A Proposal of Sheet Type Recognition Method and its Evaluation for Medical/Clinical Document Archiving Systems

Shunta Nakamura, Hiroharu Kawanaka
Hiroki Hayashi, Haruhiko Takase, Shinji Tsuruoka
Mie University
1577 Kurima-Machiya, Tsu, Mie 514-8507, Japan
{shunta@ip., kawanaka@}elec.mie-u.ac.jp

Koji Yamamoto
Suzuka University of Medical Science
1001-1 Kishioka, Suzuka, Mie 510-0293, JAPAN
yama-k@suzu-u.ac.jp

***Abstract*—This paper proposes the sheet type recognition method for tabular form documents in hospitals. Evaluation experiments using actual medical documents were conducted. The experimental results showed that the proposed method could extract the table structures in the documents and recognize their sheet types with high accuracy. Authors also implemented the proposed recognition method into the developed system for practical use. This paper shows the detail of the proposed recognition method, experimental results and the developed system. We also discuss the effectiveness of the proposed method and future works of this study.**

Keywords- Medical/Clinical Document Archving Systems, Sheet Type Recognition, Node Information

I. INTRODUCTION

Recently, many paper-based documents used in hospitals have been computerized because of diffusion of Hospital Information Systems (HIS) [1, 2]. On the other hand, many paper-based documents before computerization are still archived. These documents have the details of treatment approaches and past cases, thus they are valuable for medical/clinical studies. However, these documents are not still used effectively because it takes a lot of time and cost to convert them into electronic data.

Previously, we have studied document image recognition methods for medical documents and its applications [3, 4]. The developed systems can recognize table structures of the documents and convert them into electronic data such as XML. However, we have to give the sheet type of the input documents by hand before scanning. Actually there are about 1,000 kinds of documents in the hospitals, and a few thousands of documents are generated every day. Therefore, the automatic sheet type recognition method is required for medical document scanning systems.

In this paper, we discuss a sheet type recognition method for scanning of medical documents. This paper proposes the sheet type recognition method using crossover points of ruled lines called "Nodes". Evaluation experiments using actual medical documents are conducted to discuss the effectiveness of the proposed method.

II. MATERIALS

In this study, we use printed discharge summaries and medical interview sheets archived in Mie University Hospital. These documents are scanned by an optical image scanner with full color of a resolution of 300dpi. In this paper, we use nine kinds of medical documents, and the number of employed documents is 334.

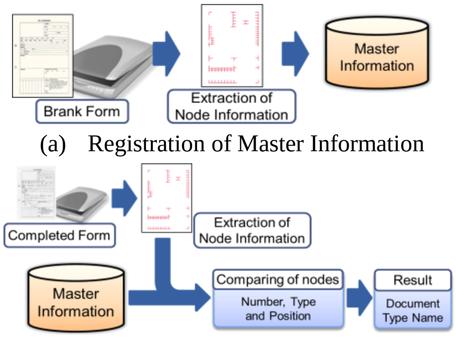
III. RECOGNITION METHODS

A. *Outline of our Method*

Figure 1 illustrates the outline of our system. The system consists of two phases. One is (a) Generating a Master Database, and the other is (b) Sheet Type Recognition. In the first phase, some features for recognition are extracted from blank sheets, and the reference database is generated. As the next phase, documents inscribed by medical staffs are scanned, and then features are extracted using the same method in (a). The features are compared with that in the reference database, and then the sheet type of the input document is recognized. After this, inscribed characters in the document image are recognized and then XML file for resemble case search is generated. The scanned document image and generated XML file are stored into the database.

B. *Extraction of Features for Recognition*

Generally, a tabular form document has at least one table, and its form and location heavily depend on sheet type. In other words, features of the table in the document would be the key information for sheet type recognition. Thus, we extract crossover points of ruled lines, this we call "Nodes" in this paper, and their positions from the document (Fig. 2). To extract the nodes from the document, n_l contiguous



(a) Registration of Master Information

(b) Sheet Type Recognition

Fig.1 Outline of our System

regions of black pixels are first extracted from the image. Next, the crossover points of them are extracted considering their shapes, and finally the node types and their positions are determined. In this study, the value of n_l was determined experimentally.

C. *Determination of Sheet Type*

Figure 3 shows the rough image of the sheet type recognition technique. First, Region of Interest (ROI) whose size is $n_R \times n_R$ pixel is set to each node. Then, we search whether the same node type exists in the same ROI of a sheet in the master database, and count the success cases. After this, the degree of coincidence to the sheet in the master database is calculated as the ratio of the number of success to the total number of nodes of that sheet in the master database. Finally, we determine the sheet type of the image as that which has the highest degree of coincidence among the master database.

IV. EXPERIMENTA L RESULTS AND DISCUSSIONS

In this paper, we implement the proposed method into the developed system to evaluate the effectiveness of the proposed method (Fig. 4). To make our system robust for the misalignment of medical records to the scanning machine, we use the ROI, but misalignment error exceeding this range may occur in practical use. Thus, we take the following three techniques into the recognition method, and examine its accuracy and the processing time.

1. Using Absolute Coordinate System Based on the Position of the Top-left Pixel (Method 1)
2. Using Relative Coordinate System Based on the Pixel Position of each Node (Method 2)
3. Using Relative Coordinate System Based on the Pixel Positions of the Top-left and Bottom-right Nodes (Method 3)

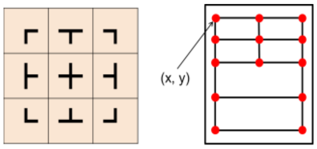


Fig.2 Node Types and Their Positions

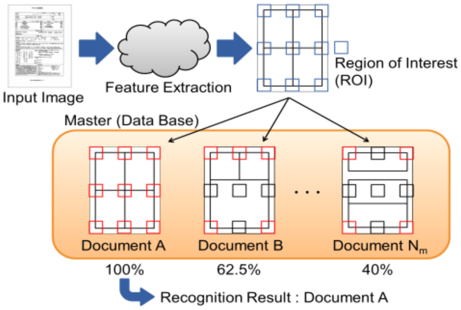


Fig.3 Rough Image of Recognition Method

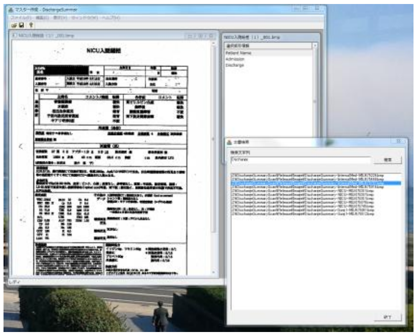


Fig. 4 Screenshot of Developed System

Table.1 Results of Sheet Type Recognition

Method	Recognition Rate [%]	Processing Time [ms/document]
Method 1	96.3	17
Method 2	100	16961
Method 3	100	17

Table 1 shows experimental results of sheet type recognition. The table shows that all documents were recognized correctly in cases of relative coordinate systems, *i.e.* method 2 and 3. But, if we use absolute coordinate system, *i.e.* method 1, the obtained recognition rate was 96.3%. However, in the case of method 2, it took a lot of calculation time because the method employed large number of nodes. Actually a few thousands of paper-based medical documents are generated in the hospital every day thus the method 2 might not be a practical solution. From these results, we can conclude the followings.

1. The methods using relative coordinate systems are effective for determining the sheet type.
2. From the viewpoint of processing time, we should use as few nodes as possible for sheet type recognition.

V. CONCLUSIONS

We proposed the sheet type recognition method using nodes in the document. Some experiments using actual medical documents were conducted. Experimental results showed that the recognition method using nodes worked well and high recognition accuracy was obtained. As future works of this study, we have to discuss the stability of a various document types.

REFERENCES

[1] S. Kuwata, “Development of Medical Record Scanning System Conformable to e-Document Legislation in Tottori University Hospital”, The 29th Joint Conference on Medical Infomatics, pp. 40-41, 2009

[2] Y. Matsumura, N. Kurabayashi, T. Iwasaki, *et.al*, “A Scheme for Assuring Lifelong Readability in Computer Based Medical Records”, MEDINFO 2010, C.Safran *et al*. (*Eds*.) , IOS Press, pp. 91 - 95, 2010

[3] H. Kawanaka, T. Sumida, K. Yamamoto, T. Shinogi, S. Tsuruoka, “Document Recognition and XML Generation of Tabular Form Discharge Summaries for Analogous Case Search System”, Methods of Information in Medicine (Schattauer), pp. 700-708, 2007.

[4] H. Kawanaka, K. Yamamoto, H. Takase and S. Tsuruoka, “Document Image Processing for Hospital Information Systems”, Information System/Book 3, C. Kalloniatis (*Ed*., ISBN 979-953-307-373-5), (In Print)