Current Comments

Has OCR Finally Arrived? Or Is It a Technology with a Lot More Problems Than Meet the Eye?

Number 19 May 7, 1979

When I was a graduate student in the early 1950s I conceived of a device that would selectively copy text from books, journals, or other printed materials. Like so many other researchers, I had spent many hours in libraries, making copious handwritten notes. Photostating was too expensive and the Xerox machine had not yet reached the market. Deciding that something had to be done to alleviate the drudgery of note-taking, I created an imaginary device called the Copywriter.[1]

From 1951-53 I was working on the Johns Hopkins University indexing project. I decided to take an evening adult education course in electronics offered at a Baltimore high school so that I could figure out how to make this device a reality. In the fall of 1953, when I returned to my hometown of New York to study library science, I heard about a novel device built for the Veterans Administration (VA) by RCA.[2] It was a reading aid for the blind. Its purpose was not to copy but, in fact, to "read" letters and translate them into bird-like sounds which blind people could learn and understand.

I borrowed one of these experimental readers and hooked it up to a modified Brush laboratory oscillographic recorder. I remember my frustrations in this very amateurish approach. I tried to control the output to the recorder in order to create a series of black and white spots on electrosensitive recording paper. It was difficult to cause the stylus on the recorder to move up and down fast enough to respond to the output from the reader. I had a lot to learn about frequency responses, resolution, and dozens of other details about facsimile recording. It took another four years of fooling around before I was able to get some professional help. This came from my friend Mal Benjamin of Bionic Instruments, Inc., in Philadelphia. Mal is a well-known biomedical engineer.

I described my later efforts to make the Copywriter dream a reality in a previous essay.[1] It was a *selective* copying device, designed to let the user extract a particular line or word of text and have it reproduced instantly on a small "printer." A prototype was developed and further refined, thanks to a contract from the Council for

Library Resources, Inc., headed by Verner W. Clapp.[3] In one of its metamorphoses, the Copywriter consisted of a hand-held reading unit with which the user scanned the text to be copied and a writing unit which reproduced the information scanned. A cathode ray tube attached to the unit displayed the lines being copied so the user could monitor the copying process.

After this prototype had been around for a while, some people expressed interest in using the Copywriter not only to selectively copy information in facsimile form, but to feed that information into a computer. For this application the device would not merely have to "copy" in "analog" form but also recognize each letter, that is, convert that letter into a "digital" signal. In 1970, I incorporated this optical character recognition (OCR) capability into one proposed version of the unit. It would allow the Copywriter to produce a machine-readable code for each letter, number, or symbol scanned.[1] In fact, this work led to a patent on a proofreading typewriter.[4]

From my early contact with reading aids for the blind, I knew that many people were considering the use of OCR for this purpose. Over the years I had been kept informed of such efforts through Eugene F. Murphy, director of the Office of Technology Transfer of the US Veterans Administration.[5] Although many blind people read Braille books, the books are expensive to produce and bulky. Audio tapes of books and articles are less expensive to produce and more convenient to use, but they too have their problems. The tape is only as good as the reader recording it. Readers sometimes stumble over or mispronounce unfamiliar words, or swallow the ends of sentences. Many of those who record for the blind are volunteers. It takes many hours of volunteer time to record a single book.

In spite of the many efforts to produce audio tapes, much material that could be of use to blind people never gets into a form they can "read." Robert Bray, the late director of the National Library Service for the Blind and Handicapped in the Library of Congress, and I discussed this problem several times.[3] I might add that most people involved in developing reading aids for the blind never believed that a substantial market would ever exist for a machine that would "read." In those days no one imagined that society would be willing to pay for expensive devices that would convert printed matter to speech for the blind. It is remarkable how rapidly the computer revolution has changed the perception of the market for such technology.

While early researchers concerned with reading aids for the blind could never get adequate support for developing OCR techniques, OCR equipment for business and government applications became the "in" thing. Eugene Murphy points out that researchers into reading aids for the blind predicted the use of reading machines in business as early as 1949.[2] The first "practical OCR scanner" was developed by David Shepard of the Intelligent Machine Research Corp. in 1951.[6] Since that time, more than thirty companies have offered OCR units.[7,8]

Yet optical character recognition has had a very rocky history. I can well remember how OCR was supposed to solve many of the data entry problems of business data processing. When one considers how much costly manual labor goes into data-entry work, it is natural to believe that reading machines have a vast potential. But like the problem of mechanical translation, there is much more to the problem of automatic data-entry than meets the eye.

Standard OCR systems scan a line or entire page of characters with a photosensitive device. The device picks up the variation in light reflected from each character and translates the light pattern into electrical signals. These signals are compared with representations of letters stored in the system's memory. If a matching signal is found in memory, the letter is identified.[7,8]

Character readers differ from other optical recognition machines in that they are more complex. You may have noticed the bar code preprinted on packages at the grocery store. This Universal Product Code tells the price of the item and can be read by a relatively simple scanning device that can read bar patterns. However, human beings can't read the code. OCR characters, on the other hand, can be recognized by the human eye as well as by machine.[9]

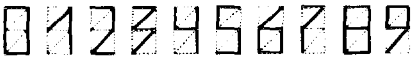
It is in the recognition process that most OCR systems demonstrate their limitations. An estimated 300 different type fonts are commonly used today. In some of these fonts, many letters are similar, making it difficult for the OCR system to discriminate between them. For example, the letter "l" and the number "1" may be identical in some type faces. And they are both similar to "i".[7]

OCR systems also have difficulty discerning a "zero" from the letter "O" and the number "5" from the letter "S".[9] Handprinting or ordinary writing is very difficult for these machines to read.[10] And if the printing of the letter isn't perfect, if the spacing between letters isn't just right, if the work is smudged—the machine could misread data to be entered.

Most OCR systems read only standard stylized OCR fonts. These fonts were developed by the OCR industry in cooperation with the American National Standards Institute (ANSI).[9] The fonts were specially designed and do not necessarily reflect the type faces that would normally be used in journals, books, or newspapers. In order for documents to be read accurately by such machines, they must be printed in one of the standard OCR fonts. An ANSI standard for handprinting was also developed. It required people preparing handprinted documents for OCR input to shape their letters in a particular way.

In the USSR, a clever system was developed for the postal service. The letter writer creates specially formed numbers in boxes on the front of envelopes. These manually created stylized letters can be unambiguously processed by the postal system's OCR equipment. Figure 1 shows how the numbers must be written. I saw no evidence that this system was actually used during my last visit to the Soviet Union, but the envelopes are widely sold.

Figure 1: Numbers in boxes on the back of an envelope from the USSR show letter writers how to form numbers for recognition by the postal system's OCR devices.



Many OCR machines read only a single standard font. These systems are used by many businesses and government agencies to enter data from preprinted forms. For example, when you pay a bill, the top half of the statement may be returned to the store with your payment. That statement is often printed in an OCR font. When the store receives the payment, the statement stubs are processed through an OCR reader which enters your name and account number into the store's computer. The amount of payment is probably keyed manually by a data entry operator.

While adequate for many business applications, limited-font OCR was not adequate for the job of reading texts for the blind. A "universal" OCR reader was needed to give the blind person access to the variety of materials a sighted person takes for granted. In 1973, after six years of study, Raymond Kurzweil, a graduate of the Massachusetts Institute of Technology, founded Kurzweil Computer Products to develop a print-to-speech machine for the blind.[11] By 1974, the machine was in its prototype form. It used OCR techniques to "read" books and other texts. A speech synthesizer converted the recognized words to voice, using over 1,000 linguistic rules stored in its memory.[12] A blind person could sit at the machine, and by placing the text face down over a light source—similar to a photocopy machine—have the book read to him. The machine read most materials at about 150 words per minute.[13] A newer model reads about 250 words per minute.

The OCR unit of the Kurzweil machine converts the letters scanned to electronic signals. A special-purpose computer then takes over. The computer enhances the image of the letter scanned, heightening the contrast between each letter and its background. If letters are contiguous, the machine separates them. If a letter is broken, the machine "fills it in." To decide which letter is being read, the machine examines the geometric properties of the letter. It makes a tentative identification by comparing the geometric features with a character definition table. The system then examines the size of the letter and its position to come up with the final identification.[14]

A built-in "learning" capability enables the machine to compare poorly printed characters with well-printed characters in the same book or manuscript. When the machine is first presented with some material, it will make occasional recognition errors. However, it corrects these errors as the material is read and the characters are learned. The Kurzweil reader originally cost $50,000. But recently the company came out with a desk-top version for about $19,000. The desk-top version will offer a "hand-tracking option." With this selective scanning device, users will be able to read complicated page formats and several columned magazines like *Time* and *Newsweek*.

A similar OCR system is being developed under a National Science Foundation grant by researchers Bob Savoie and Pat Erickson of Telesensory Systems Inc. (TSI) of Palo Alto, California.[15] TSI was founded in 1970 by Dr. James Bliss, formerly of Stanford Research Institute. TSI's OCR system will also allow the blind to selectively scan portions of text using a hand-held scanner similar to the one I incorporated in the Copywriter.

This system is designed for people who do not need to read material straight through, but must skip from section to section. It could be used, for example, while reading a magazine article or looking up information in an encyclopedia.[15] TSI is planning to combine this hand-held OCR system with its Optacon device. The Optacon directly transforms an optical image to tactile signals.[16] This device was developed by a team of researchers headed by John G. Linvill, chairman of the Department of Electrical Engineering at Stanford University. The Optacon capability will help blind people position the scanner on a page. With it, the blind person will be able to tell, for example, if tabulations run across a page or up and down it. This information would be difficult to discern with the scanner alone. As with the Kurzweil unit, a speech synthesizer will convert digital output to voice.

At ISI® , we were so excited by the possibilities that a universal OCR reader might have for us that we visited Kurzweil several years ago. At that time, the Kurzweil machine was not designed for business use. Of course, we had thought of using OCR equipment for our own data entry tasks as long as twenty years ago. As you can imagine, our data entry work-load is enormous.

We key data from more than 600,000 articles per year to produce our data base. From this data base we publish our citation indexes and other services such as *ASCA*® . *Current Contents*® itself is not based on this manual data entry operation. To input information we now use key-to-disk equipment. A data entry operator keys in the titles, authors' addresses and bibliographic citations contained in each article. A second person essentially repeats the task so that the keyed data are verified.[17]

However, when we first conducted research on methods to input data, we considered everything from OCR to dictating citations into tape recorders. From what we have been hearing lately about communicating with computers by voice, the latter approach may someday prove to be more practical than OCR.[18]

Since our visit to Kurzweil, that company has introduced a so-called universal OCR business data entry machine.[14] It differs from the reading machine for the blind in several ways. Since the system does not need to convert text into speech, the system does not use a speech synthesizer. And since commercial data entry tasks require great accuracy, the Kurzweil system has a built-in method for operator intervention. When the system is unsure about the identity of a letter or character, the letter is shown on a display screen in context with the letters around it. Its tentative identification is also displayed. The operator can agree with the machine or change the identification. In this way, accurate data go into the system.

The Kurzweil universal multifont OCR reader is a breakthrough in OCR technology. However, we may still have difficulty using this technology at ISI for a number of reasons. For instance, the system is now designed to recognize only Roman letters. We would need a system that could also recognize letters from Greek and other alphabets. It will be interesting to see how well it learns the Cyrillic alphabet. At least one manufacturer, Recognition Equipment Inc. of Dallas, has an OCR system that recognizes Cyrillic,[19] so it can be done.

There is a second and more important reason we may have difficulty using a universal OCR system. There is considerable syntactical variation in bibliographic citations. About 18 years ago we did a study of these variations. We must deal with several hundred different citation "parts of speech"—author and journal names, volume numbers, page numbers, year, etc. Our data entry operators learn to deal with these variations and must make intelligent decisions about the meanings of different numbers. For example, unless boldface or underlining is used to distinguish the volume or year from the page number, they are frequently mistaken for one another.[20]

To give you some idea of the problems we would face using even a universal OCR, just pick up a copy of the most-cited journal of them all—*Journal of the American Chemical Society (JACS).*

Let's assume we want to read the cover of *JACS.* This journal has a gray cover with the name of the journal written in white letters. The Kurzweil OCR machine works best at "reading" white bond paper printed with black ink. Therefore, it might have trouble discerning the characters printed on this journal's cover. In addition, the word "journal" is printed in very large letters. The Kurzweil OCR system treats every different type size as a new font and would have to search its disk file memory to find this size.

Alternatively, we could dispense with reading the cover. We could, as we do now, simply paste a special computer-generated label on it. label would identify the journal's title, volume number, issue number, and date. It could be typed in an easily recognizable type face so the OCR system could read it without problems. In the future, journal publishers might even print this line of information on the cover. The new postal regulations may accelerate this process if the copyright problem doesn't.

Having dealt with the cover, we turn next to the table of contents. We could pass over this page, since all the information it contains is repeated later in the journal. However, for redundancy and to double-check, we let the machine read the contents page line by line. The OCR device may have no trouble reading most of the titles and authors' names on the contents page. However, some of the titles may contain Greek letters or mathematical symbols—even an occasional integration sign. And often a title will include punctuation marks such as commas and parentheses, that may be difficult for the system to recognize unambiguously without operator assistance. These may be critical to the meaning of chemical nomenclature. Our system also would have to deal with numerous subscripts and superscripts.

Of course the OCR machine will signal the operator each time it encounters an unreadable character. The operator can then give it a little help. He or she would do exactly what one of ISI's data entry specialists has to do now—create a code that acts as a substitute for the unusual symbol.

As the system reads each title and author, it will come across an asterisk after certain authors' names in the *JACS* table of contents. The system must "know" that this is the author who gets the reprint requests. Some journals do not identify the reprint author at all while others use symbols like a dagger or double dagger. All of these common details will have to be programmed for each of the thousands of journals we index. Fortunately, many of them would apply equally to a group of journals, but even American Chemical Society journals have their own idiosyncrasies.

Next we turn to the first article in the journal. The machine reads the title printed in one type face and the authors' names printed in another. The authors' addresses may be listed in yet a third type style.

In *JACS*, authors' names and addresses are listed below the article's title, but this is not necessarily the case for other journals. In *Science*, addresses for lead articles appear at the bottom of the first page as a footnote. In the case of technical reports in *Science*, addresses appear at the end of the paper. Our OCR system would have to be programmed to recognize an address when it encounters one in each journal. Otherwise the operator will have to provide such information or it will have to be inserted during the stage we call "pre-edit."

Since we are not now interested in storing the entire texts of articles, but only the relevant bibliographic and citation information, we move on to the references. Each article in *JACS* has an average of 27.3 references the machine must read. *JACS* has references that appear at the end of the article in a section labeled "References and Notes." Some of the items listed there contain other information besides a bibliographic description of the article referenced. How exactly does our OCR machine know when the actual reference begins? It will take human intervention to help the computer decide when a citation is about to begin. Other journals, for example, the *Australian Journal of Chemistry*, put references at the bottom of each page. These too would be difficult for the unassisted OCR machine to recognize.

Let's assume that the machine has in some way been programmed to recognize that the author's name always comes first in a reference. Will it misread citations by anonymous authors? How will it deal with book chapters? Will it recognize the various type styles used within citations themselves?

As indicated above, we would have to program the machine to recognize each journal's citation style. In *JACS*, the year follows the pagination, but in the *Journal of Biological Chemistry*, the year follows the author's name. This sort of "trivia" is what makes this application of OCR a non-trivial problem.

To be economical, our OCR machine would probably have to read characters much faster than an operator could key them. It would also have to be highly accurate. At ISI we cannot tolerate an error rate which might be acceptable at some other institution. If an operator has to stand by while the machine is trying to resolve the ambiguity of badly formed characters or to create codes for unfamiliar characters, the machine may not be cost-effective. On the other hand, it may significantly increase an individual operator's productivity. This will be an important factor as inflation increases the cost of labor.

Of course the universal OCR approach to reading text described above is anything but "selective." One could take the pages to be scanned and mount them in such a way that the reading of unnecessary passages is eliminated. Or we could "mask" parts we didn't want the system to read. This is a form of pre-editing that is very expensive. It points out an OCR-based Copywriter's possible value for data entry.

I suspect that eventually ISI will wind up with a hybrid OCR/key-to-disk system. Several of these are on the market today but they do not feature universal OCR capabilities.[21] Then we could key in titles as we do now, but read in references that are printed in an OCR format. A Keysave-type system could be used to reduce the amount of keying required, insure accuracy, and eliminate manual verification.[17] Although progress has been made towards standardizing references, getting all journal publishers to agree on a single citation format does not seem feasible. But as more journals use computerized typesetting methods, they may be able to let us have their input tapes. These digital records could bypass the need for data entry and OCR altogether. We are trying to experiment with such alternatives with a few large journal publishers. But there is a lot of programming work involved in such an effort.

For OCR, many hurdles need to be jumped. Yet the advances made by the manufacturers of reading aids for the blind will have tremendous impact on business and government. ISI will continue to follow the developments in this exciting field to see if one day we can finally use a combination Copywriter and universal OCR in our own operation. Meanwhile, we salute all of those who worked so hard developing reading aids for the blind. I have no doubt that if the VA could have been supported in this research we would have seen an earlier resolution of the problem. And like so much other basic research the payoff for the public would be enormous. If we could only put a fraction of our expenditures on armaments into such R&D, imagine the benefits to all mankind.

REFERENCES

1. **Garfield E.** Introducing the *Copywriter* and ISI's subsidiary Selective Information Devices, Inc. (SID) *Current Contents* (18):5-8, 2 May 1973. (Reprinted in: **Garfield E.** *Essays of an information scientist.* Philadelphia: ISI Press, 1977. Vol. 1 p. 438-41.)
2. **Zworkin V K, Flory L E & Pike W S.** Letter reading machine. *Electronics* 22(6):80-6, 1949.
3. **Garfield E.** Library of Congress. Part 1. Looking back. *Current Contents* (9):5-13, 26 February 1979.
4. ----------. *Character recognition selective copying and reproducing apparatus.* US Patent 3, 512, 129. May 12, 1970. 7 p. Int. Cl.606r 9/100.
5. **Freiberger H & Murphy E F.** Reading machines for the blind. *IRE Trans. Human Factors in Electronics.* 1:8-19, 1961.
6. **Schantz H F.** Optical character recognition (OCR). The machines that read to computers. *OCR Today* 2(3):10-4, November 1978.
7. Datapro Research Corporation. *All about optical readers.* Delran, NJ: Datapro, 1978. 30 p.
8. *Auerbach Peripheral and Data Handling Reports.* Pennsauken, NJ: Auerbach Publishers, Inc., 1979. Monthly.
9. **Potter R I.** Character recognition (Lapedes D N, ed.). *McGraw-Hill encyclopedia of science and technology.* New York: McGraw-Hill, 1977. V. 3, p. 11-5.
10. **Tersoff A I.** Man-machine considerations in automatic handprint recognition. *IEEE Trans. Syst. Man. Cybern.* 8:279-96, 1978.
11. KRM milestones. *Kurzweil Rep.* 1(2):4, Summer 1978.
12. Kurzweil honored for inventing omni-font OCR technology. *Amer. Printer and Lithographer* 182(4):30, January 1979.
13. **Barnes B.** Robot reader for the blind. *Wash. Post* 2 July 1978, p. C3.
14. Omni-font reader handles everything. *Data Manage.* 16(9):38-9, September 1978.
15. **Savoie R & Erickson P.** *Experimental simulation of an optical character recognition. speech output reading machine for the blind.* Palo Alto, CA: Telesensory Systems, Inc. 12 December 1977. 5 p.
16. **Linvill J G & Bliss J C.** A direct translation reading aid for the blind. *Proc. IEEE* 54:40-51, 1966.
17. **Garfield E.** Project Keysave—ISI's new on-line system for keying citations corrects errors. *Current Contents* (7):5-7, 14 February 1977.
18. **Robinson A L.** Communicating with computers by voice. *Science* 203(4382):734-6, 23 February 1979.
19. **Schantz H F.** Personal Communication. 11 April 1979.
20. **Garfield E.** Style in cited references. *Current Contents* (1):5-12, 13 March 1978.
21. **Klein P L.** Key to disk and OCR. *Data Manage.* 16(9):17-8, September 1978.