Proposal of ASLA Which Is a Segmentation and Labeling Tool for Document Images Based on Deep Learning

Kanta Kakinoki*, Tetsuro Katayama*, Yoshihiro Kita†,
Hisaaki Yamaba*, Kentaro Aburada*, and Naonobu Okazaki*
** Department of Computer Science and Systems Engineering, Faculty of Engineering, University of Miyazaki, 1-1 Gakuen-kibanadai nishi, Miyazaki, 889-2192 Japan*
†Department of Information Security, Faculty of Information Systems, Siebold Campus, University of Nagasaki 1-1-1 Manabino, Nagayo-cho, Nishi-Sonogi-gun, Nagasaki, 851-2195 Japan
E-mail: kakinoki@earth.cs.miyazaki-u.ac.jp, kat@cs.miyazaki-u.ac.jp, kita@sun.ac.jp, yamaba@cs.miyazaki-u.ac.jp, aburada@cs.miyazaki-u.ac.jp, oka@cs.miyazaki-u.ac.jp

Abstract

This paper proposes a prototype of ASLA, segmentation and labeling tool for document images based on deep learning, to reduce the time required for region segmentation and label generation. To evaluate the usefulness of ASLA, we have compared the time required for region segmentation and label generation using ASLA and by hand, and then confirmed the reduction in time. We also have confirmed that the rule-based region redividing method achieves a high recall and precision.

Keywords: Region segmentation, Labeling, Document image, Rule-based region redividing

1. Introduction

Because storage space is needed to store paper documents, maintaining and managing their condition are costly. Electronic documents are used to solve this problem [1]. The use of electronic documents enables the reduction of paper and the cost of the maintenance and management. Therefore, the use of electronic documents is becoming more widespread.

The current situation of the electronic documents is only a substitute for paper. As a new way to utilize electronic documents, we focus on dividing electronic documents into regions by their elements and generating keywords and sentences as labels from the contents of the elements. In addition, the movement of the reader's line of sight when they read the electronic document is recorded as coordinate information. By realizing these features, it is possible to improve the efficiency of sales activities by capturing the level of interest and concern of customers, and to strengthen compliance by checking whether important matters are properly explained to subscribers of products or services.

To achieve the above, the following two tasks are required.

- Region segmentation that divides the components of an electronic document into some regions.
- Label generation that analyzes the contents of a region and then generates labels according to the contents of the region.

However, these tasks are time-consuming and labor-intensive in the manual. This paper proposes a prototype of ASLA (Automatic Segmentation and Labeling tool using AI), segmentation and labeling tool for document images based on deep learning, to reduce the time required for region segmentation and label generation.

2. ASLA

2.1. *Functions*

ASLA takes a single column or two columns Japanese document image as input. ASLA outputs segmentation image and class identification on the input image. Table 1 shows the definitions of class and region in ASLA. The output is a segmentation image and an analysis result file. The segmentation image is a JPEG format image. Each region of the image is color-coded and visualized as red for the text class region, blue for the figure class region, and green for the table class region. The analysis result file is a CSV (Comma Separated Value) format file. In that file, the top-left x-coordinate, top-left y-coordinate, bottom-right x-coordinate, bottom-right y-coordinate, class name, and label pair in each region are recorded per line.

Table 1 Definition of classes and regions in ASLA.

Class Name	Target	Definition of regions
Text	Text and title, etc.	Smallest rectangle surrounding each element
Figure	Figure	Smallest rectangle surrounding the figure
Table	Table	Smallest rectangle surrounding the table

2.2. *Dataset Preparation*

ASLA uses Cascade R-CNN [2] as the object detector for region segmentation. In addition, ASLA uses LayoutLMv3 [3] that is a pre-trained model specific to document AI tasks as a training model for the object detector. To improve the performance of region segmentation for Japanese document images, we fine-tune the model by using a dataset of Japanese document images prepared in advance.

The ASLA prototype uses 103 document images of the doctoral dissertation of the Graduate School of Educational and Developmental Sciences of Nagoya University provided by the Nagoya University Academic Repository [4] and 87 document images of the paper presented at the 2021 Kyushu Branch Joint Conference of the Institutes of Electrical and Related Engineers [5] as document images dataset for fine-tuning.

2.3. *Structure*

Fig. 1 shows the structure of ASLA. It consists of the following three processing parts.

- Region Segmentation Part:
  Region Segmentation Part at first generates a trained model by fine-tuning LayoutLMv3 with a previously prepared dataset. Next, it divides region by using Cascade R-CNN from the input document image with the generated trained model. This generates the top-left and bottom-right xy-coordinates and class names for each region. Thereafter, it redivides the region to modify the misalignment of regions. The rule-based region redividing is described later in Section 2.4. Finally, based on the results of region segmentation, a segmentation image is output. Since this, text class regions are referred to as text region, figure class regions as figure region, and table class regions as table region, respectively.
- Label Generation Part:
  Label Generation Part generates labels that are all nouns included in the text region for the text region. For the figure and table regions, it generates a label that is the text of each caption. Here, it extracts characters from Japanese text using Tesseract that is an OCR tool, and it generates labels from the extracted characters using MeCab which is a morphological analysis tool.

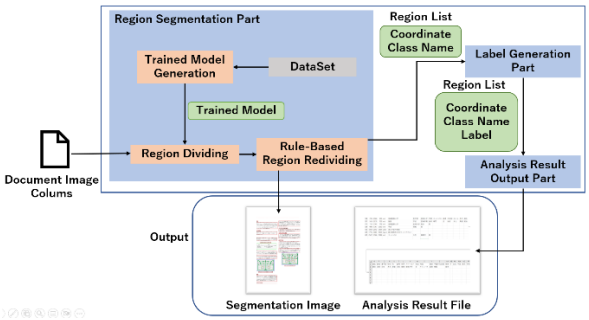


Fig. 1 The structure of ASLA.

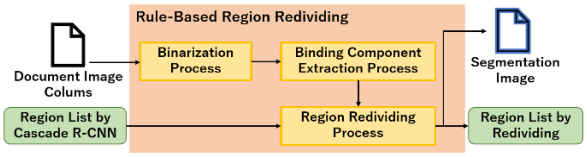


Fig. 2 The Process of rule-based region redividing.

- Analysis Result Output Part:
  Analysis Result Output Part outputs an analysis result file as a CSV format file based on the analysis results obtained by Region Segmentation Part and Label Generation Part.

2.4. *Rule-Based Region Redividing*

There are several studies that use object detectors to divide regions on document images [3], [6], [7]. The purpose is to roughly divide each region in a document image, such as layout analysis of a document image. Therefore, the existence of small misalignments between the divided regions and the actual regions is not a major problem. However, region segmentation in ASLA extracts the characters within a region after the region is divided. Therefore, if there is even a small misalignment, some characters cannot be extracted. This causes the problem that correct labels cannot be generated. Another problem with segmentation by the object detectors exists that it is difficult to detect small regions [6], [7].

Therefore, ASLA performs rule-based region redividing. This eliminates the misalignment of regions with region segmentation by the object detector, and also divides smaller regions that were not detected before. Fig. 2 shows the flow of the rule-based region redividing. In the region redividing process, each binding component extracted by the binding component extraction process determines which region is the same as the region that was divided using the object detector. Alternatively, binding components that are far from any region are divided as a new region. Fig. 3 shows an example of the region redividing process in rule-based region redividing. In Fig. 3, the region redividing process redivides the region to eliminate misalignment of text region segmented by Cascade R-CNN.

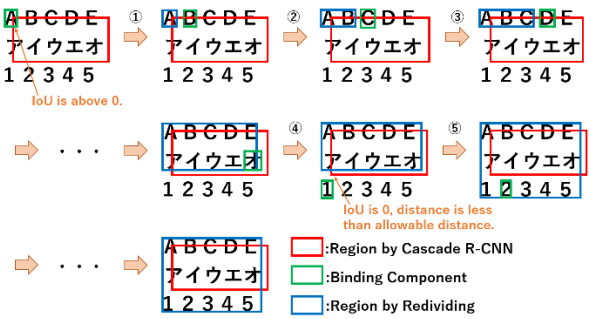


Fig. 3 An example of the region redividing process in rule-based region redividing.

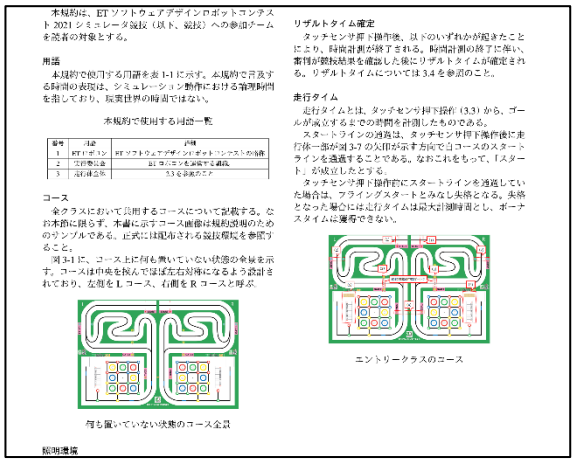


Fig. 4 Input document image sample

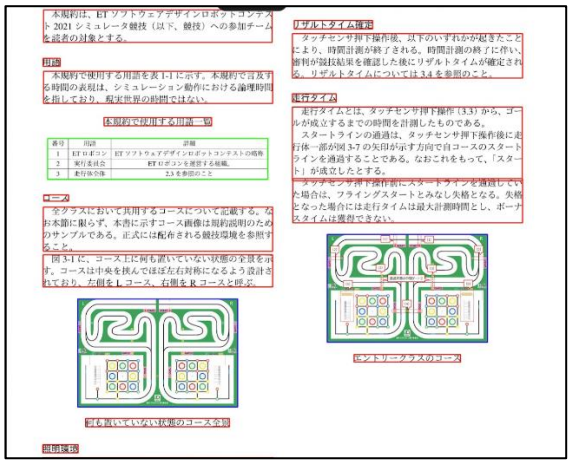
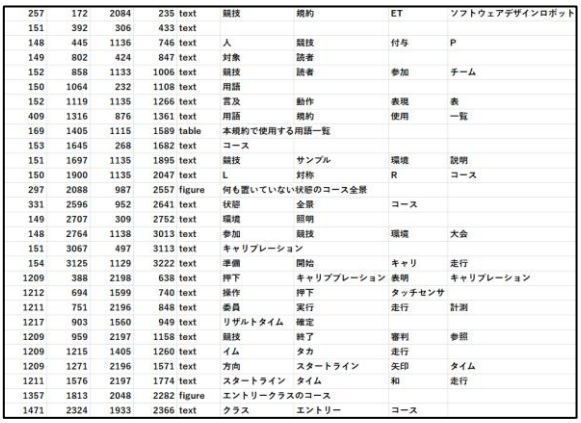


Fig. 5 Segmentation image

2.5. *Input/Output Example*

Fig. 4 shows an example of a document image to be analyzed by ASLA. Fig. 5 shows the segmentation image generated by applying the document image in Fig. 4 as input for ASLA. Fig. 6 shows an image of an analysis result file. From Fig. 5, we have confirmed that ASLA can divide the document image into text, figure, and table regions, as defined in Table 1.



257	172	2084	235	text	競技	規約	ET	ソフトウェアデザインロボット
151	392	306	433	text				
148	445	1136	746	text	人	競技	付与	P
149	802	424	847	text	対象	読者		
152	858	1133	1006	text	競技	読者	参加	チーム
150	1064	232	1108	text	用語			
152	1119	1135	1266	text	言及	動作	表現	表
409	1316	876	1361	text	用語	規約	使用	一覧
169	1405	1115	1589	table	本規約で使用する用語一覧			
153	1645	268	1682	text	コース			
151	1697	1135	1895	text	競技	サンプル	環境	説明
150	1900	1135	2047	text	L	対称	R	コース
297	2088	987	2557	figure	何も置いていない状態のコース全景			
331	2596	952	2641	text	状態	全景	コース	
149	2707	309	2752	text	環境	照明		
148	2764	1138	3013	text	参加	競技	環境	大会
151	3067	497	3113	text	キャリブレーション			
154	3125	1129	3222	text	準備	開始	キャリ	走行
1209	388	2198	638	text	押下	キャリブレーション	表明	キャリブレーション
1212	694	1599	740	text	操作	押下	タッチセンサ	
1211	751	2196	848	text	委員	実行	走行	計測
1217	903	1560	949	text	リザルトタイム	確定		
1209	959	2197	1158	text	競技	終了	審判	参照
1209	1215	1405	1260	text	イム	タカ	走行	
1209	1271	2196	1571	text	方向	スタートライン	矢印	タイム
1211	1576	2197	1774	text	スタートライン	タイム	和	走行
1357	1813	2048	2282	figure	エントリークラスのコース			
1471	2324	1933	2366	text	クラス	エントリー	コース	

Fig. 6. Result output file image

Table 2. Comparison of the time required for ASLA and manual region segmentation.

	Single column	Two columns	Average
Manual (Average)	2m6s	4m8s	3m7s
ASLA	5s	6s	6s

Table 3. Comparison of the time required for ASLA and manual labeling.

	Time
Manual (Average)	8m40s
ASLA	5s

3. Evaluation

We evaluate the usefulness of ASLA developed in this paper. It is evaluated in terms of the execution time and performance of region segmentation.

3.1. *Evaluation of the Execution Time*

We compare the time required for region segmentation by using ASLA and manually. Two document images are used for region segmentation. They are single column document image and two columns document image. In the manual region segmentation, five subjects divide regions on document images, and their working time is measured. Table 2 shows the comparison of the time required for region segmentation by using ASLA and manually. From Table 2, we have confirmed that ASLA reduces the time required for region segmentation by about 3 minutes (97%) per document image.

We compare the time required for label generation by using ASLA and manually. In the manual label generation, five subjects generate labels on document images that had been divided regions in advance, and their working time is measured. This labeling task involves extracting all nouns within each region. Table 3 shows a comparison of the time required for ASLA and manual label generation. From Table 3, we have confirmed that ASLA reduces the time required for label generation by about 8 minutes (99%) per document image. Based on the above, it can be said that ASLA helps reduce the time required for label generation.

Table 4 Comparison of precision and recall for Cascade R-CNN and Mask R-CNN with and without rule-based region redividing, respectively.

		Precision			Recall		
	IoU	Text	Figure	Table	Text	Figure	Table
Only Mask R-CNN	0.8	0.768	0.846	0.4	0.651	0.846	0.455
	0.9	0.348	0.154	0.2	0.295	0.154	0.182
Mask R-CNN and rule-based region redividing	0.8	0.906	0.929	0.9	0.955	1.0	1.0
	0.9	0.877	0.929	**0.9**	0.924	**1.0**	**1.0**
Only Cascade R-CNN	0.8	0.808	0.846	0.4	0.691	0.918	0.818
	0.9	0.337	0.317	0.2	0.305	0.308	0.182
Cascade R-CNN and rule-based region redividing	0.8	0.945	0.942	0.9	0.955	1.0	1.0
	0.9	**0.936**	**0.942**	**0.9**	**0.955**	**1.0**	**1.0**

3.2. *Evaluation of Region Segmentation Performance*

In addition to ASLA, there are other studies that divide region on document images [6], [7]. These studies use Mask R-CNN [8] as an object detector for region segmentation. Here, we evaluate the performance of region segmentation for each of the Cascade R-CNN used in ASLA and Mask R-CNN. The dataset used for performance evaluation consists of 8 document images. These document images consist of 4 single column document images and 4 two columns document images. The evaluation is based on the calculation of IoU for the regions and class names defined in Table 1 to calculate the precision and recall. Furthermore, to evaluate the usefulness of rule-based region redividing, we calculate the precision and recall with and without rule-based region redividing.

Table 4 shows the calculated precision and recall. From Table 4, we have confirmed that the combination of Cascade R-CNN and rule-based region redividing used in Region Segmentation Part of ASLA has the highest values for both precision and recall. We have confirmed that the usefulness of ASLA for region segmentation. In addition, the rule-based region redividing also has achieved the highest precision and recall when the IoU threshold is set to a high value of 0.9. We have confirmed that the usefulness of rule-based region redividing.

4. Conclusion

This paper has proposed a prototype of ASLA that is a deep learning tool for region segmentation and label generation, in order to reduce the time required to divide document images and generate labels. The proposed ASLA performs region segmentation and class identification on the input document image, generates labels each region, and outputs the segmentation image and the analysis result file.

After applying the document images to ASLA, we have confirmed that ASLA works correctly. Furthermore, we have compared the time required for region segmentation with the use of ASLA and manually. As a result, we have confirmed that the time required for region segmentation per document image could be reduced by about 3 minutes (97%). We also have compared the time required for label generation by using ASLA and manually. As a result, we have confirmed that the time required for label generation per document image could be reduced by about 8 minutes (99%). Furthermore, we have evaluated the performance of region segmentation. As a result, we have confirmed that the rule-based region redividing achieves high precision and recall even with a high IoU threshold value.

The future issues are as follows.

- Support for document images with non-white backgrounds.
- Automatic generation of sentences or nouns that represent the contents of figures or tables.
- Handling of text regions that do not make sense as Japanese.

References

1. Hirohito Shibata, “Paper vs. Electronic Media: Work Efficiency and Environmental Impact” NIP & Digital Fabrication Conference, Vol.27, pp.7-10, 2011.
2. Zhaowei Cai and Nuno Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” Proc. of 2018 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 6154-6162, 2018.
3. Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei, ”LayoutLMv3: Pre training for Document AI with Unified Text and Image Masking”, MM '22: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4083-4091, 2022.
4. “Nagoya Repository”, https://nagoya.repo.nii.ac.jp/ (Accessed 2023-12-14)
5. “The 74th Joint Conference of Electrical, Electronics and Information Engineers in Kyushu”, https://sites.google.com/jceee-kyushu.jp/2021/ (Accessed 2023-12-14)
6. Canhui Xu, Cao Shi, Hengyue Bi, Chuanqi Liu, Yongfeng Yuan, Haoyan Guo, and Yinong Chen: "A page object detection method based on Mask R-CNN", IEEE Access, Vol.9, pp.143448-143456, 2017.
7. Sanket Biswas, Pau Riba, Josep Llados, Umapada Pal, “Beyond document object detection: instance-level segmentation of complex layouts”, IJDAR 24, pp.269-281, 2021.
8. Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross B. Girshick: "Mask R-CNN", International Conference on Computer Vision(ICCV), pp.2961-2969, 2017.

Authors Introduction

Mr. Kanta Kakinoki

He received the Bachelor's degree in engineering (computer science and systems engineering) from the University of Miyazaki, Japan in 2023. He is currently a Master's student in Graduate School of Engineering at the University of Miyazaki, Japan. His research interests include software development support method and image processing.

Dr. Tetsuro Katayama

He received a Ph.D. degree in engineering from Kyushu University, Fukuoka, Japan, in 1996. From 1996 to 2000, he has been a Research Associate at the Graduate School of Information Science, Nara Institute of Science and Technology, Japan. Since 2000 he has been an Associate Professor at the Faculty of Engineering, Miyazaki University, Japan. He is currently a Professor with the Faculty of Engineering, University of Miyazaki, Japan. His research interests include software testing and quality. He is a member of the IPSJ, IEICE, and JSSST.

Yoshihiro Kita

He received a Ph.D. degree in systems engineering from the University of Miyazaki, Japan, in 2011. He is currently an Associate Professor with the Faculty of Information Systems, University of Nagasaki, Japan. His research interests include software testing and biometrics authentication.

Dr. Hisaaki Yamaba

He received the B.S. and M.S. degrees in chemical engineering from the Tokyo Institute of Technology, Japan, in 1988 and 1990, respectively, and the Ph D. degree in systems engineering from the University of Miyazaki, Japan in 2011. He is currently an Assistant Professor with the Faculty of Engineering, University of Miyazaki, Japan. His research interests include network security and user authentication. He is a member of SICE and SCEJ.

Kentaro Aburada

He received the B.S., M.S, and Ph.D. degrees in computer science and system engineering from the University of Miyazaki, Japan, in 2003, 2005, and 2009, respectively. He is currently an Associate Professor with the Faculty of Engineering, University of Miyazaki, Japan. His research interests include computer networks and security. He is a member of IPSJ and IEICE.

Dr. Naonobu Okazaki

He received his B.S, M.S., and Ph.D. degrees in electrical and communication engineering from Tohoku University, Japan, in 1986, 1988 and 1992, respectively. He joined the Information Technology Research and Development Center, Mitsubishi Electric Corporation in 1991. He is currently a Professor with the Faculty of Engineering, University of Miyazaki since 2002. His research interests include mobile network and network security. He is a member of IPSJ, IEICE and IEEE.