2003 International Symposium on Advanced Intelligent Systems
September 25-28, 2003, Jeju, Korea

Document Structure Understanding on Subjects Registration Table

Yuichi Ito, Masanaga Ohno, Shinji Tsuruoka, Tomohiro Yoshikawa and Shinogi Tsuyoshi

Department of Electrical and Electronic Engineering, Faculty of Engineering,
Mie University, 1515 Kamihama, Tsu-city, Mie 514-8507,JAPAN
email : ito2@ip.elec.mie-u.ac.jp

Abstract - This research is aimed to automate the generating process of the database from paper based table forms like this work. The registration table has so complicate table structures, and, in this research we used the registration tables as an example of general table structure understanding. We propose a table structure understanding system for some table types, and it has some steps. The first step is that the document images on paper are read from the image scanner. The second step is that a document image segments into some tables. In the third step, the character strings is extracted using image processing technology and the property of the character strings is determined. And the structured database is generated automatically.

The proposed system consists of two systems. "Master document generation system" is used for the table form definition, and it doesn't include the handwritten characters. "Structure analysis system for completed table" is used for the written form, and it analyzes the table form filled in the handwritten character.

We implemented the system using MS Visual C++ on Windows, and it can get the correct extraction rate 98% among 51 registration tables written by the different students.

I. Introduction

Table documents is popular for usual life. The document structure understanding for the table by computer is required for some cases. For example, database generation needs the human key-in labor and a lot of time. Table documents such as a check and an invoice, include various items such as the number of a company name, a brand name, and goods.

II. Outline of Document Structure Understanding

This system consists of a "Master document generation system" and a "Structural analysis system for completed table"(Fig.1). By this system, the information what is written where of a registration table is first defined as master information, and it creates from an incomplete

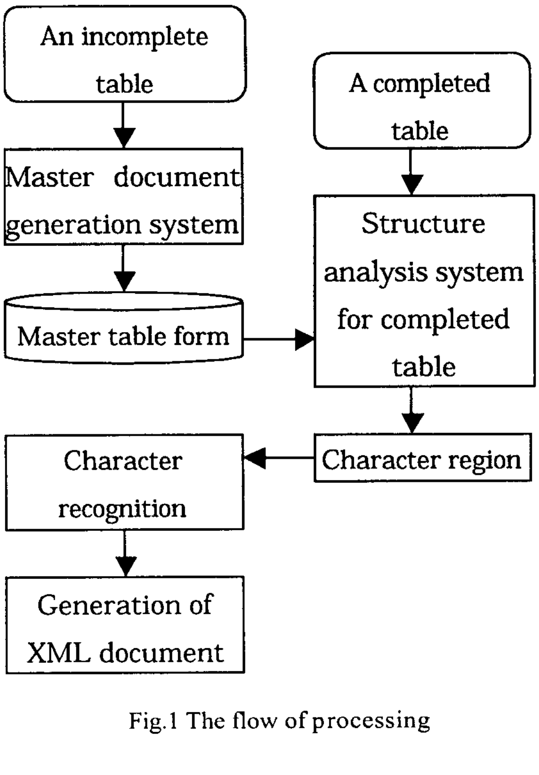
An incomplete table
A completed table
Master document generation system
Structure analysis system for completed table
Master table form
Character region
Character recognition
Generation of XML document

Fig.1 The flow of processing

registration table in a Master document generation system. The master information created at once is treated as known information in future processings. In a Structural analysis system for completed table, matching is performed from master infomation and the actually acquired information, and required information is started.

Then, it is made an XML document in order to make it the form which is easy to utilize as a database. In this research, it inquired using the registration table of the common education of Mie University, and the registration table of the faculty of technology.

III. Master Document Generation System

In a Master document generation system, an incomplete registration table on real table is scan first (Fig.2). Next, it is binarized the region of as a preprocessing. And table is divided into some isolated tables using the size of connected components (Fig.3).

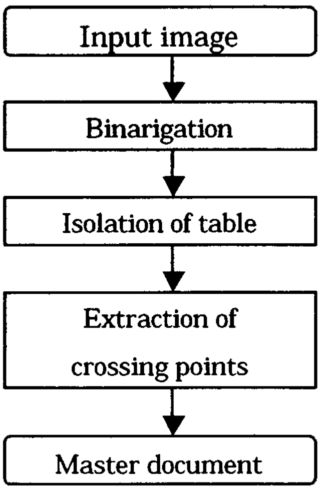

Fig.2 Master document generation system

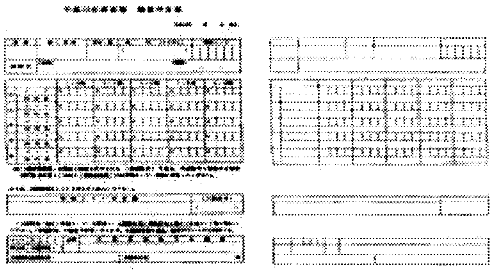

Fig.3 An incomplete table(left)
Ruled line extraction image(right)

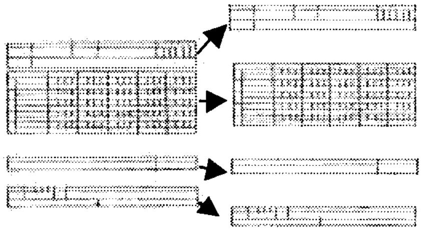

Fig.4 Domain division result

At this time, the label of table assigned by the top-left position of each table(Fig5).

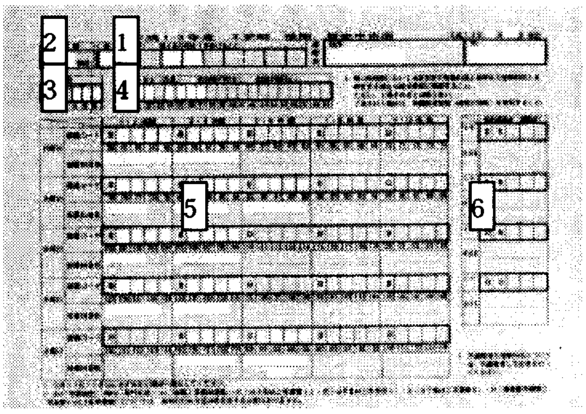

Fig.5 The label of table

For each table, a horizontal and perpendicular projection histogram is obtained for the extraction of crossing point of ruled line and the positions the crossing point x, y-coordinates are stored for each table (Fig.6)[1].

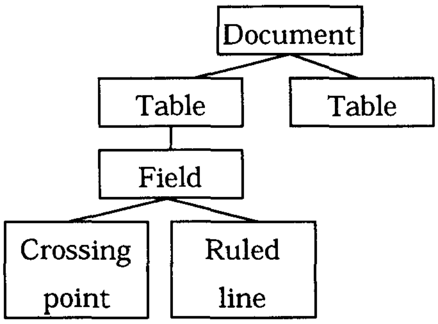

Fig.6 Hierarchy of registration form

Master information is created by specifying manually the positions of the crossing point, the ruled line, and the property of field surrounded by four ruled line (Fig.7). Master information has the hierarchy of input document. A document is the set of table, and a table is the set of field specified by the crossing points ruled lines, and property.

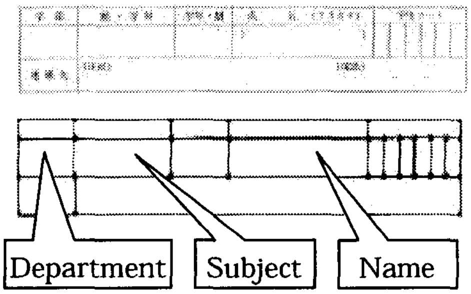

Fig.7 Input image(top) and master information(bottom)

IV. STRUCTUAL ANALYSIS SYSTEM FOR COMPLETED TABLE

Structural analysis system include the binarization, the isolation, of table the extraction of crossing points which are the some process as "Master document generation system" (Fig.8). Our system can recognize two type of registration form, which are that of the faculty of Engineering (Fig.9) and the common education of Mie University presently.

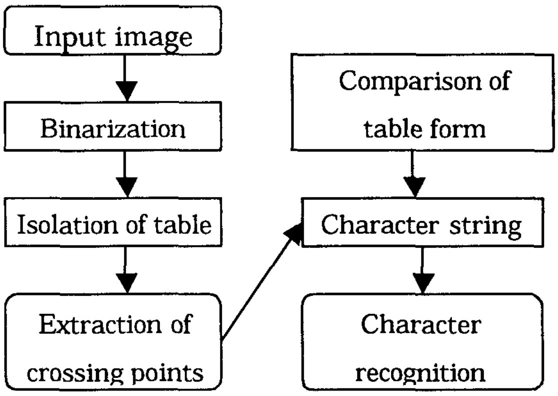

Fig.8 Structural analysis system for completed table

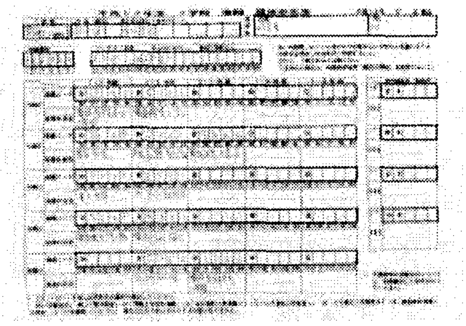

Fig.9 Completed registration table

A. Segmentation of handwritten character region

This system classifies an input table image into asset of table ruled line land some character region by the size of connected component. The character region is surround by asset of the table ruled lined. The table ruled line is extracted by the size of connection ingredient is large as a table. Then, since the character region is recognized in the inside of each table, the character region is transferred to character recognition process.

The extraction of a handwritten character region is performed to a table. The position of character region is modified using the property of field in Master document. The table assigned to the registered table by the upper left position of a table in master information.

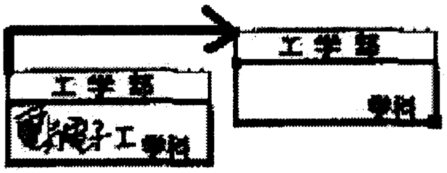

Fig.10 Modification of the point position

After the modification of the position point ,the field in the table is determined, and the position of the coordinates of the field is assigned. The field is segmented.

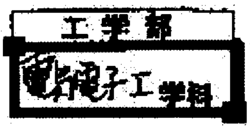

Fig.11 The comparison of the position in master information

Fig.12 Segmented field

TABLE I

Accuracy of segmentation for character region in a table

	Field	Correct
Common education	4	4
Faculty of technology	47	46

B. Ruled line removal

In the segmented field (Fig. 12), since a ruled line is surely contained, a ruled line is removed.

First, a ruled line is removed using the position of the ruled line in master information. Next, a surrounding ruled line is removed using the position of ruled line in master table form.

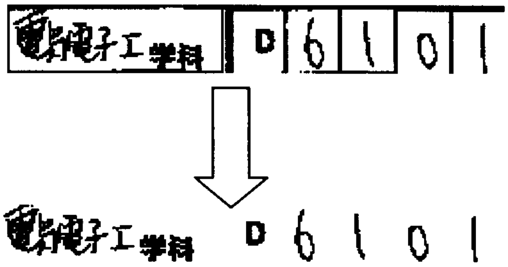

Fig.13 A character region after removing a ruled line

Character recognition is performed for the character region in Fig. 13.

C. Generation of XML document

The recognized character strings are changed into an XML document based on a table form (DTD in XML) in master document[2].

V. Conclusion

We propose a document structure understanding system for subjects registration table. The following problems are remained.

(1) Character recognition to the extracted character region.

(2) Examination for other forms

References

[1] Toru Tanaka, Shinji Tsuruoka, "Table Form Document Understanding using Node Classification method and HTML document generation", proc. of 3rd IAPR Workshop on Document Analysis Systems (DAS198), pp, 157-158,1998

[2] Shinji Tsuruoka, Chihiro Hirano, Tomohiro Yoshikawa, Tsuyoshi Shinogi, "Image-based structure analysis for a table of contents and conversion to XML documents", proc. Of document lay out interpretation and its application (DLIA2001), pp.59-62,2001