Accelerating Production Book Scanning

Joseph S. Czyszczewski, James T. Smith, and Hong Li
International Business Machines
Boulder, Colorado

Abstract

The global market created by e-commerce enables sales growth for low volume book titles. This provides an opportunity for digital production printing because offset printing remains prohibitively expensive and causes titles to go out of print. While the printing process in end-to-end digital solutions is cost effective, the problem is that most low volume titles only exist in hardcopy and digitization remains slow and difficult. We describe a streamlined and scalable process which enables end-users to quickly and easily scan black and white book blocks. This advance in book scanning sets the future direction for production operations by accelerating the process while maintaining quality and consistency.

Introduction

Digital production books are generally printed on high volume 600dpi bi-level toner based printers. Offset printed hard copy originals are scanned, and halftone images are segmented and de-screened to minimize moiré patterns in the printed copy without blurring text. The resulting grayscale images are generally converted to halftones in the scanner and the digital books are stored as 600dpi bi-level images.

While error diffusion halftones provide printer independent images, print quality is not adequate for books. As a result, the segmented grayscale images are generally re-screened with printer specific halftones in the scanner. While this approach is automated, optimizes image quality, and optimizes print file processing, it is printer specific. The problem with printer specific books is they must be rescanned to support new printers or leverage new print technologies.

Printer independence and optimal image quality require original halftone images to be stored as gray scale. Faithful reproduction for printing, rather than repurposing, allows text and line art to continue to be stored as bi-level. The problem with creating this mixed format in production scanning is that original halftone image segmentation and de-screening are generally manual processes which result in significant cost increases and production risks. A secondary problem with this approach is additional print file processing due to larger files and the need to halftone books before printing.

Production scanning thousands of books per month requires both automation and device independence. Automation is required to minimize manual processing costs and production risks and must support customization to integrate with scanned book acquisition and manufacturing processes.

Device independence is required for scanning and printing. Scanner independence is required to enable use of general purpose high volume scanners rather than specialized devices. Printer independence is required to enable use of general purpose high volume printers and leverage quality and performance improvements from new print technologies without rescanning books.

Production Book Scanning

The process described here begins with removal of the book spine followed by scanning pages on general purpose high volume scanners with dual optics for single pass operation. Operating scanners in parallel provides scalability as well as redundancy for high availability.

The scanners are controlled with a production scanning software interface to enable data streaming, process integration, and user interface customization. Scanned pages are processed with a software based image processing toolkit. An open platform with standard and custom plug-in filters enables efficient process optimization and integration. A pipeline architecture is required to deliver the performance advantages of data streaming for in-line or batch operation.

The scanned page stream is automatically segmented into multiple streams based on content type. Each content stream is processed independently and then re-joined into a page stream which is stored in a mixed gray scale and bi-level format. The books may also be stored in a print ready format with halftones applied in a print cache based on print file processing capacity and production risk tolerance.

The black and white book blocks are printed on general purpose high volume duplex printers and color covers are printed in a parallel process. Even though the run length is as short as one copy, printing is generally done in batches to optimize production, and workflow software is used to manage batches and match covers with book blocks.

Image Processing

Since common high volume scanners generally support a maximum resolution of 300dpi, the books are scanned in 8 bit gray-scale mode to optimize image quality and a scanner calibration process is established to optimize process quality and consistency.

Each page image is first processed with a de-skew filter to improve image quality and to condition the data for automatic segmentation accuracy.

The next step is to detect scanned page edges and make horizontal and vertical alignment and cropping adjustments to compensate for scanned page registration variations. Scanning with a black background increases accuracy and is only required on one side with a scanner that has dual optics but introduces image quality variations with thin original pages.

Alignment is followed by automatic classification and segmentation of the page into three types of original content. The content is identified as text and line art, halftone images, or remainder images. The remainder images represent background or misclassifications.

Text and line art is sharpened and converted into 600dpi bi-level compressed images. Halftone images are processed by first detecting the screen frequency of the image and then applying a dynamic de-screen filter. The resulting grayscale images are compressed and stored as 8 bit 300dpi images.

The remainder is processed with a de-speckle filter to remove background noise and may include misclassifications. Misclassifications must be identified and preserved as original grayscale images for subsequent automatic or manual inspection and processing.

The final step is to reassemble the content into pages and create a mixed format file suitable for viewing, editing, and printing.

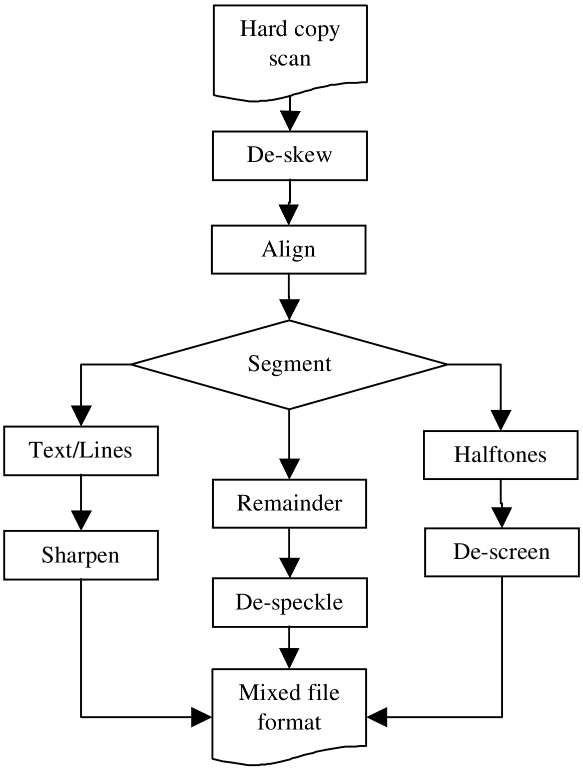

Image Processing

Misclassifications

While machine segmentation accuracy continues to improve, a variety of misclassifications may result. Examples of misclassifications in text areas are the loss of diacritic or punctuation marks. Line art used for dividers and text callouts which originate in original halftone images may also result in misclassifications. Halftone image misclassifications may result from small images such as a row of circular areas used for section breaks.

Since production scanning must be based on predictable and repeatable processes, strategies are required to handle misclassifications based on cost and quality tradeoffs. One strategy is to identify pages with misclassifications based on a quality threshold. The objective is to minimize costly proofing and manual processing based on the quality requirements of the job. For example, the threshold may be set to accept a small number of minor misclassifications for a novel while a more demanding threshold is set for technical books.

Another strategy is to automate reassembly of misclassified content. The objective is to minimize costly manual reassembly when appropriate. For example, an optional bias may be set to include misclassified content as bi-level data or as gray scale data based on the type of content on the page or the overall book.

Conclusion

The use of automatic segmentation to produce mixed gray scale and bi-level representations of hard copy books enables a breakthrough in production book scanning by balancing volume, quality, consistency, and value. The key challenge is misclassifications and ongoing research enables more complex books to be processed with less manual processing.

Acknowledgments

The authors thank their colleagues Dan Chevion, Ehud Karnin, Gerry Thompson, Asaf Tzadok, and Chai Wu in IBM Research for image processing insight and algorithms.

Biographies

Joseph Czyszczewski is a Senior Technical Staff Member with IBM's Printing Systems Division in Boulder Colorado. He joined IBM in 1977 and holds an MS degree in Electrical Engineering from the University of Texas at Austin.

Jay Smith has been a software engineer in the field of Image Processing for the last six years with the Printing Systems Division of IBM.

Hong Li is a color imaging specialist with IBM’s Printing Systems Division in Boulder Colorado. She worked at Xerox for several years before joining IBM. She holds an MS degree in Imaging Science from the Rochester Institute of Technology.