

Document Recognition and XML Generation of Tabular Form Discharge Summaries for Analogous Case Search System

H. Kawanaka¹, T. Sumida², K. Yamamoto³, T. Shinogi¹, S. Tsuruoka¹

¹Graduate School of Engineering, Mie University, Mie, Japan

²Faculty of Engineering, Mie University, Mie, Japan

³Dept. of Medical Informatics, Mie University Hospital, Mie, Japan

Summary

Objectives: This paper discusses and develops a document image recognition, keyword extraction and automatic XML generation system to search analogous cases from paper-based documents. In this paper, we propose the document structure recognition method and automatic XML generation method for the tabular form discharge summary documents. This paper also develops the prototype system using the proposed method. Evaluation experiments using actual documents are done to discuss the effectiveness of the developed system.

Methods: The developed system consists of the following methods. Paper-based summary documents are scanned by a scanner using 300 dpi first. Noise and tilt of the image are reduced by pre-processing, and the table structures are identified. Characters in the table are recognized and converted to text data by the OCR engine. XML documents are automatically generated using obtained results.

Results: In this paper, patient discharge summary documents archived at Mie University Hospital were used. The results show that XML documents can be automatically generated when standard tabular form documents are input into the developed system. In this experiment, it takes about 20 seconds to generate an XML document using the general personal computer. This paper also compares the developed system with a commercial product to discuss the effectiveness of the present system. Experimental results also show that the accuracy of table structure recognition is high and it can be used in a practical situation.

Conclusions: This paper showed the effectiveness of the proposed method to recognize the tabular form document images to generate XML documents.

Keywords

Document structure recognition, analogous case search, tabular form documents, patient discharges, computer-assisted image processing

Methods Inf Med 2007; 46: 700–708

1. Introduction

Recently, a lot of medical and clinical systems are computerized because of the diffusion of clinical information systems (CIS) [1-3]. There are, however, a lot of paper-based documents in hospitals and these documents have been archived without being computerized first. These documents require a large area to archive and it takes a lot of time to search analogous cases from them. As a result, these documents are only stored and not used effectively today. In these years, many of the medical staff are ambitious towards the development of a system that converts these documents to digital data such as XML. Particularly, the need for systems, which extract contents from paper-based documents, classify and analyze them automatically and support medical staff to search analogous cases, is drastically growing.

In this study, we discuss and develop the document image recognition, keyword extraction and automatic XML generation system to search analogous cases from paper-based documents. In the proposed system, the document structure is recognized from input document images scanned by an optical image scanner and XML documents are automatically generated using the information. The proposed method also extracts keywords for analogous case search automatically. In this study, discharge summary documents of actual hospitals are employed. Contents of the documents, i.e. results of medical treatment and/or medication, are extracted by a document structure recognition method. These extracted contents are analyzed by text mining methods and keyword tags for searches can be

entered automatically. It is expected that effective and flexible searches are realized by the proposed method. In the first step of this study, the authors propose the document structure recognition method and automatic XML generation method for the tabular form discharge summary documents. This paper also develops the prototype system using the proposed method. Evaluation experiments using actual documents are done to discuss the effectiveness of the developed system. The results show that XML documents can be automatically generated when standard tabular form documents are used as the input for the developed system. This paper also compares the developed system with a commercial product to discuss the effectiveness of the proposed method. Finally, this paper describes future works of this study.

2. Employed Discharge Summary Documents

In this study, we employed printed summary documents that consist of tables and are archived at Mie University Hospital. The main reasons we used these documents are the following:

- 1) In the hospitals, many printed summaries are still archived.
- 2) It is easy to discuss the effectiveness of the developed system due to its simple architecture.

A lot of printed summary documents that are not saved as digital data (such as MS-Word, PDF and so on) are archived at the hospitals. Handwritten summary docu-

ments are also archived. We have to appropriately extract character regions from these images for character recognition when these documents are set as the processing objects of this study. It is, however, too difficult to extract the regions appropriately using current image processing techniques and this makes the architecture of the developed system more complex. Moreover, most of the discharge summary documents used in the hospital are the regular tabular form documents.

3. Methods

3.1 Effectiveness of a Master Information

Generally, employment of the master information will improve the reliability of the proposed method when the processed objects are regarded as stylized documents such as standard tabular form documents. In this paper, master information means the positions or contents of each element in the table. The master information is generated from the master image and it is obtained by scanning a blank discharge sheet. Using the master information clarifies the XML structure and makes the extraction of strings inscribed by users easy. Furthermore, no advanced processes such as natural language processing techniques and meaning understanding methods are required. Because of these reasons, this paper employs the master information to the proposed system aggressively. In the case of handwritten summary documents, it is difficult to extract keywords and generate the master information before natural language processing techniques are applied to them. In the case of the proposed method, these processes can be added without drastically changing the concept and sequence of the proposed method.

3.2 Outline of Methods

The flow of the proposed method is illustrated in Figure 1. As the first step, paper-based discharge summary documents are scanned by an optical image scanner, and

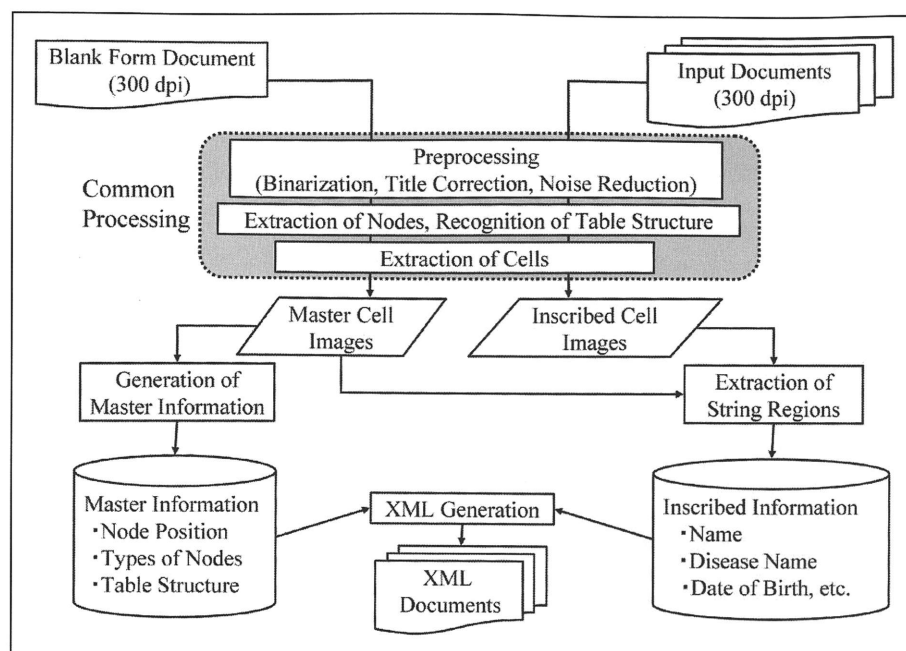


Fig. 1 Flow of proposed method

gray scale images with a resolution of 300 dpi are obtained. The proposed method uses these images as the input images. These images have some factors that make accuracy of the following processes deteriorate, e.g. noise, tilt and so on. These factors are reduced or removed by the pre-processes, and the table structures are identified to extract the regions inscribed by users. Characters in each region are recognized and converted to text data and XML documents are generated using these results. The details of each process will be described in the following sections. In Figure 1, crossover points of ruled lines in the documents and the table structures have to be extracted automatically when a master information is not generated. In this case, the master information is generated using character recognition and natural language processing techniques from the regions inscribed by users. The authors are also discussing this method and it will be applied and implemented to the proposed system [3, 4].

3.3 Pre-processing

In this study, binarization, tilt correction and noise reduction processes are applied to the

input images as the pre-processing. In the binarization process, Ohtsu's method using discriminant analysis was employed [5, 6]. In his method, the threshold for binarization is determined by using a density histogram of the input image. The pixels are divided into two groups (white pixels and black pixels) by discriminant analysis. Therefore, a fixed threshold for each image is not required, as thresholds considering the features of input images are automatically determined.

In the tilt correction process, the Local Projection Profile (LPP) Method is used to correct the tilt of the images [7]. Figure 2 shows the outline of the LPP method. In the LPP method, the target image, i.e. the input image, is divided into n sub-regions and marginal distributions of each region are obtained. In this case, horizontal projection histograms are used as the marginal distributions. Next, correlations between each region (α_k) are calculated by

$$\begin{aligned}
 \alpha_k &= \max_{-\beta < x < \beta} \left[\sum_k P_k(j) P_{k+1}(j-x) \right] \\
 &= \sum_k P_k(j) P_{k+1}(j-\alpha_k) \quad (1) \\
 (k &= 1, 2, 3, \dots, n)
 \end{aligned}$$

where $P_k(j)$ means the j -th value of the horizontal projection histogram of the k -th sub-region and β means the range of calculation. These values indicate misaligns of the phases in each region, which are equivalent to the ratio of the tilt. As the result, the tilt angle of the paper is given by

$$\theta = \tan^{-1} \frac{\alpha_m}{S_h} \quad (2)$$

where α_m and S_h mean the average of α_k and the width of sub-regions, respectively. The LPP method can detect the tilt of images with high accuracy and only little calculation. In the present method, the theoretical detection accuracy is 0.06 degree when the image size is 1024×1024 pixels, however, the detection range is from -10 degrees to 10 degrees. In the proposed method, only the PLL method is used as tilt correction processing because images tilted by more than 10 degrees do not occur frequently in practice.

As the final step of the pre-processing, a median filter was applied to the images to reduce speckle noise and salt and pepper noise.

3.4 Recognition of the Table Structure

3.4.1. Extraction of Ruled Lines

Ruled lines in the input images have to be detected and appropriately extracted to recognize the document structures, i.e. the positions of the above lines and types of these crossover points have to be recognized with high accuracy. In this paper, ruled lines in the input images are extracted using black pixels forming a straight line. When there is a horizontal connected component that consists of n black pixels, it is regarded as a horizontal solid line. The same process is also applied to extract vertical ruled lines. In this paper, the value of n is decided experimentally as 50. When the resolution of the input images is 300 dpi the length of 50 pixels is equivalent to about 4.2 mm.

Figure 3 shows a result of ruled line extraction. It indicates that all ruled lines in the input images are extracted, but partial lines of characters or underlines in the

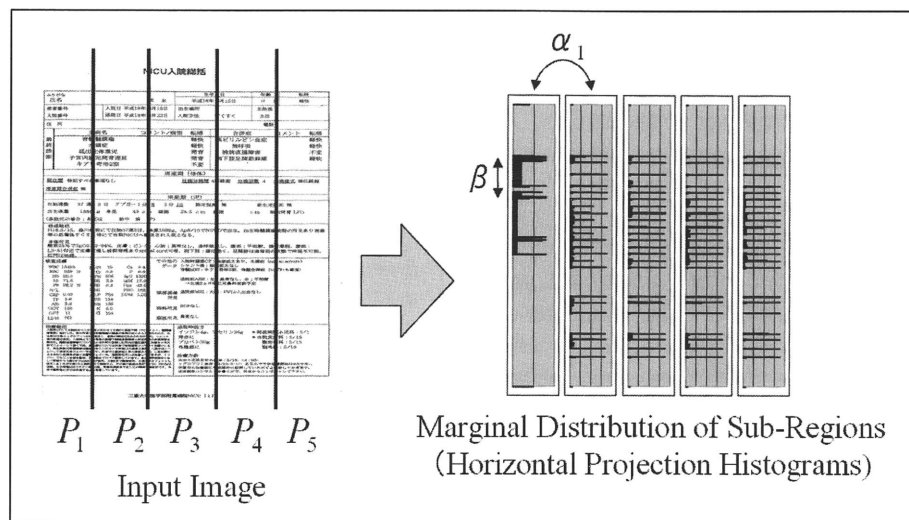


Fig. 2 Outline of LPP method

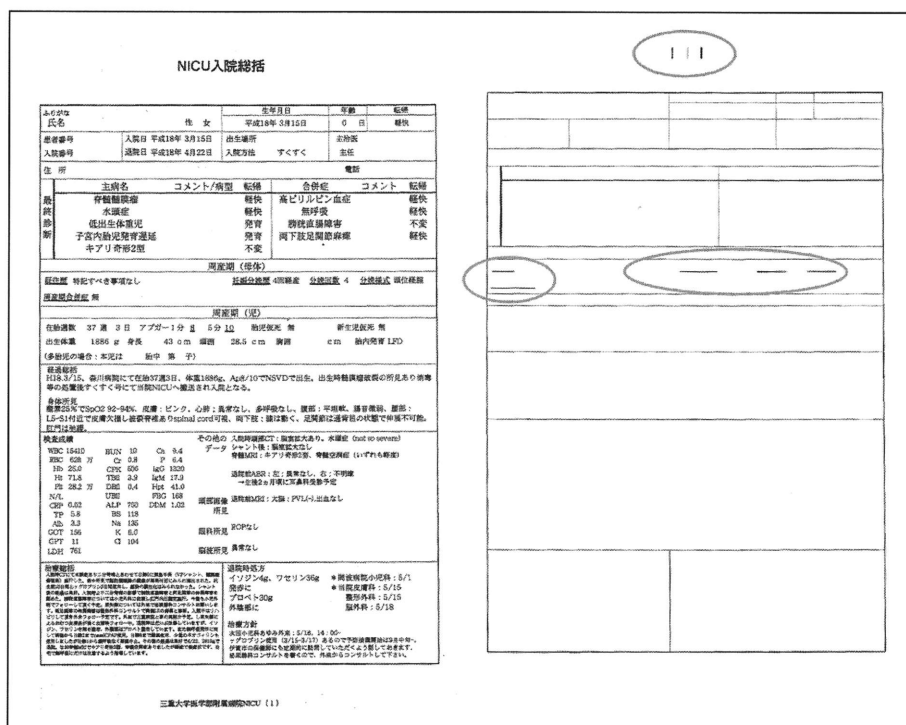


Fig. 3 Result of ruled line extraction

image are also obtained shown as the circular parts in the figure. These parts may influence the following processes, but in the proposed method, the detection of crossover points can remove these surplus lines. Therefore, this does not become a significant problem. As a matter of fact, it is easy to remove these surplus lines by adjusting the value of n .

3.4.2 Finding and Classifying the Nodes

The types and positions of crossover points have to be decided for recognition of the table structures in the input images. The majority of the crossover points, however, consist of more than one black pixel. In the proposed method, the rectangle that consists of candidate pixels, i.e. the region in which

ruled lines cross, is obtained first. The gravity point of the region is regarded as the position of the crossover point.

Next, the crossover points are classified using the above information and existence of solid lines. Figure 4 shows the outline of the method. Generally, a table consists of nine types of crossover points, which are called a “node” in this paper, and non-crossover points [8-10]. Thus, this paper expresses the table in the document using these features. In the proposed method, ruled lines around the target nodes are searched first. In the case in Figure 4, if a ruled line exists above the target node, then, nodes #1, #2, and #3 are excluded as candidates. Next step, ruled lines are also searched for on the left, right and bottom of the target node. As a result, the target node in the figure is identified as node type 4. The same process is applied to all target nodes in the image.

In this study, the matrix using the extracted nodes' number, which is called “node matrix”, is obtained. Figure 5 illustrates the generation process of the node matrix. The node matrix expresses the structure of the table, and elements in the table can be easily extracted by using the matrix and the positions of these nodes.

3.5 Character Recognition and Generation of XML Documents

3.5.1 Cell Extraction Using Node Matrix

In this paper, elements of the table are called “cells”. Figure 6 shows the outline of the cell extraction method. The node located on the top-left of the matrix is set as the starting point of the extraction. The matrix is scanned from the start point left to right until the nodes with a downward element, i.e. nodes 1-6 in Figure 5, appear. In this figure, node 2 appears first as the node with a downward element. The node is regarded as the top-left point of the cell and the matrix is scanned from this point to the bottom again. When the nodes with a left element such as nodes 2, 3, 5, 6, 8 and 9 appear, the node is regarded as the bottom-right of the cell. The same process is repeated until the start point appears again, and the cell in the table is

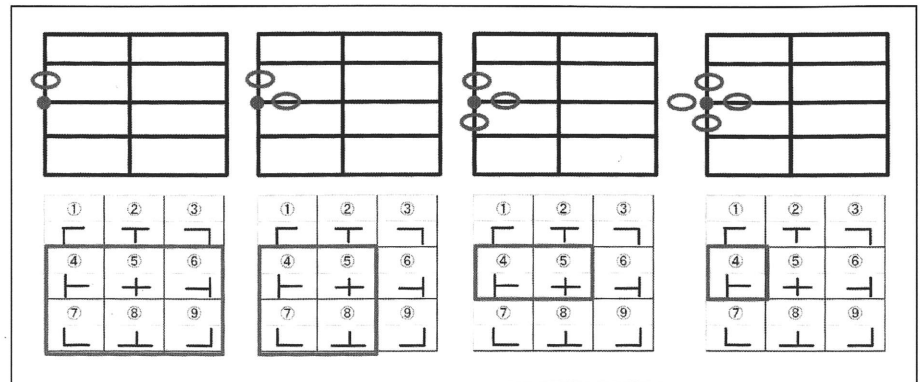


Fig. 4 Node classification method

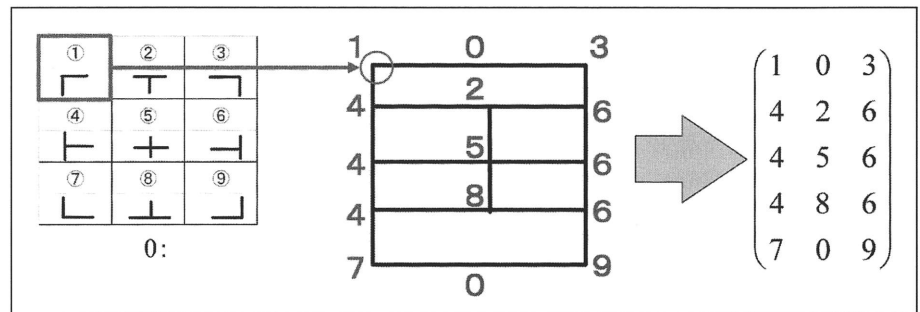


Fig. 5 Expression of the table using node matrix

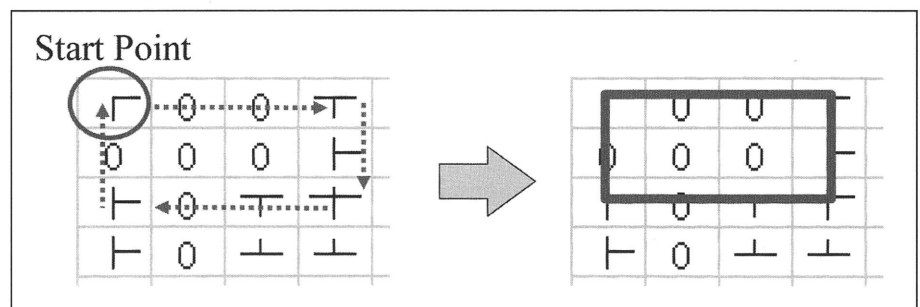


Fig. 6 Extraction of cell using node matrix

extracted. In this paper, the same process is applied to all nodes in the matrix and all cells in the table are extracted.

3.5.2 String Extraction Using the Master Image

String regions in all the cells have to be extracted to recognize characters and generate an XML document. The proposed method extracts the regions using the master

information described in 3.1. In this paper, the cell image extracted from a blank table is called the “master cell image”, and the one from a table inscribed by users is called “in-scribed cell image”, respectively. The string regions inscribed by users in each cell are extracted by a subtraction process between the master cell image and the inscribed cell image. The methodology, however, cannot extract these regions appropriately if the position of the master cell image does not

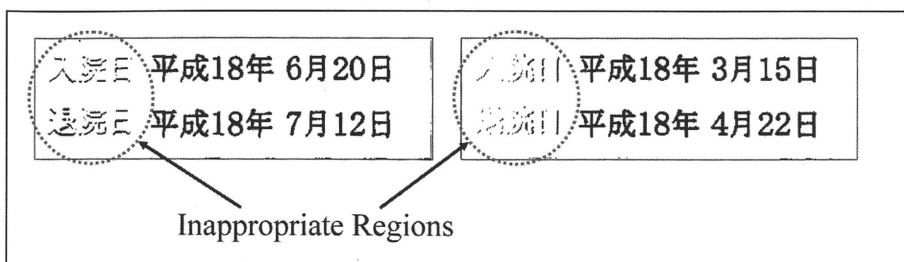


Fig. 7 Extraction results of string regions

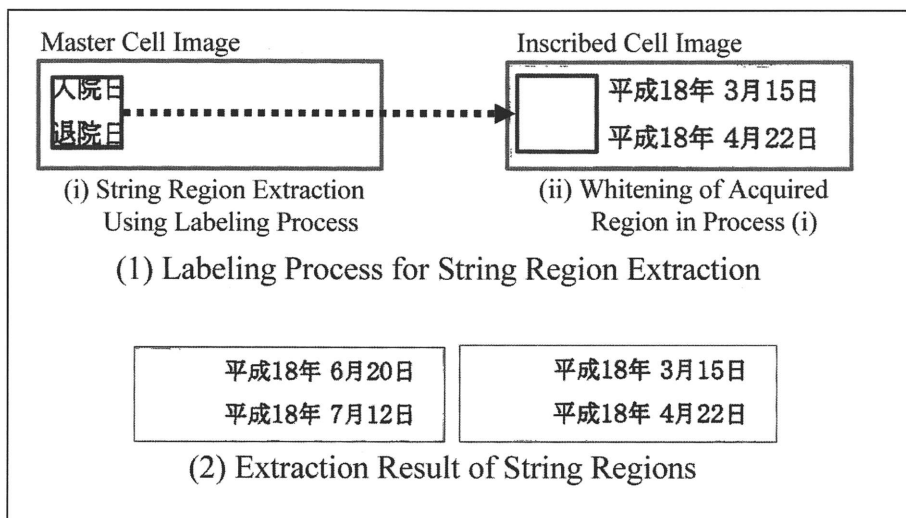


Fig. 8 Extraction of string regions (improved proposed method)

match that of the inscribed cell image. Therefore, the proposed method calculates the ratio of difference between these images first, and then the position for the subtraction process is determined to solve the above problem. In this process, the ratio of difference is calculated by the sum of the number of pixels with different values in each pixel. After this, the string regions in the cell image are extracted by the subtraction process (Fig. 7 shows the outcome of the string extraction process). The result indicates that the inappropriate regions not inscribed by users are also extracted as well as the string regions inscribed. It is considered that these results are caused by slight differences of tilt or input conditions between the master cell image and the inscribed cell images. It is, however, difficult to eliminate these differences completely.

Therefore, the proposed method was changed to solve the above problem and im-

prove extraction accuracy. Specifically, the labeling process that is shown in Figure 8 (1) was added. As the first step of the procedures, the labeling process is applied to the master cell image and the black pixels belonging to the large connected components changed to white. After this, the same subtraction process is done again. Figure 8 (2) shows an example result of the improved proposed method. It is evident that characters in the master cell image are erased completely and strings inscribed by users are appropriately extracted compared to the result in Figure 7. The extraction accuracy of the improved proposed method depends on the accuracy of the labeling method. In the case of the printed documents employed in this paper, variations of character size and distance between characters are not significant, thus, it is considered that the accuracy of the improved proposed method is high. In preliminary experiments, false extraction of

string regions such as in Figure 7 was not detected and as a result the above problem was solved.

3.5.3 Character Recognition and Generation of XML Documents

Characters in extracted strings have to be recognized and converted to text data by an Optical Character Reader (OCR) engine. Previously, we employed "Smart OCR", a free OCR software developed by a Japanese Company, as the OCR engine of the proposed system [11]. It is, however, extremely difficult to combine it with our system due to limitations of the software. As the result, the commercial OCR library, developed by "Panasonic Solution Technology, Inc." is used in this paper [12]. This library has the function to recognize table structures, but this was not employed here. As the proposed method can recognize table structures and extract the string regions appropriately, it has enough accuracy to utilize it in the practice. Moreover, the authors are planning to develop a system that can recognize various discharge summary documents with more complex tables. As the result, we developed a system that does not depend on an external library.

The table structure and characters acquired by the previous processes are used to generate an XML document. In this paper, an XSL, defining the table structure of the document is generated from the acquired node matrix first. The table structure is defined by "<table>" tags in the XSL. In the next step of the process, an XML document is generated using the XSL and converted text data corresponding the contents of each cell.

4. The Developed System

This paper developed the prototype system using the proposed method. In the system development, Microsoft Visual C# .NET was mainly used. Figure 9 shows a screenshot of the developed system. The developed system employed the "wizard form" to improve usability of the system (the top-left child window in Fig. 9). The wizard in the

system consists of three components such as "Image Input", "System Configuration" and "Generated XML Viewer". The image input component supports various input methods, not only image files but also scanner devices. For example, users can input document images from TWAIN devices or image files such as Bitmap, JPEG and so on. In the system configuration component, users can set system parameters easily if they enter them in accordance with the guidance messages. When the processes are finished, the generated XML viewer window appears and the generated XML document is shown in the window. Additionally, the generated XML documents have high compatibility with relational database systems, and users can search keywords from these XML documents easily. In CIS, a lot of database systems are employed. It is considered that the XML documents generated by the proposed method are easily imported and will be utilized effectively in CIS.

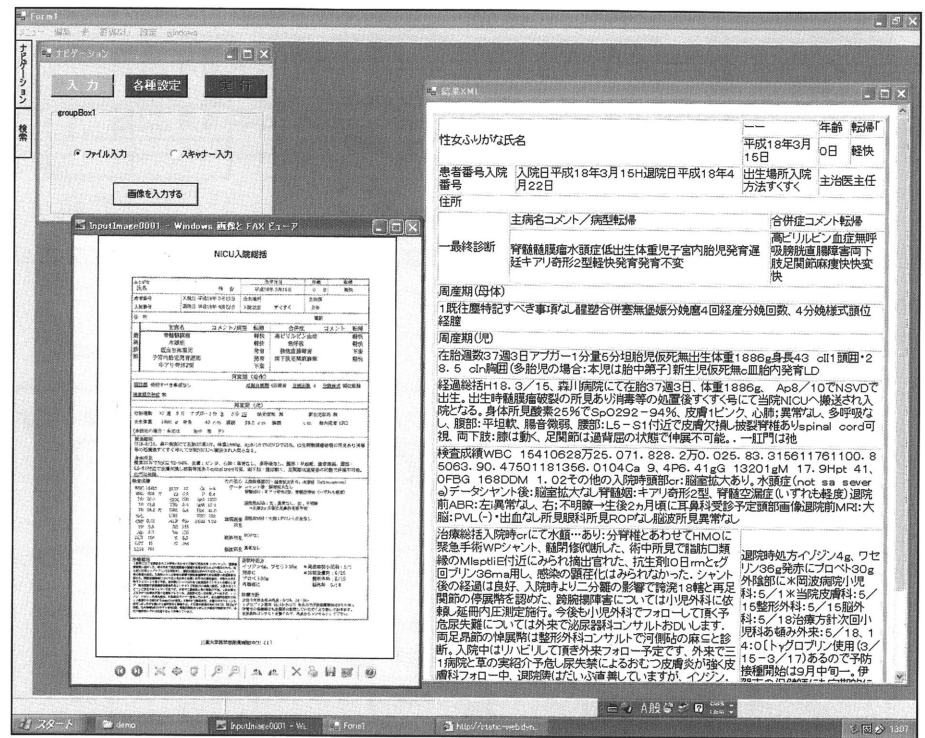


Fig. 9 Screenshot from the developed system

5. Experimental Results and Discussion

5.1 Evaluation of Proposed System

In this study, experiments for evaluation of the proposed method (and developed system) were done using actual discharge summary documents archived at Mie University Hospital. Discharge summary documents with typical tabular form such as Figure 5

were used in this experiment. Figure 10 shows an example of XML documents generated by the developed system. This document was created from the input image in Figure 3. The result indicates that the table structure of the input document was converted accurately. Moreover, string regions were appropriately extracted and characters could be converted to text data. In this experiment, no misrecognition of table structure was detected, however, some characters

in the tables were incorrectly recognized and converted to other characters. The following reasons contributed to the above results.

- 1) The resolution of the input images was low.
- 2) Redundant ruled lines adversely affected the OCR engine.

In addition to these reasons, these errors may be derived from the OCR engine itself.

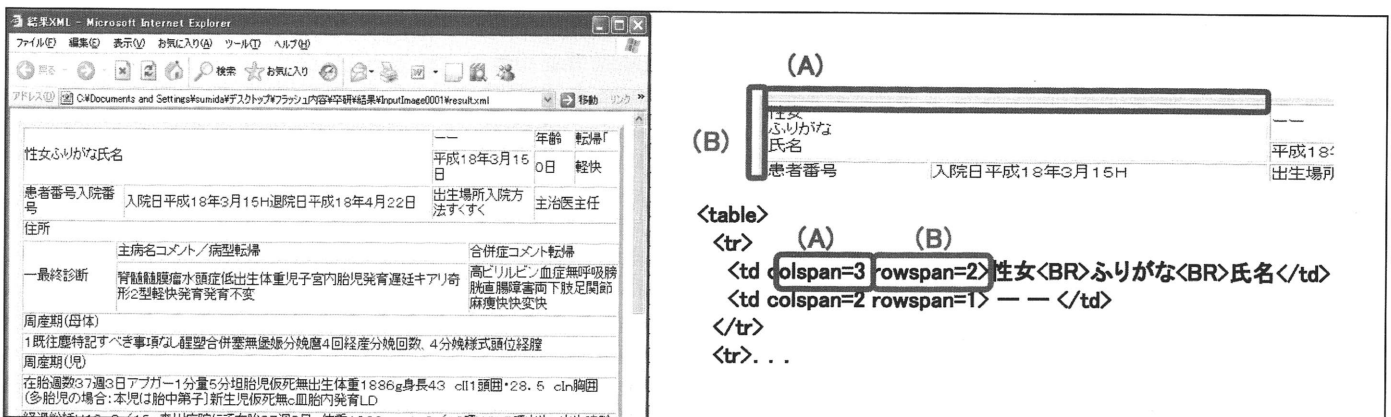


Fig. 10 Generated XML document and code. Right: source of XML documents

看護経過記録						
氏名		殿				
H 月日	時刻	問題 #	<S:主観的情報> <O:客観的情報>	<A:判断・評価> <P:計画>	<I:実施>	サイン
7/7				食思不振の原因の把握 嘔気時20%TZ20ml + プリンペラ ン1A静 腹痛時20%TZ20ml + プスコバン 1A静		
	14:00				ソルデム3AG500mg) 点 ビタメジン1V 生食100mg ベントリン2g) 点	

Dotted Line

Fig. 11 Example of an unsuccessful XML generation process (input image)

	SmartOCR	Panasonic OCR Engine	Difference of Recognition Rate	検査成績	その他の 入院時検査CT: 脳腫瘍あり、水頭症 (not so severe) データ シャント後: 脳腫瘍なし 骨髄MRI: キアリ奇形2部、骨髄空腔性 (いずれも軽微)
Document 1	88.4	93.7	5.4	WBC 15410 BUN 10 Ca 9.4	
Document 2	87.8	95.8	8.0	RBC 628 万 Cr 0.8 P 6.4	
Document 3	87.0	97.9	10.9	Hb 25.0 CPK 506 IgG 1330	
Document 4	87.9	96.1	8.2	Ht 71.8 TBI 3.9 IgM 17.9	通院前ABR: 左: 異常なし、右: 不明瞭 →生後2ヵ月頃に耳鼻科受診予定
Document 5	93.7	97.2	3.6	Plt 28.2 万 DBI 0.4 Hpt 41.0	
Document 6	91.7	95.2	3.5	N/L UBI FBG 168	頭部画像 通院前MRI: 大脳: PVL(-), 出血なし
Document 7	94.9	97.5	2.5	CRP 0.02 ALP 750 DDM 1.02	所見
Document 8	86.1	94.6	8.5	TP 5.8 BS 118	眼科所見 ROPなし
Document 9	91.0	97.6	6.7	Ab 3.3 Na 136	脳波所見 異常なし
Document 10	91.8	97.2	5.4	GOT 156 K 6.0	
Average of Recognition Rate	90.0	96.3	6.3	GPT 11 Cl 104	
				LDH 761	

Fig. 12 Recognition rate in each document

As the way of reducing these errors, a correcting method using medical dictionaries will work effectively.

In this experiment, it took about 20 seconds to generate an XML document using a personal computer with a Pentium D (2.66 GHz) CPU and 768 MB memory. At the time of the experiment this processing speed was not taken into consideration. Enhancement of devices and distributed computing techniques such as grid computing technology will be effective in reducing the calculation time of the processes. Currently, we are on the opinion that it is more important to generate XML documents with high accuracy than to be able to do it within a short time. Thus, the improvement of recognition accuracy and usability such as the correcting function for misrecognized characters, and enhancement of keyword search function will have to be achieved in the next step of this study.

This paper also applied the proposed method to other kinds of tabular form documents and discussed its efficacy. Six types of tabular form documents were employed. The proposed method could properly recognize the table structure and reconstruct XSL documents but not generate XML documents. Figure 11 shows an example of an input image from which the proposed method could not generate an XML document. The input image in Figure 11 has dotted lines as auxiliary ruled lines. The proposed method recognized each dot in the image as a character, and these were converted to "Dot Mark". As a result, this document could not be appropriately converted to an XML document. In the hospitals, there are a lot of discharge summary documents with dotted lines or broken lines such as Figure 11, thus, the processing methods for these documents will have to be added. In the proposed method, the information of a control vocabulary (the expected word sets

on each cell in the table) was not used. Generally, the use of control vocabulary significantly improves the accuracy of character recognition. This vocabulary, however, depends on which division (e.g. internal medicine, surgery etc.) issue those forms. The authors think the features of the word sets used in each division can be acquired easily once the employed words are accumulated. Thus, a terminology depending on each division will be automatically constructed using this information, and it is expected that the system managing the control vocabulary database automatically can be developed in the next step of this study.

5.2 Comparison with Commercial OCR Products

This paper discussed the accuracy of the OCR engine. In this section, "Smart OCR" was employed as a free OCR engine and "Panasonic System Solution's OCR Library" as a commercial OCR engine. The result of the experiments shows that the accuracy of the commercial OCR engine was about 6% higher than that of the free OCR engine shown in Figure 12. In the case of documents #3, #4, and #8, a significant difference of accuracy appeared between these OCR engines. One of the typical cell images is shown in the right of the figure. This cell contains the result of requested test procedures and the doctors' comments for each patient. In this case, blood test results, findings of CT and MRI tests etc. are written in the cell. The general dictionary of the OCR engine does not have specialized medical terms. It is considered that this caused the results shown as Figure 12.

This paper also discussed the accuracy of the table structure recognition function. Figure 13 shows an experimental result of table structure recognition with the commercial OCR engine used in this study. In this figure, the rectangles with dotted line mean the extracted string regions. The result shows that most of the string regions cover more than one cell in the table. Generally, a table consists of some cells and each cell corresponds to a meaning such as "name", "address", "patient number" and so on. In the case where the proposed method is em-

played, the result shown in Figure 13 is not acquired because the table structure in the document is recognized first and string regions in the table are extracted using the table structure information. By the above results, it is considered that the proposed method works effectively when tabular form documents are converted to digital data such as XML.

5.3 Related Works

Studies for document image analysis have been reported [13-21]. As related researches to ruled line extraction, the detection methods using the Hough transform technique are reported by [13-17]. Particularly in [13] and [14], complex line shapes can be extracted using a pattern-matching method and Hough transform. [15-17] propose and discuss the detection methods for character patterns, general curving lines, quadratic curving lines, circular patterns using the concept of [13] and [14]. These methods may have high extraction accuracy compared with the proposed method, but require a lot of calculation time because of these complex algorithms. In the practical situation, processing time is one of the most important factors to evaluate systems. Therefore, it is not realistic to employ these methods in the cases where a large number of documents is processed.

As for related methods for document layout (or structure) recognition, [18] reports the table structure recognition method based on the block segmentation method and [19] tries to extract the contents from printed document images using model checking. However, the method described in [18] depends on the output of commercial OCR systems. The proposed method identifies table types using a node matrix. This matrix can be acquired easily by using the extracted ruled lines – these are obtained by simple image processing techniques. Thus, the proposed method does not depend on an external library. In the case of [19], only the logical structures in the documents are detected using image analysis but the system is not developed to reuse the information.

氏名		性 女	生年月日	平成18年 3月15日	年齢	0 日	転院	軽快
患者番号	入院日	平成18年 3月15日	出生場所		主治医			
入院番号	退院日	平成18年 4月22日	入院方法	すくすく	主任			
住 所		電 話						
最終診断	主病名	コメント/病型	転院	合併症	コメント	転院		
	脊髄髄膜瘤		軽快	高ビリルビン血症		軽快		
	水頭症		軽快	無呼吸		軽快		
	低出生体重児		発育	膀胱直腸障害		不変		
	子宮内胎児発育遅延		発育	両下肢足関節麻痺		軽快		
	キアリ奇形2型		不変					
周産期 (母体)								
既往歴 特記すべき事項なし			妊娠分娩歴 4回経産 分娩回数 4 分娩様式 帝王切開					
周産期合併症 無								

Fig. 13 Result of table structure recognition (commercial product)

In another field, the techniques for analyzing cultural heritage documents are reported by [20]. This literature describes the application of document analysis techniques for the purpose of preserving and archiving cultural heritage documents. [21] reports a prototypical document image analysis system for journals. Most of these studies mainly describe the methodology and processing of typical business letters. According to the authors' survey, only a few articles propose the document image recognition method for medical documents, such as patient discharge summaries to search analogous cases.

6. Conclusions and Future Works

This paper discussed a document image recognition, keyword extraction and automatic XML generation system to search analogous cases from paper-based documents. The proposed method could recognize the table structures from the input document images and XML documents were automatically generated. In the paper, the authors developed the prototype system using the proposed method. Evaluation experiments using actual documents were done to discuss the effectiveness of the proposed method. This paper also compares the proposed system with commercial products to

discuss the effectiveness of the proposed system. The results of the experiments showed the proposed system could convert tabular form discharge summaries to XML documents with high accuracy.

There are a lot of different tabular form documents in the hospitals. In some hospitals, tabular form documents with ornamental ruled lines or without ruled lines are also used. As future works in this study, an algorithm to recognize the table structures and extract the string regions from these tables has to be developed. In addition, we will develop a system recognizing printed document images as well as handwritten documents in the near future.

References

1. Friedman HH (ed). Problem-Oriented Medical Diagnosis. 5th edition. 1991.
2. Seto K, Kamiyama T, Matsuo H. An Object-Modeling Method for Hospital Information Systems. The 9th World Congress on Medical Informatics 1998.
3. Lowe HJ, Antipov I, Hersh W, Smith CA, Mailhot M. Automated Semantic Indexing of Imaging Reports to Support Retrieval of Medical Images in the Multimedia Electronic Medical Record. *Methods Inf Med* 1999; 38 (4): 303-307.
4. Kawanaka H, Otani Y, Yoshikawa Y, Yamamoto K, Shinogi T, Tsuruoka S. Tendency Discovery from Incident Reports with Free Format Using Self Organizing Map. *Japan Journal of Medical Informatics* 2005; 25 (2): 87-96.
5. Otani Y, Kawanaka H, Yoshikawa Y, Yamamoto K, Shinogi T, Tsuruoka S. Keyword Extraction from Incident Reports and Keyword Map Generation

- Method Using Self Organizing Map. Proc of 2005 IEEE International Conference on Systems, Man and Cybernetics 2005. pp 1024-1029
6. Otsu N. A Threshold Selection Method from Gray-Level Histograms. IEEE Trans. Systems, Man, and Cybernetics 1979; SMC-9 (No. 1). pp 62-66.
 7. Otsu N. Discriminant and Least Squares Threshold Selection. Proc. of 4IJCPR 1978. pp 592-596.
 8. Akiyama T, Masuda I. A Segmentation Method for Document Images without the Knowledge of Document Formats. The IEICE Transactions on Information and Systems 1983; Vol. J66-D: pp 111-118
 9. Tanaka T, Tsuruoka S. Table Form Document Understanding Using Node Classification Method and HTML Document Generation. Proc of third IAPR Workshop on Document Analysis Systems 1998. pp 157-158.
 10. Ito Y, Ohno M, Tsuruoka S, Shinogi T. Document Structure Understanding on Subjects Registration Table. Proc of the fourth International Symposium on Advanced Intelligent System 2003. pp 571-574.
 11. Tsuruoka S, Hirano C, Yoshikawa T, Shinogi T. Image-based Structure Analysis for a Table of Contents and Conversion to XML Documents. Proc of Document Layout Interpretation and its Application 2001. pp 59-62.
 12. Smart Reading. Official Web Site of "Smart OCR Lite Edition". <http://www.smartread.biz/>.
 13. Panasonic Solution Technologies Co., Ltd. Color OCR Library. Color OCR Library: "Yomitori Kakumei" SDK. <http://panasonic.co.jp/pss/pstc/products/colorocrlib/index.html>.
 14. Casasent D, Krishnapuram R. Curved Object Location by Hough Transformations and Inversions. PR 1987; 20 (2): 181-188.
 15. Krishnapuram R, Casasent D. Hough Space Transformation for Discrimination and Distortion Estimation. CVGIP 1987; 38: 299-316.
 16. Pao D, Li M. F, Jayakumar R. Detecting Parametric Curves Using the Straight Line Hough Transform. Proc of ICPR-B 1990. pp 620-625.
 17. Fujimoto K, Iwata Y, Nakata S. Parameter Extraction of Second Degree Curve from Hough Plane. The IEICE Transactions on Information and Systems 1991; J74-D2 (9): 1184-1191.
 18. Yan J, Agui T, Nagao T. A Complex Transform for Extracting Circular Arcs and Strait Line Segments in Engineering Drawings. The IEICE Transactions on Information and Systems 1992; J75-D2 (8): 1338-1345.
 19. Kieninger TG. Table structure recognition based on robust block segmentation. Proc Document Recognition V, SPIE 1998; 3305: 22-32.
 20. Aiello M. Document Image Analysis via Model Checking. AI*IA Notizie 2002; XV (1): 4-48
 21. Ogier J, Tombre K, Madonne. Document Image Analysis Techniques for Cultural Heritage Documents. Proc. of International Conference on Digital Cultural Heritage 2006; Madone Project, <http://l3iexp.univ-lr.fr/madonne/>
 22. Nagy G, Seth S, Viswanathan M. A Prototypical Document Image Analysis System for Technical Journals. IEEE Computer 1992; 25 (7): 10-24.

Correspondence to:

Hiroharu Kawanaka
 Graduate School of Engineering
 Mie University
 1577 Kurima-Machiya, Tsu
 Mie 514-8507
 Japan
 E-mail: kawanaka@elec.mie-u.ac.jp