DOCUMENT RECOGNITION SYSTEM WITH LAYOUT STRUCTURE GENERATOR

Yoshitake TSUJI, Hiroyuki KAMI*, Masaaki MIZUNO,
Toshiyuki TANAKA**, Haruhiko TANAKA**,
Masao IWASHITA, Tsutomu TEMMA

C&C Information Technology Research Laboratory, NEC Corporation
* EDP System Engineering Division, NEC Corporation
** NEC Scientific Information System Development Ltd.

4-1-1 Miyazaki, Miyamae-ku, Kawasaki, Kanagawa 213 JAPAN

1. Abstract

A document input system, with character recognition technique, is used for converting printed matter, such as books and magazines, into code-format information. In order to improve this document input system's performance, an appropriate document structure analysis technique is indispensable[(1)-(4)]. When storing data from general printed documents into a database, it is necessary to represent the document structure. Therefore, a document layout structure generation method is especially important[(5)(6)]. For this purpose, the authors have developed a document image structure analysis method to generate a layout structure, as well as to detect such document elements as characters, pictures and figures. This method was developed on a personal computer. Its usability is described in this paper.

2. Document Structure Representation

Let us assume the following hypothesis. (1) Document elements (character, figure, picture, table and so on) and their sets of elements can be represented in a rectangular shape. (2) Each block has only three relations (inclusion/exclusion relation, vertical relation, horizontal relation). Then, the document layout structure is represented with these three relations in tree-shaped structures. For example, as shown in Fig.1, the character blocks B_6 and B_7 satisfy the vertical relation. The character line in blocks B_5 and B_4 satisfy the horizontal relation. An imaginary block, such as B_3, that is made up from several kinds of elements, is regarded as one block, and it satisfies the inclusion relation. This document structure representation involves features that can represent such positional relations as horizontal relations or vertical relations, compared to the conventional representation used previously[(5)].

Next, the basic philosophy of the document structure analysis method is explained. In printed documents, each document element is normally printed regularly, mostly in horizontal or vertical lines. With this features, the split detection method[(4)], which analyzes a document image, separating the whole image into smaller regions thereof, is applied effectively. For freely printed matter, the conventional block separation method, using projection for one direction, does not produce good results. For example, in order to segment character lines from a multi-columned document, projection features for both horizontal direction and vertical direction, as well as positional information are needed. Therefore, the authors introduced a feedback ability to reconstruct block relations from the features in the split detection method, and to calculate vertical and horizontal direction projection. The document image features are easily extracted from the projection profile. This method uses projection profile per pixel and obtains individual block features and block relations.

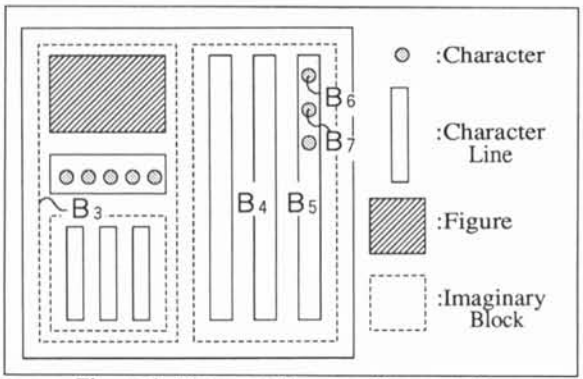

Figure 1. Document Structure Representation

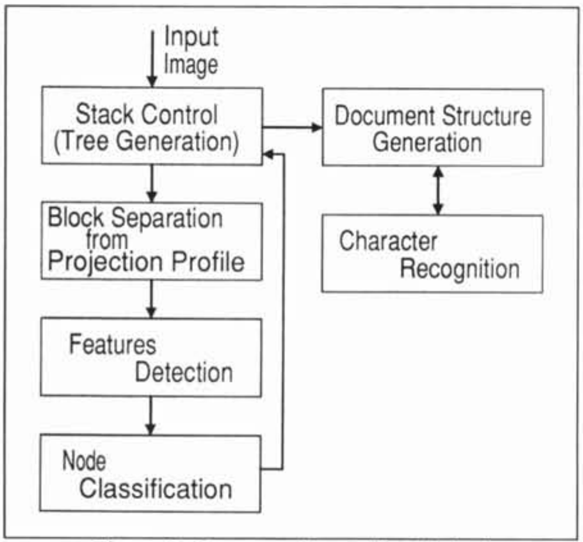
Input Image
Stack Control (Tree Generation)
Document Structure Generation
Block Separation from Projection Profile
Character Recognition
Features Detection
Node Classification

Figure 2. Document Recognition Method with Layout Structure Generator

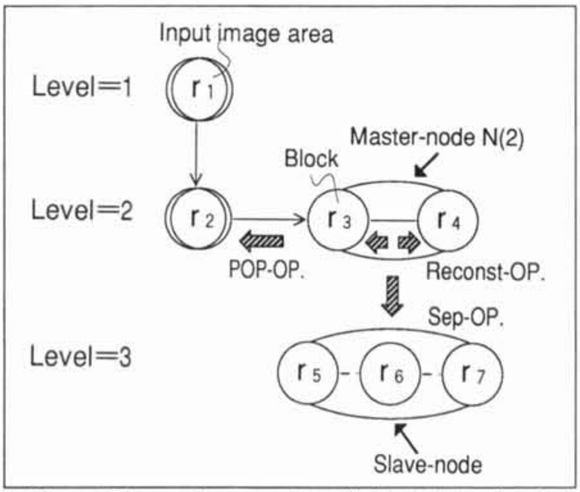

Figure 3. Image Segmentation with Depth First Search

3. Overview of the Document Recognition Method with Layout Structure Generator

Figure 2 shows an abstract diagram of the document structure analysis process. The document elements separation process is controlled using stack structure. Control action consists of three operations. (i) Split operation, that separates one block into several blocks. (ii) Feedback operation, that reconstruct obtained blocks and re-split them. (iii) Element detection operation, that controls the split operation and searches for the next block to be analyzed.

Document layout structure is generated as follows. First, with these three operations, a tree-shaped document structure representation, whose leaf nodes are document elements (such as characters, picture, figures and so on), is generated. Next, this tree-shaped document structure representation is evaluated, in order to search for document element blocks (e.g. character line block, text block). If required, the document structure representation is reconstructed on a bottom-up approach. Finally, the fixed document layout structure representation is generated. Character image is easily converted into character codes, using this layout structure representation.

4. Block Separation Using the Split Detection Method

Table I. Node Attribute

Attributes		Node Features
	S_0	Line candidate
S_C	S_1	Character/Line candidate (larger than S_2)
	S_2	Character candidate
	S_3	Character/Line candidate (smaller than S_2)
	S_4	Undefinded
	S_{C2}	Block set of S_c , that contains S_2
	S_{CC}	Block set of S_c , that doesn't contain S_2
	S_{C4}	Block set of S_c , that contains S_4
	S_{44}	Set of undefined blocks
		⋮

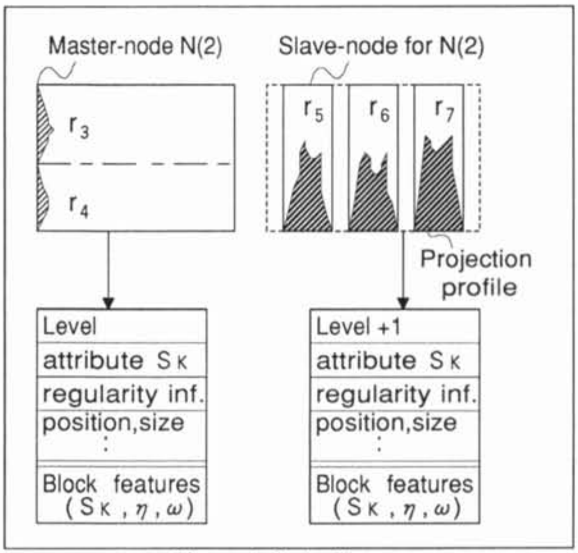

Figure 4. Node Features

Document image skew affects block separation results. To eliminate this effect, the authors introduced the distortion ratio as a normalized value for evaluating separation between blocks[4]. From this normalized value, obtained from the projection profile, the following features are obtained.

(a) Regularity between blocks

Select a region where several successive and regularly placed blocks have a high distortion ratio. From these features are obtained the values for mean block width, minimum width and maximum width of blocks, as well as mean distance between blocks.

(b) Character pitch estimation[7]

If a block is a character line candidate, estimate the character pitch from the block. This character pitch estimation method is based on minimum variance criterion. This criterion uses the mean distance between blocks and evaluates this mean distance with the least squares error method.

(c) Block attributes

Assign attributes to each block, using projection profile for one direction (horizontal or vertical direction). Attributes are, roughly, sorted into three categories, as shown in Table I. For example, a line candidate block is selected according to aspect ratio and pixel density evaluation result.

Figure 3 shows the block separating process with split detection. In Fig.3, a circle represents a block, and a set of circles, connected with an arc, represent a node. The arrow represents the positional relation. In Fig.3, a state is shown, after separating the master-block into three blocks r_i (i=5,6,7). The feature values are obtained from projection profile in the vertical direction. Figure 4 shows the features for master-node and slave-nodes in Fig.3. These nodes have the following several elements ; separation level, attribute, regularity , position, size, and other features, including that between blocks.

For each master-node N(L), one of three separation operations takes place by the node classification process.

(i) Split operation (Sep-OP.)

Separate slave-nodes and generate each node on level L+1. In this operation the master-node is regarded as an imaginary node.

(ii) Feedback operation (Reconst-OP.)

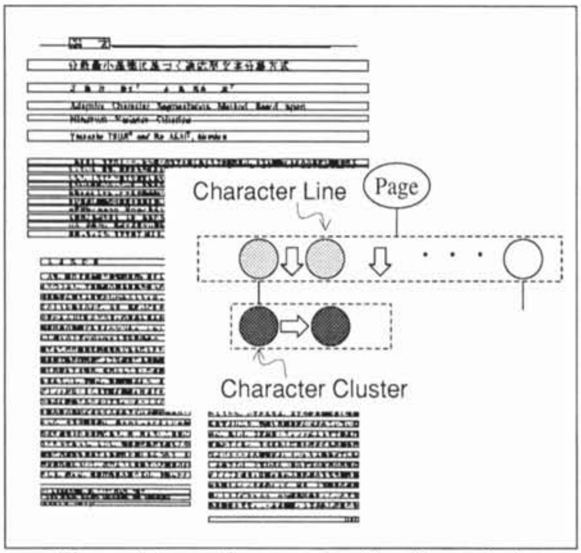

Figure 5. Image Segmentation (Tree Generation)

Separate the master-node into several blocks and generate individual slave-node, in level L. The master-node is rejected.

(iii) Element detection operation (POP-OP.)

This operation occurs, when the master-node seems to be a document element. The classification name (line, character line, picture, figure, table, etc) is given to the node and recorded.

The node classification process sorts attributes into tables and decides upon the next operation, using the block features and tables result from this sorting process. This process uses both projection profile features in horizontal and vertical directions. For these processes, details were explained in the previous paper(6).

5. Document Layout Structure Generation

After the separation process, described in Section 4, the analyzed result is given in tree-shaped structure representation, whose leaves are document elements, such as characters or their blocks, pictures, tables, figures, and so on. This tree-shaped structure representation has the following features. Each stage has positional relations, either vertical or horizontal directions. The positional relation representations come in turn. One example of this tree-shape structured result is shown in Fig.5. The lefthand part represents the character line separation result. The arrows represent the positional relation, as well as in Fig.3. The document positional-structure is generated from this tree-shaped representation, by reconstruction upon a bottom-up approach, as follows.

[*Reconstructing Process*]

[STEP 1] Character line direction detection

Evaluate the regularity value in each level and detect the character line direction. This result affects the analyzing direction sequence.

[STEP 2] Document layout structure generation

Search for the node and its master-nodes, that have a maximum separation level. Repeat this process, changing this separation level, up to the upper level. When the initially analyzed block (page node in Fig.6) is detected, this process is terminated.

[STEP 2.1] Character segmentation

First, detect symbol characters (period, comma, etc) and space characters from the estimated character pitch. Next, separate each character region, with the dynamic programming method, using character spaces distortion, which can separate touched characters. This process also uses the block position information to handle such documents, where characters are printed in proportional pitch.

[STEP 2.2] Node reconstruction

Compare leaf-nodes' classification name and, if necesscary, reconstruct them into one block and give them an appropriate classification name.

[STEP 2.3] Text block generation

This process generates a new text block, when the following conditions are satisfied ; if the spaces between the blocks is smaller than the mean pitch of characters, this block

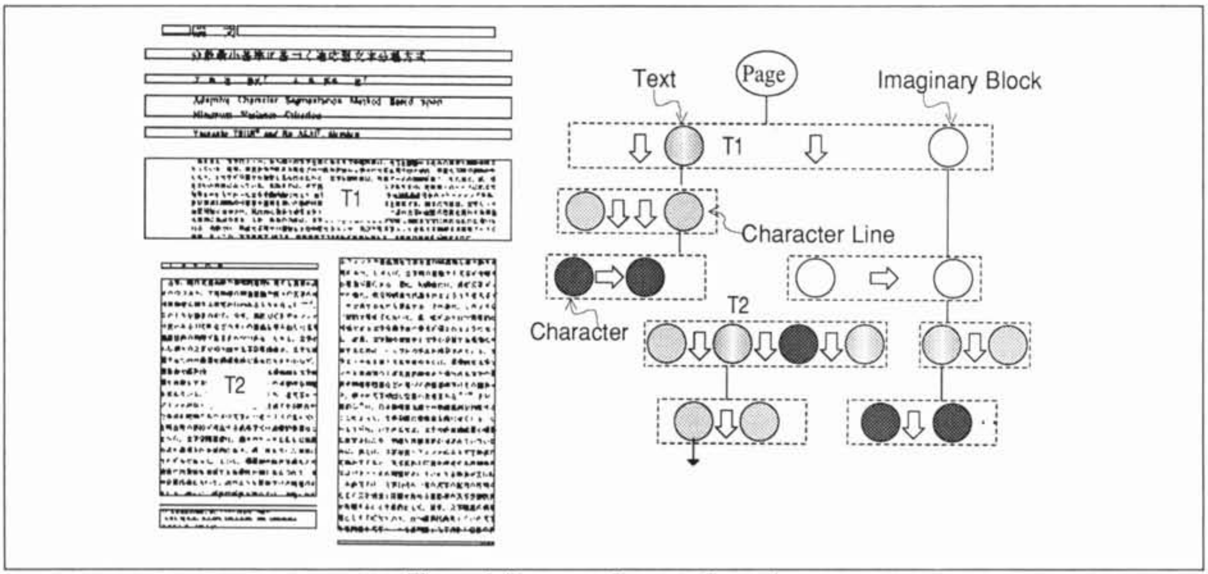

Figure 6. Document Structure Generation

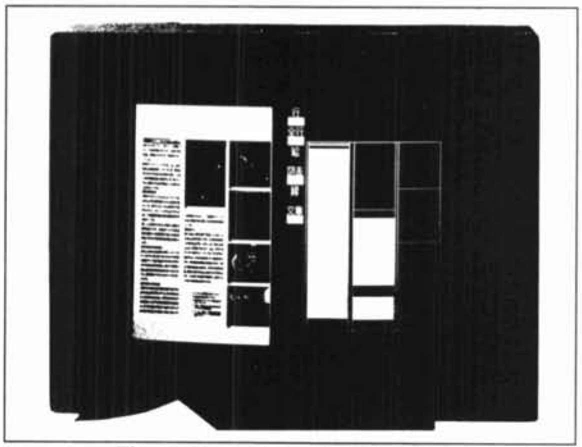

Figure 7. Layout Recognition Result

is regarded as a text block. In this process, new text blocks are generated by combining this text block with neighboring text blocks. This new text block has the same separation level as other character blocks, including this new text block. From this feature, the relations with neighboring blocks are changed and the document layout structure is reconstructed in bottom-up approach.

Figure 6 shows the document layout structure obtained from Fig.5. The lefthand part is an example of experiments. That part shows the layout structure for document elements, including text blocks. The arrows in Fig.5 and Fig.6 represent the positional relation. It is clearly shown that text block T_1 exists on the imaginary block (expressed as an imaginary block in Fig.6) and that the lines exist under text block T_2, according to the arrows.

6. Making to Document Recognition Prototype System

The authors developed a prototype document recognition system, on a personal computer. This system consists of a personal computer (NEC PC series), image scanner, and several ImPP (Image Pipeline Processor) boards[8]. The ImPP board is a general purpose pipeline processor. This prototype system has no special hardware, except for the ImPP board.

The layout analysis was carried out on a personal computer. The character recognition was executed on the ImPP boards. The layout analysis processes are all described in software, C language and assembler. The character recognition method uses the vector pattern matching approach. This method requires a large amount of comparison and summation operations. ImPP board executes these operations effectively. A four vector direction pattern is transposed from the host machine to the ImPP board in one phase, and the calculation is executed in parallel. With this approach, the character recognition speed becomes much faster. The character recognition speed is about 28 characters per second, under the condition that four ImPP boards are used. The more ImPP boards that are used, the faster the recognition speed becomes.

Kanji characters (about 3300 characters), alphanumerics and other special characters are recognized. This prototype system has a multi-font dictionary and reads multifont Kanji characters. This system automatically adjusts the character pitches and cuts out character images from text line image. Half-size characters (usually in alphanumerics) are cut out correctly with the result of character pitch estimation method[7]. Therefore, half-size characters are also readable, as well as full-size characters, even when appearing in the same text lines.

In the author's experiment, the following scores were achieved. For an A4 size document image, it took about 18 seconds on an average to analyze document structure, tested on personal computer PC-H98 model 70. These document images have several blocks in them, all of them correctly separated. Some images includes picture areas, and this algorithm worked as well. Character recognition speed depends on the number of ImPP boards. With one board only, the character recognition speed is 10 characters per second. With two ImPP boards, about 18 characters per second character recognition is achieved.

The mean character recognition rate is about 98%, using a multi-font dictionary. The total recognition time, including document layout analysis and character recognition, is about two minutes for a typical A4 document image, including about 3000 Kanji characters.

The authors also developing this character recognition system on NEC EWS4800 series, software only. Details will be explained in a forth paper.

7. Conclusion

The authors developed a prototype of a compact document recognition system. This system automatically analyzes the document structure, chooses text region and recognizes characters. The authors made the prototype of this document recognition system on a compact personal computer and achieved a high performance level. This system helps in reading documents or data input process into computers.

Acknowledgement

The authors would like to thank the members of the pattern recognition research laboratory for their guidance and encouragement.

References

(1) S.N.Shihari and G.W.Zack,"Document Image Analysis," *8th ICPR*, pp.434-436

(2) K.Y.Wong, et al.,"Document analysis system," *IBM J. Res.&Dev.*, vol. **26**, No.**6**, pp.647-656, 1982

(3) I.Masuda, et al.,"Approach to a smart document reader system," *Proc. CVPR*, pp.550-557, 1985

(4) Y. Tsuji et al,"Document Image Analysis for Reading Books," *SPIE* Vol. **804**, pp.237-244, 1987

(5) J.Higashino, et al.,"A Knowledge-based segmentation method for document understanding," *Proc. 8th ICPR*, pp.745

(6) Y. Tsuji,"Document Image Analysis for Generating Syntactic Structure Description," *Proc. 9th ICPR*, pp.744-747, Nov. 1988

(7) Y. Tsuji and K. Asai,"Character Image Segmentation, Based Upon Minimum Variance Criterion," *NEC Res.&Dev.*, No. **78**, pp.23-30, 1985

(8) T.Temma, M.Iwashita et al,"Data Flow Processor Chip for Image Processing," *IEEE Trans. Electron Devices*, vol. **ED-32**, pp.1784-1791, 1985