AUTOMATIC TEXT EXTRACTION, REMOVAL AND INPAINTING OF COMPLEX DOCUMENT IMAGES

YEN-LIN CHEN
Department of Computer Science and Information Engineering
National Taipei University of Technology
No. 1, Sec. 3, Chung-hsiao E. Rd., Taipei 10608, Taiwan
ylchen@csie.ntut.edu.tw

Received August 2010; revised January 2011

Abstract. This study proposes a novel system that extracts text lines and restores text-removed images from various types of complex document images of mixed and overlapping text, graphics and pictures, which may contain text lines with different illumination levels, sizes and font styles. The proposed system first decomposes the document image into distinct object planes to separate homogeneous objects, including textual regions of interest, non-text objects (such as graphics and pictures) and background textures. A knowledge-based text extraction and identification method accurately detects and extracts text lines with different characteristics from each object plane. Afterwards, a computationally efficient text removal and inpainting process, based on an effective adaptive inpainting neighborhood adjustment scheme, is applied to the obtained text-line regions to produce a clear non-text restored background image. Experimental and comparative results demonstrate that the proposed system can provide accurate extraction of text-line regions of interest with diverse illumination levels, sizes and font styles from various complex compound document images, and can effectively and computationally produce clear and well-preserved non-text background images with satisfactory visual quality for further applications.

Keywords: Document image processing, Text extraction, Text removal, Image inpainting, Complex compound document images

1. Introduction. The extraction and manipulation of objects from digital media is essential for understanding, editing and retrieving information contained in digital media. Textual objects from various forms of digital media consist of extremely crucial information. Examples of these applications include the extraction and analysis of text information, text categorization, semantic and syntax analysis, optical character recognition, document retrieval, editing, compression, archiving and publishing [1-7]. Despite the recent adoption of information distributed via electronic media, paper-based published media, such as documents, books and magazines, continue to be popular. However, because paper-based media are less convenient than electronic media when archiving, modifying and retrieving, the practical applications of object extraction, analysis and manipulation for both digitalized paper-based documents and books receive considerable attention. Of the aforementioned applications, textual information extraction and processing is the most essential task in document image analysis and digital archiving. Conversely, textual objects in digitalized document images are not always meaningful and sometimes do not contain useful information for users. An example of this includes indirect advertisements. Removing unnecessary and indirect advertisements and preserving valuable image content are also critical problems for digital archiving and publishing applications. As for the issues of text removal and restoration of document images, due to the large number