SUGISAKA MASANORI

Journal of Robotics, Networking and Artificial Life Vol. 10(4); March (2024), pp. 362–367
ISSN (Online): 2352-6386; ISSN (Print): 2405-9021
https://alife-robotics.org/jrnal.html

Research Article

ASLA: Automatic Segmentation and Labeling by Deep Learning for Document Pictures

Kanta Kakinoki[1], Tetsuro Katayama[1], Yoshihiro Kita[2], Hisaaki Yamaba[1], Kentaro Aburada[1], Naonobu Okazaki[1]

[1]*Department of Computer Science and Systems Engineering, Faculty of Engineering, University of Miyazaki, 1-1 Gakuen-kibanadai nishi, Miyazaki, 889-2192, Japan*
[2]*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*

ARTICLE INFO

Article History

Received 24 November 2023
Accepted 22 July 2024

Keywords

Document picture
Area segmentation
Label generation
Rule-based area correction

ABSTRACT

In this paper, we propose ASLA, a segmentation and label generation system for document pictures. ASLA reduces the duration needed for separating document pictures into areas and label generation. By using the application example, we have verified that ASLA operates properly. We have evaluated the usefulness of ASLA in regard to time and accuracy. We have assessed the efficacy of the rule-based area correction method. As a result, we have verified that ASLA is useful.

1. Introduction

As the quantity of printed texts grows, their storage becomes increasingly challenging. The maintenance and management of print-based records require significant expenses. Moreover, there's always the possibility for documentation to be misplaced or unlawfully removed. Digital documentation offers solutions to these issues. Transitioning to digital formats can substantially decrease the costs associated with maintenance and management of documentation. Digital files also have many advantages, such as the easy sharing of data. Additionally, implementing a digital document management system can reduce the time spent searching for the needed documentation. As a result, the adoption of digital documentation is gaining widespread popularity across various sectors.

A new approach to utilizing digital documentation should be proposed because they mainly serve as substitutes for printed texts at present. This approach involves separating digital documentation into areas based on their constituent elements, and subsequently generating descriptive keywords and phrases from the content of these areas. Furthermore, our approach incorporates the tracking of the reader's gaze movements, mapping it. The implementation of this approach can enable the following significant benefits.

- Enhancing sales strategies through a deeper understanding of client interests and concerns
- Enhancing compliance by ensuring accurate communication of critical information to service subscribers and product consumers

Documentation in Portable Document Format or picture format is convenient for human readers, but it is difficult to understand the layout of the document and extract information from these formats. Therefore, the following two tasks are needed to solve the above.

- Area segmentation that separates a digital document into areas based on their constituent elements.
- Label generation that generates descriptive keywords and phrases from the content of areas.

Manual execution of these processes requires time-consuming and labor-intensive. To deal with this inefficiency, we propose a prototype system named

Corresponding author 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*

ASLA, a segmentation and label generation system by deep learning for document pictures.

2. ASLA

2.1. *Functions*

ASLA is designed to process Japanese document pictures in either single or dual column layouts. ASLA generates two primary outputs. These outputs are the separated picture and analyzed file. The separated picture is rendered as a JPEG (Joint Photographic Experts Group) picture, which is the original document overlaid with color-coded rectangles. The analyzed file is a structured data file in CSV (Comma Separated Value) format. In this paper, the area is the smallest rectangle surrounding the components of a document.

Table 1 Classes and colors of rectangles.

Classes	Caption	Footnote	Formula	List-item	Page-footer	Page-header	Picture	Section-header	Table	Text	Title
Colors	violet	olive	cyan	teal	purple	orange	blue	magenta	green	red	brown

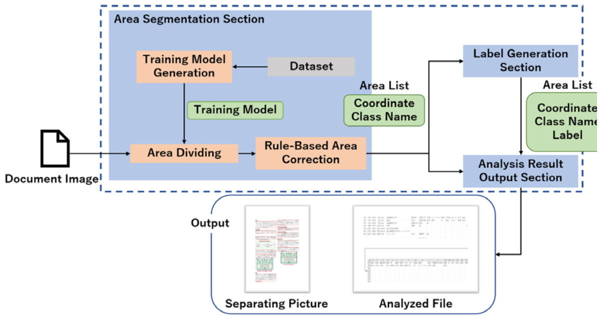

Fig. 1 ASLA's structure.

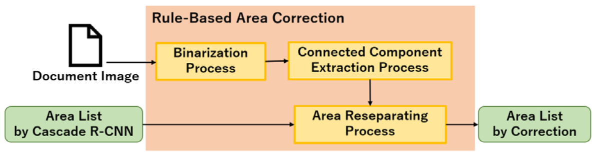

Fig. 2 The flow of rule-based area correction.

2.2. *Structure*

Fig. 1 presents ASLA's structure. ASLA is composed of the three following processing sections.

- Area Segmentation Section:
  This section adopts Cascade R-CNN [1] known as a sophisticated object detection architecture for images. To enhance its accuracy, the section integrates LayoutLMv3 [2], a specialized pre-trained model designed for document AI applications, as the basis for its object detection algorithm. First, this section generates a training model. The system refines LayoutLMv3 model by using DocLayNet [3] that is a diverse dataset containing a wide range of document formats and their corresponding layout annotations. Next, Cascade R-CNN processes the input document picture by utilizing the refined model, delimiting areas within it. For each identified area, this section determines and records the coordinates (top-left and bottom-right) along with the appropriate classification. Thereafter, the rule-based area correction is applied to adjust any misalignments in the separated areas. This process is discussed in more detail in Section 2.3 of this paper. Finally, it sends the final segmentation data to the label generation section.
- Label Generation Section:
  This section processes the areas separated by the area segmentation section to generate appropriate labels. For visual elements such as picture and table, this section extracts the associated captions and utilizes these as labels. For textual areas, the section generates all nouns included in the area as labels. For these processes, the OCR tools Tesseract [4] and the morphological analysis tool Mecab are used.
- Area Analysis Section:
  This section outputs the segparating picture and the analyzed file. The separated picture is a picture which is drawn rectangles surrounding an area in colors defined for each area on the input picture. The analyzed file is the CSV format output of the analysis results obtained from the area segmentation and labeling sections. Table 1 presents the definitions of classes and colors of rectangles in ASLA.

2.3. *Rule-based Area Correction*

Some recent researches have considered the application of object detection frameworks for document layout analysis [2], [5]. These researches primarily focus on broad separating document into areas. In such contexts, misalignments between separated and actual area boundaries are generally considered inconsequential. However, approach of ASLA differs from these previous research in a crucial aspect. ASLA requires precise character extraction within each segmentation areas. Consequently, misalignments between separated and actual area boundaries are unacceptable in ASLA. Furthermore, existing object detection architectures often struggle with the detection of smaller document elements [5].

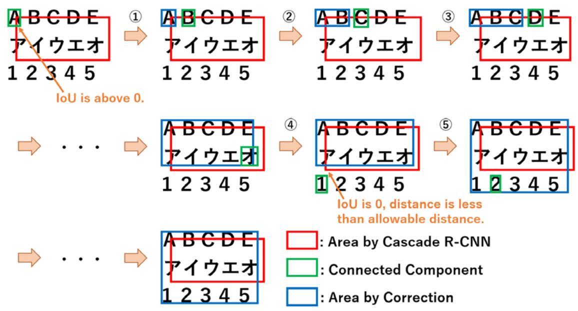

Fig. 3 The area reseparating process in rule-based area correction.

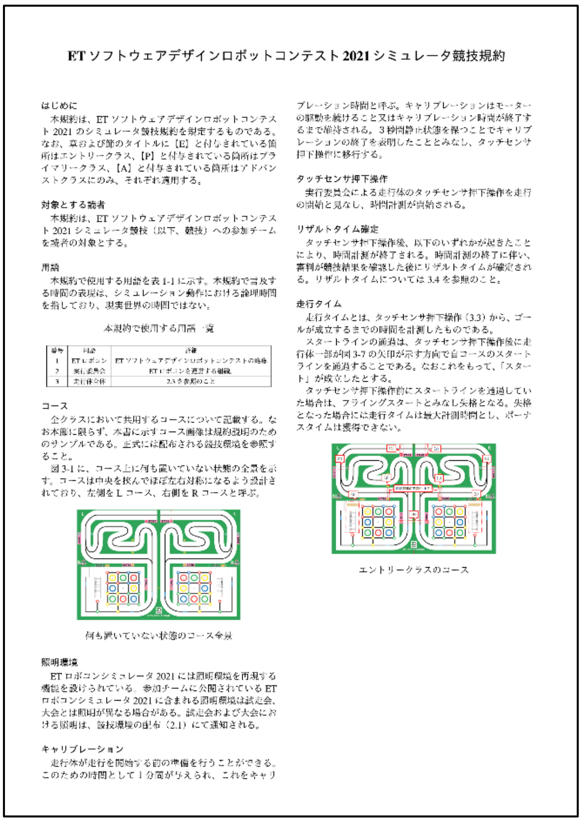

Fig. 4 Input document picture sample

To deal with these limitations, ASLA incorporates the rule-based area correction. This process serves to bridge the gap between initial object detection-based segmentation and the need for highly accurate area segmentation. Fig. 2 presents the workflow of this correction process. This process adjusts the areas initially identified by the object detection architecture based on the connected components extracted from the document areas for refinement. Fig. 3 presents a visual demonstration of this refinement process. In Fig. 3, an example of correction misalignments in text areas that were initially detected by Cascade R-CNN architecture.

3. Application Example

Fig. 4 presents a document picture that serves as our test case. Fig. 4 is a document picture containing the components of title, caption, text, section-header, picture, and table. Fig. 5 presents the separated picture generated by inputting the document picture of Fig. 4 into ASLA. Fig. 6 presents the analysis results file. From Fig. 5, we have verified that ASLA is able to separate the picture into each area, as defined in Table 1. Also, we have verified that the small areas are separated and that there are no misalignments in the segmentation areas.

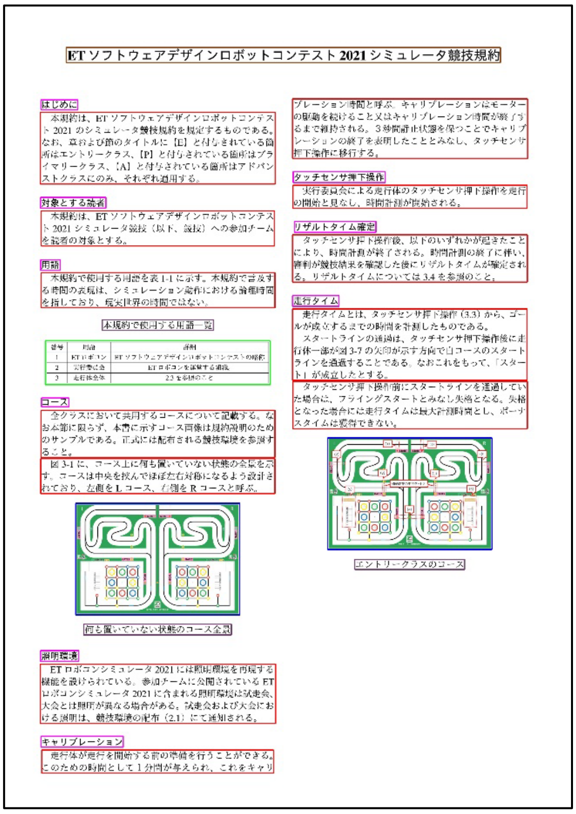

Fig. 5 Separated picture

257	172	2084	235	Title	競技	規約	ET	ソフトウェアデザイン
151	392	306	433	Section-header	はじめ			
148	445	1136	746	Text	人	競技	付与	P
149	802	424	847	Section-header	対象	読者		
152	858	1133	1006	Text	競技	読者	参加	チーム
150	1064	232	1108	Section-header	用語			
152	1119	1135	1266	Text	言及	動作	表現	表
409	1316	876	1361	Caption	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	Section-header	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	Picture	何も置いていない状態のコース全景			
331	2596	952	2641	Caption	状態	全景	コース	
149	2707	309	2752	Section-header	環境	照明		
148	2764	1138	3013	Text	参加	競技	環境	大会
151	3067	497	3113	Section-header	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	キャリブプレーション	表明	キャリブレーション
1212	694	1599	740	Section-header	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	Section-header	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	Section-header	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	Picture	エントリークラスのコース			
1471	2324	1933	2366	Caption	クラス	エントリー	コース	

Fig. 6 Analyzed file

4. Evaluation

We evaluate the usefulness of prototype ASLA. It is evaluated in the execution time, accuracy of area segmentation, and accuracy of text extraction. In this paper, we evaluate the accuracy of area segmentation and text extraction by calculating the F score and Levenshtein distance.

4.1. *Evaluation of the Execution Time*

To evaluate the efficiency of ASLA, we conducted a comparative analysis of processing times between our tool and hand-operated methods. This evaluation focused on two key tasks: segmentation and label generation.

Our segmentation evaluation utilized two document layouts that are a single-column format and a dual-column arrangement. For the hand-operated segmentation, we engaged five subjects to perform area segmentation on these documents. Table 2 presents a comprehensive comparison of the time requirements for both methods. From Table 2, We have verified a reduction in processing time of approximately 3 minutes (95%) per document. In other words, we gain a striking efficiency with ASLA in area segmentation.

Table 2 Evaluating the duration needed for ASLA versus hand-operated segmentation.

	One column	Dual column	Average
Hand-operated (Average)	2m6s	4m8s	3m7s
ASLA	7s	9s	8s

Table 3 Evaluating the duration needed for ASLA versus hand-operated labeling.

	Time
Hand-operated (Average)	8m40s
ASLA	5s

In label generation evaluation, we measured the time efficiency of label generation. The hand-operated labeling involved five subjects who were tasked with generating labels to pre-separated document pictures. Table 3 presents the comparative results of this label generation task. From Table 3, ASLA achieved a time reduction of approximately 8 minutes (99%) per document in the label generation. In other words, we gain a striking efficiency with ASLA in label generation.

4.2. *Evaluation of Area Segmentation Accuracy*

In the realm of document picture analysis, various research has explored the object detection architectures for area segmentation [5]. One of these architectures is Mask R-CNN [6] that is a sophisticated object detection architecture. Based on these previous research efforts, this evaluation aims to provide an all-inclusive comparative evaluation.

We evaluate segmentation accuracy by contrasting the performance of two architectures. Two architectures are Cascade R-CNN which forms the backbone of ASLA and Mask R-CNN utilized in previous research. To ensure assessment, we adopt DocLayNet test dataset which is a well-recognized benchmark in the document layout analysis field. Our evaluation derives IoU value for the areas and classes defined in Table 1 to calculate the F score. In addition, we calculate the F score with and without rule-based area correction to evaluate the usefulness of rule-based area correction. The equations for the F score are shown below.

$$Precision = \frac{Areas\ with\ IoU\ Above\ Threshold}{All\ Detected\ Areas} \quad (1)$$

$$Recall = \frac{Areas\ with\ IoU\ Above\ Threshold}{All\ Corrected\ Areas} \quad (2)$$

$$F\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

Table 4 presents the computed F score. In this result, we have verified that the integration of Cascade R-CNN with area correction implemented in ASLA's area segmentation section is the most favorable F score. Moreover, when the IoU threshold is elevated to a stringent 0.9, the rule-based area correction continues to exhibit superior performance, maintaining the highest F score among the compared methods. These observations provide strong evidence for the effectiveness of the rule-based area correction approach.

4.3. *Evaluation of Text Extraction Accuracy*

After area segmentation, ASLA uses OCR to extract the characters in the area and generates labels. Therefore, the accuracy of character extraction has a significant impact on the accuracy of label generation. Thus, to evaluate the accuracy of character extraction, we calculate the Levenshtein distance. The Levenshtein distance is a metric for the degree of similarity between two strings. The distance is the number of deletions, insertions, and replacements needed to convert one string into another [7]. Fig. 4 presents the test data is the document picture. We calculate the Levenshtein distance between the character strings in this document picture and the character strings generated by the Label Generation Section of ASLA.

Table 5 shows the calculated Levenshtein distances. From Table 5, we have verified that the combined use of Cascade R-CNN and area correction in the area segmentation section of ASLA gives the best accuracy in character extraction. In addition, we have verified that area correction improves the accuracy of character extraction for both Mask R-CNN and Cascade R-CNN.

5. Conclusion

This paper has presented a prototype system named ASLA, a segmentation and label generation system by deep learning for document pictures, to reduces the duration needed for segmentation document pictures into areas and label generation. ASLA operates by segmentation input document pictures into distinct areas, classifying these areas, generating appropriate labels, and subsequently producing both a separated picture and an analysis file.

Empirical testing with various document pictures has validated ASLA's functional accuracy. Our evaluation of ASLA's efficiency reveals significant time savings. Area segmentation time per document is reduced by approximately 3 minutes (95%). In addition, label generation time is reduced by approximately 8 minutes (99%). Further assessment of ASLA's segmentation accuracy demonstrates its ability to maintain high F scores, even under stringent IoU threshold conditions. Moreover, an analysis of text extraction accuracy using Levenshtein distance metrics indicates that ASLA achieves superior performance in character extraction tasks.

Table 4 Evaluating the accuracy of area segmentation by F score.

Method		Only Mask R-CNN		Mask R-CNN and area correction		Only Cascade R-CNN		Cascade R-CNN and area correction (ASLA)	
IoU		0.8	0.9	0.8	0.9	0.8	0.9	0.8	0.9
class	Caption	0.58	0.24	0.71	0.68	0.60	0.27	**0.75**	**0.71**
	Footnote	0.69	0.31	0.75	0.72	0.71	0.37	**0.76**	**0.76**
	Formula	0.61	0.27	**0.69**	**0.69**	0.61	0.33	**0.69**	0.68
	List-item	0.83	0.38	0.84	0.80	0.83	0.54	**0.88**	**0.87**
	Page-footer	0.68	0.25	0.72	0.71	0.71	0.31	**0.73**	**0.72**
	Page-header	0.73	0.29	0.76	0.73	0.74	0.35	**0.79**	**0.78**
	Picture	0.71	0.28	0.77	0.75	0.79	0.42	**0.81**	**0.77**
	Section-header	0.62	0.21	0.69	0.69	0.66	0.28	**0.71**	**0.69**
	Table	0.69	0.26	0.82	0.80	0.73	0.34	**0.83**	**0.80**
	Text	0.85	0.32	0.90	0.88	0.87	0.39	**0.91**	**0.89**
	Title	0.74	0.34	0.71	0.67	0.74	0.32	**0.78**	**0.75**
	Average	0.70	0.29	0.76	0.74	0.72	0.36	**0.79**	**0.77**

Table 5 Evaluating Levenshtein distances.

Method	Only Mask R-CNN	Mask R-CNN and area correction	Only Cascade R-CNN	Cascade R-CNN and area correction (ASLA)
Levenshtein distances	219	5	183	**3**

The future issues include the following.

- Support non-white backdrop document pictures. ASLA treats the background of document pictures with non-white backgrounds as an area. This prevents appropriate area segmentation. Therefore, it is necessary to support document pictures with non-white backdrops.
- Automatically create nouns or phrases that describe the details of pictures or tables. ASLA generates the text in the caption below the picture and above the table as the label. This means that the target of analysis includes the condition that there is a caption below the picture and above the table. In recent documentation, there are often no captions for pictures or tables. Therefore, the current ASLA lacks versatility. Therefore, it is necessary to automatically generate labels from pictures and tables.
- Consider processing of meaningless parts in Japanese.
  ASLA cannot generate appropriate labels if the text in the area does not make sense as Japanese. This means that it cannot generate labels for areas such as page numbers. Documentation often has these kinds of areas. Therefore, it is necessary to handle areas that do not make sense as Japanese in an exceptional way.
- Improve Accuracy of Optical Character Recognition. ASLA uses Tesseract in the label generation section. However, the current Tesseract cannot extract characters from pictures 100% accurately. This may result in incorrect extraction of characters. Therefore, it is necessary to improve accuracy of OCR.

References

1. 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.
2. Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei, ”LayoutLMv3: Pre training for Document AI with Unified Text and Picture Masking”, MM '22: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4083-4091, 2022.
3. Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter Staar, “DocLayNet: A Large Human-Annotated Dataset for Document-Layout Segmentation”, KDD '22: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3743-3751, 2022.

4. ”tesseract”, https://github.com/UB-Mannheim/tesseract (Accessed 2024-06-05)
5. Sanket Biswas, Pau Riba, Josep Llados, Umapada Pal, “Beyond document object detection: instance-level segmentation of complex layouts”, IJDAR 24, pp.269-281, 2021.
6. Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross B. Girshick: "Mask R-CNN", International Conference on Computer Vision(ICCV), pp.2961-2969, 2017.
7. Vladimir I. Levenshtein., “Binary codes capable of correcting deletions, insertions and reversals”, Soviet Physics, Doklady, Vol.10, pp.707-710, 1966.

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 picture 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.

Dr. 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.

Dr. 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.