Optical Character Recognition

Anne Permaloff, *Auburn University at Montgomery*
Carl Grafton, *Auburn University at Montgomery*

Optical character recognition (OCR) is a process by which printed text is detected and transformed into a computer text file. OCR consists of two basic processes: scanning and recognition. Scanning, performed with a device called a scanner, digitizes the printed page, creating a coded graphics version of the text that may be stored on disk. That coded version transforms the scanned image into pixels, and it is readable by graphics programs.[1]

The separate recognition process translates the picture of an "A" into the letter "A." A new file is created in a format determined by user instructions. That file is readable by word processor, statistics, and/or database software supported by the OCR program used.

OCR is a technique that can be useful to political scientists. For example, research notes taken from printed sources, rather than being laboriously typed, could be scanned, processed, and saved as a file readable by a word processing package. Content analysis might be almost completely mechanized. Numerical data from government reports could be scanned rather than entered by hand and then made readable by a spreadsheet, database management program, or statistics package.

Choice of Hardware

There are three basic types of scanners and one highbred combination. One basic type is hand-held. A hand-held scanner resembles a paint scraper in shape and is moved slowly across a page as it reads a strip of text approximately four inches wide. Hand-held scanners are relatively inexpensive, ranging in price from approximately $100 to $500. Their chief advantage, aside from their relatively modest cost, is that they can be used to read bound materials. With an adaptor, specific models can be attached to a portable computer for use in archival research.

The standard hand-held scanner's major disadvantage is that two or three overlapping passes per single page are required to digitize a document measuring more than four inches on a side. A more serious drawback is the problem of accuracy. OCR software is extremely sensitive to the quality of the scanned image it is asked to process. The scanner must be moved along a page with great steadiness to produce a distortion free image. Hand-held scanners have rollers that facilitate even, smooth movement across a page, but it is nonetheless difficult to produce an image readable with high accuracy by an OCR program. The widely advertised Logitech and Microtek scanners are the best-selling brands.

Flatbed scanners, the second basic type, resemble a photocopying machine. A page is placed on top of the scanner, and the entire page is scanned at once. Flatbed scanner prices start at approximately $900 and extend to $2,000. They are more accurate than their hand-held counterparts, except for the case of bound documents that cannot be flattened against the screen. Flatbed scanners are also much faster than hand-held models. Leading flatbed scanner manufacturers include Hewlett-Packard, Chinon, and Canon.

Sheet fed scanners are the third basic type. Physically, they resemble flatbed machines, but sheets are fed into a slot and guided past the scanner with mechanized rollers. Sheet fed scanners fall between hand-held and flatbed machines in price and approximate the latter in accuracy and speed. Their major disadvantage is that they can read only sheets of paper and not bound documents. The Complete Page Scanner at approximately $550 is the leading sheet fed scanner model.

A highbred scanner such as the LightScan 400P, which we use, combines the best features of the major scanner types. It can be operated in a hand-held mode, but, at 8.5 inches, its scanning width is much larger than normal. It is sold with a software controlled sheet feed mechanism that runs a page under the scanner head with an evenness difficult, if not impossible, to achieve by hand. LightScan's sheet feed handles a batch of up to ten pages. Since scanning time for a page is only about 30 seconds, little time is saved by taking advantage of this feature.

LightScan produces high quality digitized images. Its price ($599) includes scanning and graphics software but not OCR software.[2] OCR software that works with the LightScan costs an additional $100. LightScan's scanning and graphics software requires the use of Windows 3.0. This critical fact is not mentioned in some advertising for the product. LightScan's biggest disadvantage is that it is not fully supported by most major OCR software manufacturers.

When an OCR package is said to support a scanner, OCR software manages the scanning as well as the recognition process. Such OCR-scanner interaction is convenient but not necessary. Virtually all major scanners operating with their own scanning software can produce files readable by many OCR programs. The most common is the TIFF format, a major graphics file format. The procedure when working with an unsupported scanner is to load the scanning software, do the scans, save the images in TIFF files, exit the scanning software, load the OCR program, load the TIFF files, and then perform the recognition with the OCR program. When hardware and OCR software work together, there are fewer steps.

Scanner Installation

Scanner installation generally requires removal of the computer cover and insertion of a circuit board that is connected to the scanner by a cable. Installation instructions sometimes assume more knowledge about computer viscera than many users possess. The user may not know which of several open computer slots should be used. A visual comparison of the circuit board and slot connections usually resolves this problem. Computer circuit boards may not slip as easily into place as installation instructions suggest. Carefully applied force often must be applied.

The scanner will not function if it collides electronically with other computer components such as a disk drive or modem. This situation may require changes in DIP switches or jumpers either on the scanner circuit board, other circuit boards, or both. DIP switches resemble miniature wall light switches. Depending on their shape, they are most easily moved with the tip of a ball point pen or a blunted wooden toothpick. Jumpers are changed by shifting tiny plastic loops from one set of copper prongs to another. When illustrations are poorly drawn, or instructions are written by technicians not fully conversant with English, and/or instructions are inconsistent, consultation with scanner technical staff or local computer technicians may save time, trouble, and even equipment damage.

The only time scanner installation does not require opening the computer is when an adaptor is used that allows a scanner to be attached to a computer via the parallel port (usually used to connect a printer).[3] The adaptor is intended for laptop computers, and, according to the manufacturer, it does not slow the scanning process. The authors have not tested this product.

OCR Software

OCR programs vary considerably in features, accuracy, and hardware requirements. Many are available for Macintosh, IBM, and IBM compatible computers. The prospective purchaser should note the widely varying hardware requirements of OCR programs. The minimum configuration is 640K RAM and a hard disk. Most of the new Microsoft Windows based programs require 4MB, 6MB, and even 8MB or RAM and a great deal of hard disk space. Some manufacturers understate the practical memory and disk space requirements of their products.

Major Categories

OCR programs can be divided into two major categories. One, the older and much less desirable of the two, compares a scanned image with a bit-mapped, pixel by pixel standard. This works well until the program is faced with characters of a different size or font in which case it must be "trained" one character at a time to recognize what is for it a completely new alphabet (52 characters including capital letters) and set of numerals. Considering that the pages of many publications contain three or more sizes and/or fonts, the difficulty involved in such a procedure becomes apparent.

All of the programs discussed here approach optical character recognition in a newer and better way. Characters are defined not by a pixel or bit-map profile but in terms of universal characteristics that do not vary with print size or style. The software understands an "O" to be a circle of any size and width, an "A" as a triangular shape with cross bar, etc. An OCR program that uses the older approach should not be purchased regardless of how inexpensive it may be.

The OCR Programs Examined

With one exception, the OCR programs discussed below represent some of the highest quality market leaders for IBM and IBM compatible computers. All require Microsoft Windows 3.0 or later. WordScan Plus and ReadRight are also available for the Macintosh, and the manufacturer of TypeReader is scheduling the release of a Macintosh version in the spring of 1992. As has been our common experience, no exclusively Macintosh software manufacturers agreed to provide their products for our analysis.

Image-In Read, Version 2.01. Image-In, Inc., 406 East 79th St., Minneapolis, MN 55420. Tel: (800) 345-3540. FAX: (612) 888-3665. $149. Program documentation states that 1MB RAM is required, but this is not a realistic specification because of Windows' memory requirements.

OmniPage Professional, Version 2.0. Caere Corp., 100 Cooper Court, Los Gatos, CA 95030. Tel: (408) 395-7000. FAX: (408) 354-2743. $995. Program documentation states that 4MB RAM is required, but it refused to operate with that amount on two different computers. At least 8MB hard disk space is also required.

READRIGHT for Windows, Version 3.1. OCR Systems, Inc., 1800 Byberry Road, Suite 1405, Huntingdon Valley, PA 19006. Tel: (215) 938-7460; FAX: (215) 938-7465. $495. Requires 4MB RAM and 6.5MB free disk space.[4] Version for Macintosh Classic/SE and above, $195. Requires 4MB RAM and System 6.03 or higher.

TypeReader for Windows, Version 1.0. ExperVision, Inc., 3590 North First Street, San Jose, CA 95134. Tel: (800) READ-TYP. $895. Requires 4MB RAM (although 6MB is recommended) and 8MB free disk space. A Macintosh version is scheduled for release in the spring of 1992.

Wordscan Plus, Version 1.02A. Calera Recognition Systems, 2500 Augustine Dr., Santa Clara, CA 95054. Tel: (408) 986-8006; FAX: (408) 986-1440. $995. The manual specifies 2MB RAM, but the program was not tested with that relatively small amount. The authors' experiences with Windows suggest that 4MB RAM is a more realistic requirement. 6MB free disk space is also required. Macintosh version $895. A French and German language module costs an additional $195.

Program Features

Table 1 summarizes important OCR program features.

Some of the table entries require explanation. "Text rotation" refers to a program's ability to recognize text that has been scanned at a 90 degree angle across a page rather than down it or upside down or a page presented in landscape mode

TABLE 1
OCR Program Features

	Program				
Feature	IIR	OP	RR	TR	WS
Text rotation	Yes	Yes	Yes	Yes	Yes
Trainable	Yes	Yes	No	No	No
Complex formats	Yes	Yes	Yes	Yes	Yes
Dot matrix	No	Yes	No	Yes	Yes
Numeric	No	Yes	No	No	Yes
Writes to word processor[a]	No	Yes	Yes	Yes	Yes
(Number)		19	4	8	15
Writes to spreadsheet[b]	No	Yes	Yes	Yes	Yes
(Number)		3	3	2	4
Writes to desktop publishing	No	No	Yes	No	Yes
(Number)			2		2
Writes to a database[c]	No	Yes	Yes	No	Yes
(Number)		2	1		2
Writes to ASCII	Yes	Yes	Yes	Yes	Yes
(Number)	1	5	2	3	3
Number of FAX formats supported	0	7	0	0	7
Number of image files reads and writes	6	3	4	2	4
Number of scanner brands supported	9	25	8	11	11
Batch mode in addition to page	No	Yes	Yes	Yes	Yes
Context sensitive help	No	No	No	No	No
800 number of help	Yes	Yes	No	Yes	No
Multiple languages supported	Ger/Fr Add-on	Yes Many	No	No	No
Can view original image	Yes	Yes	Yes	Yes	Yes

IIR = Image-In Read; OP = OmniPage; RR = ReadRight; TR = TypeReader; WS = WordScan

[a] All programs that write to a word processor will write to WordPerfect, Microsoft Word, and WordStar.

[b] All programs that write to a spreadsheet will write to Lotus 1-2-3 and Excel.

[c] The most common database format to which OCR programs write is dBASE.

but read as a normal page. A program is "trainable" if it can be taught that certain initially unrecognized scanned images are a particular number or letter. The "complex formats" entry refers to the program's ability to process text formatted in complex ways such as multiple columns mixed with graphics. In practice, the programs vary widely in their ability to recognize text and numerical data in complex formats.

"Dot-matrix" indicates whether the program can read draft-quality dot matrix text. "Numeric" means that a program can be set to recognize only numeric information. All OCR programs will read alphanumeric materials.

The next few rows in Table 1 list the file formats to which the programs can save including word processing, spreadsheet, database management, and ASCII. "Reads/writes image file" refers to the formats in which graphic images such as scanned text images, can be stored.

Except for Image-In Read, all of the programs can process files manually (one at a time) or automatically in batches. The "Multiple languages" refers to the program's foreign language text recognition features.

All of the programs allow the user to view and edit recognized text while looking at the original scanned image superimposed on the screen in magnified form. The advantage of this feature is that one does not need to compare the original text with the OCR version by shifting back and forth between paper and screen, a surprisingly time consuming process. Unfortunately, all of the programs only allow this feature to be invoked over characters that the programs regard as suspect. If the user sees a word or number that appears to be incorrect but that the OCR program has not marked with an error symbol, the user cannot pop up the image of the original scanned version. The user must refer to the original version on paper. Comprehensive proofreading must also be done by comparing the original paper version with the OCR rendition.

Guidelines for Best Results

The success of an OCR operation is more dependent on the quality of the scanned image than any other factor. The best scanner and software working with a flawed image that appears readable to the human eye may not produce satisfactory results. An image might be too light with parts of letters slightly malformed or separated, or an image might be too dark with letters touching and loops (such as the inside of a small "e") nearly filled in.

None of the programs produced perfect results with a photocopied table of figures. A photocopy might be used because there is no other way to get the material to a scanner. Photocopying might also be used to enlarge text too small to be recognized in its original size.

A photocopy must be as dark as possible without producing spurious connections between characters. It should also be square to the page. It must be free of distorted images such as those caused by a bound original that could not be flattened on a copier. Fortunately, some new photocopiers allow bound materials to drape over the edge to permit nearly distortion free work. The OCR programs tested here are much more sensitive to overly light or dark images than they are to images that are a bit out of square or distorted.

As suggested above, hand scanning for OCR applications should be thought of as a last resort. The need for carefully overlapped images (to fit the four inch strips together within the computer) and the requirement for absolutely even movement make hand scanning uncompetitive with typing in many instances.

Testing Procedures

Test Material

All the programs were tested by scanning the following samples, typical of what would be found in ordinary work:

1. A photocopy of a table of figures from a professional journal.[5] The figures are rather small, measuring 20 characters/inch (CPI), but they are easily readable by the human eye. The quality of the photocopy is typical of what is produced by a heavily used university library copying machine.
2. An original page from *PS*.[6] The print quality is very high. Characters are black and well defined, the paper is enamel-white providing a high contrast with the characters, and regular text is of ordinary size (roughly 17 CPI). On the other hand, this page offers challenges to OCR software because its print is serif (decorative twists) and characters are frequently almost touching, endnotes are rather small (20 CPI), and seven different type styles are present.[7]
3. The same *PS* page photocopied and enlarged 141%. The enlargement is clear, but a darker image could have been achieved on a slightly better maintained copier.
4. The original copy of a memo containing 12 pitch courier type printed on a laser printer. The print is dark black on pure white paper.
5. An original page from the *Statistical Abstract of the United States.* This page mixes very small print (28 CPI) and numerical data enclosed in lined columns.[8]
6. A 141% enlarged page from *The Statistical Abstract of the United States* with print characteristics identical to the one above, but with a simpler format and fewer lines separating columns.[9]
7. A photocopied page that lists state names and numerical data. This page was produced with a typewriter and then printed rather badly on poor quality paper.

Testing Standards

OCR software manufacturers invariably report accuracy by individual characters. Thus one incorrectly recognized character in the word "registration" would produce an accuracy claim of 91.7%. However, any mistake in a word or, worse yet, a number means that the entire word or number is incorrect. For the user, one incorrect character in a word means that the accuracy is zero for that word. Our accuracy counts use this more stringent and realistic measure.

Most mistakes in word recognition can be detected with the OCR software's built-in spelling checker or that of the user's word processor. A large majority of mistakes in numerical recognition turn numbers into letters (for example, a "6" recognized as a "b" and an "8" as an "S") or some other nonnumeric symbol. Such errors are automatically identified when numerical data are read into a spreadsheet or statistics package. In Lotus 1-2-3 with default settings, letters stand out in a column of numbers, and most statistics packages generate error messages when presented with letters that are supposed to be numbers. It is when numbers are transformed into other numbers (for example, "1" into "7" and vice versa) that the user realizes the importance of accuracy.

When a scanner is first operated, the user will need to experiment with small samples of text. Scanner hardware and software are adjustable for image brightness and contrast. These adjustments are second in importance only to original image quality. The original *PS* page (Sample 2) was first scanned with default middle range hardware and software brightness and contrast settings. The best accuracy result for the main text of this perfectly printed page was a poor 96.9%; endnote accuracy was only 5%. In the main text, "G"s were recognized as "C"s, and "W"s generated the program's error symbol, an ampersand. Close inspection of the scanned image on the computer monitor revealed that the parts of some "W"s in the scanned image were almost disconnected, and the crossbars of the "G"s were almost missing.

The brightness settings were adjusted to the darkest extreme, and the image scanned again. This time, the "W"s had been completely connected, but the interiors of characters with loops such as "e," "g," and "s" were almost completely filled. Another scan done with the brightness setting placed halfway between the two previous ones resulted in correctly connected "W"s, but the loops were not filled. The subsequent recognition attempts produced drastically improved results.

Test Hardware

Recognition testing was done on a 16 Mhz 386 SX computer using a 20MB Bernoulli disk drive. By today's standards, this is a slow hardware configuration. All recognition rates discussed below could be vastly improved on a computer with a faster clock speed or disk drive. Initial testing was done with 4MB RAM; retesting was done after RAM was doubled to 8MB.

The Programs

Table 2 summarizes accuracy and speed measurements for the programs for the first four and last two samples described above. The fifth sample, the original *Statistical Abstracts* page, defeated all of the programs.

Image-In Read

Image-In Read is one module in a comprehensive image management system.[10] Image-In Read is by far the least accurate of the programs examined. Its accuracy with Sample 1 was a dismal 17%.

Image-In Read performs recognition in an interactive mode that stops at each character that it cannot recognize. The user is then given the opportunity to identify the character so that Image-In Read can learn it. Unfortunately, this learning appears to be highly specific. For example, in the table it failed to recognize the first "9" it found. Upon being informed that the character was a "9," it failed to recognize the next 9 as well. This happened throughout the recognition process with most "9"s and "8"s.

Image-In Read did better with the original PS text, scoring a 93.3% accuracy measured by word. It

TABLE 2
OCR Program Speed and Accuracy

	IIR	OP	RR	TR	WS
Sample 1: Three Columns of Numbers					
Alphanumeric mode accuracy	<50.0%	88.7%	74.0%	95.3%	96.8%
Alphanumeric mode speed	N.A.[a]	150n/m	11.5n/m	150n/m	98.4n/m
Numeric mode accuracy	N.A.	50.0%	N.A.	N.A.	98.4%
Numeric mode speed	N.A.	300n/m	N.A.	N.A.	300n/m
Sample 2: Original *PS* Page					
Accuracy on main text	93.3%	99.6%	100%	100%	99.1%
Accuracy on endnotes	<50.0%	97.6%	96.3%	100%	98.8%
Speed (alphanumeric)	N.A.[a]	120w/m	62w/m	120w/m	102w/m
Sample 3: Enlarged *PS* Page					
Accuracy on main text	97.8%	97.6%	100%	100%	98.0%
Accuracy on endnotes	<50.0%	97.6%	95.9%	100%	97.6%
Sample 4: Laserprint Page (Text)					
Accuracy	94.6%	98.6%	99.2%	100%	95.0%
Sample 5: Original Statistical Abstract Page					
None of the programs were able to handle this page.					
Sample 6: Enlarged Statistical Abstract Page					
Alphanumeric mode accuracy on text	<50.0%	Crashed	<50.0%	<50.0%	84.3%
Alphanumeric mode accuracy on numbers	<50.0%	Crashed	<50.0%	<50.0%	90.5%
Alphanumeric mode speed	N.A.[a]	Crashed	146w/m	N.A.	214w/m
Numeric mode accuracy	N.A.	<50.0%	N.A.	N.A.	80.0%
Numeric mode speed	N.A.	102n/m	N.A.	N.A.	100n/m
Sample 7: Typed Page (Data Table with Labels)					
Alphanumeric mode accuracy on text	<50.0%	<50.0%	<50.0%	95.0%	84.5%
Alphanumeric mode accuracy on numbers	<50.0%	<50.0%	<50.0%	100%	97.9%
Alphanumeric mode speed	N.A.[a]	86w/m	***[b]	59w/m	90w/m
Numeric mode accuracy	N.A.	96.0%	N.A.	N.A.	100%
Numeric mode speed	N.A.	300w/m	N.A.	N.A.	139w/m

IIR = Image-In Read; OP = OmniPage; RR = ReadRight; TR = TypeReader; WS = WordScan

n/m = numbers per minute; w/m = words per minute

[a] Speed measurement is not applicable because Image-In Read operates in an interactive mode.

[b] ReadRight ran too fast to clock, and the text generated from this sample was close to zero in accuracy.

reported 98% accuracy because it was measuring its own accuracy by character. The program's learning capabilities were fairly effective when dealing with the higher quality text on the *PS* page. At first, the program stumbled on roughly every tenth word, but as it learned, it went farther and farther between stops.

With the original *PS* text, Image-In Read completely broke down on the small print endnotes. When the text was enlarged by 141%, it did no better on the endnotes, but its accuracy on the main text improved to 97.8%.

With the extremely clear laser printer page, Image-In Read was only 94.6% accurate. It missed a "y" and a "g" completely; it simply did not see them. It was never able to learn that an "s" was an "s" even though all of them were perfectly formed. Image-In Read's accuracy on the remaining samples was 25% or lower.

Despite its low cost, we cannot recommend Image-In Read.

OmniPage Professional

OmniPage demands access to more RAM and hard disk space than any of the programs described here. If the user's computer is what the OmniPage manual blandly refers to as "minimally configured," i.e., if it contains 4MB of RAM, a permanent swap file of at least 4MB must be created on the hard disk. Such a file isolates the entire 4MB so that it cannot be used for any other purpose.[11] Thus on most machines OmniPage requires not just the 8MB of disk space described in the manual, but 12 MB for the program, data, and the permanent swap file.

The OmniPage manual claims that the program will run on a computer with 4MB of RAM. The manual's 4MB specification is not as precise as it might seem. Out of memory messages resulted when the program was run on two different machines with that much memory. More important than the amount of RAM physically present in the machine is the amount of RAM available after DOS and Windows are loaded. This figure varies among computers and computer configurations. OmniPage support personnel were unable to specify a minimum amount. In contrast, the ReadRight manual indicates that at least 2,800K must be available after Windows is loaded. (ReadRight also requires 4MB.)

RAM can be freed by modifying the DOS CONFIG.SYS file to reduce the number of files that can be open and the number of buffers available. Unfortunately, if these have been set to run other programs, the user must find a way to switch between a stripped down configuration for OmniPage and a different version for other applications. RAM can also be freed by reducing or eliminating the Windows disk cache, but this produces significantly degraded performance. The authors were able to operate OmniPage only after memory was expanded beyond 4MB.

OmniPage provides an 800 help number for users, but the caller can expect to be put on hold for 10 minutes or more.

Like Image-In Read, OmniPage can be trained to recognize flawed or nonstandard characters, but OmniPage learns more effectively than Image-In Read.

OmniPage can save in an impressive variety of formats: ASCII (several versions); dBASE; DisplayWrite; Enable; EBCDIC; Excel; Framework; GSA Navy DIF; HP Advance Write Plus; IBM Writing Assist; Lotus; Lotus Manuscript; Microsoft RTF; Microsoft Word; MultiMate; Office Writer; PeachText; PFS; Q&A; Rapid File; Samna Word; Volkswriter; Wang PC; WordPerfect; WordStar; XyWrite; and Dec Dx. It can read and write the following image file formats: PCX; TIFF uncompressed; and TIFF Compressed.

OmniPage fully supports far more scanners than the other programs represented here (with multiple models for most brands). These include: Abaton; Agfa; AVR; Brother; Canon; Complete PC; Datacopy; DEST; Epson; Fora; Fujitsu; Hewlett-Packard; Howtek; IBM; Kyocera; Lightspeed; Mectel; Microtek; Panasonic; Ricoh; Sharp; Sirius; UMAX; Visa; and Wang. OmniPage also supports the following Fax file formats: Brother; The Complete Fax; Hayes; Intel; Panasonic; and Quadram.

OmniPage can recognize text in Danish, Dutch, English, French, German, Irish/Gaelic, Italian, Norwegian, Portuguese, Spanish, and Swedish. Among the programs examined, only one other, WordScan, offers a foreign language recognition capability, but it is an added cost option that supports only French and German.

OmniPage is easy to operate. Its menu structure is logical, and the program's automatic features often provide the user with correct choices for a given recognition task. Like all of the other programs, OmniPage's on-screen help is not context sensitive. It is no better than an abbreviated printed manual that happens to be on screen.

OmniPage offers a variety of options for controlling recognition. Page features such as columns and graphics can be automatically detected or defined manually. For example, OmniPage automatically determined that Sample 1 contained three separate columns. The recognition process required one minute, and the accuracy was 88.7%. All but one of the incorrect figures were recognized as letters instead of numbers or identified with OmniPage's error symbol. So, with one exception OmniPage's errors were mechanically detectable.

OmniPage can be set to recognize only numerical data. This feature should improve accuracy and speed, but it failed to work because OmniPage recognized commas in the numbers as decimals or flagged them as potential errors. The same problem arose when OmniPage attempted to read the *Statistical Abstract* pages (Samples 5 and 6).

Working with the original *PS* page, OmniPage achieved an accuracy of 99.6% on normal text and 97.6% on small print footnotes. All of its mistakes were detected by its spell checker and would have been spotted by a word processor spell checker. It scored only 97.6% with the standard text portion of the enlarged version of the *PS* page and no better with the endnotes.

OmniPage measured a somewhat disappointing 98.6% accuracy with the perfect laser printer text. This might seem to be a high score, but on a page containing 250 words, it would mean that approximately three or four words would be incorrect. OmniPage crashed when it attempted to read the enlarged *Statistical Abstracts* page. When shifted to its numeric mode, its accuracy was less than 50% on the same page partly because it once again failed to recognize commas. When dealing with the typewritten table (Sample 7) in alphanumeric mode, OmniPage's accuracy with words was 4% and 28% with numbers. In numeric mode it scored 96% because there were no commas in the numbers. However, the numbers were badly scattered rather than being set in their original neat columns.

ReadRight

Before these OCR programs initiate the recognition process, they divide a page into regions. In the case of the *PS* page, ReadRight correctly separated columns and the endnotes within two of the columns into three regions. It failed to separate the three columns in the table in Sample 1. This was probably because ReadRight became confused by column headings and horizontal lines at the top and bottom of the table. The headings and lines seemed to draw the columns into a single unit even though the columns were much farther apart than the *PS* columns. This resulted in ReadRight combining all figures in a row into one number (for example, 964, 26, and 1010 read as 964261010).

One solution to this problem might be to crop the lines and column headings just after scanning (a quick and easy process with most scanning software). Another approach is to isolate the three columns using ReadRight's region definition tool. This was done by drawing a box on screen around each region.

Even after isolating the three columns, ReadRight's accuracy with the numbers in Sample 1 was a nearly useless 74%. Furthermore, the three columns and 50 rows of numbers with as many as four digits each took ReadRight 13 minutes to read. This is a rate of one number per 5.2 seconds or 11.5 numbers per minute, and most of the numbers contained only three or fewer digits. It took ReadRight 40 minutes to process the original *Statistical Abstracts* page. If ReadRight had been installed on a computer with a faster hard disk, these times would have been closer to 6 minutes and 20 minutes. ReadRight is, by far, the slowest program represented here, and ReadRight does not have a separate numerical mode to speed numerical processing.

ReadRight's recognition time is greatly extended when print quality is poor because the software must devote time to puzzling over problem characters. When dealing with the high quality *PS* original, ReadRight read at the rate of approximately 62 words per minute (slow compared to other programs) at an accuracy of 100% for the main body of text and 96.3% for the endnotes. Again, a faster hard disk would have brought this rate to 124 or more words per minute. ReadRight read the enlarged *PS* text at 100% (main body) and 94.9% (endnotes) accuracy. It was 99.2% accuracy with the laser print text. It nearly tied for best with OmniPage in reading the original *Statistical Abstract* page, but at just under 90%, neither did very well. ReadRight failed to produce useful results in the last two samples even when vertical lines and other extraneous material were cropped.

ReadRight can save in the following formats: ASCII Data; ASCII Text; dBASE; DCA-RFT; Excel; Lotus; Microsoft RFT; Microsoft Word; WordPerfect; WordStar; and XyWrite.

ReadRight fully supports the following scanner manufacturers (with multiple models for most): Canon; Complete PC; Epson; Hewlett-Packard; Microtek; Panasonic; and UMAX.

At one point ReadRight refused to close an image file in memory, and when an attempt was made to open another image file, the program crashed. No error message accompanied the crash, a common problem with Windows-based software. Another crash, also unaccompanied by an error message, occurred when an attempt was made to load a file that appeared to be no different from ones it had been routinely loading.

ReadRight recognizes only English. It is not trainable.

TypeReader

TypeReader is basically a full featured program, but it is not trainable, can only process English (extra price foreign language modules are forthcoming), and it does not have a numeric mode.

TypeReader can write to the following formats: ASCII; Smart ASCII; RTF ASCII; Ami Pro; DisplayWrite; Microsoft Word; Multimate; PFS Professional Write; WordPerfect; WordStar; Excel; and Lotus. It supports the following scanner brands: AVR; Canon; Xerox; Epson; Hewlett-Packard; Howtek; Microtek; Panasonic; and UMAX.

TypeReader's menu choices are extremely easy to interpret. One rarely has to use the printed manual.

TypeReader's accuracy was 95.3% with the first sample, the three-column table of numbers. It recognized the separate columns automatically. Approximately one-quarter of its mistakes were incorrect numbers and three-quarters were gross errors that a spreadsheet or many statistics packages would have identified automatically. It required only one minute to read this file.

TypeReader was 100% accurate with the main *PS* text, and an impressive 100% accurate with the footnotes. Its speed at roughly 120 words per minute, was comparable to OmniPage and twice that of ReadRight. Its manufacturer claims that it can read 800 words per minute on a computer operating at 25 Mhz. Assuming perfect text and large print, this is a credible claim. TypeReader also read the enlarged version of the *PS* page, the laser print text, and the badly reproduced Sample 7 numbers without a single error. However, the complex formatting and small print of the original *Statistical Abstract* page defeated TypeReader completely; its accuracy was well under 50%. Oddly, the simpler formatting and larger print of the enlarged *Statistical Abstract* page produced even worse results; numbers contained within columns and separated by vertical lines went completely unrecognized by TypeReader. It behaved as if that portion of the page was blank.

WordScan Plus

Versions of WordScan Plus are available for both IBM and Macintosh platforms. WordScan Plus has a set of features very similar to those of TypeReader and somewhat less impressive than OmniPage.

The WordScan installation procedure was somewhat cranky. It insisted on searching for the Windows path statement in the AUTOEXEC. BAT file, but did not search very thoroughly. Having failed to find it, the installation procedure was automatically canceled. The authors finally wrote the path statement manually in DOS, and the installation and subsequent operation proceeded smoothly.

WordScan supports ten scanner brands: Canon; The Complete PC; Datacopy; DEST; Epson; Hewlett-Packard; Microtek; Panasonic; Pentax; and ProScan.

WordScan saves in the following formats: ASCII; Database ASCII; Decolumnized ASCII; DCA/RFT; DisplayWrite; EBCDIC; Enable; Excel; Framework; GSA Navy DIF; IBM Writing Assistant; Interleaf; Lotus Manuscript; Lotus 1-2-3; Microsoft RTF; Microsoft Word; Multimate; Office Writer; PDA; Peachtext; PFS; Rapid File; Samna Word; Volkswriter; WordPerfect; WordStar and XyWrite.

WordScan is irritating to use at first owing to its nonstandard use of Microsoft Windows. For example, rather than opening a file with the universal Windows "File" menu choice, one must select "Start Processing." File save choices are also nonstandard; and there is no "Clear" command. One of Windows' virtues is that programs that adhere to its standard menu choices are relatively easy to learn. There was no reason why the designers of WordScan had to invent their own eccentric and often less than obvious menu structures.

Despite its odd design, WordScan gives the user a greater sense of control over the recognition process when problems arise than do the other programs. For example, WordScan made it easier to isolate segments of the difficult original *Statistical Abstracts* page to simplify the recognition, although in the end this ease of use did not give it an accuracy score above the others. WordScan was 96.8% accurate with the three columns of numbers (Sample 1). It automatically recognized that there were three columns. All of its mistakes were marked by the program's error symbol. It read these data at a middle range 98 words/minute.

WordScan can be set to recognize only numbers, and, unlike OmniPage, this setting generally improved WordScan's accuracy (98.4%). It also tripled the recognition speed to 300 numbers/minute. WordScan's treatment of the enlarged *Statistical Abstracts* page was quite peculiar. In its alphanumeric mode its text accuracy was 84.3% and its accuracy in recognizing numbers was 90.5%. In its numeric mode its accuracy dropped by over 10 percentage points to 80.0%. When dealing with the typewritten table (Sample 7), it recognized 84.5% of the text and 97.9% of the numbers and its alphanumeric mode and 100% of the numbers in its numeric mode. Its speed in dealing with this relatively poor quality data was only 139 numbers/minute. WordScan's numeric mode has no difficulty dealing with commas.

WordScan is clearly the best program for numerical recognition because of its speed, accuracy, and ability to deal with commas in numbers.

WordScan's accuracy for the original *PS* page was 99.1% for the text and 98.8% for the endnotes. Its recognition rate was approximately 102 words per minute. On the enlarged *PS* page WordScan's accuracy was 98% on the main text and 97.6% on the endnotes. WordScan read the laser printer file with a disappointing 95% accuracy, but it identified all of its errors.

Conclusions

Our testing narrowed the choices among these programs to OmniPage, ReadRight, TypeReader, and WordScan. ReadRight installs easily, requires less hardware space than OmniPage and TypeReader, and is less expensive than the others. ReadRight is very accurate with high quality text, but it is not as robust as OmniPage, TypeReader, and WordScan when dealing with a flawed original. It is the slowest. No foreign language modules are available or planned for ReadRight.

Among the leaders, OmniPage is the least accurate. It requires the most hardware, but its language capabilities and the great number of scanners and output formats it supports would make it the first choice for some.

TypeReader is more accurate than WordScan when dealing with alphanumeric text material, but WordScan's numeric mode gives it an accuracy margin and a speed advantage over TypeReader when dealing with numeric data and an even bigger advantage over OmniPage. WordScan is also somewhat better able to handle complex tabular formats of the sort found in *Statistical Abstracts,* although its results with the two samples from that source were only better than the others; its output was not very useful. WordScan has the additional advantage over TypeReader of offering a French and German language module for $195. TypeReader's foreign language modules are not yet available nor are future prices.

In past articles, we have often been able to locate bargain priced software that performs better than programs costing hundreds of dollars more. OCR software is not one of those cases. ReadRight, the weakest program that offers acceptable performance, costs $495, and the others are $895 and $995. WordScan with its language module is $1,190.

Notes

[1] A scanner may also be used to digitize graphic images such as photographs, which may then be included in word processing text or other computer output. That use is not discussed here.

[2] Computer Friends, Inc., 14250 N.W. Science Park Drive, Portland, OR 97229. Tel: (800) 547-3303. FAX: (503) 643-5379. The DEST scanner appears to be identical to the Lightscan, but sells for $699 through Global Computer Supplies, 1050 Northbrook Pkwy., Dept. 31, Suwanee, GA 30174. Tel: (800) 227-1246. FAX: (404) 339-0033.

[3] CAT Hand Scan Adapter LPT, manufactured by Computer Aided Technology, Inc., Dallas, TX. Tel: (214) 350-0888. $149. Supports the following scanner brands: The Complete PC; DFI; GeniScan; Logitech; Marstek; and Niscan.

[4] OCR Systems also manufactures ReadRight Personal for hand scanners. It costs $249 and does not require Windows. A minimum of 575K of RAM must be available after DOS is loaded. The program can take advantage of but does not require expanded or extended memory. ReadRight Personal's major disadvantage is that it cannot process a scanned image wider than 5 inches. ReadRight Personal supports most major brands of hand scanners, and it can process a TIFF file. A non-Windows version of ReadRight is also available at $495. Its hardware requirements are similar to those of ReadRight Personal. It claims a maximum accuracy of 99.5% compared to its Windows version claim of 99.9%.

[5] Brian Powell and Lala Carr Steelman, "Variations in State SAT Performance: Meaningful or Misleading?" *Harvard Educational Review,* Vol. 54, No. 4 (November 1984), 399.

[6] June 1990, p. 175.

[7] This is a measurement of pitch. OCR manufacturers more commonly express minimum and maximum recogizable print sizes in points, but pitch is easier for an individual to measure and more effectively conveys the size of the characters being processed.

[8] *Statistical Abstract of the United States,* 1986, p. 461.

[9] *Statistical Abstract of the United States,* 1991, p. 146.

[10] It includes Scan & Print, Plus (edits scanned photographs), Vect (converts bitmapped images to vector images), and Panorama (an image database manager).

[11] The permanent swap file acts as a de facto extension of RAM. As far as Windows is concerned, a computer with 4MB of RAM and a 4MB permanent swap file has 8MB of RAM.

Quick Notes

Minitab, Version 8.2, Minitab, Inc., 3081 Enterprise Dr., State College, PA 16801-3008. Tel: (814) 238-3280. FAX: (814) 238-4383. List price $695. Price for university faculty or students, $395. IBM PC and compatibles with 80286 or higher processor, DOS 3.0 or later, hard disk with 5.6MB available, 1MB RAM. Minitab's performance is improved with more than 1 MG RAM and a math coprocessor. It can use a mouse although it works quite well without one. Versions of Minitab are available for the Macintosh, DEC, and Hewlett-Packard as well as several mainframe computers and minicomputers.

Minitab can now be operated via menus as well as the commands with which previous users are familiar. New users will find that the menus are easy to use. Reference to the manual will usually be necessary only when confusion arises regarding statistical procedures and terminology that cannot be answered by short on-screen prompts. On those infrequent occasions when problems with menus arise, Minitab's documetation is not very helpful to the novice user because it concentrates almost entirely on the program's command structure and ignores menus.

Although Minitab offers most of the major data manipulation and statistical procedures available in other high end statistics packages, one is often disappointed by limitations. For example, it imports only ASCII and Lotus Version 2.2 files. Also, Minitab refuses to import variable labels requiring considerable retyping in a data set containing many variables.

Annoying limitations also arise with Minitab's graphics. Although the manual shows vertical histograms, we were only able to produce unattractive horizontal ones. Some graphic displays can be annotated, but others cannot. It offers an extremely narrow collection of high resolution graphics tools, and none is of presentation quality. It has no three dimensional plotting capabilities.

Minitab's statistical tools will satisfy most needs, but it lacks nonlinear regression, and its time series capabilities are very skimpy.

Potential purchasers of this program should note that it can handle no more than 100 variables, a severe limitation for some social science applications such as polling. Many far less expensive statistics packages can accommodate 200 variables and more than 30,000 observations or over 6,000,000 data cells compared to Minitab's 16,174 data cells.