Book Scanning – Technologies and Techniques

Michael L Mansfield
Engineering and Development - Ancestry.com, Genealogy.com, MyFamily.com
480 East 500 North
Lindon, UT 84042
e-mail: mmansfield@myfamilyinc.com

INTRODUCTION

This presentation will focus on the spectrum of scanning technologies and techniques specifically suited to scanning bound volumes in a large scale production environment. First, the key parameters of a book scanning project will be considered. Issues such as color-depth, dithering, DPI resolution, curve-correction, and on-device de-skew and cropping will be discussed as critical production parameters in the book scanning process. Next, industry techniques and technologies for scanning large numbers of bound volumes will be presented, each with its abilities and limitations, including cost and production efficiency. The presentation will be limited in scope to the digitization of the pages of bound volumes. Post-processing of the digital images such as OCR, compression, watermarking, archiving, on-line hosting preparation, and all other post-digitization processes will not be discussed.

EXTENDED OVERVIEW

In any large scale production scanning project the content to be digitized must be evaluated in conjunction with the project goals and requirements to determine the operational parameters of the scanning process. This analysis often dictates that certain scanning devices and methods cannot be used with the subject material due to limitations of the devices themselves. Likewise, the results of the analysis often imply that certain scanning systems must be employed in order to meet the pre-determined parameters and specifications. Common and recurring digital image issues such as bi-tonal conversion, gray-scale, color, dithering, and resolution fidelity will be presented as they relate to the specific requirements of digitizing pages of bound books.

It is imperative that this analysis be done at the beginning of the project. Experience has shown that the image-capture process is by far the most important step in ensuring that the scanning process and device settings are producing images which meet the project requirements. Post-process image enhancement and transformation algorithms are only effective at making minor qualitative changes to the digitized images.

There is no universally optimal technology for scanning bound materials. At one end of the spectrum there are domains of books and bound volumes where it may be appropriate to remove the spines of the books by physically cutting them off and running the individual pages of the books through a document scanner. This is an inexpensive and fast technique. However, these books are rendered nearly useless for continued use without expensive re-binding. Few bound materials in the Family History and Genealogy domain lend themselves to such a destructive technique but they do exist. One such example is phone and city directories where a second, sacrificial copy, often exists.

In the middle of the spectrum are bound volumes such as county and local histories, family histories, and published vital records. These are often 70 to 120 years old or more but are not considered rare or extremely fragile. These types of bound materials lend themselves well to a number of specialty book scanners which typically employ overhead downward oriented CCD arrays, integrated lighting systems, and articulating beds designed specifically for supporting the book during scanning. A number of devices exist in this classification from manually operated planetary book scanners to robotic page turners.

At the other end of the domain spectrum lie very rare, old, and fragile bound volumes which are usually restricted in access to trained professional librarians and archivists. These materials require extreme care during a scanning operation to ensure that the handling of the binding and pages is done correctly during the digitization process. Also, lighting, heat, and other environmental conditions may be of concern. The scanning projects of these types of rare, old, and fragile books often dictate higher than normal standards in image fidelity and color-depth as the volumes themselves are considered artifacts and works of art in their own right. These digitization efforts are often as much about art preservation as they are about information capture. These types of materials are best digitized by a class of scanning devices which employ computer controlled flying linear CCD arrays with specialized lighting and articulating book cradles.

The characteristics and features of different book scanning technologies will be presented with the goal of educating the audience on the issues, options, and best practices for executing various large scale book scanning projects.