High Volume Book Block Scanning

Joseph S Czyszczewski *, James T Smith *

Keywords: Book, Digital, Workflow, Scanning, Halftone

Abstract: Offset printing is prohibitively expensive for low volume books, causing titles to go out of print. E-commerce exacerbates this problem by offering immediate and unlimited selection. While digital printing delivers low volume books that are cost effective and immediately available, digitizing hard copy for digital printing remains prohibitively expensive and slow. We describe a cost effective and scalable workflow, based on two years of research, development, and production, that enables low skill operators to digitize hundreds of book blocks per week. This workflow integrates the production consistency of manufacturing processes with original image quality techniques, culminating in superior digital output.

Introduction

Consumers get frustrated when a book they really want is out of print. Publishers are frustrated with annual waste of up to 40 percent of books due to changing customer tastes. And authors who haven't sold particularly large quantities, as well as their publishers, are frustrated by being unable to economically accommodate the demand from booksellers for only a few dozen copies. Campaign managers are frustrated by the lack of flexibility to deliver individually customized books.

Looking for a way to reduce all this frustration, companies are turning to print-on-demand technology. With this technology, even single custom copies of books can be printed economically, as customer need dictates, while avoiding the warehousing costs and waste associated with large-volume output using traditional offset printing methods. And now, a book never goes out of print, as its life cycle is extended indefinitely by storing it electronically for printing at any time. Also, books become dynamic with personalized content in each copy.

*IBM Printing Systems

Print-on-demand technology also enables companies to participate in growing demand to become part of e-commerce supply chains by accepting telephone, fax and online orders from physical and Internet-based booksellers as well as publishers. Books are typically printed and delivered to the warehouse within 24 to 48 hours of receiving the order and shipped in the next batch to the bookseller. While using this technology to print books has become practical and essential in a variety of markets, it remains complementary to traditional offset printing which delivers the highest quality at the lower cost for high volume titles.

Background

The use of print-on-demand technology for book production requires two key operations to be established. First, books must be processed into a print ready digital format. Print ready books are created from both digital and hard copy originals. Digital originals are stored in both a mixed text with continuous tone image format and a raster image format. The mixed format is preserved for output device independence with full print quality. The raster image format is used to maximize print performance and insure consistency.

Hard copy originals must be first be scanned into digital format. A separate scanning process is used for color covers and for black and white book blocks. Traditionally, book blocks are only stored in raster image format. While the traditional process is currently used in production, an enhanced image scanning process is being proven on a pilot manufacturing line and is described in this paper. The new process provides many of the advantages enjoyed with digital originals by creating continuous tone images which are stored in a mixed format with text in raster image format. Unlike digital originals, the text is maintained in raster image format to maintain true hard copy original reproduction.

The print ready books are then finalized through a proofing and approval process with publishers. When the prepress work is complete, the books are indexed and stored in a document management system. The document management system is integrated with both the ordering and printing systems.

The second key operation is printing. Printing starts with selecting books from the document management system based on orders and printing the required quantity for delivery. Printing is divided into separate processes with color covers produced by one process and black and white book blocks produced by another process. Color covers are grouped into batches to minimize startup overhead and are printed on 600DPI continuous form printers. They are then laminated in preparation for binding.

Black and white book blocks are printed multiple-up on 600DPI continuous form printers with slit-and-merge post processing. This approach enables single quantity runs to be delivered with full performance and without waste or inventory. The covers and book blocks are then merged by using bar codes which are included on the covers and book blocks to automate tracking and assembly in preparation for finishing. Matched covers and book blocks are finally bound, trimmed, and prepared for shipping.

High Volume Scanning

Print-on-demand book production is most cost effective in short run applications. As a result, developing a significant print-on-demand book business requires a high volume of titles to be processed and stored in a print ready digital format. This is in contrast to traditional book printing with offset presses where long runs are most cost effective and as a result, a smaller number of titles are required.

The high volume of titles required to support short run print-on-demand book production requires more of a manufacturing prepress operation than is justified for traditional longer run print-on-demand operations. While the required technologies and techniques are similar, they are selected, optimized, and integrated to deliver consistency and productivity while minimizing time and skill requirements. Significant investment is also made in specialized tools and in a formal process to support the manufacturing operation.

To further complicate the task of preparing books for printing, publishers generally deliver hard copy rather than digital originals. This surprising fact significantly increases the volume of scanning required to create a significant library of print ready titles. It also increases the overall cost because preparing hard copy originals requires more work than digital originals. We found that over 90% of books are submitted as hard copy originals in this market. Even books which are created with digital tools are often submitted as hard copy because the digital originals are not guaranteed to be up to date and an exact replica of the published book.

We note that as more print-on-demand book production operations are established, scanning and creation of digital print ready formats is taking place independently. As a result of not pooling this work, the potential exists for a great deal of duplicate prepress effort and with it a significant business-to-business opportunity to improve the availability and cost of print-on-demand books across the industry. This unnecessary duplication even further increases the demand for high volume scanning in order to independently develop a sufficient selection of titles and to keep the digital presses at full utilization.

Book Block Scanning

Separate processes, equipment, and operators are used to scan color covers and black and white book blocks. Covers are removed from the books and scanned with a prepress scanner. The images are then edited with common applications to adjust the size of the spine and correct defects in the original hard copy cover. The images are finally color corrected, proofed, indexed, and stored in the document management system. We found that an operator requires over one hour to edit and correct a color cover.

In parallel with the cover scanning, the spine is cut from the book blocks which average 300 pages in length and are generally six inches by nine inches in size. The pages are scanned on production black and white scanners with automatic document feeds and duplex support. Multiple book blocks are scanned in parallel and more than one is scanned in parallel by each operator. The text pages are scanned as bilevel raster images and are interpolated to 600DPI in the scanner. An experienced operator can scan over 10 books per shift with two scanners.

After the book blocks are scanned, pages with images are re-scanned and replaced at separate scan stations where the images are manually zoned and separated from the text. The original images are generally halftone images. The screen frequency of the original halftone is determined and used to descreen the image in the scanner. A new halftone and screen frequency is applied in the scanner with settings for brightness, contrast, sharpness, and other common adjustments. Images are stored as 600DPI bilevel raster images. Experience results in a collection of sets of scanner settings which minimize the amount of trial and error required to establish optimal values during the proofing cycle. We found that under 5% of pages scanned in print-on-demand operations contain images and an operator can manually zone over 100 images per shift.

With text and image scanning completed, the pages are edited and cleaned up. Cleanup is done with automated tools to deskew and despeckle the pages. Manual cleanup tools are also provided for page-to-page and front-to-back page alignment. Editing consists of using tools to erase, annotate text, and cut, copy, or paste. All tools are duplex sensitive and support batch operation to process all pages or page ranges in a book block. We found that an operator requires over one hour to edit and cleanup a book block.

Since thousands or tens of thousands of books must be scanned per year and most pages do not include images, text scanning is scaled by using multiple shifts and multiple text scanning stations operating in parallel. Operating multiple shifts is dependent on minimal skill requirements. To further support

this process the scan tools are designed to maximize speed by combining multiple steps and provide simplicity by automating steps. Also, a single scan application is used to minimize cross training and eliminate time lost by switching between applications. Even with this level of optimization, image scanning, editing, and proofing currently require higher skill levels and are only done during normal shifts.

Enhanced Image Scanning

There are several problems with the traditional process described above for scanning images in book blocks. Significant experience and proofing are required to optimize output image quality. Another problem is that storing raster images based on halftones supported by the scanner does not allow finely tuned and often more complex halftones developed for print-on-demand printers to be leveraged. A third and most costly long term problem is that the stored raster images are optimized for the print engine and technology on which the proofing process was done.

As a result of these problems, an enhanced process for image scanning has been developed. This process consists of scanning the original halftone as a gray scale zone, descreening it in the scan application, and storing the result as a continuous tone image. The descreen tool automatically determines the screen frequency of the original halftone and applies an appropriate filter to produce a continuous tone image while minimizing artifacts. The resulting page is stored as a layered TIFF file where one layer contains a bilevel representation of the text page and additional layers contain continuous tone images. While the resulting digital files are not as compact as in the traditional process, the increase is small because the percentage of images in books is small and the area of the images is small. On balance, the benefits of this enhanced image scanning process far outweigh the small increase in storage size.

Storing book blocks as a mix of bilevel text and continuous tone images provides a variety of advantages which address the problems identified above with the traditional process. Storing images as continuous tones enable the print-on-demand system to deliver true digital quality. By storing continuous tone rather than bilevel images, the optimized halftones delivered with the digital printer are used for printing rather than the halftones supported by the scanner. Capturing a gray scale representation of the original halftone rather than optimizing a variety of scanner settings through proofing minimizes the scanning skill and proofing time required for scanning images. Storing continuous tone images also minimizes problems with image artifacts created when the resolution of the original, scanner, and printer are not optimized.

Using this continuous tone scanning process also enables output device independence. Device independence provides investment protection which is essential due to the volume of books scanned and their long term nature. The problem with the traditional approach of storing raster images is that they are device dependent. While this may be acceptable for the short term, it becomes an issue in the long term when new models of output engines replace existing engines or new engine technologies are adopted.

A process based on continuous tone images also improves process consistency which is a key measurement from publishers. By shifting halftoning of images from the scanner to the output device, process variations can be offset on a daily basis without the need to re-scan the stored images. Finally, shifting rasterization from the scanner to the output device enables repurposing of the data for additional applications such electronic books.

Conclusion

While the approach described here supports scanning thousands or tens of thousands of books per year and printing tens or hundreds of thousands of books per year, research continues in order to achieve greater saleability. One technique used to increase book scanning rates is automation. Currently, integrated tools are provided to enable page alignment and image zoning, but these operations remain manual. The production process is being studied and research is underway to develop tools for automatic page alignment and automatic image zoning.

Higher volumes are also supported by advances in performance and ease of use of production scanners. While new scanners continue to be evaluated and upgraded as justified to improve productivity and image quality, new scanners are also being evaluated to provide an alternative to the proposed low resolution gray scale scanning of original halftone images. The production process is being studied and research is underway to develop a high resolution bilevel scanning alternative to further improve performance and ease of use.

Finally, work continues in broadening these tools as the print-on-demand market matures. One key growth area is in electronic books. Key areas of research for electronic books include OCR and content tagging tools and techniques which enable repurposing of printable content for efficient viewing and searching.

In summary, the high volume print-on-demand manufacturing workflow described in this paper was developed to keep books from going out of print. The problem is that even with optimized processes, traditional scanning techniques are not adequate to fully protect out of print and low volume titles. The key issues are that the traditional approach does not provide device

independence and the true digital image quality. As a result, an enhanced high volume image scanning approach was developed.

The individual technologies and techniques described in this paper are not original. They have been used by prepress and print shops for many years. However, the selection, optimization, and integration of these techniques in the manner required to support high volume print-on-demand book preparation and printing as a manufacturing operation is original.

This workflow has been proven to be cost effective and practical based on several years of research, development, and production scanning and printing. It has also proven to be scalable as production volumes grow and capable of delivering consistent results required by publishers from a manufacturing process. Finally, in its enhanced form, it is unique in delivering the highest levels of image quality possible at high volumes with investment protection based on output device independence.

Acknowledgments

The authors thank our colleagues Ravi Rao and Gerry Thompson in IBM Research at Yorktown Heights, New York for their insight and image processing algorithms.

We also extend our gratitude to Rick Voytko and the production team in IBM Book Services at LaVergne, Tennessee for providing a pilot and manufacturing proving ground for these techniques.