On the Application of Voronoi Diagrams to Page Segmentation [*]

Koichi Kise, Motoi Iwata and Keinosuke Matsumoto
Department of Computer and Systems Sciences,
College of Engineering, Osaka Prefecture University
1-1 Gakuencho, Sakai, Osaka 599-8531, Japan
E-mail: kise@cs.osakafu-u.ac.jp

1 Introduction

The history of research and development on layout analysis can be considered as a series of challenges to remove restrictions on page images as well as to improve the robustness of analysis. Some of the important challenges in recent years would be as follows:

layout — from rectangular to non-rectangular layout:

Layout is a physical arrangement of document components such as text-blocks, text-lines, figures and tables in a page. The difficulty of layout analysis depends on a class of layout to be analyzed. An important and the most investigated class would be *rectangular* layout. Layout is rectangular if all document components are circumscribed by non-overlapping upright rectangles. In recent years, however, we often see pages outside this class in, for example, magazines and journals. Such pages include document components of arbitrary shape as well as tilted text-lines to make them attractive. Thus we need to develop methods which are capable of analyzing pages with non-rectangular layout.

process — from one-pass to cooperative and adaptive processing:

The process of layout analysis is sometimes divided into page segmentation and classification. Page segmentation is to decompose pages into a set of homogeneous regions regardless of types of document components, and page classification is to assign types to extracted regions. The former is further divided into steps of extraction of blocks, text-lines, words and characters. These steps are then combined to form complete layout analysis. The simplest way of the combination is either local-to-global or global-to-local. Although they allow us to deal with pages efficiently, errors at a step cause additional errors in succeeding steps and thus lower the overall accuracy. To cope with this problem, some researchers have attempted to renew the way of combination based on cooperative and adaptive processing [3, 2].

image — from binary to gray-level and color images:

Since the majority of documents are printed in the black-on-white manner, researches on layout analysis have mainly focused on binary images. In recent years, however, documents often contain color pages aiming at easy understanding of contents. In addition, it would be necessary for the analysis of pages with toned backgrounds to capture pages as gray-level images. Thus the generalization of images from binary to gray-level and color is also required to generalize layout analysis.

We are concerned here with the generalization of layout for the analysis of binary page images in the context of segmentation of text-blocks and text-lines.

For the analysis of pages with rectangular layout, rectangles play an important role for both representation and analysis. Needless to say, however, they are inappropriate for pages with non-rectangular layout.

Page segmentation of binary images is fundamentally a process of clustering black (foreground) pixels in 2D space[1]. With little loss of generality, connected components of black pixels can be utilized as primitives. This enables us to simplify the task of page segmentation: clusters are obtained by combining connected components appropriately. However, the number of all possible combinations is generally too large to explore. Thus we need a way to limit possible combinations of connected components.

It is natural to use the *distance* as a primary guide for clustering, since connected components in the same docu-

[*] This research was supported in part by Grant-in-Aid from the Ministry of Education, Japan, as well as from The Telecommunication Advancement Foundation.

[1] Based on the objects to be analyzed, existing methods can be classified into two categories: foreground analysis which is to merge black (foreground) pixels to obtain document components [7, 2, 5, 4], and background analysis which is to merge white (background) pixels[1, 6]. A discussion about these categories can be found in [5]. We focus here on the foreground analysis.

ment component are often closer to those in different document components. O'Gorman has proposed a method based on k-NN to limit combinations to be explored to extract text-lines and text-blocks[7]. Although his method is effective for pages with a variety of layout, it requires to predetermine the value of k which depends on page layout.

In order to solve the above problem, we have proposed the representations based on area Voronoi diagrams as well as methods of text-block and text-line extraction [5, 4]. In what follows, we describe an overview of our representations and methods as well as open problems for the extension.

2 Current Work

Our current work is to develop general representations of physical structure based on area Voronoi diagrams, as well as to apply them to page segmentation. Summaries of the results up to now are described below.

2.1 Area Voronoi Diagram

An area Voronoi diagram is a generalization of an ordinary (point) Voronoi diagram[8]. While a point Voronoi diagram is generated from a set of *points*, an area Voronoi diagram is generated from a set of *non-overlapping figures*.

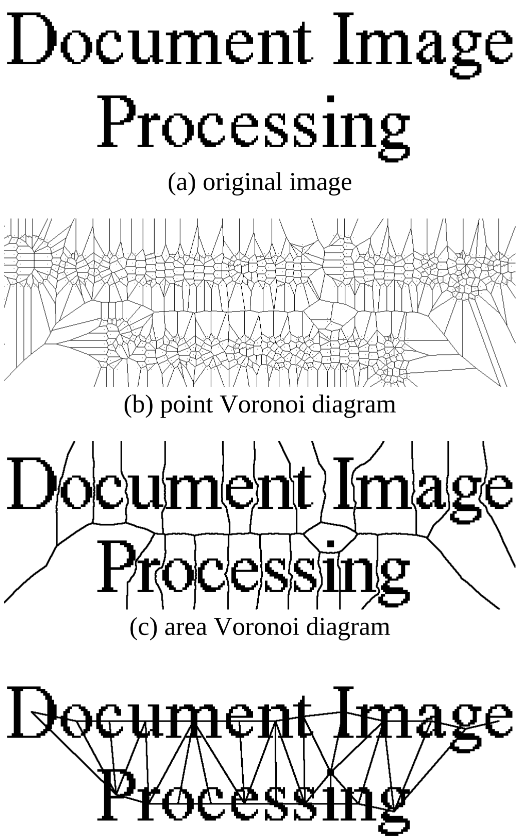
An area Voronoi diagram can be approximately constructed based on a point Voronoi diagram. Figure 1 illustrates an example. First, a point Voronoi diagram is constructed from a set of sample points (black pixels) on *contours* of connected components as shown in Fig. 1(b). Then, an area Voronoi diagram as in Fig. 1(c) is obtained by deleting from the point Voronoi diagram all edges generated from a pair of points on the *same* connected component.

An area Voronoi diagrams has the property that white pixels in each divided region are closest to the connected component in it. Since an edge in an area Voronoi diagram represents a part of boundaries of a region, a pair of connected components which share an edge on the boundaries of their regions can be considered to be *adjacent* or *neighbors* with each other. In order to represent such neighbor relations, we utilize a neighbor graph shown in Fig. 1(d). A neighbor graph is a graph in which a vertex corresponds to a connected component and an edge represents a neighbor relation between connected components.

An area Voronoi diagram and a neighbor graph have the following important properties for a general representation of physical structure:

- They can be constructed independently of layout and skew of page images.

- Unlike k in k-NN, they require no parameters determined depending on layout.



(a) original image

(b) point Voronoi diagram

(c) area Voronoi diagram

(d) neighbor graph

Figure 1. An area Voronoi diagram and a neighbor graph.

- They can be quickly constructed. Figures 2(b) and (c) illustrate an area Voronoi diagram and a neighbor graph constructed from a page image shown in Fig. 2(a). For example, the construction of Figs. 2(b) and (c) from the labeling result of Fig. 2(a) required 1.4 [sec.] with a Pentium II 450MHz computer (they were simultaneously constructed).

- Edges in an area Voronoi diagram can be viewed as potential boundaries of document components. In other words, they include correct boundaries of document components. In most cases, a neighbor graph also contains correct text-lines as its paths (i.e., sequences of edges).

- Therefore, the process of page segmentation is transformed into the selection of appropriate edges from an area Voronoi diagram and a neighbor graph. Note that the number of edges is quite small as compared to the number of all possible pairs of connected components.

2.2 Page Segmentation

We have already proposed methods of text-block and text-line extraction based on area Voronoi diagrams. Examples of processing results are shown in Fig. 2.

The method of text-block extraction works directly on an area Voronoi diagram[5]: it selects the appropriate edges from an area Voronoi diagram using the distance and the area ratio (i.e., the ratio of the number of black pixels) both of which are defined between adjacent connect components. An important point is that the method relies on the inter-column and inter-line gaps of body texts estimated from the frequency distribution of the distance between adjacent connected components. This enables us to remove thresholds which depend on page layout. However it also poses the problem in extraction of minor blocks, e.g., blocks with larger fonts such as titles.

The method of text-line extraction is, on the other hand, based on a neighbor graph[4]: it selects the appropriate edges from a neighbor graph iteratively by greedy search of an edge to be connected to each current (partial) text-line. As features for selection, we utilize linearity [2] of each path in addition to the distance between adjacent connected components. The threshold of the distance is also calculated based on the frequency distribution.

3 Open Problems

In this abstract, we have presented an overview of our current work on page segmentation based on area Voronoi diagrams. Although we consider that the representations are general enough to cover most of black-on-white pages [3], we have problems to be resolved with respect to the three points in Sec. 1.

(1) *Is layout-free layout analysis possible?*

In our methods, we utilize various estimated gaps to reduce the number of thresholds influenced by page layout. However, this approach is sometimes unsatisfactory, since it cannot relief the minorities. We should not rely solely on global estimations to realize layout-free layout analysis. In addition, we utilize a few but strong assumptions for segmentation, which sometimes cause errors. For example, we assume that the shape of text-lines is linear in the method of text-line extraction, but text-lines are not necessarily linear in modern magazines and journals. Thus we should answer the question of what the general knowledge about page layout is, and how it is represented.

(2) *Do Voronoi diagrams provide foundations of complete layout analysis for non-rectangular pages?*

Since the two methods were applied independently to the image in Fig. 2(a), a part of the results shown in Figs. 2(d) and (e) are contradictory to each other as shown in Fig. 2(f). Thus, we need to combine them cooperatively to improve the accuracy. The questions are (a) whether Voronoi diagrams provide us foundations for such a combination as they do for individual steps, and (b) how they are combined.

(3) *What are general representations of physical structure for gray-level and color images?*

In addition to the analysis of binary images, it is also required for the analysis of gray-level and color images to utilize general representations of physical structure. Although the representation based on Voronoi diagrams would be general for binary images, it cannot be directly applied to gray-level and color images. Thus the questions arise: What are representations of physical structure appropriate for these images? What properties should they have for keeping representations general?

Acknowledgments

We would like to thank Prof. Dr. A.Dengel and his colleagues for helpful comments on the method of text-line extraction.

References

[1] A. Antonacopoulos. Page segmentation using the description of the background. *Computer Vision and Image Understanding*, 70(3):350–369, 1998.

[2] K. Gyohten, T. Sumiya, N. Babaguchi, K. Kakusho, and T. Kitahashi. A multi-agent based method for extracting characters and character strings. *IEICE Trans. Info. & Syst., Japan*, E97-D(5):450–455, 1996.

[3] Y. Ishitani. Document layout analysis based on emergent computation. In *Proc. 4th ICDAR*, pages 45–50, 1997.

[4] K. Kise, M. Iwata, A. Dengel, and K. Matsumoto. A computational geometric approach to text-line extraction from binary document images. In *Proc. of Third IAPR Workshop on Document Analysis Systems*, pages 346–355, 1998.

[5] K. Kise, A. Sato, and M. Iwata. Segmentation of page images using the area voronoi diagram. *Computer Vision and Image Understanding*, 70(3):370–382, 1998.

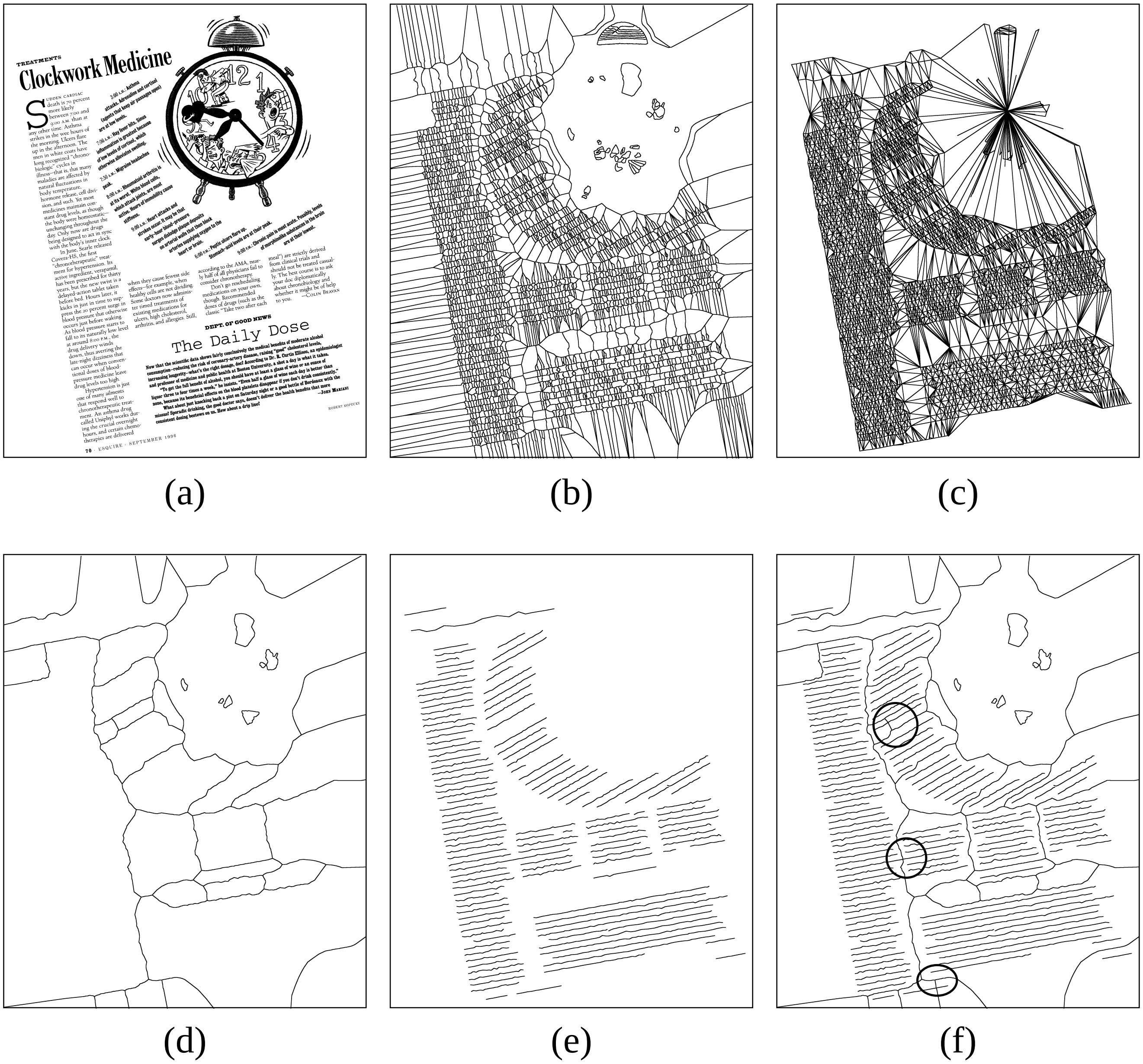
[6] K. Kise, O. Yanagida, and S. Takamatsu. Page segmentation based on thinning of background. In *Proc. of the 13th ICPR*, pages 788–792, 1996.

[7] L. O'Gorman. The document spectrum for page layout analysis. *IEEE Trans. PAMI*, 15(11):1162–1173, 1993.

[8] K. Sugihara. Approximation of generalized voronoi diagrams by ordinary voronoi diagrams. *CVGIP: Graphical Models and Image Processing*, 55(6):522–531, 1993.

[2] The linearity is defined based on the angular error between an edge and a current text-line[4].

[3] Important exceptions are pages with toned or decorated backgrounds. For the application of Voronoi diagrams, there should be the clear distinction between the foreground and the background.



(a) (b) (c) (d) (e) (f)

Figure 2. Examples of processing results. (a) page image (300dpi) tilted by 10°. It includes 5.4×10^3 connected components (1.08×10^6 black pixels). (b) area Voronoi diagram constructed from 6.0×10^4 sample points (black pixels) on contours of connected components. It consists of 8.3×10^3 edges. (c) neighbor graph. The number of edges is equal to that in (b). (d) extracted blocks. (e) extracted text-lines represented by 2.6×10^3 edges. (f) (d) and (e) are superimposed. The circles indicate the parts in which the extracted blocks and text-lines are contradictory. The total computation time required for obtaining (d) and (e) were 3.4 [sec.] and 5.2 [sec.] with a Pentium II 450MHz computer, respectively, both of which included 1.4 [sec.] for the computation of the Voronoi diagram (a) and the neighbor graph (b) (they are obtained simultaneously). The rest were consumed by the steps of labeling, feature extraction, selection of edges, file I/O, memory allocation, etc.