AIM

The Association for Automatic Identification
and Data Capture Technologies

Optical Character Recognition (OCR)

This technical paper was developed by AIM, the global trade association of providers and users of components, networks, systems and services that manage the collection and integration of data with information management systems. AIM strives to stimulate the understanding, adoption, and use of technology and member company products and services through setting standards, marketing and education, market research, advocacy and information technology industry relations.

AIM member companies were offered the opportunity to review and make contributions to the document prior to publication.

Technical papers are intended to be informational only and serve to help the manufacturer, the consumer, and the general public understand the subject contained within.

AIM, Inc., its member companies, or individual officers assume no liability for the use of this document.

Table of Contents

What is OCR?

OCR is the acronym for Optical Character Recognition. This technology allows a machine to automatically recognize characters through an optical mechanism. Human beings recognize many objects in this manner our eyes are the "optical mechanism." But while the brain "sees" the input, the ability to comprehend these signals varies in each person according to many factors. By reviewing these variables, we can understand the challenges faced by the technologist developing an OCR system.

First, if we read a page in a language other than our own, we may recognize the various characters, but be unable to recognize words. However, on the same page, we are usually able to interpret numerical statements - the symbols for numbers are universally used. This explains why many OCR systems recognize numbers only, while relatively few understand the full alphanumeric character range.

Second, there is similarity between many numerical and alphabetical symbol shapes. For example, while examining a string of characters combining letters and numbers, there is very little visible difference between a capital letter "O" and the numeral "0." As humans, we can re-read the sentence or entire paragraph to help us determine the accurate meaning. This procedure, however, is much more difficult for a machine.

Third, we rely on contrast to help us recognize characters. We may find it very difficult to read text which appears against a very dark background, or is printed over other words or graphics. Again, programming a system to interpret only the relevant data and disregard the rest is a difficult task for OCR engineers.

There are many other problems which challenge the developers of OCR systems. In this paper, we will review the history, advancements, abilities and limitations of existing systems. This analysis should help determine if OCR is the correct application for your company's needs, and if so, which type of system to implement.

History of OCR

The engineering attempts at automated recognition of printed characters started prior to World War II. But it was not until the early 1950's that a commercial venture was identified that justified necessary funding for research and development of the technology. This impetus was provided by the American Bankers Association and the Financial Services Industry. They challenged all the major equipment manufacturers to come up with a "Common Language" to automatically process checks. After the war, check processing had become the single largest paper processing application in the world. Although the banking industry eventually chose Magnetic Ink Recognition (MICR), some vendors had proposed the use of an optical recognition technology. However, OCR was still in its infancy at the time and did not perform as acceptably as MICR. The advantage of MICR was that it is relatively impervious to change, fraudulent alteration and interference from non-MlCR inks.

The "eye" of early OCR equipment utilized lights, mirrors, fixed slits for the reflected light to pass through, and a moving disk with additional slits. The reflected image was broken into discrete bits of black and white data, presented to a photo-multiplier tube, and converted to electronic bits.

The "brain's" logic required the presence or absence of "black" or "white" data bits at prescribed intervals. This allowed it to recognize a very limited, specially designed character set.

To accomplish this, the units required sophisticated transports for documents to be processed. The documents were required to run at a consistent speed and the printed data had to occur in a fixed location on each and every form.

The next generation of equipment, introduced in the mid to late 1960's, used a cathode ray tube, a pencil of light, and photo-multipliers in a technique called "curve following". These systems offered more flexibility in both the location of the data and the font or design of the characters that could be read. It was this technique that introduced the concept that handwritten characters could be automatically read, particularly if certain constraints were utilized. This technology also introduced the concept of blue, non-reading inks as the system was sensitive to the ultraviolet spectrum.

The third generation of recognition devices, introduced in the early 1970's, consisted of photo-diode arrays. These tiny little sensors were aligned in an array so the reflected image of a document would pass by at a prescribed speed. These devices were most sensitive in the infra-red portion of the visual spectrum so "red" inks were used as non-reading inks. That brings us to this generation of hardware:

Where are we today?

The advent of the array method of scanning, coupled with the higher speeds and more compact computing power, has led to the concept of "Image Processing". Image processing does not have to utilize optical recognition to be successful. For example, the ability to change any document to an electronically digitized item may effectively replace microfilm devices. This provides the user a much more convenient method of sorting images compared to handling actual documents or microfilm pictures. Image processing relies on larger more complex arrays than early third generation OCR scanners.

When these image scanners are coupled with OCR logic, they provide an extremely powerful tool for users. Image recognition can be done in an "off-line" mode rather than in "real time" - a tremendous advantage over earlier versions of OCR devices. This allows a much more powerful logic system to work over time and requires less rigorous demands on both the location of the information and the font design of the characters to be scanned. An example of this is found in the coupling of "image with convenience amount recognition" planned for the Financial Services Industry for check processing - still the world's largest paper processing application. This will be the first viable marriage of MICR with optical technology.

What are its Applications?

OCR has been used to enter data automatically into a computer for dissemination and processing. The earliest of systems was dedicated to high volume variable data entry. The first major use of OCR was in processing petroleum credit card sales drafts. This application provides recognition of the purchaser from the imprinted credit card account number and the introduction of a transaction. The early devices were coupled with punch units which made small holes to be read by the computer. As computers and OCR devices became more sophisticated, the scanners provided direct access into the CPU (computer processing unit). This quickly lead to the payment processing of credit card purchases, known as "remittance processing". These two applications are still the two major applications for OCR.

Over time, other applications evolved. They included cash register tape readers, page scanners, etc. Any standard form or document with repetitive variable data would be a candidate application for OCR. Some very imaginative applications have evolved. Perhaps the most innovative are the Kurzwell scanners which read for the blind. With these devices, the optically scanned pages are converted to spoken words.

What are its Competitors?

OCR has never achieved the success that was anticipated in the 1950's. One of the main reasons has been the introduction of on-line systems. Much of the data that could have been OCR input has gone to Point Of Sale (POS) devices. For example, the Petroleum retailers are installing POS systems in gas stations around the country which no longer require scanning of credit card sales invoices.

In the same manner, electronic cash registers have replaced the need for OCR reading of cash register rolls. The applications that were used for page readers, insurance forms, etc. have, for the most part, gone to terminal entry.

Remittance processing still continues to be the primary application for OCR. This has been a successful system, but still has limitations.

What are its Limitations?

OCR has never achieved a read rate that is 100% perfect. Because of this, a system which permits rapid and accurate correction of rejects is a major requirement. Exception item processing is always a problem because it delays the completion of the job entry, particularly the balancing function.

Of even greater concern is the problem of misreading a character (substitutions). In particular, if the system does not accurately balance dollar data, customer dissatisfaction will occur.
The success of any OCR device to read accurately without substitutions is not the sole responsibility of the hardware manufacturer. Much depends on the quality of the items to be processed.

Through the years, the desire has been:
- to increase the accuracy of reading, that is, to reduce rejects and substitutions
- to reduce the sensitivity of scanning to read less-controlled input
- to eliminate the need for specially designed fonts (characters), and
- to read handwritten characters.

However, today's systems, while much more forgiving of printing quality and more accurate than earlier equipment, still work best when specially designed characters are used and attention to printing quality is maintained. However, these limits are not objectionable to most applications, and dedicated users of OCR systems are growing each year. But the ability to read a special character is not, by itself, sufficient to create a successful system.

What does it take to make a successful OCR System?

1. It takes a complimentary merging of the input document ~ stream with the processing requirements of the particular application with a total system concept that provides for convenient entry of exception type items with an output that provides cost effective entry to complete the system. To show a successful example, let's review the early credit card OCR applications. 1. Input was a carbon imprinted document. However, if the carbon was wrinkled, the imprinter was misaligned, or any one of a variety of reasons existed, the imprinted characters were impossible to read accurately.

2. To compensate for this problem, the processing system permitted direct key entry of the fail to read items at a fairly high speed. Directly keyed items from the misread document were under intelligent computer control which placed the proper data in the right location for the data record. Important considerations in designing the system encouraged the use of modulus controlled check digits for the embossed credit card account number. This, coupled with tight monetary controls by batch totals, reduced the chance of read substitutions.

3. The output of these early systems provided a "country club" type of billing. That is, each of the credit card sales slips was returned to the original purchaser. This provided the credit card customer with the opportunity to review his own purchases to insure the final accuracy of billing.
This has been a very successful operation through the years. Today's systems improve the process by increasing the amount of data to be read, either directly or through reproduction of details on the sales draft. This provides customers with a "descriptive" billing statement which itemizes each transaction.

Attention to the details of each application step is a requirement for successful OCR systems.

Input

When installing an OCR system, the most important consideration is the manner of creating input.

1.   How do you intend to create the input? If the input is typewritten data, how many different typewriters will create the input? Will they be electronic, electric, or manual? What type styles or fonts do they have? Will the typewritten material be from a fabric or carbon ribbon? This gives you an idea of the information you need to obtain.

2.   What kind of a document will be used for the application? For most systems, the data to be scanned must occur in the same location from document to document. Guide lines or the location of data identifiers, need to be pre-printed. Do they need to be in a "non-reading" color (drop-out ink)? Where will they be printed? What size will they be? Will the form meet the requirements specified by the scanner manufacturer? Will the right data be in the right location for best digit or balancing routines to facilitate performance? Remember, your attention to detail and reviewing the "what if" possibilities before installation will save a tremendous amount of dissatisfaction later.

3.   How will input be handled both prior to preparation, after printing, and after processing? If accurate registration must be maintained, moisture-proof wrapping your pre-printed forms may be necessary. If the item is to be mailed to the processing center individually, you may want to prescribe a heavy duty envelope to prevent damage in transit. If the items are to be picked up in large quantities, a special basket or other carrier may be required to ensure documents are not damaged. Although a rubber band is a fine tool to bind a group of documents together, it is a prime cause of damage to paper documents that are to be processed in an automatic feeding device. If you require subsequent archival of the documents for retrieval purposes, proper storage containers are required. You may also need a pre-printed serial number to help research archived material.

These are but a few of the questions that need to be answered. Your OCR system manufacturer is the best source of information on individual system input requirements.

Importance of Font Design

Input, as we have seen, is very dependent on the application. This is especially true when considering the design of the font (style of characters).

For example, the first OCR device used in a commercial application read carbon imprinted credit card sales drafts. The font used is known as 7B. This font was designed by the Farrington Corp. for this type of imprinting. The characters are large enough to be embossed on a plastic card. The "lakes" (open areas) of 6,9,0, and 8 have been opened so the carbon does not fill in those areas. The numbers are distinctly different from each other to reduce the possibility of substitution. This particular font is still the standard for this application .

One of the earliest OCR devices to read input from a data processing printer was the IBM 1418. At the time this device was designed, the printer used most was an IBM accounting machine called the 407. Therefore, the 1418 was designed specifically to read the 407 font. Due to style problems in some characters the font was modified and now retains the designation 407-1. This font is still used in some applications .

IBM then introduced an OCR machine to read a full alphanumeric character set with its 1428 reader, thereby establishing the 1428 font. A modified version of this font, known as 1428E, is also available in an elongated style for imprinter applications.

These fonts formed the basis for standardized fonts established by ANSI, the American National Standards Institute. This organization is comprised of participants who have agreed to create voluntary compliance standards.

Two standard fonts established by this organization will improve the overall performance for OCR systems. The OCR A font is stylized and similar to the early 1428 font, Today, it is widely used in remittance processing billing documents where information to be scanned is on a separate line from the information to be visually read by the customer. Every scanning device manufactured today can read this font due to its proven reliability.

The OCR B font is used in applications where data to be scanned must also be read by humans. It is less stylized in appearance than OCR A/ and is used to a great extent in European Countries.

These fonts are available in three sizes. Size 1 is commonly used by high speed data-processing printers. Size 2 has been expanded for use on devices such as cash registers that use a numbering sheet type of printer. Size 3 is even larger for use as an imprinter font. These gradations in size are proportional, allowing the fonts to be electronically reduced to the same size for presentation to the logic of the readers.

Using these fonts appropriately allows users to select readers that are very reliable and cheaper than devices capable of reading intermixed, multiple fonts. For the most part, today's OCR readers recognize several fonts, although they are most efficient and successful when running documents printed with a single font at a time.

Documents

Once the application has been defined, a decision must be made regarding which paper should be used and what information should be printed prior to distribution.

Paper Considerations

The vendor of the equipment usually has a list of basic requirements that must be met for the item to be successfully processed. The basis weight is the first characteristic to be specified and will play a major role in the ultimate cost of the document. In general, most systems in place today prefer the use of a 24 Ib. sheet. (In other words, 500 sheets of paper, 17" x 22", would weigh 24 Ibs.) Paper thickness or caliper is also a consideration.

Other characteristics that may be specified include: stiffness, tear strength, bursting strength, fold resistance, porosity, etc.. Any one of these may play a part in the processability of the document. Other characteristics, such as smoothness, may be specified to provide a better printing surface. In addition, some vendors specify the cleanliness of the paper to avoid inadvertent "slime" spots that will not interfere with scanning.

Paper Color

For certain applications, it may be desirable to use colored stock to help users readily identify documents for use in different applications. Color coding is a very simple and satisfactory visual system. However, since OCR depends on the contrast between the printed characters and the background color, some color control is necessary.

Color for the most part is controlled by reflectance. Standardized tests are available to measure the relative reflectance of a sheet in comparison to absolute white (defined as 100% reflectance). For most systems 60% reflectance is the minimum that can be used. The readability of any individual character is determined by the print contrast signal it generates. This is determined by the formula, "PCS =

reflectance of the background minus reflectance of the printing divided by the reflectance of the background ."

Printing Inks

As with paper color, the color of inks used for pre-printing the document are also very important. Some data that is pre-printed, such as a serial number, may be OCR read, but other pre-printed data should be ignored. This creates a new specification for non-reading inks.

Read vs. Non-read Inks

Read inks need to contain sufficient print contrast to be easily read by the OCR system. Black is the preferred color, but even black inks with insufficient density or coverage may not be recognized. Other colored inks will work in most systems if the PCS (print contrast signal) is 30% or less.

Non-read inks are dependent on the response level of the OCR system used. For example, if it is a cathode ray scanner, it probably responds to the ultra-violet spectrum. A non-read ink for this system would be light blue. If the system used is sensitive to the infra-red region of the spectrum, then a nonread ink would be red.

Where is OCR going?

With the advent of higher computer speeds and desktop personal computers, scanners are being developed for desk use for such things as high-speed entry of articles, or other information, through imaging devices. Not all of these utilize OCR recognition, but as the logic to handle intermixed fonts is developed, the number will increase. Over time, OCR will become more powerful and less expensive. Recognition will be done from captured images rather than from the actual item. There will continue to be a need for improvement in handwritten character recognition and to reduce the fairly stringent document requirements of today's systems.

What Applications?

While many applications today use direct data entry via keyboard, more and more of these will return to automated data entry. The reasons for this include the increased incidence of operator wrist problems from constant keying and the potential hazards of video display terminal emissions. Therefore any application imaginable is a candidate for OCR.

Future of Optically-Read Handwriting

Many applications exist where it would be desirable to read handwritten entries. Reading handwriting is a very difficult task considering the diversities that exist in ordinary penmanship. However, progress is being made. Early devices, using non-reading inks to define specifically-sized character boxes, read constrained handwritten entries. This resulted in the development of a standard encouraging a certain style of handwriting. The best example of unconstrained handwriting reading was the IBM 3895. This device read the convenience amount entries from checks and then encoded the amount on the check in magnetic E13B characters. It is difficult to design a system to take care of mis-read characters. The 3895 also reads the entries from deposit listings to confirm or to prevent substitutions.

With the advent of image processing systems, this type of recognition is once again being developed. Restrictions on character size and the ability to provide target areas that are outlined in non-read inks will assist the accuracy of recognition. It would be helpful if our school systems could teach the proper manner to write numeric characters to enhance recognition.

Conclusion

What does the future hold for OCR? Given enough entrepreneurial designers and sufficient research and development dollars, OCR can become a powerful tool for future data entry applications. However, limited availability of funds in a capital-short environment could restrict the growth of this technology. It will be very difficult to identify a single application that could generate a sufficient return on investment for extensive research. Marketing professionals will have to create enough general use applications to justify these expenditures. But, given the proper impetus and encouragement, the automated entry of data by OCR is one of the most attractive, labor reducing technologies available.