**Optical Character Recognition**

By Sami Lais

Computerworld | Jul 29, 2002 12:00 am PST

Suppose you wanted to digitize the novel *Moby Dick* overnight. You could stay up all night typing and still not finish. Or you could use a high-end scanner and in minutes scan all of author Herman Melville's works into a computer using optical character recognition (OCR) technology.

This is the technology long used by libraries and government agencies to make lengthy documents quickly available electronically. Advances in OCR technology have spurred its increasing use by enterprises.

More

**[ Related:** [**Tech event calendar: Upcoming shows, conferences and IT expos**](https://www.computerworld.com/article/3313417/it-industry/tech-event-calendar-2019-upcoming-shows-conferences-and-it-expos.html) **]**

[Computerworld   
QuickStudies](https://www.computerworld.com/article/2544330/quickstudy-directory.html)  
For many document-input tasks, OCR is the most cost-effective and speedy method available. And each year, the technology frees acres of storage space once given over to file cabinets and boxes full of paper documents.

Before OCR can be used, the source material must be scanned using an optical scanner (and sometimes a specialized circuit board in the PC) to read in the page as a bitmap (a pattern of dots). Software to recognize the images is also required.

The OCR software then processes these scans to differentiate between images and text and determine what letters are represented in the light and dark areas.

Older OCR systems match these images against stored bitmaps based on specific fonts. The hit-or-miss results of such pattern-recognition systems helped establish OCR's reputation for inaccuracy.

Today's OCR engines add the multiple algorithms of neural network technology to analyze the stroke edge, the line of discontinuity between the text characters, and the background. Allowing for irregularities of printed ink on paper, each algorithm averages the light and dark along the side of a stroke, matches it to known characters and makes a best guess as to which character it is. The OCR software then averages or polls the results from all the algorithms to obtain a single reading.

**Technological Progress**

Advances are being made to recognize characters based on the context of the word in which they appear, as with the Predictive Optical Word Recognition algorithm from Peabody, Mass.-based ScanSoft Inc. The next step for developers is document recognition, in which the software will use knowledge of the parts of speech and grammar to recognize individual characters.

Today, OCR software can recognize a wide variety of fonts, but handwriting and script fonts that mimic handwriting are still problematic.

Developers are taking different approaches to improve script and handwriting recognition. OCR software from ExperVision Inc. in Fremont, Calif., first identifies the font and then runs its character-recognition algorithms.

Advances have made OCR more reliable; expect a minimum of 90% accuracy for average-quality documents. Despite vendor claims of one-button scanning, achieving 99% or greater accuracy takes clean copy and practice setting scanner parameters and requires you to "train" the OCR software with your documents.

The first step toward better recognition begins with the scanner. The quality of its charge-coupled device light arrays will affect OCR results. The more tightly packed these arrays, the finer the image and the more distinct colors the scanner can detect.

Smudges or background color can fool the recognition software. Adjusting the scan's resolution can help refine the image and improve the recognition rate, but there are trade-offs.

For example, in an image scanned at 24-bit color with 1,200 dots per inch (dpi), each of the 1,200 pixels has 24 bits' worth of color information. This scan will take longer than a lower-resolution scan and produce a larger file, but OCR accuracy will likely be high.

A scan at 72 dpi will be faster and produce a smaller file—good for posting an image of the text to the Web—but the lower resolution will likely degrade OCR accuracy.

Most scanners are optimized for 300 dpi, but scanning at a higher number of dots per inch will increase accuracy for type under 6 points in size.

Bilevel (black and white only) scans are the rule for text documents. Bilevel scans are faster and produce smaller files, because unlike 24-bit color scans, they require only one bit per pixel. Some scanners can also let you determine how subtle to make the color differentiation.

Which method will be more effective depends on the image being scanned. A bilevel scan of a shopworn page may yield more legible text. But if the image to be scanned has text in a range of colors, as in a brochure, text in lighter colors may drop out.

*Lais is a freelance writer in Takoma Park, Md.*