Document Layout Analysis as Solving Multiple Constraints Problem

Fu Chang
Document Analysis and Recognition Laboratory
Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.
fchang@iis.sinica.edu.tw

Abstract

The aim of document layout analysis is to uncover the hierarchical layout structure by means of a series of split and merge operations. The validity of such operations depends on certain parameters that need to be properly estimated. Moreover, the values of those parameters cannot be calculated independently, but are parts of the problems to be solved. The layout analysis is thus considered as multiple constraints problem whose unknown variables are the elements of the hierarchical structure, and the constraint conditions specify the relations between those elements. To solve this problem, we start with a solution that satisfies some but not all the constraints. We then gradually improve the solution by estimating the values of parameters, also appearing in the constraints, based on the temporary solutions obtained in previous stages. This solution method proves to be extremely useful for analyzing stylized but complicated documents, such as Chinese documents that are known to allow both horizontal and vertical reading orders as well complicated compositional structure even on the character level.

1. Introduction

The textual contents of documents are organized in hierarchical order (Figure 1). Individual characters are organized into textlines, which in turn are organized into paragraphs etc. In the past, many methods look for entering points in the hierarchy. For example, the recursive X-Y cuts method [1] and maximal white-rectangles method [2] seek to enter at the top of the hierarchy, each using different clues to decompose large objects into smaller ones. The run-length smearing method [3] seeks to enter at the bottom of the hierarchy, by gluing foreground pixels into connected pieces. The textline construction method [4] may be said to enter the hierarchy in the middle, by constructing textlines out of character components.

All the above methods work reasonably well in certain contexts. They would nevertheless face problems when their contextual assumptions are violated or their parameter values are incorrectly estimated. For example, the X-Y recursive method does not work at the environment where no simple cutting lines exist. The white-rectangle method and smearing method rely upon certain threshold values (the size of white rectangles for the former, and the distance of white gaps for the latter). The textline method cannot be applied without limit when both horizontal and vertical reading orders are allowed, as in Chinese documents.

2. Problem Setting

In our view, all methods for layout analysis employ

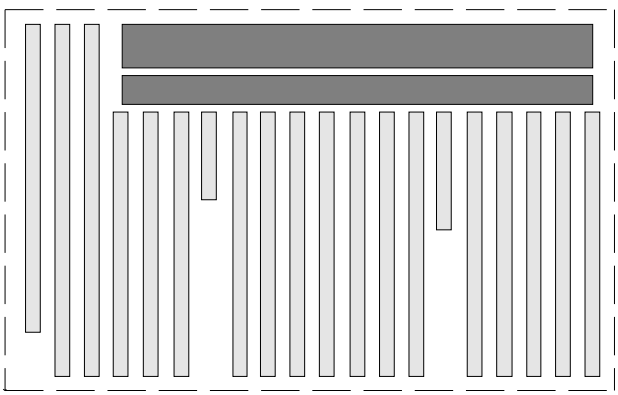

Figure 1. The layout structure of a Chinese article is displayed. Headlines (striped boxes) are horizontal. The rest of textlines (gray boxes) are vertical. Dashed lines form the borders of the article.

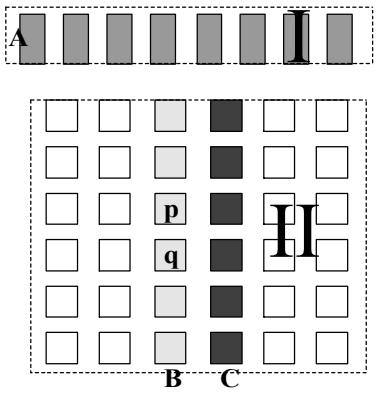

Figure 2. Two paragraphs, I and II, are enclosed in dashed lines. Paragraph I contains horizontal textline A. Paragraph II contains B, C and other vertical textlines. Each labeled textline is identified by the same fill pattern. The spacing within B (= distance between p and q) is less than that between B and C. The spacing within paragraph I (= distance between B and C) is less than that between I and II.

some *split and merge* operations. Run-length smearing, for example, is a merge operation while X-Y cut is a split operation. To get a reasonable outcome, these operations rely upon certain parameter values to function. Those values are part of the problem that needs to be solved.

Document layout analysis is a *multiple constraint problem*, where the objects (paragraphs, textlines, and characters, etc.) to be constructed or identified are the *unknown variables* built into the constraints. There are actually three sets of constraints that are met by most Chinese documents.

A. Composition Constraints: paragraphs are composed of contiguous textlines of similar sizes, which in turn are composed of contiguous characters of similar sizes; characters are composed of contiguous components, i.e., clusters of connected foreground pixels.

B. Alignment Constraints: characters contained in horizontal (vertical) textlines are properly aligned along their top (right) and bottom (left) edges; textlines contained in horizontal (vertical) paragraphs are properly aligned along their top (right) and bottom (left) edges.

C. Spacing Constraints: the spacing within a textline is less than the spacing between textlines; moreover, the spacing within a paragraph is less than the spacing between paragraphs (Figure 2).

Each set of constraints involves more than one unknown variable, and each unknown variable occurs in more than one set of constraints. The third set of constraints, moreover, incorporates some parameters (inner spacing and outer spacing) whose values depend on the attributes of other variables (textlines and paragraphs).

The methods [1-4] attempt to solve this problem in a straightforward manner. They seek a solution for some unknown variable, which in turn can be used to solve for

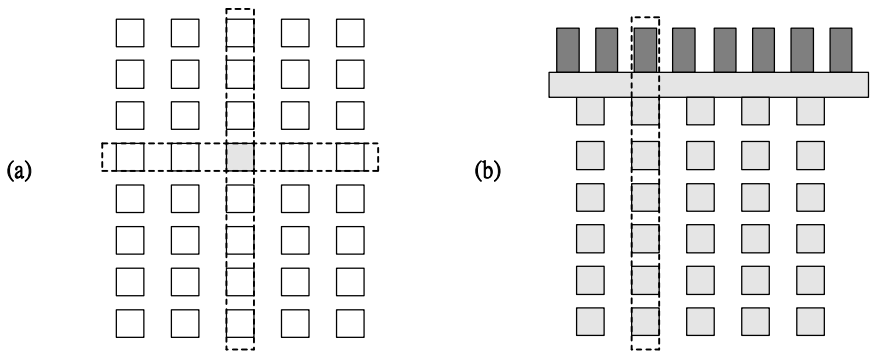

Figure 3. (a) Two textlines (enclosed by dashed lines) grow from the same (gray) component box. The horizontal line has wider inner spacing than the vertical line. (b) The vertical textline (enclosed by dashed lines) extends beyond a white rectangle (dotted box) with relatively wide spacing.

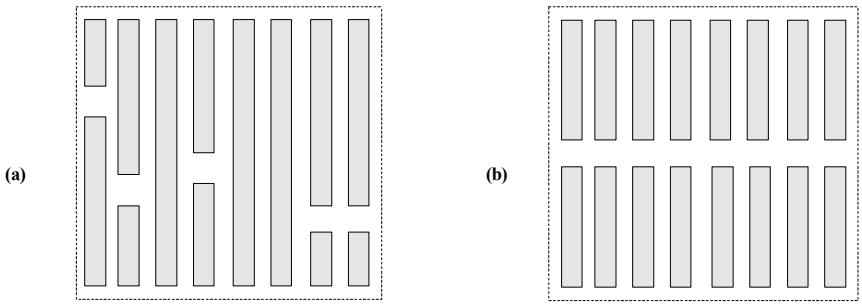

Figure 4. (a) Broken pieces falling within the same paragraph can be connected to make complete pieces. (b) Textlines belonging to different columns can be separated by means of recursive X-Y cuts.

another unknown variable, and so on. Unfortunately, for most hard constraint problems, straightforward solutions rarely exist. As an alternative, we propose the following method. We start with an approximate solution, that is, a solution satisfying some but not all the constraints. Based on this temporary solution, we then estimate the unknown parameter values appearing in other constraints. We can then derive a better solution (that is, a solution that satisfies more constraints) by taking reference of the estimated values. The solution thus obtained can still be improved if more or better knowledge is generated.

3. Solution Method

In actually solving the problem, an evolutionary path has to be specified whose aim is to gradually improve the solutions for unknown variables. At each stage, moreover, some constraint conditions are used for the construction of new solution or the refinement of current solutions. Brief descriptions of these stages are given in the following. More details are given in [5:Section 3].

Textline Construction Stage

Textlines are constructed according to the requirements embodied in the composition and alignment constraint for textlines. The composing elements of textlines are characters that are unfortunately unknown at this stage (this is the case for Chinese documents, since a Chinese character can be a composition of more than one component). But using some statistical means, we can find reliable elements to start with for the textline construction.

The textlines constructed thus far have two drawbacks. First, starting from the same initial point, it is possible to construct two textlines, one going vertically and one horizontally (Figure 3a), but only one of them can be legitimate. Secondly, some textlines may be overextended to the area of other textlines. This is most likely to occur when the two textlines have opposite reading order (Figure 3b).

Textline Consolidation Stage

By referring to the spacing constraint for textlines and for paragraphs, we are able to eliminate those textlines that are illegitimate and also prune those that are overextended (Figure 3b).

Paragraph Construction Stage

This is rather a simple operation that follows what is suggested by the composition constraint for paragraphs.

Adjustment Stage

By taking reference of the solutions obtained for the higher levels of the layout structure, it is possible to adjust the solutions for the lower level objects. For example, a merge operation can be applied to join those properly aligned textlines that fall within the same paragraph (Figure 4a). A further split operation may also be re-

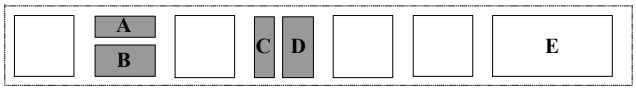

Figure 5. A horizontal textline (enclosed in dashed lines) incorporates irregular components, where A and B should merge into one box, C and D into another box, and E should be split into two smaller boxes. The unusual size of box E is caused by accidental touch of two nearby characters.

quired for separating two columns within the same article (Figure 4b). These operations are made to comply with the requirement imposed by the alignment constraint for paragraphs. Furthermore, spilt and merge operations can be applied to the components found within the same textline (Figure 5), conforming to the requirement that each textline is composed of characters of similar size.

4. Testing the Methodology

For testing the document layout methodology, we have prepared a set of samples collected from magazine and newspapers. There are approximately 1400 of them. Many samples consist of a single article, while others have more than one article. We classify these samples according to their layout structures into seven categories:

- horizontal headlines with L-shaped contents
- horizontal headlines with rectangle-shaped contents
- vertical headlines with L-shaped contents
- vertical head-lines and rectangle-shaped contents
- mixture of texts and pictures
- mixture of texts and tables
- mixture of Chinese and English characters within the same textlines

Each time when we make ready a version of layout-analysis program for testing, we randomly select a few samples from each of the seven categories, plus some "tough" samples that are always selected for the testing purpose. In the last testing we made, 50 samples were chosen and two problems were detected out of the test results. Both of the problems have to do with the assumptions being taken by the layout-analysis program being tested.

The first problem occurs at two parallel textlines, each consisting of two characters only. The spacing between the two textlines is actually smaller than the spacing within each textline. This fact obviously violates our spacing consumption for textlines. Since this kind of problem occurs rarely in regular documents, we decide to do nothing about it. The second problem has to do with the listed items that contain only two characters. Since our program assumes that a textline has to contain at least three characters, those items were not segmented as textlines. To solve this problem, care must be taken to deal with listed objects. This is a subject we would rather not go any further in this article.

5. Discussions and Unsolved Problem

We would like to emphasize that the methodology proposed here is for analyzing very stylized documents in terms of layout complexity. Moreover, the constraint conditions considered can be met by many Chinese documents, but certainly not by all of them. Some special documents (business cards, for example) may require that one or several of the conditions be removed or weakened, and that a few others be added. Moreover, the validity of those constraints has not been largely tested by non-Chinese documents.

However, one nice thing about this approach is that it can easily adjust the solution path according to changes in the constraint specification. The reason for this easy adjustment is that constraint conditions can be naturally classified into categories. (We categorized them into three sets). Thus, when one category of constraints remains unchanged, all the corresponding operations remain unchanged. The idea of modularity thus naturally grows in this solution method. However, a problem that remains to be solved is to classify the types of documents satisfying different sets of constraints.

References

1. G. Nagy, S. C. Seth, and S. D. Stoddard, Document analysis with an expert system, *Proceedings Pattern Recognition in Practice II*, Amsterdam, 1985.
2. H. S. Baird, S. E. Jones, and S. J. Fortune, Image segmentation by shape-directed covers, *Proceedings 10th ICPR*, Atlantic City, pp. 820-825, 1990.
3. F. M. Wahl, K. Y. Wong, and R. G. Casey, Block segmentation and text extraction in mixed text/image documents, *Comput. Vision Graphics Image Process*, vol. 20, pp. 375-390, 1982.
4. T. Pavlidis and J. Zhou, Page segmentation and classification, *CVGIP: graphical models and image processing*, vol. 54, pp. 484-496, 1992.
5. F. Chang, Retrieving Information from Document Images: Problems and Solutions, *Intern. J. Document Analysis and Recognition, Special Issues on Document Analysis for Office Systems*, to appear.
6. H. Baird, Anatomy of a versatile page reader, *Proceedings IEEE,* vol. 80, pp. 1059-1065, 1992.
7. A. Dengel, Initial learning of document structure, *Proceedings Intern. Confern. Document Analysis and Recognition*, Tsukuba, 1993.
8. J. Fisher, S. Hinds, and K. D'Amato, A rule-based system for document image segmentation, *Proceedings 10th Intern. Conf. Pattern Recognition,* pp. 567-572, Atlantic City, 1990.
9. H. Fujisawa and Y. Nakano, A top-down approach for the analysis of documents, *Proceedings 10th Intern.*

Conf. Pattern Recognition, pp. 113-122, Atlantic City, 1990.

10. A. K. Jain, B. Yu, Document representation and its application to page decomposition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 20, pp. 294-308, 1998.
11. L. O'Gorman and R. Kasturi eds., *Document Image Analysis,* Los Alamitos, IEEE CS Press, 1995.