Making Digital Facsimiles of Documents
Part 4: Scanning

A scanner produces a digital image from the material being scanned, which in context of these articles, usually is a page of a document. A digital image is a rectangular array of pixels ("picture elements"). A scanner makes up the image a row at a time as the scanning mechanism advances.

Two aspects of scanning translate directly into the size and quality of the image: resolution and bit depth.

Resolution determines how finely material is scanned and the detail of the result. Resolution is measured in pixels per inch (ppi); that is, the number of pixels produced for an inch of the material scanned. (Sometimes dpi, for dots per inch, is used instead; ppi and dpi mean the same thing for the purposes of scanning.) For example, at 400 ppi, one square inch of the material scanned contains $400 \times 400 = 160{,}000$ pixels. An 8" × 10" scan produces $8 \times 10 \times 160{,}000 = 12{,}800{,}000$ pixels. At 600 ppi, the number of pixels is 28,800,000. There is a significant trade-off between detail and image file size.

Bit depth refers to how many bits of computer data are stored for every pixel. Only one bit is needed for scanning text and line art; every pixel is either black (ink) or white (paper) — black / white scanning. For non-color pictures, 8 bits per pixel are needed to provide smooth gradation of shades of gray. This is referred to as *grayscale* scanning. For color, 16 bits per pixel are needed. Therefore, color scanning produces 16 times the amount of data as text scanning does.

Recommended Scanner Settings

When scanning documents for the purpose of making digital facsimiles, it is important to use appropriate scanner settings.

Although manufacturers of scanners tout capabilities for very high resolution and bit depth, scanning for the purpose of making digital documents requires very modest resolution and bit depth. More important, scanning at too high a resolution or bit depth can be very slow, produce unmanageably large files, and at best results that are no better than for appropriate settings. One way to understand this is that there is no advantage to having resolution or bit depth higher than devices that use digital facsimiles (monitors and printers) can represent.

Here are guidelines for common types of printed material [1]:

print type	*scan type*	*bit depth*	*resolution in ppi*		
			minimum	*recommended*	*maximum*
text	black / white	1	300	400	600
line art	black / white	1	300	400	600
non-color pictures	grayscale	8	150	200	250
color pictures	color	16	150	200	250

Note that the recommended settings for text and line art are the same. In the case of very detailed line art, the maximum setting of 600 ppi may be needed. Text does not need to be scanned at a resolution higher than 400 ppi, but for a page with both text and detailed line art, 600 ppi can be used.

Pictures that are printed as halftones need to be descreened to prevent a mottled appearance (the scanning process converts the dots to real shades of gray). Some scanner-control software provides descreening as an option. For others, there may be different scan types for descreening and not descreening. For example, sometimes **Document** is used to indicate grayscale scanning with descreening, while **Photograph** is used for grayscale scanning without descreening, and similarly for color.

January 11, 2003; last modified May 23, 2003

Physical Aspects of Scanning

Using a flatbed scanner requires putting the document to be scanned face down on the scanning surface.

Pages must be flat on the scanning surface to get good scans. If a portion of a page is not in contact with the scanning surface, the resulting scan may show unwanted curvature or other distortions, as well as dark areas where the page is substantially above the surface.

For unbound documents, it is relatively easy to get pages flat. The scanner lid, when lowered, usually is heavy enough to press pages flat. If the scanner lid is not heavy enough to press a page flat, weights can be place on the lid or on the page itself with the lid not lowered. A heavy book makes a convenient weight.

If a document is bound, as in a book, getting the printed portions of pages flat may be difficult or impossible. The usual problem is portions of pages near the binding. This problem may be particularly severe in books that have been rebound, which usually involves trimming pages at the margin with the result being narrow binding-side margins. Excessively heavy weights may damage the binding or even the scanner. Some techniques to deal with this problem will be covered in a future article. In any event, some curvature in scanned images may be unavoidable, short of debinding the book.

For large books, supports at the sides of the scanner may be needed to hold up the parts of the books that overhang the edge of the scanner. If there is enough space at the side of the scanner, stacks of books may be used. *Ad hoc* supports also can be fabricated from boxes and similar objects. This topic will be discussed at greater length in a future article.

If there is only room for a support at one side of the scanner, which often is the case, a book that needs side support can be rotated 180º on successive scans so that the overhang is always to one side. In this case, half the scans will be upside down, but that is easily fixed during image touch-up.

Pages with More than One Print Type

It is common for printed pages to contain more than one print type. Normally, text and line art can be scanned together with the same settings, as mentioned earlier.

When pictures occur on a page with text or line art, more than one scan is needed. One scan should cover the entire page and include text and line art. If there are pictures as well, a second scan is needed for them, but only the area covered by the pictures needs to be scanned. The same procedure applies to color pictures. In the unusual case where there are both “black-and-white” and color pictures, separate scans are needed for each. When a page requires more than one scan, the images are composited during page layout, a process that will be described in a subsequent article.

Area Selection

Scanner software allows specification of the area to be scanned. It is best, of course, not to scan more area than is needed. But since the result of scanning will be touched up, covered in a subsequent article, it is not worth the time and effort to set the scan area precisely.

To specify the area to be scanned, it is necessary to do a preview scan. Although previews are faster than final scans, they do take significant time. For example, if many pages of text and line art of the same size are to be scanned, one preview can show the maximum area for all subsequent scans.

For text and line art, the entire page should be scanned. For pictures, only the area enclosing the pictures need be scanned. For this, a preview scan is worthwhile.

Scanning More Than One Page at a Time

For documents with small page sizes, it often is possible to scan two pages at once (or possibly more if the document is unbound). This reduces the time for scanning and wear and tear on the scanner. If this is done, the scanned image for the two pages needs to be separated into two images before layout is done. This will be covered in more detail in a subsequent article.

Pages Too Large to Scan

The typical scanner has a surface area large enough to scan standard 8.5" × 11" pages. The pages of some documents are larger than this. There are large-format scanners that can scan large pages, but such scanners are expensive and have large footprints.

An alternative is to scan overly large pages in sections. The sections then can be combined ("stitched") in an image-manipulation program. A subsequent article will discuss the process.

Reference

1. Griswold, Ralph E. "Making Digital Facsimiles of Documents, Part 3: Types of Printed Material", 2003: http://www.cs.arizona.edu/patterns/weaving/webdocs/gre_dd03.pdf

Ralph E. Griswold
Department of Computer Science
The University of Arizona
Tucson, Arizona