Chapter 2
Efficient Word Segmentation and Baseline Localization in Handwritten Documents Using Isothetic Covers

Mousumi Dutt
Bengal Engineering & Science University, India

Arindam Biswas
Bengal Engineering & Science University, India

Aisharjya Sarkar
Bengal Engineering & Science University, India

Partha Bhowmick
Indian Institute of Technology, Kharagpur, India

Bhargab B. Bhattacharya
Indian Statistical Institute, India

ABSTRACT

Analysis of handwritten documents is a challenging task in the modern era of document digitization. It requires efficient preprocessing which includes word segmentation and baseline detection. This paper proposes a novel approach toward word segmentation and baseline detection in a handwritten document. It is based on certain structural properties of isothetic covers tightly enclosing the words in a handwritten document. For an appropriate grid size, the isothetic covers successfully segregate the words so that each cover corresponds to a particular word. The grid size is selected by an adaptive technique that classifies the inter-cover distances into two classes in an unsupervised manner. Finally, by using a geometric heuristic with the horizontal chords of these covers, the corresponding baselines are extracted. Owing to its traversal strategy along the word boundaries in a combinatorial manner and usage of limited operations strictly in the integer domain, the method is found to be quite fast, efficient, and robust, as demonstrated by experimental results with datasets of both Bengali and English handwritings.

DOI: 10.4018/978-1-4666-2928-8.ch002

INTRODUCTION

A handwriting portrays the characteristics of an individual, and hence has been studied in numerous disciplines including experimental psychology, neuroscience, engineering, anthropology, forensic science, etc. (Plamondon, 1993; Plamondon & Leedham, 1990; Simner et al., 1994, 1996; Galen & Morasso, 1998; Galen & Stelmach, 1993; Wan et al., 1991). The analysis of handwritings has been quite important in recent times with the advancements of document digitization (Hole & Ragha, 2011; Saba, 2011; Terrades, 2010; Zhu et al., 2009), biometric authentication (Henniger & Franke, 2004; Hoque et al., 2008; Low et al., 2009; Makrushin, 2011; Schimke et al., 2005; Vielhauer, 2006; Vielhauer & Scheidat, 2005), forensic science (Franke & Köppen, 2001; Máadeed et al., 2008; Mahmoudi et al., 2009; Pervouchine et al., 2008), etc. The result of analysis strives to interpret, verify, and recognize a particular handwritten document. The most difficult problem in the area of handwriting recognition is segmentation of cursive handwriting. The infinitude of different types of human handwritings amidst the similarities in the shapes of different characters renders the problem even more difficult. Hence, over the last few years, various works have been presented for specific domains, e.g., Bengali character recognition (Majumdar & Chaudhuri, 2007; Parui et al., 2008), text line identification (Chaudhuri & Bera, 2009), numeral recognition (Bhattacharya & Chaudhuri, 2009), check sorting (Gorski et al., 1999), address reading (Srihari & Keubert, 1997), tax reading (Srihari et al., 1996), office automation (Gopisetty, 1996), automated postal system (Vajda et al., 2009), etc.

Handwriting recognition techniques are based on either holistic or analytic strategies. In the holistic method, a top-down approach is employed where the whole word is recognized by comparing its global features against a limited size lexicon (Guillevic & Suen, 1998). On the other hand, analytic strategies adopt the bottom-up approach starting with characters, strokes, etc., eventually producing the meaningful text (Wang & Jean, 1994; Mohamed & Gader, 1996; Mao et al., 1998; Kim & Govindaraju, 1997). Clearly, in connection with handwriting recognition, it is important to extract/segment the words in a cursive writing such that the task of segregating the individual characters and strokes may be taken up. The segmented regions may be found out from the peaks of the projection profile of a gray-level image (Lee et al., 1996). Proper baseline extraction is important to segment out words correctly. To avoid the problems arising out of ascending and descending portions of the characters, information about the upper and the lower baselines are necessary. In general, the baselines are extracted from the projection profile (Guillevic & Suen, 1998).

As per the existence practice, the text line segmentation is usually done by considering a subset of the connected components in a document image (Louloudis et al., 2009). Word segmentation is achieved using the distinction of inter- and intra-word gaps using a combination of two different distance metrics.

To overcome the disadvantage of different distance measures in word segmentation, a gap metric based on the average distance is used for word segmentation (Huang & Srihari, 2008). A number of works have been reported in the literature for line and word segmentation in other languages, e.g., Chinese, Arabic, etc. Chinese word segmentation has various applications on Chinese text processing (Haizhou & Baosheng, 1998). The algorithm for detection of straight or curved baselines for Arabic handwritten text can be applied on online handwriting or off-line handwritten writing (Boubaker et al., 2009). A method for precise identification of ascending or descending parts of the words has been proposed using lexicon based search in Aida-zade & Hasanov (2009).

In our work, we have devised a novel method to segment out the words in a cursive handwriting by using the outer isothetic covers of the