Journal of Computer Science and Cybernetics, V.32, N.2 (2016), 153–167
DOI no. 10.15625/1813-9663/32/2/8546

A HYBRID PARAGRAPH-LEVEL PAGE SEGMENTATION

HA DAI TON[1], NGUYEN DUC DUNG[2]

[1]*Ha Long Gifted High School, Quang Ninh Province, Viet Nam*
[2]*Institute of Information Technology, Vietnam Academy of Science and Technology;*
[1]*hadaiton83@gmail.com;* [2]*nddung@ioit.ac.vn*

Abstract. Automatic transformation of paper documents into electronic forms requires geometry document layout analysis at the first stage. However, variations in character font sizes, text-line spacing, and layout structures have made it difficult to design a general-purpose method. Page segmentation algorithms usually segment text blocks using global separation objects, or local relations among connected components such as distance and orientation, but typically do not consider information other than local component's size. As a result, they cannot separate blocks that are very close to each other, including text of different font sizes and paragraphs in the same column. To overcome this limitation, we proposed to use both separation objects at the whole page level and context analysis at text-line level to segment document images into paragraphs. The introduced hybrid paragraph-level page segmentation (HP2S) algorithm can handle difficult cases where the purely top-down and bottom-up approaches are not sufficient to separate. Experimental results on the test set ICDAR2009 competition and UW-III dataset show that our algorithm boosts the performance significantly comparing to the state of the art algorithms.

Keywords. Page segmentation, text-lines, homogenous regions, separation objects, paragraphs, evaluation result.

1. INTRODUCTION

Document layout analysis is one of the main components of any OCR (optical character recognition) system. The task of structural analysis includes automatically detecting image zones on a document image (physical structure analysis) and classifying them into different types such as: text, images, tables, header, footer... (logical structure analysis). The results of page segmentation are used as an input to the process of recognition and automatic data entry of image processing systems in general.

Compared with the analysis of the logical structure analysis, the physical structure analysis (page segmentation) has attracted more attention due to the diverse and complex structures of different types of document. Not only the specific types of document (books, newspapers, magazines, reports...) but also the other factors of a page such as editors and font size, layout, alignment constraints... affect detection and segmentation accuracy of the algorithm.

Document layout analysis algorithms are primarily divided based on their order of processing into three approaches: bottom-up, top-down and hybrid.

Bottom-up algorithms are both the oldest [17] and more recently published [6, 7, 13] algorithms. They classify small parts of the image (pixels, groups of pixels, or connected

components), and gather those of the same type together to form regions. The key advantage of bottom-up algorithms is that they can handle arbitrarily shaped regions with ease (rectangular or non-rectangular). But the fact that they are really sensitive to the measure used to form higher-level entities is the key disadvantage; this often leads to error of over-segmentation in the page with many changes in font sizes and styles, especially in the titles with large or extra-large font size.

Top-down algorithms, e.g. [5, 12], cut the image recursively in vertical and horizontal directions along white spaces that are expected to be column boundaries or paragraph boundaries. Although top-down algorithms have the advantages that they start by looking at the largest structures on the page, they are not really able to handle free-formed layouts that often occur in magazines, such as non-rectangular regions and cross-column headings that blend seamlessly into the columns below.

A third type of algorithm [14] is based on bottom-up method to find delimiters such as rectangular whitespaces (the Fraunhofer method [2]), tab-stops (the Tab-Stop method [16]). This reduces top-down structure. And then, it is to use a combination of bottom-up method and top-down structure to detect text regions. So, these approaches can overcome over-fragmentation error of bottom-up algorithms as well as perform better with non-rectangular regions. However, it is not trivial to detect exactly delimiters by many following reasons: text regions are very close to each other, text regions are not left aligned or right aligned, large space of connected components is also the reason to lead misidentified separations,... The current hybrid algorithms almost failed when facing to these difficulties. Besides, variations in character font sizes, text-line spacing, and layout structures have made them difficult to group connected components into text blocks.

In this paper, the authors propose a new page segmentation algorithm, called Hybrid Paragraph-level Page Segmentation (HP2S), that can not only handle text blocks of any shape and variation in font sizes, but also split document images at the level of paragraph. Firstly, character-like connected components are grouped into text-lines. Before merging the text-lines into regions, under-segmented and over-segmented text-lines are treated specially using local context information like size of characters and global white columns. The processed text-lines are then grouped into homogeneous text-regions of similar font-size characters. To reduce the under-segmentation errors, the text-lines located at the beginning and ending of each paragraph are found. These local separation text-lines are finally used in combination with the global white columns to split the discovered text-regions into paragraphs. As a result, HP2S can segment page images of very hard and complicated structure. Experiments on the test set ICDAR2009 competition and UW-III dataset show that HP2S algorithm achieves the highest performance compared to the state of the art algorithms.

This paper is organized as follows. Section 2 describes in detailed the HP2S algorithm. Section 3 gives experimental results and analysis. Finally, conclusion is given in section 4.

2. PAGE LAYOUT ANALYSIS VIA SEPARATION OBJECT DETECTION

The proposed page segmentation algorithm is based on the mix of the bottom-up approach and the top-down approach. Figure 1 outlines two stages and main processing steps of the proposed HP2S algorithm. The first stage aims to group connected components into homogeneous text regions. The second stage divides homogeneous text regions into para-