Document Layout Structure Extraction Using Bounding Boxes of Different Entities

J. Liang, J. Ha, and R.M. Haralick

Department of Electrical Engineering
University of Washington
Seattle, WA 98195

I.T. Phillips

Department of Computer Science
Seattle University
Seattle, WA 98122

{jliang, haralick, yun}@george.ee.washington.edu

Abstract

This paper presents an efficient technique for document page layout structure extraction and classification by analyzing the spatial configuration of the bounding boxes of different entities on the given image. The algorithm segments an image into a list of homogeneous zones. The classification algorithm labels each zone as text, table, line-drawing, halftone, ruling, or noise. The text-lines and words are extracted within text zones and neighboring text-lines are merged to form text-blocks. The tabular structure is further decomposed into row and column items. Finally, the document layout hierarchy is produced from these extracted entities.

1 Introduction

The goal of document understanding is to convert existing paper documents into a machine readable form. It involves estimating the rotation skew of each document page, determining the geometric page layout, labeling blocks as text, math, figure, table, halftone, etc., determining the text of text blocks through an OCR system, determining the logical structure, and formatting the data and information of the document in a suitable way for use by a word processing system or by an information retrieval system.

The document layout analysis is a specific instance of a more general problem group in image analysis, that of image segmentation, classification and representation. The segmentation problem discovers various objects of interest in an input document image. An object is a homogeneous rectangular region in a document image that corresponds to one type: glyph, word, text-line, text block, table, graph, picture, etc. The classification problem identifies the detected objects' types. The representation problem expresses the spatial relationships among the various objects of interest. The classification method must complement the segmentation approach chosen [1]. It should utilize the data and measurements produced during segmentation before performing further measurements when computing features. Accessing the image data again is time consuming and should be avoided wherever possible.

There are two major approaches of segmentation and classification. In the first approach, the document image is decomposed into a set of zones enclosing either textual or nontextual content portions using the global spatial properties. The statistical and textural features are extracted and used to label each zone as one of the predefined categories [8]. In the second approach, the low-level segments (connected components [9], line segments [7], etc.) are extracted and classified as text or non-text. Then the text segments are grouped into text blocks.

This paper presents a hybrid technique for page layout structure extraction and classification by analyzing the spatial configuration of the bounding boxes of different entities (connected components, words, textlines, etc.) on a given document image. We segment the image into a list of zones by analyzing the projection profiles of connected component bounding boxes. Then, the text-line and word entities are extracted within each zone. We classify each zone into one of the categories: text, table, line-drawing, halftone, and horizontal or vertical line, using the bounding box information of connected component, word, and textline entities. By analyzing the alignment of text-line boxes and word boxes, text blocks are extracted from text zones, and tabular row-column structures are extracted from table zones. Since our segmentation algorithm manipulates only the bounding boxes of different entities, and the classification algorithm utilizes

0-8186-7620-5/96 $5.00 © 1996 IEEE

Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV '96)
0-8186-7620-5/96 $10.00 © 1996 IEEE

the data produced during segmentation, it does not require intensive computation.

Section 2 describes the hierarchical layout structure of a document and defines the layout structure extraction problem. In Section 3, our algorithm for extracting layout structure is discussed in detail. Experimental results on some scientific/technical documents are shown in Section 4.

2 Layout Structure

This section provides a formal definition of the *layout structure model.* The layout structure extraction problem is presented in term of this definition.

We define a hierarchical structure, called a *Polygonal Spatial Structure,* to capture the information about a document image.

- A polygon and the divider around it is called a Polygonal Spatial Structure (PSS).
- A basic Polygonal Spatial Structure, which is not further divided, carries a content, and the nature of the divider.
- A composite Polygonal Spatial Structure consists of one or more non-overlapping Polygonal Spatial Structures which are either basic or composite Polygonal Spatial Structures.
- We denote by Θ the set of content types (text-block, text-line, word, table, equation, drawing, halftone, handwriting, etc.).
- We denote by $\mathcal{D}$ the set of dividers (spacing, ruling, etc.).
- *Polygonal Area*

  We denote by $\mathcal{A}$ the set of non-overlapping homogeneous polygonal areas on document image. Each polygonal area $A \in \mathcal{A}$ consists of an ordered pair (θ, I), where $\theta \in \Theta$ specifies the content type and I is the area. A polygon is homogeneous if all its area is of one type and there is a standard reading order for the content within the area.

We frame the document layout analysis problem in terms of the abstract PSS definition. A layout structure of a document image is a specification of the geometry of the polygons, the content types of the polygons, and the spatial relations of these polygons. Formally, a layout structure is $\Phi = (\mathcal{A}, \mathcal{D})$, where $\mathcal{A}$ is a set of homogeneous polygonal areas, and $\mathcal{D}$ is a set of dividers.

3 The Algorithm For Layout Structure Extraction

In this section, we present the algorithm for layout structure extraction and classification using the bounding boxes of different entities. The flow diagram is shown in Figure 1. The algorithm first computes the connected components of black pixels on the input image and produces the bounding box for each of the connected components. Next, the algorithm performs the horizontal and vertical projections of these bounding boxes. The projection profiles of the bounding boxes are then analyzed to decompose the document image into a set of rectangular zones. Inside each zone, the connected component entities are classified into large component, small component, vertical or horizontal line, and noise, according to their sizes and locations on the page. The projection profiles of the small connected component bounding boxes are analyzed to extract text-line and word entities. Each zone is labeled as textual or non-textual according to the distribution of text-line and word entities within the zone. By finding the peaks from the vertical projection profile of word bounding boxes, the tabular structure can be identified. Within the text zone, the neighboring text-lines are merged to form text-blocks based upon the inter-text-line spacing statistics and also based on the changes in the text-line justification. The non-textual zones are further classified as line-drawing or halftone by computing the black pixel density.

3.1 Page Segmentation Using Bounding Boxes of Connected Components

The document page segmentation roughly divides the image into a list of zones. It can be accomplished by analyzing the spatial configuration of connected components in a document image. We choose to work with the bounding boxes of the connected components rather than the pixels. The *bounding box* of a connected component is defined to be the smallest upright rectangle which circumscribes the connected component. The bounding box projection approach has many advantages over the pixel projection approach. It is less computationally intensive. It is possible to infer from projection profiles how bounding boxes (and, therefore, primitive symbols) are aligned and/or where significant horizontal and vertical gaps are present.

Now we briefly describe the page segmentation algorithm in a step-by-step manner. A detailed description of the algorithm may be found in [2]. The input of the algorithm is a binary document image. We assume that the input document image has been cor-

Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV '96)
0-8186-7620-5/96 $10.00 © 1996 IEEE

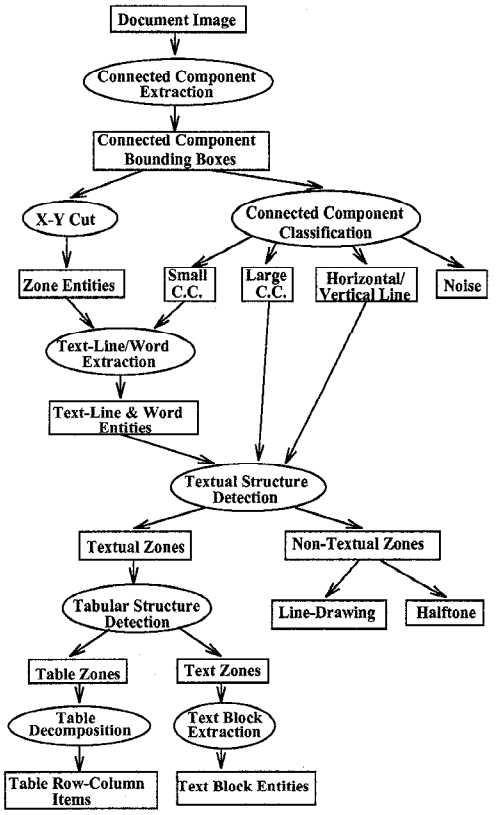


Figure 1: Flow of layout structure extraction and classification algorithm.

rectly deskewed.

Obtaining bounding boxes of connected components

A connected-component algorithm [3] is applied to the black pixels in the binary image to produce a set of connected components. Then, for each connected component, the associated bounding box is calculated. A bounding box can be represented by giving the coordinates of the top-left and the bottom-right corners of the box. Each bounding box is considered as a smallest entity on the page.

Figure 2(a) shows a segment of an English document image (taken from the UW English Document Image Database I, page id "L006SYN.TIF") and Figure 2(b) shows the bounding boxes produced in this step.

Projections of bounding boxes

Analysis of the spatial configuration of bounding boxes can be done by projecting them onto a straight line. Since paper documents are usually written in the horizontal or vertical direction, projections of bounding boxes onto vertical and horizontal lines are of particular interest. For brevity, the projection of bounding boxes onto a vertical straight line is called horizontal projection. The vertical projection is defined analogously. By projecting bounding boxes onto a line, we mean counting the number of bounding boxes orthogonal to that line. Hence, a projection profile is a frequency distribution of the bounding boxes on the projection line.

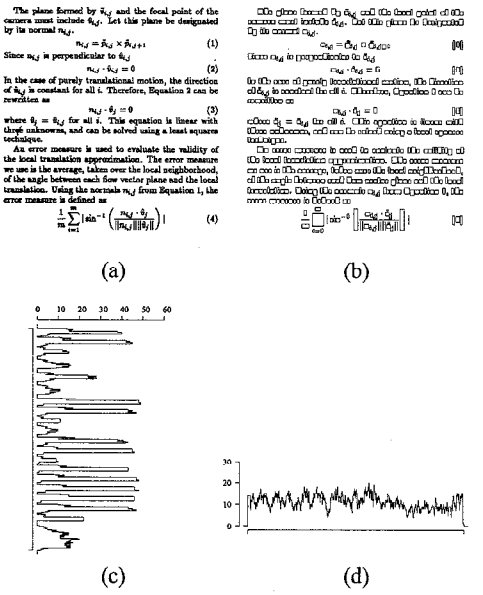


Figure 2: (a) an English document, (b) bounding boxes of connected components of black pixels, (c) horizontal projection profile, (d) vertical projection profile.

Figures 2(c) and 2(d) show the horizontal and vertical projection profiles of the bounding boxes in Figure 2(b).

Recursive X–Y cut

A document image can be segmented using a recursive X–Y cut procedure based on bounding boxes. At each step, the horizontal and vertical projection profiles are calculated. Then a zone division is performed at the most prominent valley in either projection profile. The process is repeated recursively until no sufficiently wide valley is left.

3.2 Classification of Connected Component Entities

Connected components within each zone are determined to be large connected components, small connected components, horizontal or vertical lines, or noise, according to the physical properties of the components.

If the width/height ratio of a connected component

Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV '96)
0-8186-7620-5/96 $10.00 © 1996 IEEE

is lager than a certain threshold, it is labeled as a horizontal or vertical line. The graph and figure regions usually include irregular lines, curves or shapes which form connected components with a larger size than those of individual characters. The table usually has rectangular boxes or separate lines and characters. Some text zones, such as underlined text, text with an enclosing frame, include big connected components. We label each connected component as large component if its size (either in the horizontal direction or in the vertical direction) is larger than a threshold, otherwise, a small component. The threshold size is empirically determined based on the size of the characters in the text. Some components are considered as noise if their sizes are extremely large or small, or they are very close to the boundary of page. These attributes of connected components will be used in the following structure extraction steps.

3.3 Text-Line and Word Extraction

Inside each zone generated from the page segmentation process, we analyze the projection profile of bounding boxes to extract text-lines. During this step, we only compute the projection profile of small connected components. The spatial configuration of the bounding boxes within each of the extracted text-lines is analyzed to extract words. By analyzing the distribution and alignment of text-lines and words within a zone, we can classify the zone as textual (text or table) or non-textual.

Extraction of text-lines

From Figure 2(c) and 2(d), it is clear that the text-lines can be extracted by finding the distinct high peaks and deep valleys at somewhat regular intervals in the horizontal projection profile. Since the bounding boxes are represented by a list of coordinates of the two corner points, the bounding boxes of text-line entities are easily extracted. The result is shown in Figure 2(f). Other important features, such as frequency distribution of text-line heights and inter-text-line spacings, can also be deduced from the horizontal projection profile.

Extraction of words

The vertical projection profile for each text-line is computed. The algorithm considers each such profile as a one-dimensional *gray-scale image*, and thresholds it at 1 to produce a binary image. During the binarization a symbol (or a broken symbol) with multiple bounding boxes may be merged into one. Consequently, adjacent symbols whose bounding boxes overlap each other are also merged. However, this will not cause any problem as we merge symbols' bounding boxes to form words. Figure 2(e) shows projection profiles within text-lines.

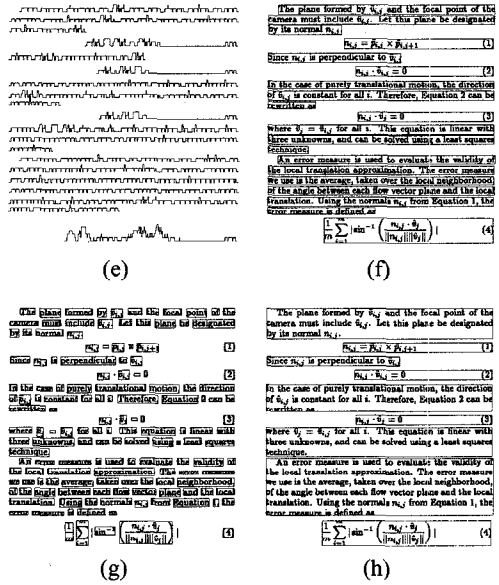


Figure 2: (e) vertical projection profiles, (f) text-line bounding boxes, (g) word bounding boxes, (h) text-block bounding boxes.

The binarization is followed by a morphological closing with a structuring element of appropriate length to close spaces between symbols, but not between words. The length is determined by analyzing the distribution of the run-lengths of 0's in the binarized profile. In general, such a run-length distribution is bi-modal. One mode corresponds to the inter-character spacings within words, and the other to the inter-word spacings. The bottom of the valley between these modes gives the desired structuring element length. If there are more than two types of spacing, this method can be recursively applied to the extracted segments. Figure 2(g) shows the results of this step.

Textual and non-textual classification

According to the properties of extracted text-line and word entities, we classify each zone as textual (text or table) or non-textual (line-drawing or halftone). The properties are the text-line density ratio (total text-line area/zone area), text-line height ratio (total text-line height/zone height), text-line height variance, and justification of text-lines (left, right, center, or justified). The justification of text-lines can be found by detecting the peaks from the vertical projection profile of text-line bounding boxes. The textual zones usually have a regular alignment of text-lines.

Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV '96)
0-8186-7620-5/96 $10.00 © 1996 IEEE

3.4 Detection of Table Structure

For each textual zone, we detect if it has a tabular structure. A table is a systematic arrangement of data usually in rows and columns for ready reference. The LaTeX tabular environment produces a box (visible or invisible) consisting of a sequence of rows of items, aligned vertically in columns. Each column has a list of item which are left-aligned, right-aligned, or centered. A single item can span a number of columns.

The tabular structure can be identified by analyzing the projection profile of word bounding boxes. In this, our method is different from [6], which detects peaks from a white space density graph. A table image is shown in Figure 3. We extract the text-line and word boxes using the method described in Section 3.3. From Figure 3(a) and 3(b), it is clear that the table columns and column separators can be detected by finding the distinct high peaks and deep valleys in the vertical projection profile of word boxes. For each textual zone, if such a structure is found, the zone is a table; otherwise, it is not a table. We then detect the row and column structure of the table by applying the X-Y cut algorithm based on word bounding boxes (See Figure 3(c)).

The itemized list is another example of tabular structure. This algorithm can be applied to detect the list structure, too.

3.5 Extraction of Text Blocks

For all textual zones which do not have tabular structure, we merge text-lines into text block. The beginning of a text block, such as a paragraph, math zone, section heading, etc., are usually marked either by changing the justification of the current text-line or by putting extra space between two text-lines: One from the previous paragraph and the other from the current one or by changing text height. So when a significant change in text-line heights, inter-text-line spacings, or justification occurs, we say that a new text block begins. The distributions of text-line heights and inter-text-line spacings together with the horizontal projection profile give us cues for text block segmentation. Figure 2(h) shows the text block bounding boxes for our example text. The font size can be estimated roughly from the text-line height. In the future, we will try to do the font detection, font style detection and other style information extraction. Then we can classify each text block into paragraph, section heading, title, etc., from these information.

For the non-textual zones, we compute the black pixels density d. If d is larger than a certain threshold, the zone is classified as halftone; otherwise, linedrawing.

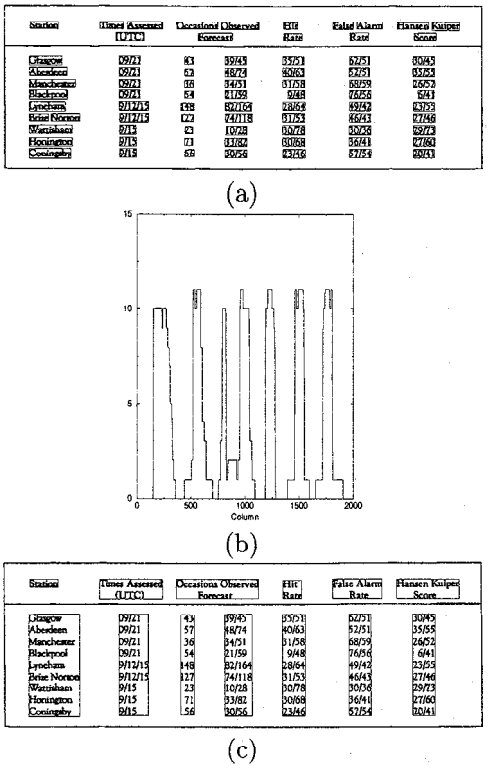


Figure 3: (a) a table image and the extracted word bounding boxes, (b) vertical projection profile of word boxes, (c) table column items.

4 Experimental Results

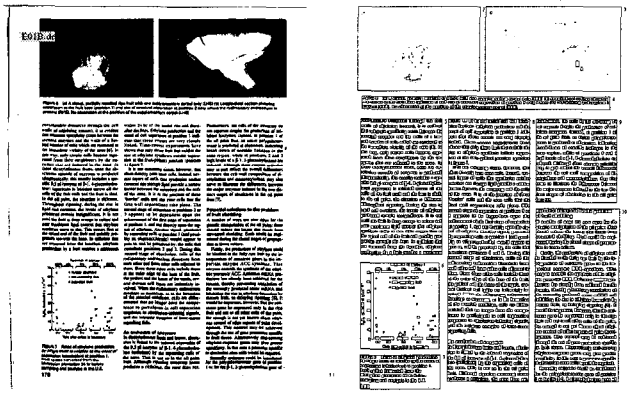
The method presented in this paper is part of a complete document understanding system we are currently developing in the Intelligent Systems Laboratory at the University of Washington. Experiments on layout structure extraction and classification were carried on several hundred document pages taken from UW English Document Image Database I, II and III [5]. The results are satisfactory on the scientific/technical documents with Manhattan page layout.

A systematically evaluation is important and necessary. So far, we have only done a qualitative evaluation. We are developing a quantitative performance evaluation and characterization scheme for the layout analysis and for the whole document image understanding system. The performance of our method on the large data sets will be reported in [4].

An example of layout structure extraction process on a document image (E01BBIN.TIF from UW-I database) is shown in Figure 4. The results of connected component extraction and zone segmentation are shown in Figure 4(b). The different structures:

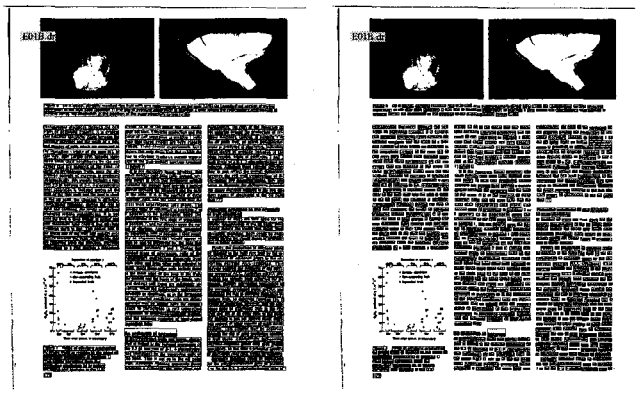
Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV '96)
0-8186-7620-5/96 $10.00 © 1996 IEEE

text, line-drawing, and halftone, are correctly identified. Figure 5(c) and (d) show the results of text-line and word entities extraction.



(a) (b)

Figure 4: (a) a document image, (b) the connected component bounding boxes and extracted zone boxes.



(c) (d)

Figure 5: (c) extracted text-line entities, (d) extracted word entities.

5 Discussion

We have presented a method for using the bounding boxes of different entities to construct hierarchical layout structure and to classify nodes in the hierarchy according to their contents. The experiment results have shown that the method is applicable to a variety of documents. In contrast to the previous approaches, once the connected components are extracted, there is no need to access the pixels of the document image.

A small number of rules are introduced to the detection of layout structure. For each rule, a reasonable threshold value has to be determined. One approach is to require the user to specify the value using his knowledge about the data. Another approach is to have the algorithm statistically learn the value by analyzing the samples which have been correctly identified in the database. A third solution is to have the algorithm calculate the needed thresholds by gathering the statistical data from the entities which have already been extracted. For example, the threshold value for the word and character spacing can be dynamically computed from the distribution of the connected component spacing within each text-line.

References

[1] A. Antonacopoulosi and R.T. Ritchings, "Representation and Classification of Complex-Shaped Printed Regions Using White Tiles," *Proc. 3rd Int. Conf. on Document Analysis and Recognition,* pp. 1132-1135, Montreal, 1995.

[2] J. Ha, R.M. Haralick, and I.T. Phillips, "Document Page Decomposition using Bounding Boxes of Connected Components of Black Pixels," *Document Recognition II, SPIE Proceedings,* Vol. 2422, pp. 140–151, San Jose, February 1995.

[3] R.M. Haralick and L.G. Shapiro, *Computer and Robot Vision,* Volume I, Addison-Wesley, 1992.

[4] J. Liang, I.T. Phillips, and R.M. Haralick, "A Performance Evaluation Protocol for Document Layout Analysis," *ISL Technical Report,* U. of Washington, 1996.

[5] I.T. Phillips, S. Chen and R.M. Haralick, "CD-ROM English Document Database Standard," *Proc. 2nd Int. Conf. on Document Analysis and Recognition,* pp. 478-483, Japan, 1993.

[6] D. Rus and K. Summers, "Using White Space for Automated Document Structuring," *Proceedings of the Workshop on Principles of Document Processing,* Seeheim, 1994.

[7] S. Tsujimoto and H. Asada, "Major Components of a Complete Text Reading System," *Proc. of the IEEE,* Vol. 80 no. 7, pp. 1133-1149, July 1992.

[8] F.M. Wahl, K.Y. Wong, and R.G. Casey, "Block Segmentation and Text Extraction in Mixed Text/Image Documents," *Computer Graphics and Image Processing,* Vol. 20, pp. 375-390, 1982.

[9] S.Y. Wang and T. Yagasaki, "Block Selection: A Method for Segmenting Page Image of Various Editing Styles," *Proc. 3rd Int. Conf. on Document Analysis and Recognition,* pp. 128-135, Montreal, 1995.

Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV '96)
0-8186-7620-5/96 $10.00 © 1996 IEEE