Lower costs and higher reliability in optical character recognition bring data entry systems into a new developmental phase that utilizes marketing, mathematical, and engineering skills

OCR System Design Benefits from Technological Advances

Richard K. Dove

Ball Computer Products, Incorporated
Oakland, California

Technological advances have recently enabled optical character recognition to enter a new phase of development. These advances offer the system designer considerably lower costs and higher reliability and accuracy—a new cost-effective solution to old data entry problems.

Furthermore, as part of this phase, both systems and product designers have learned that they cannot assure success by simply applying the latest state-of-the-art components and techniques. They now realize that success is most likely to result from an integrated design team that combines the skills of engineers, mathematicians, and marketing people.

This phase-transition results from advances in paper transports, in scanners and digitizers, and in recognizers—as well as in system integration. Transports bring the document into position for data entry; scanners and digitizers transform data on the document into patterns of electrical pulses; and recognizers interpret those patterns and translate them into conventional standard codes for further processing.

Advances in Paper Transports

One of the problems that mechanical designers live with has been aptly expressed: "There is no such thing as mechanical hi-fi." These mechanical limitations must be tolerated in many electronic systems, including optical character recognition (OCR) devices, where serious difficulties arise from the mechanics of paper movement, the source of most machine failures. For example, a machine cannot scan documents any faster than it can unstack and restack them. Although it can work faster with heavy paper or card stock, it should be able to accommodate less expensive and lighter papers, including common bond, which have irksome aerodynamic qualities.

Vacuum and friction belts, previously not feasible in many designs, can now be useful if they are made from certain new materials that are based on synthetic fibers impregnated with natural or synthetic rubber. In particular, rubberized nylon is such a material; another is neoprene with dacron. Both of these are very strong, and have good friction characteristics and long wear.

For example, one simple paper-feeding mechanism uses four ½-in. wide belts, side by side, that extend into the top margin area of a stack of forms placed facedown directly on the scanner (Figs. 1 and 2). After the bottom sheet has been read, the friction belt pulls it out of the scanning area, through a gate adjusted to let a single sheet pass but not two. This belt mechanism functions as both picker and transporter; documents need not be moved before scanning. The weight of the stack itself, with perhaps a cap to hold down the last few sheets, is the force that creates the friction against the belt. If the transport mechanism fails, documents may be manually placed on the scanner; unfortunately, if the scanning mechanism fails, the transport is useless.

Other mechanisms feed the top sheet from the stack, which must rest on a pallet that rises as the stack diminishes in height (Fig. 3). In general, these mechanisms require that the sheet be moved into position for scanning, and then moved out to make room for the next sheet.

Nevertheless, moving paper, even at low speeds, makes the registration problem difficult. To maintain rigid and positive control over a moving piece of paper,

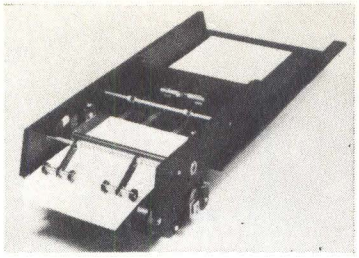

Fig. 1 Simple transport assembly. Success of this method of moving paper in an optical character reader is due in part to a new high-friction material used in the four belts that carry the paper

mechanisms must be designed with such close tolerances that the incidence of mutilated documents rises sharply.

Since any device that moves documents is subject to jamming, the product designer must not only make paper-jam clearing easy, but must also keep such jams and other single-point failures from seriously affecting the system's throughput. Most downtime of electronic systems is attributed to its mechanical portions—in an OCR system, the paper mover. As OCR equipment cost continues to decrease, instant service response by resident engineers grows less economical. Thus, to protect installations that cannot tolerate downtimes of more than 15 minutes or so, an OCR system's paper-transport mechanism should be designed so that, if it fails, documents can be manually placed for scanning.

One way to bypass some paper handling problems is to keep the paper stationary and move the scanner. Although scanners in the past have required solid mounting, the advent of plastic lenses, which make optical systems much lighter, as well as other weight savings that are possible with the newer solid-state electronic devices, have made the use of moving scanners feasible.

However, moving the scanner introduces a new problem: finding a way of getting electrical signals from the moving scanner to the rest of the system, which remains stationary. This can be done using flat cables, developed to withstand many million flexings before failure. These cables are used, for example, in IBM's 3886 and in the Ball Computer Products OCR 7600, both of which use mechanically moving scanners.

Advances in Scanners and Digitizers

In its simplest form, a scanner consists of a light source, a means of directing the light over the two dimensions of a document, an optical subsystem to form an image from the reflected light, and a light sensor to detect the resulting image. The scanner must also be capable of correlating each sample of reflected light with the coordinates of the position on the scanned document from which it was obtained. The source must provide enough light in the short time allowed, limited by the instantaneous character recognition rate, to insure a good signal-to-noise ratio in the sensor and to permit adequate resolution in the digitizer, while not exceeding the amount of heat that is allowable.

In recent years, beginning with the invention of the laser, many new components have been developed that improve scanner technology. The most interesting advances have come in solid-state technology, including silicon junction diodes, self-scanned linear arrays, light-emitting diode arrays, TTL-compatible phototubes, silicon photodiodes with modified spectral response curves, and hybrid packaging of photosensing devices. Others, however, are not to be discounted, such as fiber optics, high resolution vidicons, and optical shaft encoders. These developments increase scanning speed, improve

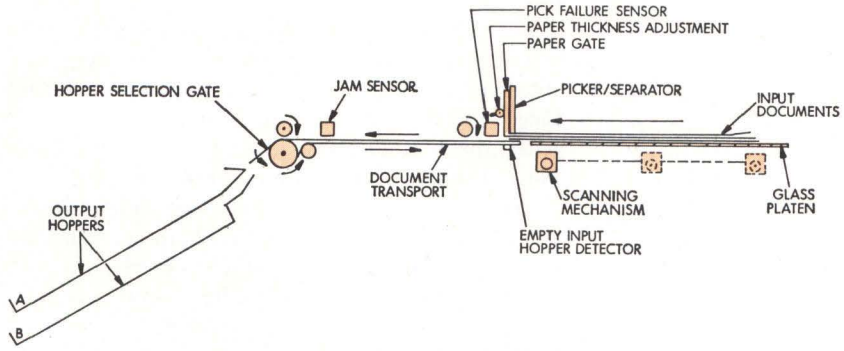

Fig. 2 Paper transport mechanism. Cross-section of assembly pictured in Fig. 1 shows how documents are scanned before being moved. Stack of documents is placed facedown on glass platen; scanner moves under glass to read data from bottom document. After scanning, picker pulls bottom document from stack; transport deposits it in one of two output hoppers, usually depending on whether it was read successfully

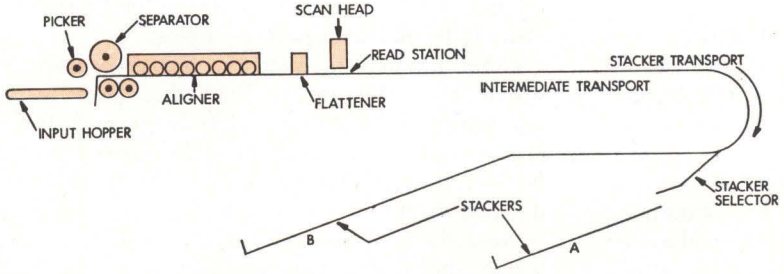

Fig. 3 Alternative transport method. Some OCR machines remove top document from stack, pull it under a scanner (read station), pause for reading, then drop it into one of two or more output hoppers

resolution, and reduce the size and weight of the scanner—the last opening up possibilities for moving subassemblies. Eventually, volume production of these components will considerably reduce associated manufacturing costs.

One of the primary problems in choosing a light source is obtaining enough illumination without generating too much heat. A significant advance in this respect, the laser, produces an extremely narrow beam of coherent light and does not waste power; it concentrates the full power of the light source on that portion of the document which is being illuminated, losing none to the surrounding area. Additional advantages of the laser are that the light beam itself can be directed to effect the actual scanning; the usual red light is well suited to silicon junction sensing, thus using energy efficiently.

Primary drawback of the laser is the danger that its beam will enter someone's eye, causing severe damage or blindness; consequently, it can generate an uncomfortable feeling in the minds of operators and other personnel near the machine, even when the beam is shielded. Lasers also tend to make production machines difficult to ship and install.

Flash tube technology has been refined to a state that may lead to a significant breakthrough. For instance, a flash tube could be used in conjunction with a high resolution vidicon in a high speed OCR system. As documents are moved continuously past a scan area, a high speed strobe unit would illuminate the entire document; the vidicon would capture all the data instantly, and scan it electronically while the next page is moved into position.

Tungsten filament bulbs have become brighter, smaller, and lighter, all concurrently; their output wavelength is suitable for sensing in a silicon junction. Their small size makes them suitable for use in mechanical scanners previously mentioned.

Light-emitting diodes produce wavelengths well matched to those sensed by silicon junctions and, like lasers, use energy efficiently. They are available in linear arrays, which can be coupled with a single sensor [as in the IBM 3886 (Fig. 4)] to achieve a scanning digitization that is the inverse of the digitization of an array of photodiodes coupled with a single light source (as in the Ball Computer Products OCR 7600).

Scanning the Document

Light can be sensed from various points on a 2-dimensional document in any of several ways. In some applications, the problem is considerably simplified if only one or two lines on each document must be read. If these documents contain many lines of printing or graphic material from which the alphabetic characters must be distinguished, the data to be read must be identified in some way that is intelligible to the OCR machine. This identification may be as simple as precise placement of the characters on the document or the appearance of a special character at the beginning of readable lines; or it may require a servo-controlled search.

If, however, many lines or a whole page must be read *in toto*, a true 2-dimensional scan is necessary (Fig. 5). The scanner may be fixed, and the documents moved past a read station; difficulties in moving documents have already been discussed. Second, the scanner may move as a unit, going from point to point over a fixed document, presenting both mechanical and interconnection problems. Third, a combination of these is

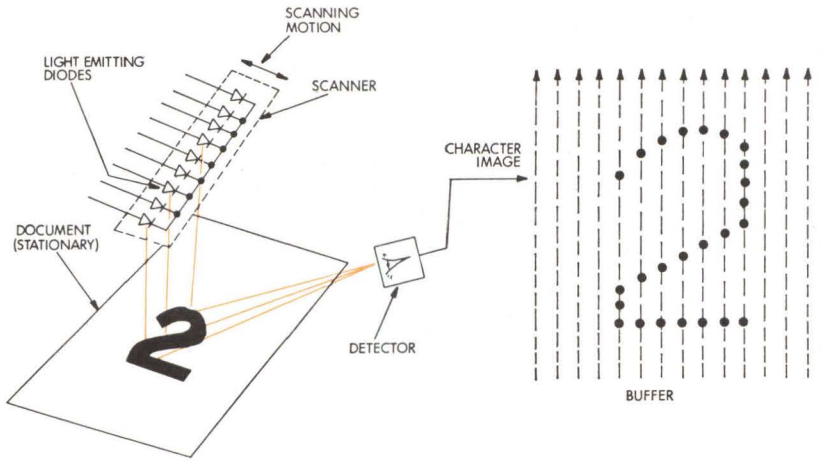

Fig. 4 Multiple light source scanner. One method of scanning documents is to move an array of sources (light-emitting diodes in this example) back and forth across the document, in the X direction, turning on one source at a time to scan at different levels in the Y direction. Output is an image of the character in the buffer

possible, with the scanner moving in the X direction and the document in the Y direction, or vice versa. Fourth, the direction of the optical imaging path may change—an alternative that also presents problems, because optical paths that change direction usually also change length, and therefore must be refocused.

However, at least one optical reader curves the document along a cylindrical arc and deflects the imaging path along the same arc without refocusing it. Another approach is to make the focal length of the optical path very long relative to the degree of deflection, so that focusing changes caused by deflection are negligible. However, this requires either a physically large (but not necessarily heavy) scanner, or a convoluted optical path with many mirrors, prisms, and lenses.

With fiber optics, a fifth alternative becomes possible: a bundle of fibers can be mechanically transported across a document, providing an imaging path of constant length from source to document.

One manufacturer has put a series of self-scanned linear photosensing arrays end to end, with a total length equal to the full width of a page. Documents move past this array, which reads everything in its field across the whole width of the page, in a single sampling. Data are shifted out of the array serially between samples. While this design is imaginative, it is also extremely expensive, compared to alternative ways to use the same technology. Linear arrays cost $300 per half inch, and this design ties up $5100 in sensing devices alone in a scanner for 8½-in. wide pages. Whether or not this particular design will be successful in the market depends on future cost reduction in arrays, which may or may not bring the price to a competitive level.

Vacuum-Tube and Solid-State Sensors

The sensor, which is more or less independent from the scanning method, was originally a photomultiplier. However, silicon photosensors, self-scanned arrays, and spectral response modification offer reduced weight and size, TTL compatibility, and a better spectral response, leading to smaller packaging and lower costs; they also contribute to other possibilities such as the mechanical scanning previously mentioned.

Photomultiplier tubes, which respond primarily to ultraviolet light, are also sensitive to visible light that is emitted by paper and ink, some types of which fluoresce under ultraviolet light. This spurious radiation is a source of noise, and papers emitting it were forbidden in old OCR systems. The problem was largely

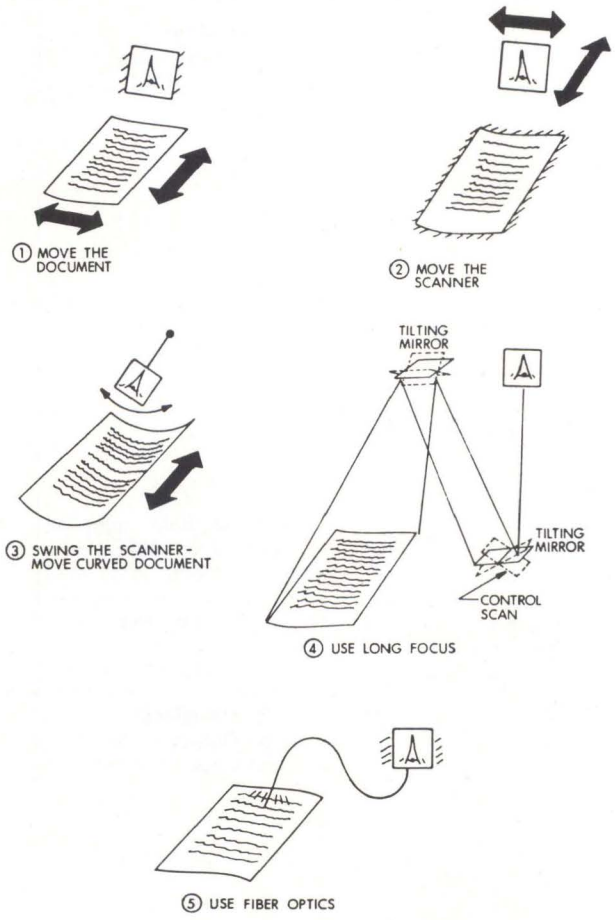

Fig. 5 Scanning techniques. In some machines, the scanner (1) is fixed while documents move beneath it; in others (2) documents are fixed while the scanner moves. A third technique (3) swings the scanner (or a mirror) back and forth over a quasi-cylindrical document that increments along the axis. Focusing problems are minimized (4) by using a long beam path with small deflections. Fiber optics (5) keep both document and scanner motionless, directing light through flexible "pipe"

eliminated by the introduction of silicon photosensors which respond to infrared light. Recent developments in silicon technology have produced photosensors that respond to shorter wavelengths; their spectral response curves are said to be green- and blue-enhanced. These devices, offered by several manufacturers, provide product designers with a much wider choice of characteristics and media interaction.

Photomultiplier tubes are necessarily rather bulky, while silicon photosensors can be closely packed into complex linear and planar imaging arrays, arranged either in a straight line or in a square array. However, silicon photosensors, singly or in arrays, do have disadvantages, some of which have been overcome by recently developed types of vacuum phototubes. Some phototubes, for example, are compatible with solid-state circuitry. In others, new cathode materials offer a much wider selection of spectral wavelength sensitivity than that obtained with silicon technology or the classical photomultiplier. Other advantages include higher resistance in the dark, higher sensitivity and linearity, and long-term stability.

All sensors react to incoming light to an extent that varies with its intensity—that is, they are analog rather than digital. Therefore, their outputs must be converted into digital form—usually with rather low resolution. Four bits provide 16 levels of resolution, which is adequate for most documents. Successive 4-bit units correspond to the black and white areas of the document crossed by a narrow stripe across a character or line of characters swept during a scan. (The reflectance of whole characters varies too little from one to another in samples of equal quality, and too much between samples of different quality, to be reliably recognized as such. A whole character, or a whole line of characters, is therefore assembled during the recognition process from several successive scans across small parts of characters.) These 4-bit units are generally easily distinguishable from one another, unless the document being read is unusually faded or dirty; therefore, they are replaced by 1-bit equivalents. For example, on a perfect document, printed with dense black ink on bright white paper, black may be recorded as 1111 and white as 0000, which are replaced by 1 and 0, respectively. The same document, using faded black ink on old newsprint, might be recorded as 1101 and 0011. Although these numerical equivalents to intensity of reflected light are closer together than in the perfect case, they are also easily distinguishable, and are replaced by 1 and 0. An isolated flyspeck might be picked up in either case, recorded as a dark (1100), and replaced with an isolated 1 in a field of 0's, but the recognition algorithm would ignore it. A problem might arise with a coffee stain, which might show up as an intermediate 0111 amid all the 0000 and 0001 for white paper and the 1101 and 1110 for black ink; the white and the black could be easily classified, but the stain might be put into the wrong class and confuse the algorithm.

From 1000 to 1500 of these successive equivalent bits, obtained from scanning a single character, are stored in a buffer (Fig. 6). A recognition algorithm acts on data in the buffer, transforming them into individual characters in computer code, usually seven or eight bits per character.

Advances in Recognizers

Matrix matching provides the most straightforward method for recognizing characters. Basically, this technique compares an image of an unknown character with a set of idealized patterns, identifying the unknown character with the one pattern to which it corresponds most closely. At one time, optical matching with masked cathode-ray tubes was tried; more recently, idealized patterns have been made with resistor arrays which are faster and can compensate for misalignments more easily.

In such an array, resistors are connected to various stages in a long shift register (Fig. 7) into which the bits of an individual character are shifted from the scanner/digitizer. Certain combinations of bits correspond to individual characters that the system can recognize; when those combinations line up with the corresponding resistors, a current pulse appears in the line corresponding to that character. Smaller pulses appear from time to time in other lines as the bits propagate through the register, but only one pulse is big enough to unequivocally identify the character. The diagram is a vastly simplified representation of such an array, which in a practical system would have several hundred shift register stages, one horizontal line for each character in the set it can recognize, which might number several dozen, and thousands of resistive interconnections—60 to 100 for each character. However, one resistor array must be built for every character in the font to be recognized; consequently, the method is both bulky and costly.

These factors become especially significant in multifont machines—those capable of recognizing characters in any of several fonts—and have stimulated the use of semiconductor memories to store the matrix patterns. Read-only, programmed read-only, and read-mostly memories have all been used this way, maintaining high speed while achieving a new degree of flexibility.

Recognition can be performed in software as well as in hardware. Large general-purpose computing systems have been doing this for many years, working with patterns much more complex than simple alphanumeric characters. Similar techniques are now possible with minicomputers, subject to the drawback of low speed that is intrinsic to software. While many OCR applications are well suited to these speeds, others are being served with hardware-augmented software recognition methods. Such implementations relegate high speed digitization totally to hardware, using software strictly for making decisions.

Some optical character recognition systems now on the market can read a limited set of hand-printed characters—usually the ten decimal digits, a few letters of the alphabet, and some special symbols. Minicomputers offer a potentially cost-effective means to extend the recognizable sets, possibly to all 26 letters, common punctuation marks, and business and technical symbols.

Further decrease in cost can be expected when microprocessors are applied to OCR. Although they are not fast enough yet, they have already shown attractive cost reductions in other applications.

Software techniques and hardware-augmented techniques now available include versions of the fast Fourier

transform. These transforms can be processed quickly and inexpensively using currently available hardware subassemblies.

Eventually, character recognition using holographic techniques will be developed. This will represent the closing of a full circle, because it will be a return to the original optical matrix-matching technique. A holographic mask containing a coded "A," for example, used with a laser, could pinpoint the location of all A's on a given document simultaneously. Coordinates of these locations would correspond to memory locations into which A's would be loaded. Then the procedure would be repeated with other masks corresponding to other characters. Whether or not holography can be used with degraded print quality may be the primary technical problem.

System Integration

Whether holographic techniques or any other technology will find their way into successful commercial machines depends on a total system design approach rather than specific component selection. This approach requires system designers to select a component technology that is consistent with successful integration of the recognition system with its expected environment, including operator interaction, serviceability, and data input and output.

Selecting correct component technologies will not in and of itself guarantee a successful design because many human factors must also be considered. Operators of some previous OCR systems, which had bright flashing lights and laser beams, whirling gears, moving belts, flying paper, and loud bumps and clunks that shook the floor, were best recruited from the engine room of a battleship. Today's system designers might consider packaging their equipment in a physical configuration that resembles card readers or office copiers, to allow operators to use familiar document loading techniques and ordinary stop and go buttons.

High speed applications, such as mail sorting and credit card processing, may well continue with today's physical monstrosities simply because they store documents efficiently after recognition. There are, however, many prospective medium and low speed reading applications that will require reasonably compact devices resembling standard input and output data processing peripherals. A large market exists in business environments for document-to-OCR-to-magnetic tape, where the tape will be processed later at a different location.

For such remote data capture, businesses cannot afford specially qualified operators. A secretary who can operate a Xerox machine should also be able, with equal ease, to perform data entry functions on an OCR machine.

The large numbers of machines that are designed and installed in this new phase will require service from time to time. Inevitably, some of this service will be rendered by less than fully capable service people, or by third party service organizations. Another critical factor is the loss of data entry redundancy that will arise as one OCR machine replaces from five to 12 keypunches; while loss of a single keypunch machine in such an installation only slightly degrades the data entry process, downtime on the OCR machine can stop the whole system.

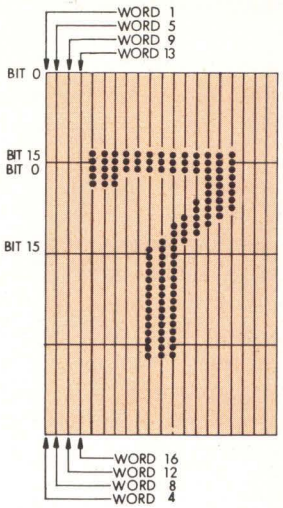

Fig. 6 Scanned character in buffer. After analog sensor output is converted to digital form and then into individual 1's and 0's, the bits are stored in computer words which, if arranged in the proper order and rendered visible, would resemble the original character. This representation is input to the recognition algorithm

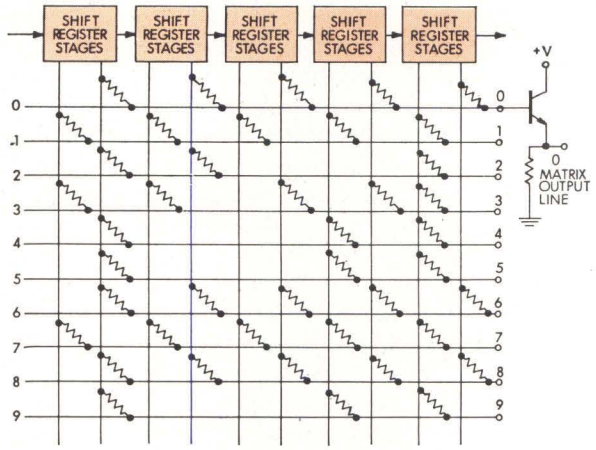

Fig. 7 Resistive recognition. Bits representing an individual character propagate through the shift register (top) in different combinations for different characters. Each bit causes a current pulse in resistors that connect the register stages to the horizontal character lines; for a particular character, the resistors form an adder that accumulates a larger pulse in the line corresponding to that character than in any other line. In practice, the register has hundreds of stages, dozens of character lines, and thousands of resistors

For these reasons, machines should be designed for maintenance by exchanging modules, and should require no tools, simple tools, or perhaps built-in tools. Furthermore, systems should be capable of self-diagnosis that can isolate the most common failures. OCR designs based on mini- and microcomputers can easily incorporate sophisticated self-diagnosic features at small additional cost.

Similarly, machines must be installable by field personnel rather than factory engineers, without critical alignments; and must permit easy crating for shipment by ordinary transport without sustaining damage to

critical parts. These criteria make laser systems less and less desirable.

Throughput, Accuracy, and Error Correction

The only meaningful measure of speed in a character recognition system is the throughput achieved at the output interface. It should be stated in terms of the number of forms output per hour across the interface for a given input quality and kind of document type. This throughput figure has little to do with instantaneous scanning speed or recognition speed, because paper movement generally accounts for a significant fraction of system operating time. In many systems, degraded input copy requires lines or even whole documents to be rescanned, further reducing the throughput. Finally, the size of the document can be critical, as can placement of characters on that document; they affect, respectively, the ease of handling by the mechanism and the time required for the scanner to find the beginning of the data that are to be read.

Although accuracy is a function of the scanning and recognition process, it is measurable only at the output interface, where it can be subjected to real-time correction processes. Quoted reject and substitution rates (proportion of characters not recognized or incorrectly recognized) are virtually meaningless, because they are contingent upon the quality of the input documents. The only meaningful measure of these rates is a sufficiently large sample of real documents run through the machine, followed by an output analysis.

Some currently available machines offer sophisticated error detection and correction techniques. When the machine encounters an unrecognizable character, it can display it on a cathode-ray screen, perhaps with the context. An operator at a console can often figure out what the rejected character should be and enter it through a keyboard. Some systems may stop and wait for such corrections; others store rejected lines on a magnetic tape or disc to wait for more leisurely correction while continuing to process other documents.

Substitution errors in textual input—an occasional misspelled word or even a name—carry enough context to be recoverable, whereas the misrecognition of numbers can cause devastating problems through inaccuracy. Nevertheless, numerical substitutions can be recovered through a few simple techniques. Check digits, for example, can be used with serial numbers, part numbers, invoice numbers, or similar input, and verified by the machine. Forming part of the number, the check digit is related to the other digits by an explicit formula. Other numbers, such as prices on an invoice, cannot have check digits, but can be checked for accuracy by arithmetic techniques. For example, an invoice ordinarily lists both the prices for the individual items and the total bill; individual prices as read by the OCR system can easily be added up and compared with the total as read. If the recognized sum and the internally computed sum are different, a substitution can be presumed somewhere along the line. For both types of numbers, the substitution is detected, but not corrected. It can only alert the operator or the system that an error has occurred.

Scanner resolution also affects accuracy. A recognizer working from digitized characters in a 7 x 9 matrix, over the long run, will give considerably different results from a similar recognizer working with a 20 x 32 matrix. Here again, accuracy or departure from accuracy manifests itself at the output interface.

Input Is Important

The data input interface is more important than it may seem. Customers buy optical character recognition systems to serve as tools in processing important documents. They are interested in the scannable documents, not in the machine that processes them. A corollary to this is that an OCR system, whenever possible, must process existing forms in their present formats and on their present paper stock, using the same ink. Trying to sell a customer a system that works only with redesigned forms is almost sure to fail.

On the other hand, selling the customer a system with the assurance that it will work with his present forms, when in fact it would work better with redesigned forms, is likely to create ill will. His forms may contain data that the machine should ignore, or preprinted information that it should pick up along with variable data printed in a different ink. With OCR registration (accurate placement) of printed data can be important. With such considerations as these in mind, forms redesign could make the difference between success and failure.

At least two machines now on the market can describe the format of a scannable document and differentiate scannable from nonscannable areas in a flexible manner.

Rapid technological advances in the computer and peripheral industry can make a machine purchased today obsolete tomorrow. To combat obsolesence, an OCR system's fixed mechanical and electronic components can be built with general-purpose control capability, with readily alterable operational characteristics stored in memories, floppy discs, or other interchangeable media, so that the user can upgrade his system at minimum cost. Such machines are most flexible when controlled by a minicomputer or microprocessor. At least one machine now on the market maintains all its operational characteristics on a single floppy-disc cartridge—including recognition font algorithms as well as the document processing and scanning control systems. As new fonts or more sophisticated error-correction techniques become available, this machine's users will benefit from them without replacing the machine.

Richard K. Dove holds a BSEE degree from Carnegie-Mellon University, Currently director of marketing at Ball Computer Products, where he is responsible for three product lines, his background includes experience in systems programming and in systems analysis.