor that can be used to expand the data set, but this method will make it difficult to further process the recognized objects.

Therefore, the only adequate augmentation method is to shift the frames of detected objects. We applied this method, increasing the existing dataset by 2 times.

4. Experiments

Three neural network models showed different learning outcomes. For a more accurate assessment of the results of the model, a deferred test sample was used, on which the model was not trained. The deferred sample size was 20% of the total data set.

The model with a single neural network showed the worst results in training. The model did the best job of identifying tables: The map for this class was 0.338. However, all other classes have a map value below 0.3. This problem may have occurred due to the problem of merging some objects with text, which was often defined by the model as the background.

The model with two neural networks showed approximately the same results for logotypes, tables, tables without borders and text. The map for these classes varies around 0.45. However, it is worth noting that this model did the best job of identifying a table without borders. Slightly worse, the model recognized prints and images. The map values for these classes are 0.276 and 0.289. The model identified the signatures the worst.

The model with four neural networks showed a higher variation in the detection of different classes. The best results of detecting a model with four neural networks were obtained when applied to the "tables" class. The map for this indicator was 0.625. This model also recognized logos types with a map score of 0.600 well. The model did the worst with tables without borders: the map for this class was 0.396 (Table 1). However, in general, this result should be considered positive, especially given the complexity of recognizing objects of this class.

Table 1. results of object detection using various neural network models

	AP		
Object class	**Model with one neural network**	**Model with two neural networks**	**Model with four neural networks**
Pictures	0.241	0.276	0.437
Logos	0.289	0.460	0.600
Signatures	0.134	0.213	0.401
Seals	0.257	0.289	0.423
Tables	0.338	0.456	0.625
Tables without borders	0.297	0.466	0.396
Text	0.196	0.442	0.456

Note: authors ' results.

In addition, introduction of additional neural networks in the model has significantly improved the quality of detection of various classes. Especially significant was the allocation of the text class for recognition by a separate model, which allowed to reduce noise and improve the quality of recognition not only for text, but also for other classes. Also, the positive effect of allocating the "images" and "logos" classes to a separate neural network allowed us to improve the recognition of these classes. It is worth noting that the use of a model with 4 neural networks has significantly improved the detection of tables that are especially important in financial statements.

The proposed method can be compared with other modern studies. For example, a paper published by Hu and co-authors (Hu H. et al., 2018) on the relationships between objects in the detection process suggests a different approach to object detection. The authors propose to evaluate not the objects, but the relations between these objects, which can be an addition as the fifth neural network in our study, which would participate in the analysis of the relations between the objects of financial statements.

In another review article, Zou Z. et al. (Zou Z. et al., 2019) they analyze all the research on object detection over 20 years of research (from 1998 to 2018). The authors consider the invented methods, highlighting their shortcomings. Thus, one of the disadvantages of using RetinaNet is the imbalance between the foreground-background class when training dense detectors, which is the central cause. This paper does not distinguish studies based on the use of several neutron networks for the detection of different types of objects.

In another paper (Cai Z. et al., 2016) devoted to the analysis of the use of neural networks for object detection, a single deep neural network is proposed, denoted by the multiscale CNN (MS-CNN). The MS-CNN consists of a suggestion subnet and a discovery subnet. In the proposed subnet, the detection is performed on several output layers, so that the receptive fields coincide with objects of different scales. This approach is similar to the authors ' approach but differs in that the authors of the article use independent neural networks for each type of object, rather than a single deep one.

Thus, most approaches to the analysis of object detection are based on the use of a single neural network with deep learning. Our approach is distinguished by the use of several neural networks for detection, which allows us to independently evaluate each of the types of objects.

We compared various solutions for detecting objects of a similar class according to the AP criterion. The results of the comparison are presented in Table 2.

Table 2. comparison of the effectiveness of methods by the AP indicator

Method	AP			
	Text	Images and logo	Table	Seals
Fast (Girshick, R., 2015)	0, 042	0,463	0,582	-
Faster (Ren S. et al., 2016)	0, 116	0,503	0,611	-
SSD (Liu W., 2015)	0,04	0,347	0,435	-
Hao, et al. (Hao, Leipeng, et al., 2016)	/	/	0,701	-
Sokerin et al. (Sokerin, Volkova, Kushnarev, 2021)	0,456	0,6	0,625	0,423

Note: Yi, X., Gao, L., Liao, Y., Zhang, X., Liu, R., & Jiang, Z. (2017) and authors ' results.

From Table 2, it is clear that according to the AP criterion, the authors ' results are superior to similar studies of object detection in financial documents.

5. Conclusion

To develop an optimal adaptive approach to segmentation, we completed the following tasks: we analyzed the work of our predecessors, collected an initial sample of scanned documents from open sources in English and Russian (divided into training and test documents in the proportion of 80% and 20%, respectively), performed an augmentation procedure by shifting the boundaries of detected objects, then built three neural network models using the RetinaNet architecture with different classes (one neural network for recognizing all seven classes of objects and models, including two and four neural networks, respectively), the evaluation of the constructed models and their comparative analysis was carried out. As an indicator that reflects the quality of segmentation of models, we selected the mean average precision indicator.

As a result of comparative analysis, we came up to the conclusion that the approach, based on the inclusion of four neural networks in the model to focus each of them on segmentation of specific types of objects is advantageous, showed the best result on the test sample in the context of most detection classes. In the future we plan to apply a similar approach to table segmentation, as classes will be the table areas that differ in functions: title, cell with a name, cell with a data, empty cell.

References

[1]Agne S., Rogger M., Rohrschneider J. Benchmarking of document page segmentation //Document Recognition and Retrieval VII. – International Society for Optics and Photonics, 1999. – T. 3967. – C. 165-171.

[2]Ale L., Zhang N., Li L. Road damage detection using RetinaNet //2018 IEEE International Conference on Big Data (Big Data). – IEEE, 2018. – C. 5197-5200.

[3]Antonacopoulos A., Ritchings R. T. Flexible page segmentation using the background //Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5). – IEEE, 1994. – T. 2. – C. 339-344.

[4]Bertelli L. et al. Kernelized structural SVM learning for supervised object segmentation //CVPR 2011. – IEEE, 2011. – C. 2153-2160.

[5]Bradski G., Kaehler A. Learning OpenCV: Computer vision with the OpenCV library. – " O'Reilly Media, Inc.", 2008.

[6]Brahmbhatt S. Practical OpenCV. – Apress, 2013.

[7]Cai Z. et al. A unified multi-scale deep convolutional neural network for fast object detection //European conference on computer vision. – Springer, Cham, 2016. – C. 354-370.

[8]Can Y. S., Kabadayı M. E. CNN-Based Page Segmentation and Object Classification for Counting Population in Ottoman Archival Documentation //Journal of Imaging. – 2020. – T. 6. – №. 5. – C. 32.

[9]Cattoni R. et al. Geometric layout analysis techniques for document image understanding: a review //ITC-irst Technical Report. – 1998. – T. 9703. – №. 09.

[10]Dong J. et al. Low-level cursive word representation based on geometric decomposition //International Workshop on Machine Learning and Data Mining in Pattern Recognition. – Springer, Berlin, Heidelberg, 2005. – C. 590-599.

[11]Feng Z. et al. Deep retinal image segmentation: a FCN-based architecture with short and long skip connections for retinal image segmentation. //International Conference on Neural Information Processing. – Springer, Cham, 2017. – C. 713-722.

[12]He D. et al. Multi-scale multi-task fcn for semantic page segmentation and table detection //2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). – IEEE, 2017. – T. 1. – C. 254-261.

[13]Henderson P., Ferrari V. End-to-end training of object class detectors for mean average precision //Asian Conference on Computer Vision. – Springer, Cham, 2016. – C. 198-213.

[14]Hönes F., Lichter J. Layout extraction of mixed mode documents //Machine vision and applications. – 1994. – T. 7. – №. 4. – C. 237-246.

[15]Hu H. et al. Relation networks for object detection //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2018. – С. 3588-3597.
[16]International Conference on Neural Information Processing. – Springer, Cham, 2017. – С. 713-722.
[17]Jain A. K., Zhong Y. Page segmentation using texture analysis //Pattern recognition. – 1996. – Т. 29. – №. 5. – С. 743-770.
[18]Kisantal M. et al. Augmentation for small object detection //arXiv preprint arXiv:1902.07296. – 2019.
[19]Laganière R. OpenCV Computer Vision Application Programming Cookbook Second Edition. – Packt Publishing Ltd, 2014.
[20]Li K. et al. A maximal figure-of-merit learning approach to maximizing mean average precision with deep neural network based classifiers //2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2014. – С. 4503-4507.
[21]Lin T. Y. et al. Focal loss for dense object detection //Proceedings of the IEEE international conference on computer vision. – 2017. – С. 2980-2988.
[22]Nagy G. Twenty years of document image analysis in PAMI //IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2000. – Т. 22. – №. 1. – С. 38-62.
[23]Perner P., Imiya A. (ed.). Machine Learning and Data Mining in Pattern Recognition: 4th International Conference, MLDM 2005, Leipzig, Germany, July 9-11, 2005, Proceedings. – Springer Science & Business Media, 2005. – Т. 3587.
[24]Randriamasy S., Vincent L. A region-based system for the automatic evaluation of page segmentation algorithms //Proceedings of the International Association for Pattern Recognition Workshop on Document Analysis Systems DAS94. – 1994. – С. 29-41.
[25]Review: RetinaNet — Focal Loss (Object Detection). Towards Data Science. URL: https://towardsdatascience.com/review-retinanet-focal-loss-object-detection-38fba6afabe4 (режим доступа: 17.01.2021)
[26]Setitra I. et al. Text Line Segmentation in Handwritten Documents Based on Connected Components Trajectory Generation //International Conference on Pattern Recognition Applications and Methods. – Springer, Cham, 2017. – С. 222-234.
[27]Szegedy C., Toshev A., Erhan D. Deep neural networks for object detection //Advances in neural information processing systems. – 2013. – С. 2553-2561.
[28]Taheri S. et al. OpenCV. js: Computer Vision processing for the open Web platform //Proceedings of the 9th ACM Multimedia Systems Conference. – 2018. – С. 478-483.
[29]Wang Y. et al. Automatic ship detection based on RetinaNet using multi-resolution Gaofen-3 imagery //Remote Sensing. – 2019. – Т. 11. – №. 5. – С. 531.
[30]Wood S. L., Marks J. P., Pearlman J. A segmentation algorithm for ocr application to low resolution images //Conference Record of the Fourteenth Asilomar Conference on Circuits, Systems and Computers. – 1980. – С. 411-415.
[31]Wei H. et al. Evaluation of SVM, MLP and GMM classifiers for layout analysis of historical documents //2013 12th International Conference on Document Analysis and Recognition. – IEEE, 2013. – С. 1220-1224.
[32]Yi X. et al. CNN based page object detection in document images //2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). – IEEE, 2017. – Т. 1. – С. 230-235.
[33]Zeng N. RetinaNet Explained and Demystified [Электронный ресурс]. 2018 URL: blog.zenggyu.com/en/post/2018-12-05/retinanet-explained-and-demystified
[34]Zhang H. et al. Cascade retinanet: Maintaining consistency for single-stage object detection //arXiv preprint arXiv:1907.06881. – 2019.
[35]Zou Z. et al. Object detection in 20 years: A survey //arXiv preprint arXiv:1905.05055. – 2019.