Chapter XV

Mathematical Morphology-Based Automatic Restoration and Segmentation for Degraded Machine-Printed Character Images

Shigueo Nomura, Kyoto University, Japan
Keiji Yamanaka, Federal University of Uberlândia, Brazil
Osamu Katai, Kyoto University, Japan
Hiroshi Kawakami, Kyoto University, Japan
Takayuki Shiose, Kyoto University, Japan

ABSTRACT

This chapter presents a morphological approach (AutoS) for automatic segmentation with feature vector extraction of seriously degraded machine-printed character images. This approach consists of four modules. The first detects and segments natural pitch characters based on the vertical projection of their binary images. The second uses an algorithm based on the vertical projection and statistical analysis of coordinates to detect fragments in broken characters and merge them before the eventual segmentation.
The third employs a morphological thickening algorithm on the binary image to locate the separating boundaries of overlapping characters. Finally, the fourth executes a morphological thinning algorithm and a segmentation cost calculation to determine the most appropriate coordinate at the image for dividing touching characters. By the automatic running of a suitable segmentation module for each problem, the AutoS approach has been robust, flexible and effective in obtaining useful and accurate feature vectors concerning degraded machine-printed character images.

INTRODUCTION

Mainly in the case of degraded images, character image segmentation is fundamental to such user interfaces as automatic reading systems based on optical character recognition (OCR), which perform on individual characters. Segmentation consists of an essential step for avoiding incorrect character recognition due to inappropriate input data, such as degraded images. Though problems like broken and touching characters are responsible for the majority of errors in automatic reading of machine-printed text (Casey & Nagy, 1982), segmentation of degraded character images is all too often ignored in the research community. The processing required for information capturing and extraction is still in its infancy (Marinai, Gori, & Soda, 2005).

In particular, the automation of character image segmentation and feature vector extraction is an essential stage for processing degraded images acquired from real-time imaging systems and used in our experiments.

Since most OCR systems already work with binary images, our proposed approach falls within the scope of these images.

RELATED WORKS AND TOPIC RELEVANCE

Essentially, existing techniques (Jung, Shin, & Srihari, 1999; Lu, 1995; Tan, Huang, Yu, & Xu, 2002; Taxt, Flynn, & Jain, 1989) are based on heuristics, whereby the text contains horizontal lines and the characters are proportionally sized and uniformly well-separated. In the case of degraded characters, recognition-based segmentation algorithms (Arica & Yarman-Vural, 1998; Lee & Kim, 1999; Nomura, Michishita, Uchida, & Suzuki, 2003), which require a recognizer to validate the segmentation process, are used. Under this concept, if the character is recognized then the segmentation is accepted, otherwise the segmentation is re-applied. These recognition-based segmentation algorithms are not only time consuming, but their outputs are heavily dependent on the character recognition process. In other words, existing techniques have not yet been able to simultaneously and appropriately treat such fragmenting, touching and overlapping character problems. Furthermore, they are not able to accurately extract feature vectors from real-world degraded images. On the other hand, automatic reading systems such as Anagnostis (Bourbakis, 1998) scan the text pages in order to extract characters in the form of binary strings.

Thus, approaches for providing those image restoration and segmentation abilities to the automatic reading systems still consist of a relevant topic in the field of segment processing.