

Title:	What Do You Need to Know About OCR?
Created:	2/2/2005
Scanner Models:	All
Operating Systems:	Windows 98 / ME / 2000 / XP

The right OCR product can save you time and money. Buying the wrong product will waste your time and money, and you'll ultimately be dissatisfied and quit using it. To buy the right OCR product, you need to know what OCR is and what it can do for you. This document provides some important information to help you evaluate your OCR options.

- **What is OCR?**

OCR stands for Optical Character Recognition. Very simply stated, OCR means converting an electronic picture of text (such as a letter) into a form your text-based applications-such as word processors, DTP, spreadsheets, and databases-can use.

Technically speaking, OCR products look at a picture of a character and convert it into an ASCII or ANSI character that applications programs can utilize. This conversion process is called recognition.

- **Who's Using OCR and Why**

Over 65% of the time spent at the keyboard is spent retyping existing material that's been typed at least once already! With OCR, you save this valuable time.

Busy professionals are using OCR in many ways. For example, exchanging contracts between clients and attorneys involves many rounds of fax, retype, fax, retype, etc. OCR can easily convert the faxed document into WordPerfect™, for example. Columns of numbers from financial reports no longer need to be rekeyed into Lotus™ or Quattro™. OCR can put them directly into useable spreadsheet format.

There are lots of other examples, like storing articles and abstracts, revising manuals produced on obsolete or dedicated word processing equipment, and capturing data from forms. Furthermore, you can easily do things you would never have attempted.

For example, you can create a correspondence database, or file resumes on-line for easy applicant/position match-up. Or even analyze your phone bills by scanning them and putting the data into a spreadsheet, so you can easily total up the cost of calls to particular phone numbers or cities.

- **What Scanners Do**

Scanners are nothing more than fancy cameras. They simply take a picture of a page, then pass it to the PC in electronic format as a bit-mapped image file. Taking an electronic picture is called scanning.

For graphics applications, an electronic picture is just fine. Using a program like PC Paintbrush", you can easily clip the picture or illustration that you want to put in a report, newsletter, or brochure, for example.

Fax machines include a scanner; they take an electronic picture, then send it over telephone lines. So, if you are using OCR on a fax file, whoever sent the fax has done the scanning for you. What you receive on a fax board is a scanned image which can also be converted to text using OCR.

- **What Scanners Don't Do**

Scanners do not give you useable text. That's what OCR is all about.

- **How Do You Buy OCR?**

Generally, scanners and OCR are sold separately. Some specialized OCR products have both a scanner and an OCR engine built in. This means that you don't have to buy the components separately. Usually these types of systems are designed for production/high-volume applications and are more expensive than a system where the components are sold separately.

Your volume and throughput requirements will determine the solution most appropriate for you.

- **Beyond OCR: Getting it all Right**

There's more to OCR than recognizing the characters on the page. You've got to get it all and get it all right. Look at a few pages of text, and you'll begin to see some of the problems that simple OCR cannot address: multiple columns spanned by large headlines, a combination of type styles and faces, text wrapped around illustrations, indented lists, numbered lines, headers and footers, tables, and a number of other characteristics that would become a jumbled mess if the characters on the page were simply recognized and placed into lines of text.

If you've ever had to use a word processing program to reformat a document that someone typed "typewriter style" (spaces in place of tabs and centering, hard returns at the end of every line, and so

forth), you understand how important intelligent analysis and formatting can be.

The term document recognition means extracting everything that is important from a document. This ability separates highly competent OCR products from simple OCR.

The end results of document recognition can only be evaluated in your word processor, spreadsheet, or DTP.

- **Reading the Document Right**

Think about the types of documents you need to read. Are they all first-generation, clean and crisp, black on white pages? Or (more likely), are they photocopied pages, computer-printed pages, and pages that have been circulated, written on, and corrected? Do you need to read both typewritten and typeset documents?

Some OCR systems are severely limited by the types of document imperfections they can handle. Others are limited by what they consider to be a recognizable character. You should look for products that will handle the types of pages you need to read. Check for the ability to handle these kinds of pages:

- Typeset documents in a range of point sizes (footnotes can be as small as 6 points; big headlines sometimes range up to 28 points)
- Laser printed documents (whose characters tend to bleed together more than typeset documents)
- Typewritten documents
- Second, third, fourth ... generation photocopies
- Dot-matrix documents including draft-quality typical of spreadsheet output (some vendors claim to recognize dot-matrix, but they generally refer to "near letter quality" printing)
- Line printer listings
- Fax images from PC fax boards
- Paper faxes from a fax machine
- Keep the future in mind, and think about how others in your company can benefit from OCR.

If you need to read printed faxes or dot-matrix spreadsheets, for example, check to see if the product has a setting for dot-matrix. Run a sample page of draft-quality dot-matrix text and check the results. Generally, a special setting for dot-matrix or mono-spaced text will significantly improve recognition on these pages.

- **Reading columns Right**

Think about documents with multiple columns and tables. OCR products that automatically decolumnize multiple-column documents will be much easier to use than those that don't.

- **Why Decolumnize?**

If you scan a multi-column document to use in your word processing program and the OCR product does not decolumnize it, you'll see multiple columns on your screen. However, you'll soon realize that the columns are made by dividing each line with spaces or tabs. Imagine trying to edit the text in the first column. As you type, the characters in all the other columns move simultaneously, even though they are supposed to be in separate columns. And when they wrap around at the end of the line, the entire page becomes a mess.

When a document is decolumnized, each column is kept together as a separate entity. You can edit the text in one column without affecting the text in other columns.

What you see on your screen and in the final output depends on the formatting abilities of your application. Most word processing programs enable you to format and print in multiple columns, so recolumnization is possible.

- **Reading Tables Right**

On the other hand, you can have too much of a good thing. You don't want your recognition product to decolumnize tables, as they become nearly impossible to reconstruct. A good recognition product will decolumnize intelligently, distinguishing between body text and tables.

Check formatting of a multiple-column document and a document that has tables. Give preference to systems that can properly decolumnize documents with no operator intervention.

- **Reading Both Sides**

Double-sided documents can pose a problem to some OCR products. Some scanners have a tray that holds a stack of pages (generally 20 to 50 pages). However, if you need to scan both sides of the page, you may have to do one sheet at a time to get both sides in the proper order.

Check for the ability to read double-sided documents by scanning all of one side, then all of the other side, ending up with the whole document in the right order.

- **Reading at the Right Time**

Scanning long documents normally requires someone feeding in pages and waiting for both scanning and recognition to complete. Most products scan a page, then recognize it, then scan another, and so forth. Scanning is usually a matter of 3 to 11 seconds per page, depending on the scanner. Recognition can take considerably more time 50 to 60 seconds on a typical 2,000-character page.

Very few products offer the ability to "scan now, recognize later." This ability may be called deferred processing, delayed recognition, batch processing, or some similar term. This means you can do the fastest part of the job (the scanning) all at once, while the slower part (the recognition) can take place unattended, at another time. You can scan 100 pages, then go to lunch while recognition proceeds without further intervention.

Check for deferred processing and choose a reliable, high-speed scanner that can keep up with your workload.

- **Reading Parts of Pages Right**

Often, you won't want to recognize the entire page. For example, if you are working with magazine or newspaper articles, forms, or invoices, you should look for a product that supports zoning or clipping. With such a product, you see a graphic representation of each page on your screen, and then can draw boxes around the text or numeric areas you want to recognize and around the graphic images you want to capture.

Zoning increases throughput because the OCR system isn't spending time recognizing unwanted text. Clipping enables you to capture illustrations in such a way that you don't have to clean up the extraneous surrounding material later.

By combining deferred processing with clipping, you can really speed through magazine articles, where you want just part of the text on each page. Clipping lets you quickly specify the areas of each page you want to process. With deferred processing, you can quickly move from page to page without waiting for recognition to complete. Recognition can be done unattended after you've clipped what you want from each page.

Check for the ability to combine deferred processing with clipping different areas from each page.

- **Reading Forms Right**

Another obvious use of zoning is forms. Some OCR products let you define zone templates for processing similar pages. A smaller subset of products let you specify an identification zone to automatically determine which template to apply to each form, You then can process a stack of different forms; the OCR system determines which template to apply to each one.

If you plan to process multiple forms, make sure the OCR product you choose has an identification zone feature.

- **Processing Multiple Jobs at Once**

Some OCR products let you put a blank page between each separate document in a stack, then process the whole stack at once, saving each of the recognized documents in a separate file. This can be a handy feature if you routinely process resumes, job applications, reader response surveys, or just want to read several documents at a time.

If you choose an OCR product that allows page job separators, be sure to choose a scanner with a large-capacity, reliable document feeder.

- **Handling Formatting Right**

You may need to reformat documents to match your company's style. If you will routinely be reformatting a particular type of document, such as in technical manual conversion, you'll need a "style sheet" feature. This is useful for removing formatting so that you can easily apply your own. With custom style sheets, you can specify your own margins, indents, line spacing, page length, font, and so forth, and automatically apply these to your finished document.

If reformatting is important to you, choose a product with style sheet features.

- **Reading the Most Type Right**

The OCR systems available today provide vastly different font recognition capabilities. (See the Glossary at the end of this booklet for definitions of unfamiliar terms.) They can be broadly categorized as follows:

- **Polyfont Recognition**

Polyfont recognition means the ability to read several fonts, in many cases, a polyfont product will only recognize specific fonts and cannot recognize others.

Polyfont systems are sufficient when you are only going to read a specific set of known documents that you can test before you buy the system. However it's unlikely that you could anticipate all the fonts you need to read at the time you buy an OCR system, and once you've bought it, you're committed to this inherent limitation.

- **Trainable Recognition**

Trainable products may seem an attractive alternative to polyfont products. You can train the system to recognize virtually any font that comes along. What you'll find when you begin to use a trainable system, however, is that each individual font and style requires a time consuming training session. Even photocopying a document can change a font enough to require retraining.

Trainable systems are most useful for reading certain unusual display fonts and foreign languages such as Cyrillic. Keep in mind, however, that the training time must be added to the overall throughput of the system when making comparisons.

- **Omnifont Recognition**

Omnifont products can recognize virtually any font that maintains fairly standard character shapes. True omnifont systems require no training or other adjustments to accommodate different fonts. If you want the most versatile and powerful recognition, an omnifont system is probably the only type that will satisfy your needs.

Check recognition on a variety of different pages you're likely to want to read. Look for a system that gives good results and different pages without constant operator intervention.

- **Handling Poor Quality Documents**

Many OCR products offer adjustments that can improve things like page contrast for maximum accuracy. This can be helpful, especially if you have poor quality documents that have been photocopied several times.

- **Proofing the Document Right**

Some OCR products provide built-in or optional editors for Proofing a document after recognition. Only a very few use a text editor that can show you the actual image of a character or word just as the scanner saw it. By using image "pop - ups," you can dramatically cut your proofing time (often by up to 50%), because you won't need to constantly refer to the original paper document. You can see right away the correct character or word image, and if the text counterpart is misrecognized, correct it and move on.

Check for a built-in editor that includes pop-up images for easy verification.

- **There's No Substitute for Experience**

There is more to recognition accuracy than statistics and numbers. When you read a printed page, you are drawing upon years and years of experience and stored memories. Once you master reading, you no longer see individual characters on a page, you see patterns that merge into words, sentences, and ideas.

- **Dictionaries Help**

An effective OCR product must be able to overcome the limitations of seeing individual characters by evaluating groups of characters to see if they can make sense (by forming actual words, for example). This requires a recognition dictionary that checks for words as an integral part of the recognition process. The recognition device can do a much better job of "reading" a page if it has some "experience," in this case, a dictionary of proper word spellings. Products that check spelling after recognition cannot be as effective as products that end up with the correct recognition because they have already checked to see if the recognized character shapes form actual words.

Check for a dictionary or word list that is applied during recognition. Do not accept a spelling checker as a substitute for a recognition dictionary.

- **There's No Substitute for Accuracy**

Recognition accuracy has to do with how many characters are misrecognized on a given page. For example, on a typical 2000-character page, 98% accuracy means 20 errors; 99% accuracy means 10 errors. Think of this as a 100% difference in the number of errors, not a 1% difference in accuracy.

Beware of products that check and display their own recognition accuracy. The numbers displayed by these products are based on the number of failed attempts at recognition, not on the number of actual misrecognized characters.

- **There's No Substitute for Performance**

When choosing a recognition product, speed is certainly important. However, many other factors affect overall recognition performance. To understand the importance of performance, think about the amount of time you'll have to spend proofing and correcting your document if it contains lots of formatting errors, inappropriate decolumnization,

extraneous text and images, misrecognized characters and words, and so forth. Overall performance measurements must include:

- Scanning time
- Recognition time
- Verification time
- Revision time
- Reformatting time

Think about the amount of time it takes to go from printed page to a correctly formatted, correctly spelled on-line document. After all, the object of OCR is to save you time!

- **Is It Easy to Use?**

Recognition should not be complex. After all, the computer should do the hard work, leaving you to make simple selections.

How easily can you change settings or read just for different kinds of jobs? Remember that an OCR product should help you get your work done faster.

- **What About Support and Upgrades?**

As with any major purchase, you should look at companies that offer both the products you need today and the products you'll need tomorrow. Look for a company that specializes in OCR, not tape backup systems, spell checkers, or graphics tools. When you make the decision to put document recognition into your office, you'll find an ever increasing range of opportunities to use it. So look for a broad product line with upgrade options to allow a smooth migration to higher-performance products.