MVA'90 IAPR Workshop on Machine Vision Applications Nov. 28-30,1990, Tokyo

RECOGNITION OF DOCUMENT STRUCTURE ON THE BASIS OF SPATIAL AND GEOMETRIC RELATIONSHIPS BETWEEN DOCUMENT ITEMS

QIN LUO, TOYOHIDE WATANABE, YUUJI YOSHIDA AND YASUYOSHI INAGAKI

Department of Information Engineering,
Faculty of Engineering, Nagoya University,
Furo-cho, Chikusa-ku, Nagoya 464-01, JAPAN

ABSTRACT

This paper introduces a new method to extract and classify the meaningful information from documents automatically. The basic idea in our method is to utilize the spatial and geometric relationships between document items. Our approach is adaptable even if the layout structures are modified more or less, because the coordinate values of positions, sizes, lengthes and so on are not specified directly. Additionally, some experiments for typical documents such as library cataloging cards, name cards and letters are shown concretely.

1. INTRODUCTION

Many printed documents to be manipulated currently in office environments have their own peculiar layout structures. For instance, name cards, letters, application forms, account books and so on are the typical examples. If we use the information about layout structures of documents effectively, we can interpret the contents of documents easily. Some top-down methods using the layout structures have been suggested lately to identify the meaningful information from documents automatically[1, 2].

These methods can not only extract the characters, but also classify the extracted character-strings into individual items, in comparison with the methods based on the character recognition techniques. However, they have two disadvantages. First, it is more or less permissible to use the sizes, lengthes or coordinates of processing objects in the form description. Therefore,a change for the predefined form structures will disturb the recognition procedure from identifying each physically-specified object successfully. Second, it is troublesome and difficult to describe the detail information attended to the objects perfectly.

In this paper, we propose a more powerful method to recognize the layout structures of documents. Our method is not to specify physical coordinate values about the layout structures at all, but is based on the spatial and geometric relationships among neighboring rectangular segments. Namely, in our approach, documents are looked upon as hierarchical structures of rectangular segments which included meaningful items. The top-down recognition approach based on the spatial and geometric relationships among segments are adaptable to the stepwise interpretation of various types of documents.

In the following sections, our recognition method is first addressed. Then, the interpretive recognition ability is shown experimentally through library cataloging cards, name cards, letters and so on.

2. FRAMEWORK OF OUR METHOD

The layout recognition with a view to the issue of the document understanding is important. Although we observe many printed documents in our environments, they can be always classified into peculiar document types, in accordance with their own utilization objectives. Of course, some variations must be permissible even if individual documents are classified into the same type. The layout structures assign particular data items to appropriate locations over 2 dimensional paper areas. In general, a document can be divided into some partitions including one or more data items. The boundaries among these partitions are lines, spaces, indentations of data items, distinction of font sets, relative locations and so on. Additionally, these partitions may be repeated hierarchically. We show the layout structure conceptually in Fig.1. In order to recognize the layout structures of documents, it is very convenient to use the knowledges concerning the layout structures.

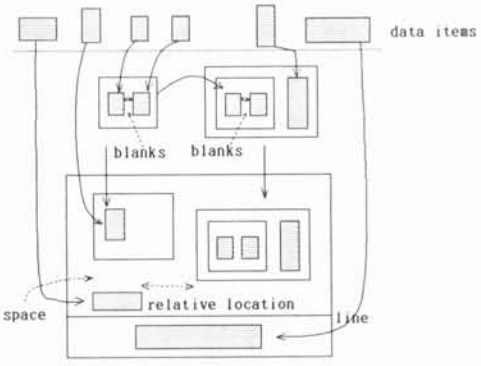


Fig.1 Conceptual structure of documents.

From such a spatial structure point of view, we concentrate our attention on the mutual connective relationships among neighboring rectangular segments including data items, but not on the positional relationships among data items. Namely, we do not extract (the areas of) data items directly, but distinguish the connections of partitions to include the neighboring data items: horizontal connection and vertical connection. In our approach, the

layout structures of documents are looked upon as a set of hierarchically patched blocks based on neighboring rectangular segments. Therefore, the positional coordinate values are not always specified explicitly. This view characterizes our approach as the recognition method based on the spatial and geometric aspects for document structures.

3. FORM DESCRIPTION FOR LAYOUT STRUCTURE

The knowledge related to the layout structure can be conceptually represented by a tree structure. In our tree structure, the nodes represent compositive segments and the branches point out the mutual connective relationships between the neighboring rectangular segments. The root node corresponds to a whole document and the terminal nodes distinguish basic elements as areas of data items. Non-terminal nodes define the horizontal/vertical neighboring relationships between their descendant nodes, in addition to keeping the information about their own segments.

For example, the representation of a layout structure about library cataloging cards is illustrated in Fig.2. As shown in Fig.2, individual nodes are characterized by symbols "h", "v", "or" and "T". They define nodes' types. Moreover, non-terminal nodes accompany additional strings such as "OP1", "OP2" and so on, which stand for segmentation operators. On the other hand, the numbers which indicate particular item names to be finally interpretive for the segments are assigned to terminal nodes.

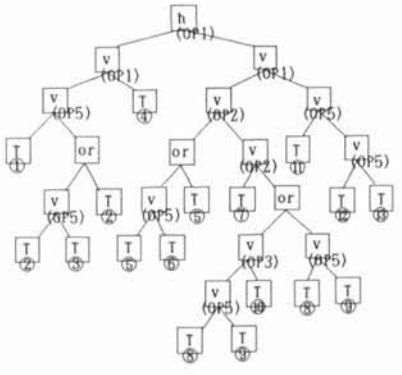
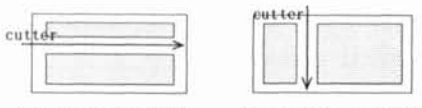


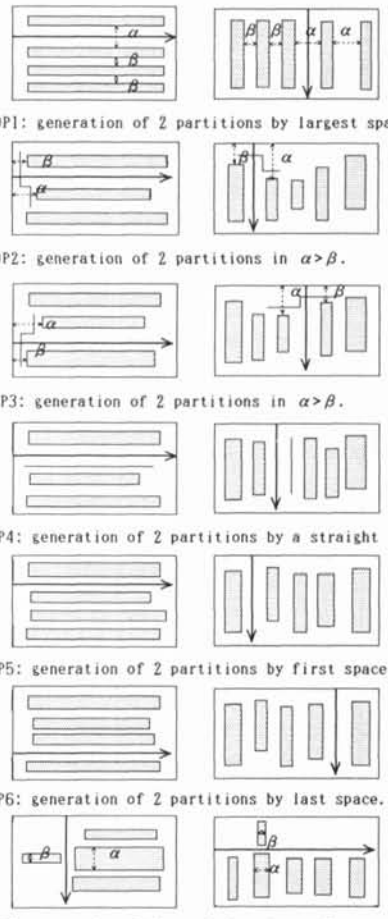
Fig.2 Tree representation for layout structure of cataloging cards.



(a) Vertical node(v). (b) Horizontal node(h).

Fig.3 Non-terminal nodes "h" and "v".

The node "h", which represents the horizontal mode, indicates that one processing object must be divided into the left-to-right partitions by the vertical cutter, while the node "v" divides one processing object into the top-to-down partitions by the horizontal cutter in the vertical mode as Fig.3. The segmentation operators point out cutting ways for the cutters. For example, some segmentation operators used in our examples are shown graphically in Fig.4. The node "or", which stands for the selection mode, indicates several candidate strategies for the next segmentation process. Namely, this informs the strategy procedure of the possibility of different division ways without performing the practical segmentation. In addition, there is another non-terminal node "rp", not appearing in Fig.2, it expresses a repeated layout structure in the vertical direction by uncertain times. The terminal node "T" indicates the end of the segmentation process.



(a) OP1: generation of 2 partitions by largest space($\alpha > \beta$).

(b) OP2: generation of 2 partitions in $\alpha > \beta$.

(c) OP3: generation of 2 partitions in $\alpha > \beta$.

(d) OP4: generation of 2 partitions by a straight line.

(e) OP5: generation of 2 partitions by first space.

(f) OP6: generation of 2 partitions by last space.

(g) OP7: generation of 2 partitions in $\alpha > \beta$.

Fig.4 Segmentation operator.

The individual non-terminal nodes hold 4 information cells: (MOD, SNUM, OP, CO). The field "MOD" represents the node type as explained above. The field "SNUM" counts the number of the following lower nodes. The field "OP" represents the practical segmentation operator based on the spatial/geometric relationships among neighboring segments. Finally, the field "CO" accommodates the coordinate values for the segment sizes, and also controls the next segmentation operator so as to be acceptable for the 2-dimensional area indicated by coordinate values.

4. LAYOUT RECOGNITION PROCEDURE

Our recognition procedure interprets the layout struc-

tures of documents owing to a tree structure as addressed in the previous section. The segmentation operations are repeatedly continued until all terminal nodes are reached. As a result, compositive segments including particular data items can be uniquely identified.

Each node in the tree structure is traversed in the prefix order. The segmentation operations for non-terminal nodes except "or" are practically performed according to the field "OP" after the connective and ordinal relationships among the lower nodes have been recognized by means of "MOD". The operator informs "CO"'s, in the lower nodes, of the coordinate values about newly partitioned areas. So, the next segmentation operators are only available within the area restricted by its own "CO". While, in the node "or", first one child node is selected. If the currently selected segmentation operation failed, the branch from this node will be cut and another node at the same level will be selected. Such the selection mechanism can be effective until a segmentation operation succeed. Additionally, because the node "or" does not take the role of "segmentation", the values in "CO" are only transfered to the lower nodes as they are. The information in "MOD" and "OP" is provided in advance as the layout knowledge. However, the information in "CO" is generated automatically while image-type documents are being processed; of course, the root node is exceptional. In our recognition procedure, we extract areas including document items on the basis of their mutual neighboring relationships instead of document items oneself. Therefore, the recognition procedure works successfully without difficulty although there are variations in the geometric configuration because the lengthes and numbers of data items are variable.

5. EXPERIMENTS

In this section, we show several experiments of documents, attended inherently typical layout structures.

5.1. LIBRARY CATALOGING CARDS

The cataloging cards accommodate various kinds of data items. These data items are arranged on the basis of the Nippon Cataloging Rules in case of Japanese cards. However, they are usually composed under different layout structures even in the same library. For example, the University Library of Nagoya University, there are 3 types of cataloging cards due to a difference in the form structure. We have already shown the layout structure by a tree of Fig.2. Fig.5 illustrates a recognition result about the cataloging card. As shown in Fig.5, the type of this card is recognized as "A" and 10 kinds of data items are extracted. Its practical tree structure is presented explicitly in Fig.6. We can observe in Fig.5 that some segments accommodate several data items yet. Our layout recognition method is adaptable to 2-dimensional processing objects. Therefore, we need another method to classify data items in 1-dimensional areas. The segments in Fig.5 are logically interpretable as 1-dimensional processing object. Another method concerning cataloging cards has been reported in our paper[3].

Fig.5 Recognition result of a cataloging card.

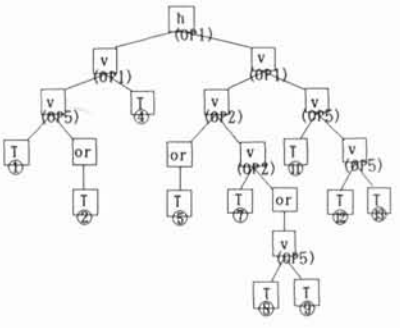


Fig.6 Tree representation for layout structure of the example in Fig.5.

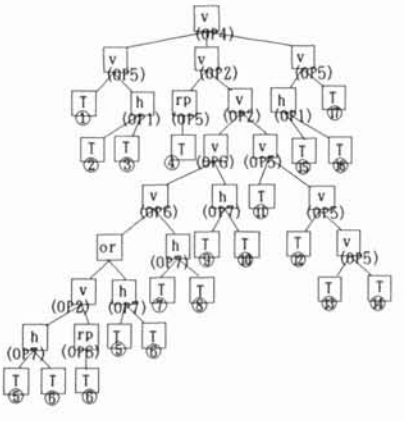


Fig.7 Tree representation of OKUZUKE.

5.2. OKUZUKE(book directory)

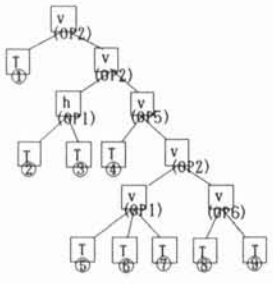
Next, we consider the OKUZUKE page, located to the last page in Japanese books. This page corresponds to "cataloging in publication data"-page in English books. These pages include the bibliography information about books directly. Each publishing company has its own fixed layout structure for the OKUZUKE page. We show one of them by a tree structure in Fig.7. A successful recognition result is shown in Fig.8.

5.3. LETTERS AND NAME CARDS

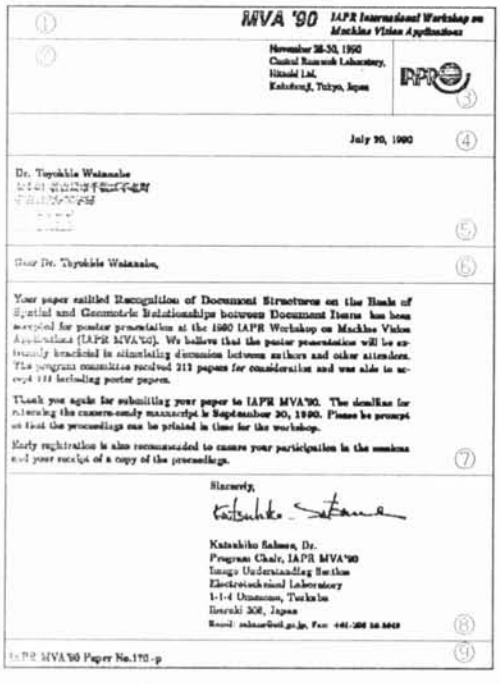
English letters and Japanese name cards are good examples whose contents can be analyzed easily with help of

the layout structures. We show tree structres and recognition results for each example, respectively in Fig.9 and Fig.10. In the former, the segmentation operation OP2 is utilized, while in the latter OP5 is successful.

Fig.8 Recognition result of OKUZUKE.

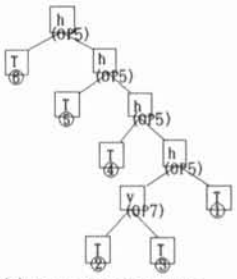


(a) Tree representation.

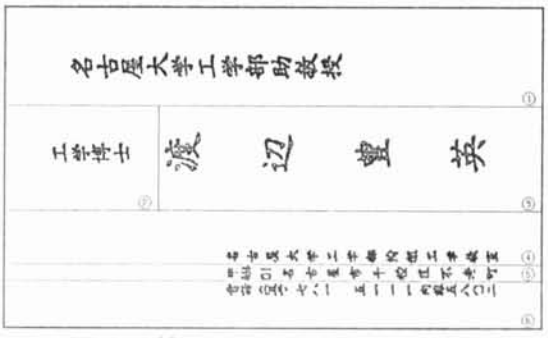


(b) Recognition result.

Fig.9 Layout structure of a letter.



(a) Tree representation.



(b) Recognition result.

Fig.10 Layout structure of a Japanese name card.

6. CONCLUSION

In this paper, we reported our experimental approach to identify the meaningful information from documents automatically based on the spatial and geometric relationships between document items. The experiments of this method for various types of documents have been shown concretely through library cataloging cards, name cards, letters and so on. Our method can apply to the documents whose layout structures can be known clearly. However, it is a problem for our method to identify documents without clear layout structures today. Therefore, we must investigate flexible methods to be appropriate to more complex documents with a view to the document understanding in the future.

ACKNOWLEDGEMENTS

The authors would like to thank Mr.Hiroyuki NARUSE and Mr.Yuuichi KANAI who carried out various experiments cooperatively for this study.

References

[1] J.HIGASHINO : "A Knowledge-based Segmentation Method for Document Understanding", IEEE, pp. 745-748 (1986).

[2] A.DENGEL & G.BARTH : "High Level Document Analysis Guided by Geometric Aspects", Int'l Journal of Pattern Recognition and Artificial Intelligence, Vol.2, No.4, pp. 641-655 (1988).

[3] T.WATANABE, Q.LUO et al. : "Automatic Extraction and Classification of Data Items from Library Cataloging Cards by a Knowledge-based Approach", Proc. of MIV' 89, pp. 67-71 (1989).