 DOI: 10.3217/978-3-85125-603-1-12

Page Segmentation and Region Classification Based on Region Bounding Boxes

Thomas Lang[1], Markus Diem[1], Florian Kleber[1] and Robert Sablatnig[1]

***Abstract*— We present an approach for the segmentation and classification of digital document images of newspapers or similar types of documents which are described by a set of partially correct text and image bounding boxes. These flawed region descriptions are used to locate the bounding boxes of the true page component boundaries in the documents. The resulting regions are classified as text, images, charts or tables. In addition to the individual evaluation of the segmentation and classification steps, the combined physical layout analysis system is evaluated and compared to the page segmentation results of an open-source document analysis software.**

I. INTRODUCTION

Page segmentation is a crucial step in document analysis, as it is a requirement for other tasks like OCR to have isolated document regions containing only one type of content and for determining the structure of a document. Layout analysis is an active research field and methods continue to improve, as is evident from the biannual competitions organized by PRImA [2]. The general aim is to locate page components and to classify them according to their content type and meaning. We present a method for the more constrained task of page segmentation and component classification for digital document images with existing partially correct annotations of text and image regions. Such region descriptions may be the result of layout analysis of PDF documents, which was studied by Chao and Fan [1]. Even though the text and image components can be accurately extracted from PDF documents, the embedded raster images are often broken up into smaller parts when the file is created by publishing software. As a result, the appearance of the images in the document is unchanged, but when the attempt is made to extract image boundaries from the PDF file, instead of the actual image borders, only smaller image segments are obtained. In addition, if clipping masks are not properly extracted along with the images, only rectangular bounding boxes are obtained, even if the image portions displayed in the document have more complex shapes.

In this paper, we use a dataset containing newspaper pages in the form of raster images and corresponding sets of rectangular image and text region descriptions, which are known to have been extracted from PDF files by the provider of the dataset. Unfortunately, this dataset is not publicly available. The image region boundaries show all the described problems resulting from PDF extraction. An example of such problematic image region descriptions can be seen in Fig. 1. Additionally, sets of manually annotated ground-truth chart and table regions are available to us (see section III-A). Since these annotations also contain only rectangular region boundaries (bounding boxes), the proposed segmentation method produces rectangular regions as well, even if the underlying document region has a more complex shape. However, the method could easily be adapted to detect the exact boundaries of document components, as will be explained in section II-A. The page segmentation and classification method has previously been published, along with the evaluation results of the classification step [5]. For the sake of completeness, we first provide a short summary of both steps and of the classification results. We then present measures for the evaluation of the segmentation step. Results are shown for the segmentation step, for the full system and for a comparison to a different page segmentation system.

Fig. 1: The image region boundaries included in the dataset describe only segments of the actual image, overlap each other and falsely include surrounding text segments. The goal is to find the bounding box containing only the image.

II. METHODOLOGY

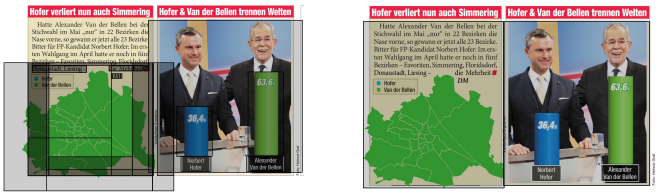
We first segment the partially incorrect image regions to obtain region rectangles that fit visible document components. These regions are then classified as being of either the chart or image class, and the text regions are classified as text or tables. It is assumed that all document images are not skewed or warped in any way.

A. Image Region Segmentation

The segmentation of image regions begins by removing the text from the gray-scale document image using a simple rule-based approach. For each of the text region rectangles, the

*This work was supported by APA-IT

[1]Thomas Lang, Markus Diem, Florian Kleber and Robert Sablatnig are with Computer Vision Lab, Institute of Visual Computing & Human-Centered Technology, Faculty of Informatics, TU Wien, Vienna, Austria {tlang,diem,kleber,sab}@cvl.tuwien.ac.at



(a) Before segmentation (b) After segmentation

Fig. 2: Example of rectangular image regions before and after segmentation.

respective document image segment binarized and the more frequent value is treated as the background. Pixels deviating from this background value by more than a specified threshold are considered foreground pixels. Connected component analysis is performed on the binary foreground mask. Two criteria are used to eliminate non-text pixels from this mask. First, connected components (CCs) are removed if their pixel area deviates from the area of all other components by more than a certain relative threshold. Second, CCs are removed if the color of most of their pixels differs from the dominant text color, which is found using k-means clustering on the a* and b* channels of the image in L*a*b* color space. As a last step, to account for smooth text borders, the image region mask is dilated. Afterwards, the mask is used to replace all text pixels in the gray-scale image with the background value.

What remains after the text removal procedure is a document image containing vector graphics, separator lines and other layout elements, which can be used in combination with the available image region rectangles to locate the boundaries of the image components. First, adjacent and overlapping (clustered) image region rectangles like the ones shown in Fig. 1 are grouped together. As in the text removal step, a global threshold is computed to find the background value of the page. Afterwards, a foreground mask is created containing all pixels which are darker than the background value (for pages with a bright background). For each group of image region rectangles, a "cluster mask" is created, containing all its pixels. The logical AND operation between a cluster mask and the document foreground mask defines a cluster foreground mask. As explained previously, we only use the bounding boxes of these components. An example of the result of the segmentation can be seen in Fig. 2.

If the goal was to obtain exact regions, their boundaries are defined by the cluster foreground masks. The text regions would have to be reduced to the borders of the text masks used to remove the text from the document image.

B. Region Classification

The available text and image regions (not segmented) are used along with manually annotated ground-truth rectangles to generate feature descriptors for each region class (text, image, chart, table). First, each region is sampled down to the highest image pyramid level (smallest size) which is still larger or equal to 64×64. Afterwards, one or more HOG features [4] are computed by sliding a 64×64 window over the region rectangle with a step size of 32 in both directions. Therefore, in each region dimension with the size dim, $\frac{dim-64}{32}$ shifts are performed. Each feature is associated with the class of the region it was computed from. Finally, the collected features from the training set are used to train a random forest classifier.

For the classification, the features are computed in the same way on the test set regions, after which they are classified by the random forest. However, since we are interested in the classes of the regions and not the individual feature samples, we add the votes of all bagged trees of the random forest for each window position inside the same region in order to find a decision for the whole region.

III. EVALUATION

The segmentation and the classification steps are first evaluated separately. We then present results for the complete system. Additionally, these results are compared to a different page segmentation method, which is part of the Tesseract OCR engine[1] developed by Google [8]. All evaluation steps are performed on a dataset of raster images of contemporary newspaper pages containing text region bounding boxes along with partially correct image region rectangles.

A. Classification

For the evaluation of the classification, the chart and table regions in 6211 newspaper pages have been manually annotated. As a result, 891 pages contain at least one chart or table. We use 70 % of this dataset for training and the remaining pages as test set (624 training pages, 267 test pages). The training set is balanced by reducing the number of feature descriptors of each class to the minimum class size. The test sets are also balanced, by classifying the same number of regions for each class. The random forest predictor included in OpenCV 3.3[2] is used with a tree depth limit of 25 and a maximum number of trees of 150. Text regions are classified as text or tables and image regions are classified as images or charts. Therefore, we train and evaluate two random forest classifiers: one for the distinction between text and tables and one for the distinction between images and charts. For the computation of text and image feature descriptors, the available (partly incorrect) region rectangles are used. We rely on the assumption that the image rectangles contain image regions for the most part, even if some of them overlap and don't always fit the actual image parts in the document. The confusion matrices in Tables I and II show the classification results. The rows represent the actual classes; the columns are the predictions. For the classification of text regions as text or tables, 99 regions are wrongly classified, which is equal to an overall error rate of 0.05 (2168 regions in total). For the image/chart classification, the overall error rate is 0.1, with 73 wrongly classified regions out of 702 in total.

[1] https://github.com/tesseract-ocr/tesseract/
[2] https://opencv.org/

	Text	Table	Re.
Text	1048	36	0.97
Table	63	1021	0.94
Pr.	0.94	0.97	

TABLE I: Text/Table confusion matrix

	Image	Chart	Re.
Image	328	23	0.93
Chart	50	301	0.96
Pr.	0.87	0.93	

TABLE II: Image/Chart confusion matrix

B. Segmentation

For the segmentation evaluation, all image regions contained in 70 of the 267 pages of the test set have been manually annotated. Since the image rectangles resulting from the segmentation are classified either as images or charts, their correctness is measured by comparing the segmented rectangles to all ground-truth image and chart regions. The ground truth set contains a total of 389 image and chart region rectangles, compared to 461 which are produced by the segmentation. Two kinds of evaluation are performed.

First, the amount of area overlap between computed and ground-truth region rectangles is computed. The ratio of total intersection area to total area of computed or ground-truth regions is the precision or recall respectively. The overall precision of the segmented regions is 0.94 and the recall is 0.77. If the same measurements are taken without the chart regions in the ground-truth set, the precision decreases to 0.9, while the recall increases to 0.88. This already shows that large parts of the ground-truth charts are not or only partly matched by segmented image regions.

The second evaluation aims to find not only the region overlap, but the number of region rectangles which are segmented correctly. For each page, there are two sets of regions S_C and S_{GT} containing the segmentation results and the ground truth. We define two regions r_1 and r_2 as *fitting* if their Jaccard index

$$J(r_1, r_2) = \frac{\text{area}(r_1 \cap r_2)}{\text{area}(r_1 \cup r_2)}, \tag{1}$$

is greater than $1-T$, where T is a tolerance value. Both region sets S_C and S_{GT} contain subsets of regions for which a fitting region exists in the other set:

$$F_C = \{r_C \in S_C \mid \exists r_{GT} \in S_{GT} : \text{fits}(r_C, r_{GT})\}, \tag{2}$$

$$F_{GT} = \{r_{GT} \in S_{GT} \mid \exists r_C \in S_C : \text{fits}(r_C, r_{GT})\}. \tag{3}$$

Furthermore, a region r of one set $S_C \setminus F_C$ or $S_{GT} \setminus F_{GT}$ is *covered* by regions $r_1^*, r_2^*, ..., r_n^*$ of the other set S_{GT} or S_C if it fits their union area, meaning that $J(r, r_1^* \cup r_2^* \cup ... \cup r_n^*) > 1-T$. This concept of covered regions is similar to merges and splits in the evaluation measures described by Clausner et al. [3]. The sets of covered computed and of covered ground truth regions are called C_C and C_{GT}. A region is *matched* if it fits one or more regions in the other set: $M_C = F_C \cup C_C$, $M_{GT} = F_{GT} \cup C_{GT}$. The ratio of matched regions in the computed or ground-truth set can be interpreted as the region matching precision Pr_M and recall Re_M respectively. Table III shows the numbers of regions in each set summed over all documents in the dataset. The F_1 score is defined as the harmonic mean of the recall and precision values.

T	$\sum\lvert F_{GT}\rvert$	$\sum\lvert F_C\rvert$	$\sum\lvert C_{GT}\rvert$	$\sum\lvert C_C\rvert$	$\textbf{Re}_M$	$\textbf{Pr}_M$	F_1
0.05	156	156	1	0	0.4	0.34	0.37
0.1	169	169	1	0	0.44	0.37	0.4
0.15	172	172	1	0	0.44	0.37	0.41
0.2	181	181	1	0	0.47	0.39	0.43
0.25	187	186	1	2	0.48	0.41	0.44
0.3	188	187	3	2	0.49	0.41	0.45

TABLE III: Segmentation region matches (images and charts)

T	$\sum\lvert F_{GT}\rvert$	$\sum\lvert F_C\rvert$	$\sum\lvert C_{GT}\rvert$	$\sum\lvert C_C\rvert$	$\textbf{Re}_M$	$\textbf{Pr}_M$	F_1
0.05	154	154	1	0	0.58	0.33	0.42
0.1	166	166	1	0	0.62	0.36	0.46
0.15	169	169	1	0	0.63	0.37	0.46
0.2	177	177	1	0	0.66	0.38	0.49
0.25	181	181	1	1	0.68	0.39	0.5
0.3	182	182	2	1	0.68	0.4	0.5

TABLE IV: Region matches of image regions only

It can be seen that depending on the tolerance value T, the recall Re_M varies between 0.4 and 0.49, and the Pr_M lies in the range 0.34 to 0.41. Compared to the area-based evaluation results with a recall of 0.77 and a precision of 0.94, the values are significantly lower. This shows that some of the region bounding boxes in each of the sets (computed or ground truth) only partly intersect the region rectangles in the other set, but do not match them exactly. The amount of regions covered by a set of other regions (C_{GT} and C_C) is rather low (1.6 % and 1.1 % for $T = 0.3$), but it does occur.

We again take a second measurement without the chart regions in the ground truth set, which reduces its size from 389 to 269. The results are shown in Table IV. With the ground truth set containing only image regions, the recall increases significantly (by 0.19 on average). However, the number of matched regions is almost unchanged. The increased recall value is mainly explained by the smaller size of the ground truth set (30.8 % decrease). This means that most ground-truth chart bounding boxes are not matched by segmented image regions. The almost unchanged precision values Pr_M show that many of the segmented image regions lie outside the ground-truth image regions. This can be explained by them being parts of chart regions, which often contain, but do not fully consist of images.

C. Complete System

For the evaluation of the complete layout analysis system, both methods from the segmentation evaluation (measuring intersecting areas and counting exact matches) are reused, but for each class individually. The results represent the performance of the full segmentation and classification procedure. The results of the area-based evaluation in Table V again show that the method generally fails to produce correct chart regions. For the text regions, we reuse the available region bounding boxes. Therefore, the errors result only from text regions being falsely classified as tables. Table VI

	Text	Image	Chart	Table
Re	0.98	0.8	0.15	0.81
Pr	0.99	0.92	0.37	0.79
F_1	0.98	0.86	0.21	0.8

TABLE V: Results of the area-based evaluation of the complete system.

T		0.05	0.1	0.15	0.2	0.25	0.3
Text	Re_M	0.98	0.98	0.98	0.98	0.98	0.98
	Pr_M	0.99	0.99	0.99	0.99	0.99	0.99
	F_1	0.99	0.99	0.99	0.99	0.99	0.99
Image	Re_M	0.53	0.57	0.58	0.61	0.62	0.63
	Pr_M	0.33	0.36	0.36	0.38	0.4	0.4
	F_1	0.41	0.44	0.45	0.47	0.48	0.49
Chart	Re_M	0.01	0.02	0.02	0.03	0.03	0.03
	Pr_M	0.03	0.05	0.05	0.08	0.1	0.1
	F_1	0.01	0.03	0.03	0.04	0.05	0.05
Table	Re_M	0.63	0.63	0.64	0.65	0.67	0.67
	Pr_M	0.6	0.6	0.6	0.6	0.61	0.61
	F_1	0.61	0.61	0.62	0.62	0.64	0.64

TABLE VI: Results of the match-based evaluation of the complete system.

again shows that almost no chart regions are matched. Since the text regions are not segmented, but only classified, the results are about as high as in the area-based evaluation. The image results have decreased because of segmented region rectangles not matching the ground-truth ones. The F_1 scores of table regions lie between 0.61 and 0.64.

D. Comparison to Tesseract

We use a converter developed by PRImA Research[3] to store the page segmentation output of the Tesseract engine (version 3.04) in the PAGE format [7]. Since Tesseract produces non-rectangular region polygons, in order to compare them to our method, it is necessary to use just their bounding boxes. An example of region polygons produced by Tesseract can be seen in Fig. 3. Directly adjoining region polygons are merged to a single region. Table VII shows the area-based evaluation results of the Tesseract regions compared to our method. The "proposed method" columns show the same numbers as in section III-C. Since Tesseract does not detect chart regions, this class is left out. For the "Image*" row, all chart regions are treated as image regions. Tesseract produces a large amount of table regions, which causes its recall value to be very high (0.81), but the precision to be low (0.4), resulting in an F_1 score of only 0.54 compared to 0.8 with our method.

[3] http://www.primaresearch.org/tools/TesseractOCRToPAGE

Fig. 3: Example of image-class polygons produced by Tesseract.

Type	Precision		Recall		F_1	
	Tess.	Prop.	Tess.	Prop.	**Tess.**	**Prop.**
Text	0.92	**0.99**	0.86	**0.98**	0.89	**0.98**
Image	0.8	**0.92**	0.78	**0.8**	0.79	**0.86**
Image*	0.87	**0.94**	0.73	**0.77**	0.79	**0.85**
Table	0.4	**0.79**	0.81	0.81	0.54	**0.8**

TABLE VII: Comparison between Tesseract and the proposed method.

IV. CONCLUSION

We have seen that the proposed page segmentation and classification method based on region bounding boxes provides satisfactory results in some respects. The classification based on HOG and random forests achieved low error rates and the overall F_1 results are higher than those of Tesseract. For a more general use case, it would be interesting to extend the method to generate more exact (non-rectangular) region descriptions, which would in turn also require more exact (e.g. polygonal) ground-truth information for evaluation.

The evaluation results show that even with simple methods, data extracted from PDF files can be used to obtain a viable page layout description. However, it has also become clear that the method fails at detecting chart regions, because they consist of multiple image and text elements. Such complex objects would have to be addressed in a different way, like considering combinations of page elements as possible regions.

REFERENCES

[1] H. Chao and J. Fan, "Layout and content extraction for pdf documents," in *International Workshop on Document Analysis Systems*. Springer, 2004, pp. 213–224.

[2] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "ICDAR2017 competition on recognition of documents with complex layouts - RDCL2017," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, nov 2017.

[3] C. Clausner, S. Pletschacher, and A. Antonacopoulos, "Scenario driven in-depth performance evaluation of document layout analysis methods," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1404–1408.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[5] T. Lang, M. Diem, and F. Kleber, "Physical layout analysis of partly annotated newspaper images," in *Proceedings of the 23rd Computer Vision Winter Workshop*, 2018, pp. 63–70.

[6] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[7] S. Pletschacher and A. Antonacopoulos, "The page (page analysis and ground-truth elements) format framework," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 257–260.

[8] R. Smith, "An overview of the tesseract ocr engine," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2. IEEE, 2007, pp. 629–633.