Document Page Decomposition by the Bounding-Box Projection Technique

Jaekyu Ha & Robert M. Haralick

Dept. of Electrical Engineering, FT-10
University of Washington
Seattle, WA 98195

Ihsin T. Phillips

Dept. of Computer Science
Seattle University
Seattle, WA 98122

Abstract

This paper describes a method for extracting words, textlines and text blocks by analyzing the spatial configuration of bounding boxes of connected components on a given document image. The basic idea is that connected components of black pixels can be used as computational units in document image analysis. In this paper, the problem of extracting words, textlines and text blocks is viewed as a clustering problem in the 2-dimensional discrete domain. Our main strategy is that profiling analysis is utilized to measure horizontal or vertical gaps of (groups of) components during the process of image segmentation. For this purpose, we compute the smallest rectangular box, called the bounding box, which circumscribes a connected component. Those boxes are projected horizontally and/or vertically, and local and global projection profiles are analyzed for word, textline and text-block segmentation. In the last step of segmentation, the document decomposition hierarchy is produced from these segmented objects.

1 Introduction

The printing process is the transformation of the logical hierarchy of a given document into the physical hierarchy. The process must follow the set of rules or protocols which prescribe the physical document layout requirements at the time of production. The requirements may include the font type, size and style for each symbol, the column format (including the number of columns and column width), the header, the footer and margin dimensions. Also, there are also intrinsic spacing protocols for the symbols and words as well as for textlines, text blocks and text columns. In almost all cases, spacings between symbols are much smaller than spacings between words within the same printed document. Similarly, spacings between textlines are smaller than spacings between text-blocks and/or text-columns. This tendency has been used as prior knowledge in most OCR and document image analysis algorithms.

This paper describes a technique for extracting words, textlines and text blocks by analyzing the spatial configuration of the bounding boxes of symbols in a given document page. In particular, the 'bounding boxes' of the connected-components of black pixels are used as the basis of such extractions.

The remainder of this paper is organized as follows: In Section 2, we describe the decomposition algorithm in a step-by-step manner. Section 3 discusses experiments on the UW English Document Image Database I. The concluding remarks are given in Section 4.

2 Text Zone Delineation

Now we describe the page decomposition algorithm in a step-by-step manner. We assume that the input document image has been correctly deskewed.

2.1 Bounding Boxes of Connected Components

Let I denote the input binary image. A connected component analysis algorithm [2] is applied to the foreground region of I to produce the set of connected components. Then, for each connected component, its associated bounding box – the smallest rectangular box which circumscribes the component – is calculated. A bounding box can be represented by giving the coordinates of the upper left and the lower right corners of the box.

Figure 1(a) shows a segment of an English document image (taken from the UW English Document Image Database I, page id "L006SYN.TIF") and Figure 1(b) shows the bounding boxes produced in this step. Note that, the number of bounding boxes are

0-8186-7128-9/95 $4.00 © 1995 IEEE

Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE