TOWARD A PRACTICAL DOCUMENT UNDERSTANDING OF TABLE-FORM DOCUMENTS: ITS FRAMEWORK AND KNOWLEDGE REPRESENTATION

Toyohide Watanabe
Qin Luo
Noboru Sugie

Reprinted from PROCEEDINGS OF THE SECOND INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION, Tsukuba Science City, Japan, October 20-22, 1993

Toward a Practical Document Understanding of Table-form Documents: Its Framework and Knowledge Representation

Toyohide WATANABE, Qin LUO and Noboru SUGIE

Department of Information Engineering,
School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-01, Japan

***** Abstract *****

The document structure recognition and document understanding are one of interesting subjects today from a viewpoint of practical application. The objective is to extract the meaningful data from document images and also classify them as the predefined data items automatically. In comparison with the traditional image-processing-based approaches, the knowledge-based approaches are currently being investigated as more applicable and flexible methods. In this paper, we propose a totally integrated paradigm for understanding table-form documents.

1. INTRODUCTION

The document structure recognition and document understanding are one of interesting subjects today from a viewpoint of practical application[1]. The objective is to extract the meaningful data from document images and also classify them as the predefined data items automatically. In comparison with the traditional image-processing-based approaches, the knowledge-based approaches, which make use of various kinds of knowledge in order to interpret structural/constructive features of documents, are currently being investigated as more flexible and applicable methods. The approach that interprets document images with the knowledge about document-specific applications, composition rules, layout structures and so on is applicable to various kinds of documents though the traditional approaches, based on the image processing techniques, were effective to only very limited document structure. Namely, the knowledge-based approach supports a global processing paradigm in comparison with the local processing methods of image-processing-based approaches.

The methods proposed in the knowledge-based approaches are roughly divided into two classes in point of knowledge specification means: rule-oriented methods[2-4] and frame-oriented methods[5-8]. These methods are selective for document types, which characterize the applications/forms of documents[1]. The rule-oriented method is applicable to some kinds of documents: e.g. newspapers[9], in which the composite items are allocated by logical structures. While, the frame-oriented method is successful for documents, in which the composite items are controlled strictly by geometric layout structures: the examples are name cards, business letters, official documents, library cataloging cards and so on. However, these approaches had not always attached to research subjects, which were attended inherently to document structure recognition and document understanding, though they made the fundamental framework clear. It is important to investigate various problems, that have never been assessed sufficiently in practice.

In this paper, we propose a totally integrated paradigm of document understanding with respect to various kinds of document knowledge, and also address the recognition/understanding method of table-form documents from an architectural point of view.

2. PARADIGM OF DOCUMENT UNDERSTANDING

Documents may be classified into several document types, depending on the mutual relationship between logical and geometric structures[1]. Each knowledge-based document understanding system must be designed, depending on the application of document, and also the currently proposed methods were applicable to application-specific documents. The document class recognition plays an important role in distinguishing documents with different layout structures. We illustrate such a framework of document understanding conceptually in Fig.1.

This framework is organized as an enhanced version for our three-layer recognition paradigm: layout recognition, item recognition and character recognition[10, 11]. The layout recognition process identifies the geometric and spatial relationships among item blocks, which are sets of meaningfully allocated item areas, in 2-dimensional space. The item recognition process distinguishes individual items from the item blocks in 1-dimensional space. Finally, the character recognition process extracts each character code from character patterns, which

0-8186-4960-7/93 $3.00

compose individual items, in 0-dimensional space. The new framework in Fig.1 was refined progressively with respect to the flexibility, applicability and functionality for recognition of various classes of documents.

The document class recognition process classifies various kinds of documents into individual document classes, which can be interpretatively identified by the same knowledge about layout structures. For example, consider two different table-form documents in Fig.2. The geometric structures are different, but the logical structures are the same. The document class is defined as a set of documents whose layout structures can be uniquely identified by the same knowledge. Therefore, two table-form documents in Fig.2 should be determined as the same document class if the applications are the same. Until today, the researches about document recognition/understanding have never attached to this subject. This is partly because this recognition subject is very difficult, and partly because the current technical subjects focus on the understanding only application-specific documents. However, the issue on the document class recognition is important to manage many different documents selectively because documents may be available in many application-specific forms.

3. KNOWLEDGE OF TABLE-FORM DOCUMENT

Table-form documents are geometrically designed on the basis of layout structures whose individual item areas are always surrounded with vertical and horizontal line segments[12, 13]. These item areas are not independent with each other. So, the hierarchical structure and repeating structure, as parts of tables, are compositively adaptable to several meaningfully interdependent item areas. Thus, table-form documents are well specified through the layout structures in comparison with other kinds of documents because each item area is predefined rigidly. In analyzing the document forms it is better to extract vertical and horizontal line segments firstly and then identify each item by interpreting the relationships among line segments.

Many researches about document understanding and document structure recognition make use of the physical information about document forms, as knowledge. The knowledge, which is composed of the physical information such as locations, sizes, lengths and so on, is not always useful for reduced/expanded documents, irregularily transformed documents, and documents which are inconsistent to the geometrically predefined structures. It is very important that the knowledge should be

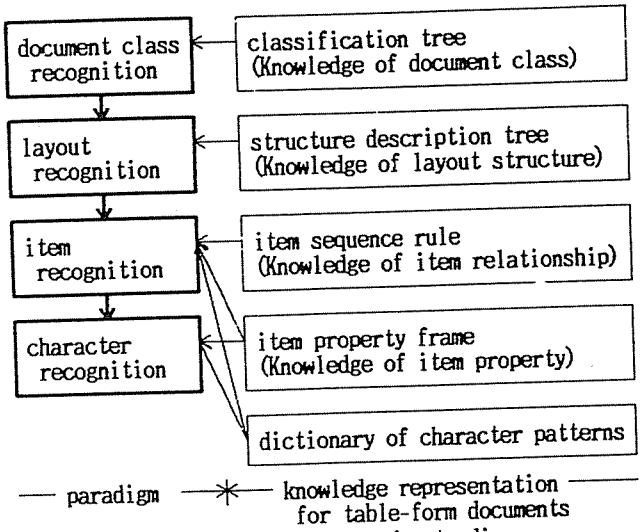


Fig.1 Framework of document understanding

支出官殿	請求者	所属	官職	氏名		命令権者
				番号	印	
概算額	精算額	追給額	返納額			

月日	出発地	到着地	車賃		鉄道賃		
			定額	実費	路程	運賃	その他
合計							
請求額							
備考							

(a) Tab.1

支出官殿	請求者	所属	官職	氏名		命令権者
				番号	印	
概算額	精算額	追給額	返納額			

月日	出発地	到着地	車賃		鉄道賃		
			定額	実費	路程	運賃	その他
合計							
請求額							
備考							

(b) Tab.2

Fig.2 Examples of table-form documents

specified by only the logical information in order to be applicable to various documents, which are consistent to the topological relationships among item areas[1]. For example, two table-form documents shown in Fig.2 are different because their coordinate values do not match well from a viewpoint of geometric structures. However, two table-form documents are the same, concerning the logical structures.

In our framework, four kinds of knowledge is

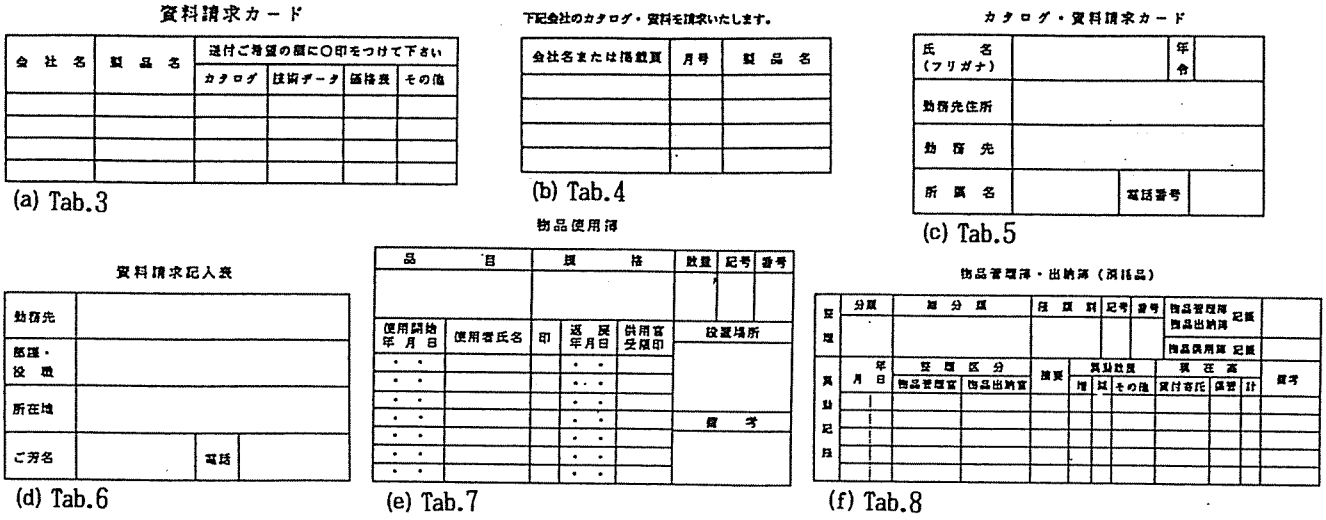


Fig.3 Examples of other table-form documents

useful under the knowledge representation means. Fig.1 also shows the corresponding knowledge.

(1) Knowledge of document class

The knowledge can be represented with a multi-ways tree in our approach. We call this tree as the classification tree. The node corresponds to each document class, while the edge represents the parent-child relationship among document classes. Namely, child document classes are derived stepwisely from the parent document class. For example, in Fig.4 we illustrate the classification tree for several table-form documents, shown in Fig.2 and Fig.3. The marked nodes indicate document classes for 8 kinds of table-form documents in Fig.2 and Fig.3. Tab.3 and Tab.7 are derived from Tab.4, respectively.

This tree grows up when a table-form document, which does not correspond to the existing document classes, must be manipulated. The node for the new table-form document is generated so as to be attached to the most similar existing document class. Namely, nodes in our classification tree correspond to a collection of rectangularily partitioned blocks, when some blocks are furthermore separated by the longest vertical/horizontal line segments, which connect to the edges of blocks[14]. A block division process is illustrated in Fig.5. In Fig.5, the right side is furthermore partitioned in comparison with the left side. The left side is transformed into the upper node, while the right side is done into the lower in our classification tree. Of course, this division process generates various branches, according to the location of longest vertical/horizontal line segments.

(2) Knowledge of layout structure

The knowledge for table-form documents is

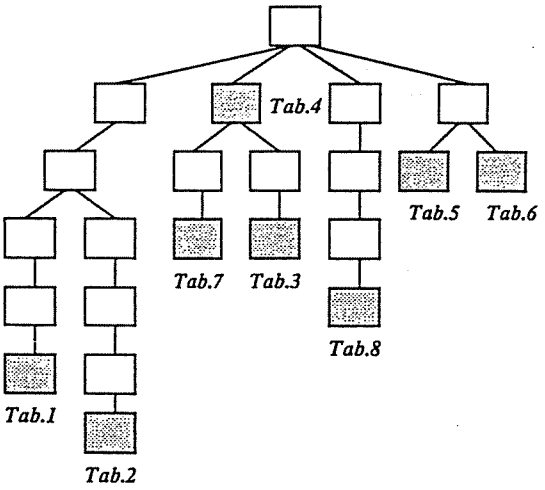


Fig.4 Classification tree

represented with the structure description tree(as a binary tree). Our structure description tree indicates the adjacent relationships among item blocks and the connective relationships among item areas, but does not represent the coordinate values of items. Namely, this tree deals with only logical information about document structure. The structure description tree is divided furthermore into the global structure tree and local structure trees.

The global structure tree represents the global feature for the whole layout structure of table-form documents: repeating structure, hierarchical structure and adjacent structure among item blocks. The nodes point out individual item blocks, and the edges correspond to the neighboring relationships among item blocks. While, the local structure tree represents the detail layout structure for individual item blocks, which are specified by the global structure tree.

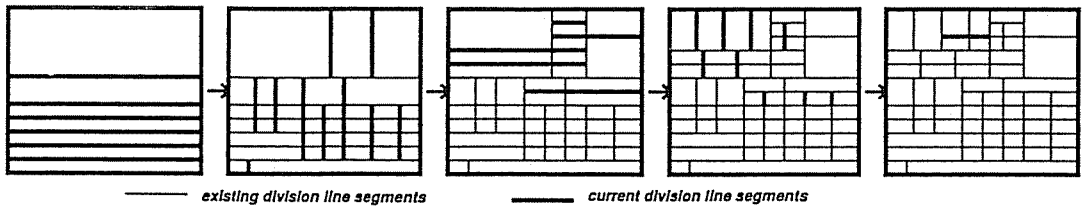


Fig.5 Block division process

Namely, the local structure trees are attached to each node in the global structure tree, and represent the connective relationships among item areas. The nodes indicate individual item areas, and the edges correspond to the connective relationships among item areas. Fig.6 shows the relationship between the global structure tree and local structure trees, conceptually. For example, consider a table-form document in Fig.2(a). Fig.7(a) is the global structure tree, and Fig.7(b) is the local structure trees. This structure description tree in Fig.7 is also applicable to Fig.2(b) though it was meaningfully generated from Fig.2(a), because this tree represents the logical structure.

(3) Knowledge of item relationship

The knowledge is represented by the item sequence rule. The item sequence rule must be applicable to the separation of individual items, because each item area does not always contain single item but may be composed of several compound items. For example, we consider a data "Furo-cho, Chikusa-ku, Nagoya 464-01"(This represents our university address). This data is composed of several item data, which are separated basically by the symbol ",", and they have the predefined left-to-right sequence.

The general syntax form is as follows:

<rule>::=<rule-name>:<item-seq>
<item-seq>::= {<item-seq>| } <item>
<item>::=<item-name> {<property>}
<property>::="optional" | "mandatory"

Here, "optional" denotes that the attended term may be abbreviated, and "mandatory" does that the term must be always assigned. In table-form documents, this item sequence rule is very simple, because each item area includes only single item in many cases.

(4) Knowledge of item property

This knowledge is represented by the record structure, called as the item property frame. The item property frame accommodates character sets, length of character strings, occurrence, keywords as separators, candidate data values, decision condition and so on in its composite slots, respectively. For example, consider "464-01" as the zip-code. The item property frame is shown in Fig.8. The item sequence rule is cooperated effectively with this item property frame in the item recognition process.

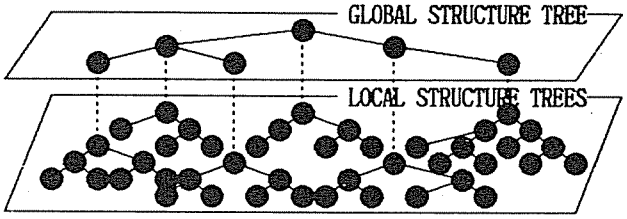
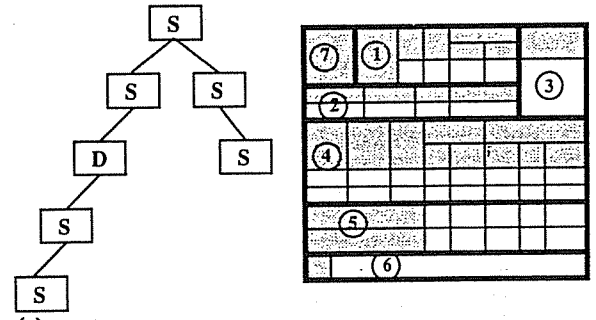
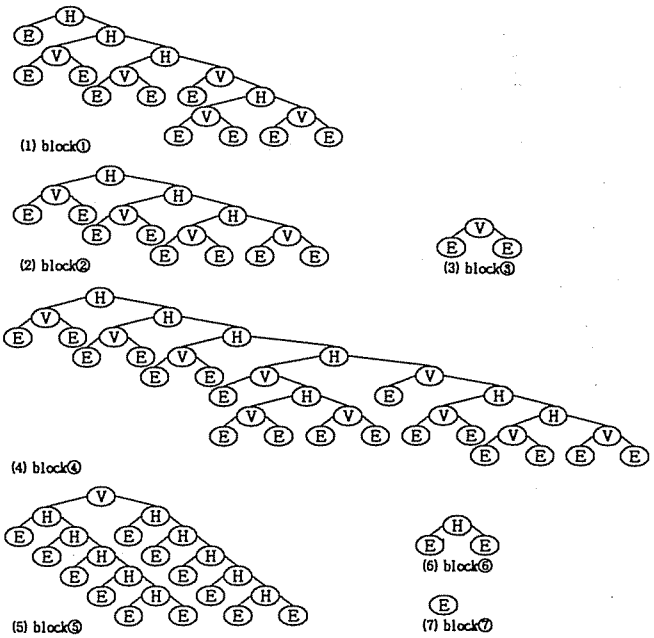


Fig.6 Structure description tree



(a) global structure tree



(b) local structure trees

Fig.7 An example of structure description tree

4. RELATIONSHIPS AMONG VARIOUS KNOWLEDGE

4-level knowledge constructs a hierarchical structure, corresponding to the interdependent

relationship among 4-layer recognition processes. For example, the document class knowledge is an upper-level knowledge to apply an appropriate layout structure knowledge to table-form document images. The layout structure knowledge is an upper-level knowledge to distinguish individual items by knowledge of item relationship. Similarly, the knowledge of item relationship is on the upper-level for knowledge of item property, and the knowledge of item property is on the upper-level for dictionary of character patterns.

(1) Relationship between classification tree and structure description trees

The classification tree distinguishes document classes, and manages various kinds of layout structure knowledge. The nodes in the classification tree are organized systematically on the basis of the physical characteristics, as shown in Fig.5. Therefore, two table-form documents Tab.1 and Tab.2 in Fig.2, which are designed under the same logical structure, are not always classified into the same node, like Tab.1 and Tab.2 in Fig.4. Of course, our structure description tree in Fig.7 is applicable to Tab.1 and Tab.2, because the structure description tree represents the logical structure of table-form documents on the basis of adjacent/connective relationships among item areas.

In order to manage the storage for knowledge representation effectively, such duplication must be avoided. Therefore, it is necessary to check up all structure description trees when a new node is added to the classification tree. This procedure works easily and rapidly because the classification tree is based only on the number and length of vertical and horizontal line segments. The first is done to the global structure tree; and the second is to the local structure trees. If two logical structures are different globally, the checking procedure is rejected in the global structure tree. Fig.9 shows the relationship between classification tree and structure description trees.

(2) Relationship between structure description tree and item sequence rules

The structure description tree distinguishes individual item areas, which may contain one or more item data. The item sequence rule is applicable to item sequence in such partitioned item areas. Of course, when the partitioned item areas contain only one item, the item sequence rule is not assigned. The item sequence rules are attached to the nodes in the local structure trees. Fig.10 shows the relationship between structure description tree and item sequence rules.

NAME	zip-code
SEPARATOR	","
LENGTH	(3,2) or (3,0) by "-"
CHARACTER SET	numeric
OCCURRENCES	single
CANDIDATES	------
LOCATION	after "city"
DICISION CONDITION	LENGTH & NUMERIC --

Fig.8 Item property frame "Japanese zip-code"

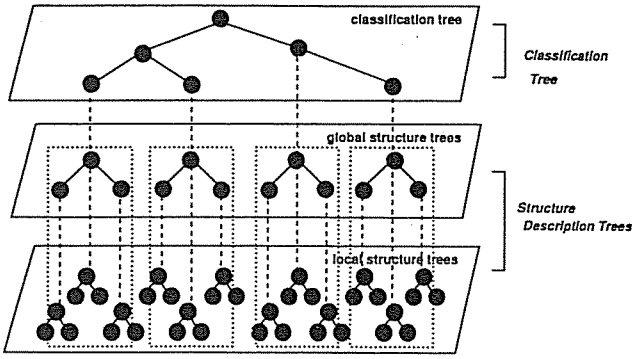


Fig.9 Relationship between classification tree and structure description tree

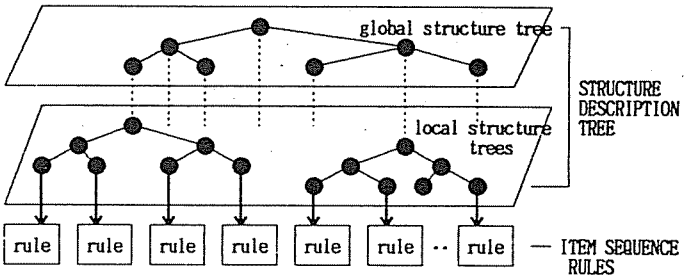


Fig.10 Relationship between structure description tree and item sequence rule

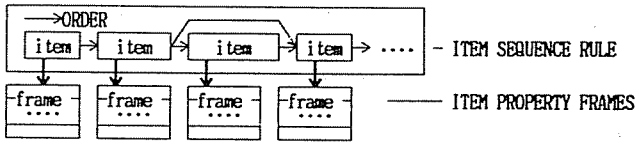


Fig.11 Relationship between item sequence rule and item property frame

(3) Relationship between item sequence rule and item property frames

The item sequence rule represents the constructive sequence of meaningful items, while the item property frame indicates the characteristic attributes of item forms. The item property frames are attached to items, which compose the item sequence rule, one by one, as shown in Fig.11.

(4) Relationship between item property frame and character pattern dictionary

The item property frame contains several

candidate character strings in the candidate slot, if possible, and specifies several constraints in the slots such as the length, character set and occurrence. These information can assist the selection method when character recognition process extracts an appropriate character pattern.

5. CONCLUSION

In this paper, we proposed a framework of 4-layer recognition processes for understanding documents and addressed the knowledge representation method, adaptable to the understanding of table-form documents. Although Nakano et al. looked upon the recognition of multi-kinds of table-form documents as an important subject from a practical point of view[5], they could not report any successful approach because their knowledge was based only on the physical coordinate data. In our approach, this recognition issue was solved, using both the classification tree based on the physical characteristics and the structure description tree based on the logical characteristics. At least, it is not so difficult to classify various kinds of documents into appropriate document classes since table-form documents are well designed on the basis of vertical and horizontal line segments. However, it is not easy in the case of the other documents because the geometric and spatial characteristics of documents are not well specified. It is necessary to investigate the application techniques for the other documents from a viewpoint of the knowledge representation.

Acknowledgements -- We are grateful to Prof. T. FUKUMURA of Chukyo University, and Prof. Y. INAGAKI, Prof.J.TORIWAKI of Nagoya University for their perspective remarks. We also wish to thank Ms.K.SUGINO and our research members for their eager cooperations and discussions.

References

1) T.WATANABE, Q.LUO & T.FUKUMURA: "A Framework of Layout Recognition of Document Understanding", Proc. of 1st DAIR, pp.77-95 (1992).
2) D.NIYOGI & S.SRIHARI: "A Rule-based System for Document Understanding", Proc.of AAAI-86, pp.789-793.
3) F.ESPOSITO, D.MALERBA, G.SEMERARO, E. ANNESE & G. SCAFARO: "An Experimental Page Layout Recognition System for Office Document Automatic Classification: An Integrated Approach for Inductive Generalization", Proc. of 10th ICPR, pp.557-562(1990).
4) J.L.FISHER, S.C.HINDS & D.P.D'AMATO:"A Rule-based System for Document Image Segmentation", Proc.of 10th ICPR, pp.567-572(1990).
5) Y.NAKANO, H.FUJISAWA, O.KUNUSAKI, K.OKADA & T. HANANO: "Understanding of Tabular Form Documents Cooperating with Character Recognition", EIC trans., Vol.J69-D, No.3, pp.400-409(1986) (in Japanese).
6) K.KISE, K.MOMOTA, M.YANAKA, J. SUGIYAMA, N. BABAGUCHI & Y. TEZUKA: "Model Based Understanding of Document Images", Proc. of MVA'90, pp.471-474.
7) A.DENGEL & G. BARTH: "High Level Document Analysis Guided by Geometric Aspects", Int'l J. of Pattern Recognition & Artificial Intelligence, Vol.2, No.4, pp.641-655(1988).
8) Q.LUO, T.WATANABE, Y.YOSHIDA & Y. INAGAKI: "Recognition of Document Structure on the Basis of Spatial and Geometric Relationships between Document Items", Proc. of MVA'90, pp.461-464.
9) Q.LUO, T.WATANABE & N. SUGIE: "A Structure Recognition Method for Japanese Newspapers", Proc.of 1st DAIR, pp.217-234(1992).
10) T.WATANABE, Q.LUO & N.SUGIE: "A Cooperative Document Understanding Method among Multiple Recognition Procedures", Proc. of 11th ICPR, pp.689-692(1992).
11) T.WATANABE, Q.LUO, Y.YOSHIDA & Y. INAGAKI: "A Stepwise Recognition Method of Library Cataloging Cards on the Basis of Various Kinds of Knowledge", Proc.of 10th IPCCC, pp.821-827 (1990).
12) T.WATANABE, H. NARUSE, Q. LUO & N. SUGIE: "Structure Analysis of Table-Form Documents on the Basis of the Recognition of Vertical and Horizontal Line Segments", Proc.of 1st ICDAR, pp.638-646(1991).
13) Q.LUO, T. WATANABE & N. SUGIE: "Structure Recognition of Table-form Documents on the Basis of the Automatic Acquisition of Layout Knowledge", Proc.of MVA'92, pp.79-82.
14) H.KOJIMA & T.AKIYAMA: "Table Recognition for Automated Document Entry System", SPIE, Vol.1384, pp.285-292(1990).