

# Document image analysis with cooperative interaction between layout analysis and logical structure analysis

Yasuto Ishitani

Toshiba Corporation

1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki 210-8582, Japan

yasuto.ishitani@toshiba.co.jp

## 1. Introduction

When a printed document is to be input to a computer system, the document must be converted to a computer-readable format, e.g., ASCII, PDF, RTF, CSV, or SGML/XML/HTML-tagged data. In order to obtain these data formats from a printed document, it is necessary to extract from the printed document as much information as possible, i.e., layout structure (layout objects and their hierarchical relationship), logical structure (logical objects and their reading order), and contents (text as OCR results).

Logical structure analysis is concerned with attaching logical meanings to the layout structure and determining the reading order. The logical structure may be extracted hierarchically from the layout structure that is extracted by the layout analysis process from a document image. In this hierarchical analysis, some ambiguities cannot be reduced in the result obtained by a single functional process due to insufficient information. For example, sparse, irregular, or unconfined layout styles such as unpredictable non-text elements, program lists, mathematical expressions, or tables without ruled lines might not be analyzed correctly using only geometric information in the layout analysis process. If processes are connected sequentially in an entire system, system performance deteriorates due to accumulation of the errors of all the processes.

In this paper, the author proposes a new document image analysis method with cooperative interaction between layout analysis and logical structure analysis to resolve the above problems. The proposed method has the following three advantages compared with other methods.

(1) It can extract logical structures from imperfect layout structures due to large variation in formats, noise, or erroneous feature detection. Erroneous layout structure is automatically detected and is resolved by cooperation between layout analysis and logical structure analysis.

(2) It breaks through previous limitations in layout analysis capabilities, because high-level logical information can be used in layout analysis. After this layout analysis, the

global column structure of a document is extracted from the logical structure.

(3) It can extract the logical structure of a document accurately, because the information to be used for logical structure analysis is obtained from the orderly column structure and is made more accurate by the cooperative interaction between layout analysis and logical structure analysis.

## 2. System Overview

The proposed document analysis system consists of layout analysis system including OCR and logical structure analysis system. Each system can communicate with other system, which means there is an interaction between systems as shown in Fig.1(a).

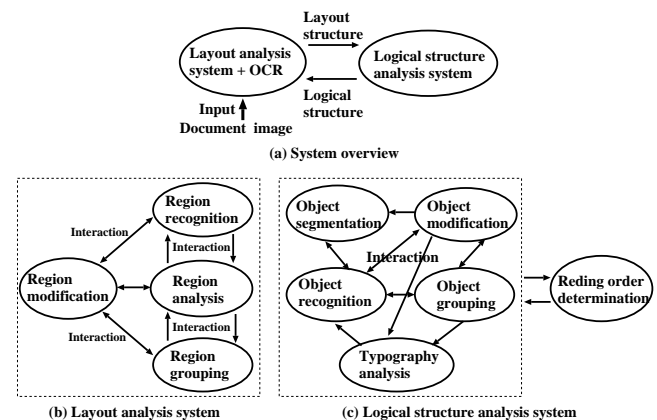
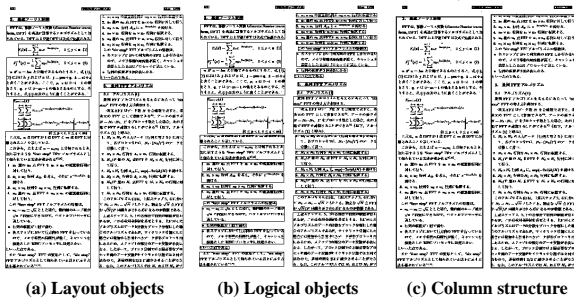


Figure 1. Document analysis system

At first, basic elements in a document image such as characters (connected components), line segments, and black blobs are given to the layout analysis system. By the layout analysis, layout objects such as texts, tables, graphics, and photographs and their hierarchical relationship are extracted by the perceptual grouping process of these elements as a global layout structure from a document image (Fig.2(a)).

When the layout analysis process stops, these layout objects are stored and sent to the logical structure analysis system.



**Figure 2. Document analysis process**

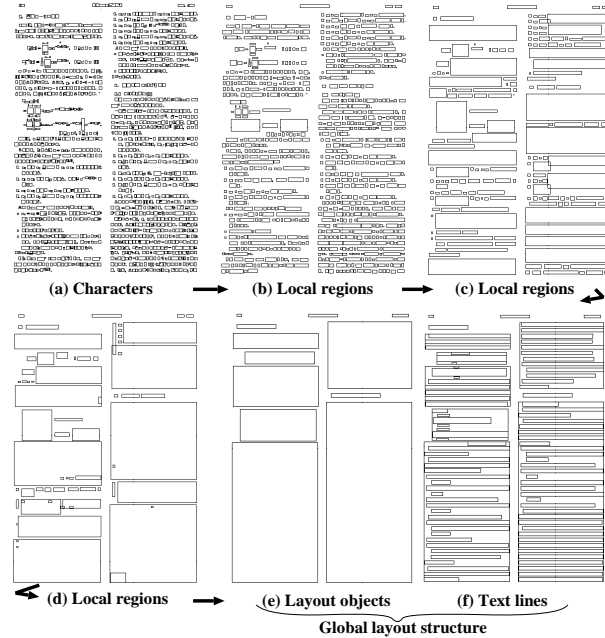
The logical structure analysis copes effectively with various logical structures using various types of information about the objects, e.g., typographical information, contents, geometric and logical relationships among objects, and logical consistency among objects. After the logical structure analysis, a set of primal logical objects such as headlines, headers, footers, captions, notes, programs, titles, paragraphs, lists, and formulas, as shown in Fig.2(b), and their reading order are extracted from a document image. The primal logical object means the minimum logical object which has no low-level object.

After logical structure analysis, a consistent global logical structure consisting of a set of logical objects is sent to the layout analysis system. Then, another global layout structure such as the column structure of a document, shown in Fig.2(c), is organized by the object grouping process based on logical information consisting of logical meanings and the reading order of logical objects. The cooperation between layout analysis and logical structure analysis is repeated while modifications to objects are required. If the results of the layout analysis process are identical to the previous stored results, the layout analysis system terminates to fix its results. These fixed layout objects are sent to the logical analysis system again to extract the hierarchical logical structure from a set of primal logical objects. Finally, the relationship between the global layout structure and the hierarchical logical structure are described using a tree structure.

### 3. Geometric Layout analysis

The layout analysis system consists of four basic modules: region grouping, region analysis, region recognition, and region modification as shown in Fig.1(b). At first, similar and nearby basic elements of an input document are grouped into a local region, in layout analysis. Next, the following region analysis and recognition are applied to each local region: geometric text analysis, object classification,

and character segmentation/recognition. The geometric text analysis consists of the estimation of the geometric parameters of a region, e.g., text line orientation, font size, character pitch, and text line spacing. These geometric parameters of each local region become the boundary conditions for the perceptual grouping processes. The perceptual grouping process based on Gestalt principles using geometric information for text objects is carried out by the following procedure (see Fig.3).



**Figure 3. Layout analysis based on perceptual grouping process**

*Procedure.* If two objects satisfy all of the following conditions, they are merged into one object. This grouping process is applied to each pair of objects while modifications to objects are required.

*Condition A.* The positional relationship between the two objects is adjacent. This geometric relationship between two objects is defined mathematically in reference [1].

*Condition B.* The object that is formed by the grouping process of the two objects must not cross any field separator.

*Condition C.* The two objects have the same text line orientation.

*Condition D.* The distance between the two objects is less than the threshold  $th_1$ , where  $th_1$  is automatically calculated using the geometric parameters of two objects. How to calculate this threshold is described exactly in reference [1].

*Condition E.* The font sizes of the two objects are almost equal.

These conditions are considered to be Gestalt principles for perceptual grouping process in visual information pro-

cessing [4]. For example, condition A and D (C and E) correspond to the principle of proximity (similarity) in Gestalt principles. The interaction between region grouping and analysis extracts good text regions and their geometric parameters. The interaction between region recognition and modification extracts text regions with good reliability. Finally, the global layout structure with good consistency of all layout objects is extracted by the interaction among modules. Details of the layout analysis process are described in reference [1].

#### 4. Logical structure analysis

The logical objects, which are the subject of extraction in the proposed method, are roughly categorized into the following two types:

*Type 1*: headlines, headers, footers, captions, notes, and programs

*Type 2*: titles (chapters or sections), paragraphs, lists, and formulas

All objects except for programs, which are categorized as type 1, can be isolated from other objects on a page. However, some objects categorized into type 2 are often combined into one layout object or one text block because they may have similar geometric parameters such as font size, character pitch, and text line spacing. On the other hand, a program list has a sparse, unconfined, or irregular layout format with some large gaps between text lines or words.

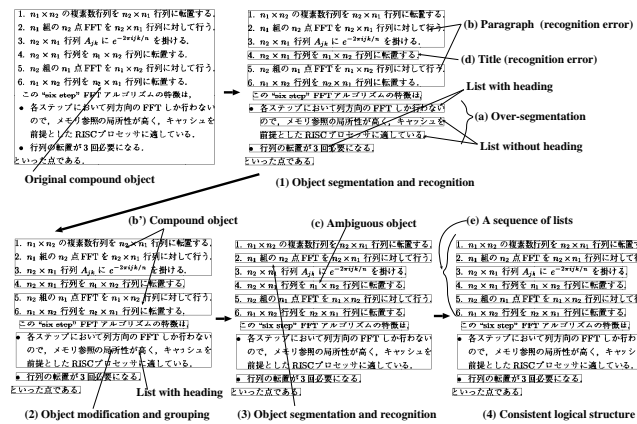


Figure 4. Logical structure analysis process

The logical structure analysis system consists of five basic modules: typography analysis, object recognition, object segmentation, object grouping, and object modification, as shown in Fig.1(c). Simple cases such as isolated objects categorized as type 1 are recognized by the object recognition module in an early process using either their contents or typography. After that, the difficult cases such as com-

pound objects, as shown in Fig.4, are automatically detected because such objects could not be identified by the object recognition module. Then, such objects are sent to the object segmentation module and are divided into different pieces based on their typographical information (Fig.4(1)). Each segmented object is sent to the object recognition module again and is classified into a certain logical class (Fig.4(1)). If segmentation errors or recognition errors occur, the object grouping or modification module recovers these respective errors (Fig.4(2)). The interaction among modules is repeated while the objects are modified. Finally, a set of consistent logical objects (Fig.4(4)) is organized from the system as a result of this interactive computation. If an erroneous layout structure (Fig.5(a)) or a sparse layout structure (Fig.5(d)) is sent to the logical analysis system, ambiguous objects are divided into smaller parts and are grouped according to their logical validity. Then, the correct global logical structure (Fig.5(b)) is extracted from such an irregular layout structure. Details of the logical structure analysis process are described in reference [3].

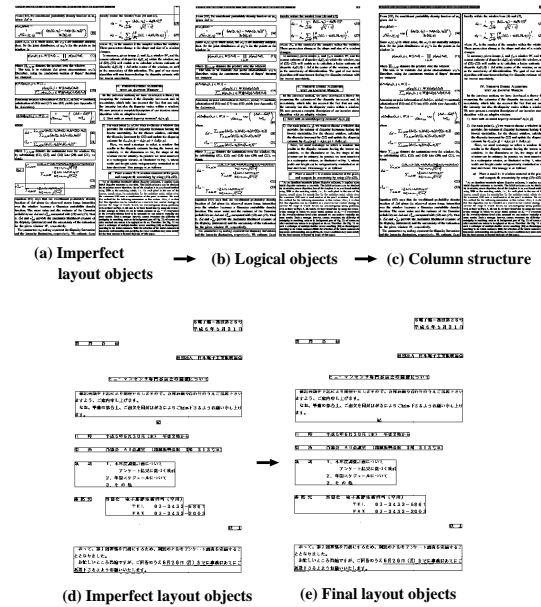


Figure 5. Improvement of results by cooperative interaction between systems

After logical object extraction, the reading order of the logical objects is determined by the process outlined in reference [2]. Details of the process for the reading order determination are not discussed here due to lack of space.

#### 5. Cooperative interaction between layout analysis and logical structure analysis

The interaction between systems overcomes the limitations in the capability of each system, and solves ambiguities

due to the shortage of information in each system as follows. The higher-level logical information of a document such as logical meanings and the reading order of logical objects can be used in the layout analysis process by the procedure below. According to this procedure, sparse or irregular layout styles with large gaps between text lines or words are analyzed correctly as shown in Fig.5(e), because such objects are merged using the logical information.

**Procedure.** If two logical objects satisfy all of the following conditions, they are merged into one layout object. When a layout object consists of a set of logical objects, the following conditions are applied to two adjacent logical objects having a minimum distance between the logical object in one layout object and the logical object in the other layout object. This grouping process is applied to each pair of objects while modifications to objects are required.

**Condition 1.** The two logical objects must satisfy Condition A, B, and C.

**Condition 2.** The two logical objects have the same logical class or have one of the logical classes categorized as type 2.

**Condition 3.** The two logical objects are arranged in the reading order.

If the reading order of logical objects is incorrect, the column structure of a document is not formed by this procedure. In this case, these logical objects can be grouped into a global layout structure if condition 3 is replaced by the following condition 3'. This condition assumes that two objects are arranged in the reading order when the distance between them is very short.

**Condition 3'.** The distance between the two objects is less than the threshold  $th_2$ , where  $th_2$  is defined as  $th_2 = \min(ch_1, ch_2) \times \alpha$ , where  $ch_1$  ( $ch_2$ ) is defined as the average font size of object1 (object2) and  $\alpha$  is a constant.

The interaction between systems improves the result of logical structure analysis as well as that of layout analysis as follows. (1) In the logical structure analysis, the logical objects of a document are extracted accurately using the typographical information that is extracted from the orderly column structure. (2) The logical object, which is divided into two or more columns or pages, can be unified and classified correctly because the reading order of objects in each page is utilized. (3) The reading order of logical objects is determined accurately because the search space of the determination of the reading order is reduced in the global column structure.

## 6. Experimental results

Experiments were carried out using 150 documents taken from magazines, newspapers, books, scientific and technical journals, business documents, business letters, manuals, patents, government files, and official documents that were

printed in either Japanese or English. Some of the test documents and their experimental results are shown in Fig.6. The rate of correct logical structure analysis was on average 96.3% for all logical objects.

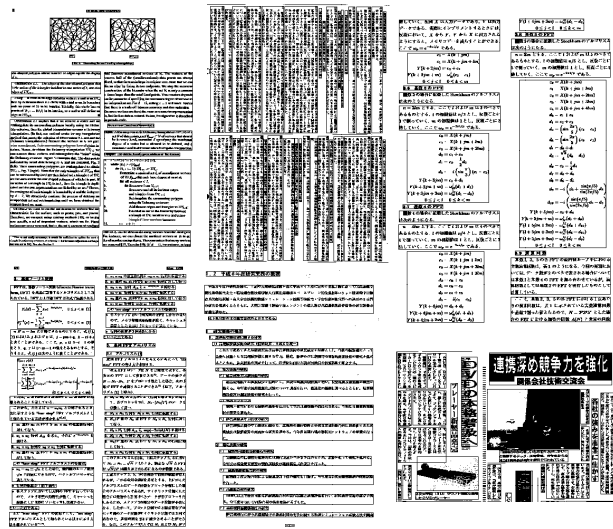


Figure 6. Experimental results

## 7. Conclusions

A new method for document image analysis with cooperative interaction between layout analysis and logical structure analysis has been proposed. According to the proposed method, the parameters of layout analysis and logical structure analysis are automatically formed, and are adjusted by the interaction between them, for effective global document structure analysis. Global system behavior that is expected to be adaptable to various document styles is established by the adaptive system configuration based on the interactive computation. Furthermore, the limitations in layout analysis capabilities due to insufficient information are overcome by the interactive computation because higher-level logical information can be used in lower-level processes. As a result, the proposed system can handle a wide variety of documents and complex document structures and achieves high performance with regard to total throughput.

## References

- [1] Y. Ishitani. Document layout analysis based on emergent computation. *Proc. 4th ICDAR*, 1:45–50, 1997.
- [2] Y. Ishitani. Document analysis system based on emergent computation. *Proc. DAS'98*, pages 205–208, 1998.
- [3] Y. Ishitani. Logical structure analysis of document images based on emergent computation. *Proc. 5th ICDAR*, 1999.
- [4] K. T. Spoehr. *Visual information processing*. W. H. Freeman and Company, 1982.