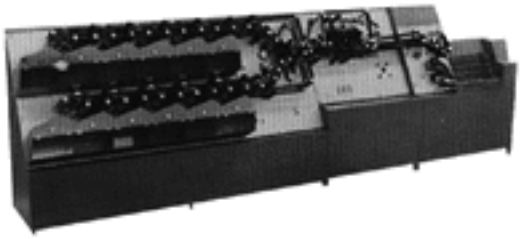
Optical Character Recognition (OCR)
What You Need to Know

By Phoenix Software International®

History of Optical Character Recognition

Optical character recognition (OCR) is the process of translating scanned images of typewritten text into machine-editable information. In the early 1950s, David Shepard was issued U.S. Patent Number 2,663,758 for "Gismo," the first machine to convert printed material into machine language. Shepard then founded Intelligent Machines Research Corporation (IMR), which produced the first OCR systems for commercial operation. Reader's Digest installed the first commercial system in 1955. The United States Postal Service has been using OCR machines to sort mail since 1965.



1966 OCR letter sorting machine

Today, OCR technology incorporates high-speed scanners and complex computer algorithms to increase speed and data accuracy. OCR systems no longer require training to read a specific font. Current systems can recognize most fonts with a high degree of accuracy and some are capable of outputting formatted text that closely approximates the printed page.

Types of Recognition Engines

Optical Character Recognition (OCR)

OCR engines turn images of machine-printed characters into machine-readable characters. Images of machine-printed characters are extracted from a bitmap. Forms can be scanned through an imaging scanner, faxed, or computer

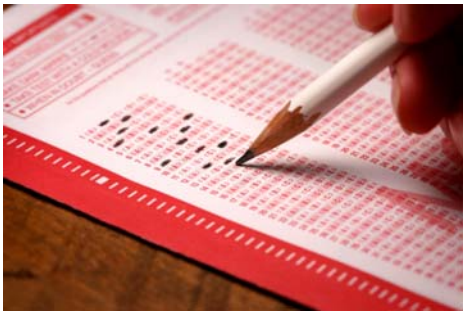
generated to produce the bitmap. OCR is less accurate than optical mark recognition but more accurate than intelligent character recognition.

Intelligent Character Recognition (ICR)

ICR reads images of hand-printed characters (not cursive) and converts them into machine-readable characters. Images of hand-printed characters are extracted from a bitmap of the scanned image. ICR recognition of numeric characters is much more accurate than the recognition of letters. ICR is less accurate than OMR and requires some editing and verification. However, proven form design methods outlined later in this paper can minimize ICR errors.

Optical Mark Recognition (OMR)

OMR technology detects the existence of a mark, not its shape. OMR forms usually contain small ovals, referred to as 'bubbles,' or check boxes that the respondent fills in. OMR cannot recognize alphabetic or numeric characters. OMR is the fastest and most accurate of the data collection technologies. It is also relatively user-friendly. The accuracy of OMR is a result of precise measurement of the darkness of a mark, and the sophisticated mark discrimination algorithms for determining whether what is detected is an erasure or a mark.



The College Board SAT uses OMR technology

Magnetic Ink Character Recognition (MICR)

MICR is a specialized character recognition technology adopted by the U.S. banking industry to facilitate check processing. Almost all U.S. and U.K. checks include MICR characters at the bottom of the paper in a font known as E-13B. Many modern recognition engines can recognize E-13B fonts that are not printed with magnetic ink. However, since background designs can interfere with optical recognition, the banking industry uses magnetic ink on checks to ensure accuracy.

E-13B control characters

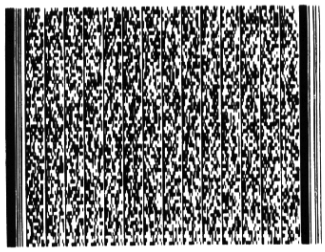
E-13B numerals

Barcode Recognition

A barcode is a machine-readable representation of information. Barcodes can be read by optical scanners called barcode readers or scanned from an image using

software. A 2D barcode is similar to a linear, one-dimensional barcode, but has more data representation capability.

1D barcode
Universal Product Code (UPC)



2D barcode
The Gettysburg Address

The Benefits of Electronic Paper

When paper documents are scanned and translated into electronic data, companies benefit from faster and more reliable access to vital information. They have an efficient format for storing and distributing documents. Companies can keep the information in a central place and gain control over how the documents are accessed and altered by personnel. Following is a list of some of the key benefits:

Access

- An entire company with multiple sites can access documents on a central server.
- Robust database applications can manage electronic documents, performing searches based on document location or content.
- A content management program can add value to an electronic storage system, allowing users to store additional information with the document.
- Finding information within a long document is easier and faster.
- Multiple users can access an electronic document simultaneously.
- Users can easily and instantly distribute documents to a number of people at once via email.

Control

- Electronic document storage systems prevent documents from being misfiled or erroneously deleted.
- A central content management system serves as a single source for up-to-date information.
- The system can keep track of document revisions maintaining a record of the who, when, and what of every change made.
- Security is easier to maintain in an electronic environment. Administrators can control who can see, read, modify, or destroy a particular document.

- Electronic documents can easily be stored offsite as part of a disaster recovery program.

Resource Efficiency

- Electronic documents use less office space than traditional paper files
- Paper documents that need to be retained can be moved off-site to storage facilities.
- Electronic filing systems save human resources because users can access files on their own rather than requiring the help of support staff.

Best Practices

OCR systems can translate into significant savings for companies but only if they are able to achieve a certain level of speed and accuracy. Following are some guidelines to help companies get the best possible OCR results.

Document Scanning

Before OCR software can be used on a paper document, the document must be scanned and saved in an electronic format. Scanners come with their own software to accomplish this task. System requirements will vary from one OCR product to another but here are some general guidelines for scanning images suitable for character recognition:

- Use a high-end, hopper-fed, high-speed professional scanner designed specifically for document scanning.
- Use a scanner that allows you to drop out one color from your scan to enable "form removal." This is discussed in more detail in the *Form Design* section of this paper.
- Scan documents at a resolution of 300 dpi or greater. Optimal resolution will depend on whether the content is machine or hand-printed.
- Use a lossless file format such as TIFF. Use LZW compression to reduce file size and save space.
- Make sure scanned images are aligned properly. Some OCR products will compensate for alignment problems by de-skewing images to a certain degree.

Form Design

Most companies implement an automated forms processing solution to reduce labor costs. More accurate OCR/ ICR means more data captured without the need for human intervention resulting in greater cost savings.

Recognition of hand-printed forms can be challenging. ICR (hand-printing recognition) engines work best when companies can design their own forms. Form design can be vital to ICR accuracy. In some cases, a properly redesigned

form will result in the elimination of virtually all ICR errors while reducing the number of characters that require verification.

Over the years, certain form design practices have proven to yield the best results. A properly designed form encourages the user to fill out the form completely and accurately. Forms should be visually appealing and well organized. The form must be easy to fill out and allow the user to write normally where possible. But the most critical element to ICR accuracy is the separation of characters. A good form keeps users from running characters together.

Here are some form design tips:

- When forms must be processed with a number of other forms, use a bar code or special mark to so the system can identify the form. Assure the identifier is not in the dropout color.
- Define fields to encourage answers in the correct format such as mm/dd/yyyy for date fields and (xxx) xxx-xxxx for telephone numbers. OCR/ICR can then use the format to interpret the characters correctly.

Encourage proper date formatting

- Add "internal use" boxes for clerks to rubber stamps or comments on the original paper form. Make sure the box is large enough to allow for any misalignment.
- Design your form with all lines and labels printed in a drop-out color. The scanner uses a colored light that eliminates one color on the resulting image file. Refer to your scanner manual for colors defined as drop-out colors and how to use them.
- If you cannot use a drop-out color, assure that labels identifying fields do not constrain the area in which the respondent will write. Ideally, use rule lines to separate the label from the field.
- Use drop-out boxes or tick marks (often called combs) to constrain answers. These methods encourage users to separate each letter and are used for the most critical fields. Supply an example of how to fill in those fields if necessary.

Drop-out boxes

Drop-out tick marks

- Make check boxes designed for OMR processing large enough and far enough apart to keep marks in one box from spilling over into the next.

- To avoid confusion, use as few methods as possible to collect the information. These methods include multiple choice questions, yes/no questions, constrained answers and unconstrained answers.
- The signature field should be large enough, and positioned so the user is not forced to mark nearby fields when signing the document.
- Detect the presence of a signature using OMR technology.
- Limit ICR fields to numeric characters, which can be recognized with greater accuracy.
- Avoid putting key fields where they are likely to get creased either before or after the form is completed (such as folds to fit into an envelope).
- Print registration marks to enable precise de-skewing.

Data Validation

When data is captured, the engine algorithms assign a "confidence" level to each character. If that value is above a certain user-defined threshold, the character is considered to be correct. Decreasing character confidence levels can increase the number of errors found, but also increases the amount of operator intervention required. Administrators will likely find the optimum level can vary depending on the particulars of different applications.

However, it is possible to reduce the number of errors in a field to almost zero without increasing the need for manual review by using various methods of validation. Professional data entry software can provide built-in validation routines, as well as the means to program custom routines to suit your application.

Some examples of validation include checksum routines for credit card numbers, databases for checking customer or social security numbers, postal code look-up tables, mathematical checks for invoice totals, ranges for date fields, etc. Some systems include routines to test the engine's second best guess. For example, the best guess of "1" might fail, while the second best guess of the letter “l” might pass the validation routine, thus avoiding manual review of the field.

For important fields, validation routines can use data from the form itself. For example, users can be asked to put in their age and their birth date. The system can compare the two fields and require operator review if the fields do not match. Methods that require extra user effort should be kept to a minimum.

Manual Data Entry

Data entry systems can be set up to require operator verification of some or all OCR fields. Using the double-key verify method, an operator types the data from an image without seeing the original OCR input. If the operator input doesn't match the recognition engine results, the operator is alerted and can make corrections as necessary.

To complete the forms processing operation, all data not read by the recognition engine must be handled efficiently while ensuring maximum data accuracy. Key-from-image and key-from-paper software along with data validation can be

implemented to achieve this goal. For high-volume data processing, indexing modules are generally not sufficient, since these systems can not keep up with professional data entry typing speeds.

Summary

For certain types of clear, consistent paper documents, OCR technology provides fast, automated data capture. Proper scanning, good form design, sufficient data validation, and targeted manual review deliver accurate results with huge savings over manual processes. Errors can be avoided through solid planning and good follow-through.

About Phoenix Software International

Phoenix Software International, Inc. is a major systems software development company providing advanced software solutions to enterprises around the world. Our diverse products support IBM and compatible mainframes, personal computers, and local and wide area networks. PSI customers range from small entrepreneurial companies to major federal and state agencies to Fortune 500 leaders in the telecommunications, automotive, and insurance industries, among others. PSI has been providing software solutions since 1979.

Phoenix Software offers Falcon64™, a complete data entry application with key-from-image capabilities and an optional optical character recognition (OCR) engine.

Some key features include:

- Input conversion for popular image formats (TIFF, JPEG, PDF)
- Optical character recognition (OCR)
- Zone definition—define recognition zones and key-from-image zones
- Option to put the image window on top and fill the screen for a heads-down data entry experience
- Robust keyboard orientation control as demanded by data entry experts
- Split-second response rate for displaying the next image
- Workflow management system—supervisors organize, prioritize and assign data entry tasks to streamline production.

More information can be found on our website at www.phoenixsoftware.com.

Phoenix Software International®
831 Parkview Drive North
El Segundo, California 90245-4932
1 (310) 338-0400
http://www.phoenixsoftware.com