[*Translating and the Computer 7*. Proceedings of a conference… 14-15 November, ed. Catriona Picken
(London: Aslib, 1986)]

Developments in OCR for automatic data entry

Julie Harnett

Editorial Consultancy and Services, Twickenham, UK

A team of researchers at International Resource Development predicts that optical character recognition (OCR), linked to online signature verification, will revolutionise the use of credit cards in the future. The US report predicts rapid movement of OCR and other image-processing technologies into a variety of specialised markets, including credit card verification, cheque processing, legal searches and so on. The team foresees a dramatic decline in credit card fraud and theft as a result of the new technology.

But if such research is going on into the development of new products we read very little about them in the press, even less about optical character recognition for full-page document input for word-processing applications.

Yet one might expect companies to be crying out for such technology, thereby encouraging manufacturers to progress such developments with greater urgency. Speed and efficiency are the selling points of word processors and computers – once the data is entered into the system. Keying it in, however, takes a long time. Re-keying it when it has already been created once on a typewriter is a tremendous waste of skilled resources.

For example, when companies install computers, they very often need to create their database from existing documents. It is estimated that the labour costs of keying that information are about 50 pence per record. Assuming an average database of 25,000 to 50,000 records, that overall cost will probably be far more than the computer system; and while staff are tied up re-keying existing documents, they are obviously not devoting any time to generating new business. That is equally true whether it be in the accounts office, which handles numerical data, or the translator's office, which handles words.

Document reading by machines using optical character recognition techniques would also appear to have a significant part to play in the day-to-day support of word processing. In their simplest form, present systems read text produced on a standard typewriter, in a stylised, but generally acceptable font (such as OCR-B) and record it on storage devices (such as floppy disks) to be edited by a word processor, if need be, prior to printing. This permits every typewriter to serve as an entry point to word processing, and reserves the word processor for editing, at which it is most cost-effective.

Just let me explain a moment the technique for those of you who are new to OCR. It is a technique for scanning a page of text with a light source to form a series of images. These are interpreted by software which tries to match the patterns it reads against a series of patterns from a pre-defined font of characters stored in its memory. If a good match is obtained, the appropriate computer code is generated and stored. If not, there are various techniques used, from outright rejection to recommended best matches. A main need for OCR is to handle external documents whose format, quality and type form cannot always be predicted.

Translators, of course, have to contend not only with different typefaces but with different languages, and machines which can recognise more than one language are few and far between. Not many machines, either, can recognise bold characters or underlining.

But 'All such problems can be overcome with proper planning,' as Michael Mason, Managing Director of Interlingua/TTI Group says. Interlingua bought a Compuscan Alphaword OCR machine back in 1979 and it was supplied capable of understanding English, French and German. Naturally, three languages were not enough for such a large translation company, so they asked the suppliers to adapt the machine to recognise twelve languages – which they did. The system has proved to be a good workhorse ever since.

Interlingua employs around 500-600 freelance translators every month who may not be the same people each time, but Interlingua needed the work each person produced to be input to the CPT word processor for editing and correcting. They could have equipped all of the freelancers with an IBM PC, but that would have been tremendously expensive. The alternative was to supply them with an OCR-B golfball element which the Alphaword OCR could recognise and input to the word processor, thus avoiding the necessity to re-key all the text produced.

Mr Mason did have some good advice to pass on – if a machine cannot understand underlining, for example, then translators should manually underline words in red, which the system cannot see. If there are corrections to be made, again, mark in red so only the word processing operator can see where to make the necessary changes. In other words, most problems can be overcome with imagination and patience.

The American company Compuscan has, since then, developed a more capable OCR machine, the Alphaword 3 + , which is available for around £22,000 in the UK from Formscan or from Kendata. It automatically identifies and reads up to eight resident fonts, including the most popular – Courier 10, Prestige Elite, Letter Gothic and so on – at the pitch for which they were designed. It automatically adjusts to read 10 and 12 pitch and varying line spacing, and automatically inserts appropriate information-processing codes, such as decimal tabs, centring codes, paragraph codes, etc. It is claimed that it makes less than one error in over 300,000 characters. It can be connected to most word processors and computers, typesetters and modems for communications.

The latest member of the Compuscan family, the Alphaword Series 80, is said to be able to capture text and data faster than forty typists. Margins are adjustable to permit reading of non-standard text and to avoid having to edit corrections to garbage that should not be scanned in the first place. Fonts are as powerful and encompass as many typesources as the Alphaword 3+ , including all commonly used characters on the IBM Golfball such as the degree sign and fractions.

Important indices, such as underlining, tabs, accents in foreign language texts and special symbols, are read as well and converted to the desired code or commands. It can also read right-justified text and outputs with full word-processing formats.

More intelligence is also claimed for the American Typereader TR3, the most sophisticated in the Hendrix OCR family which is available in the UK from General Audio & Data Communications. It features automatic recognition of up to four common typestyles and their font variations in both 10 and 12 pitch, with underscores and in any standard line spacing. The types of text it will read include correspondence, legal, fractions, accounting and foreign fonts of all available typestyles (for example, Courier 12, Courier 72, Letter Gothic, Prestige Elite, Prestige Pica, OCR-A, OCR-B and Hendrix Gothic).

With auto font and auto pitch recognition, the TR3 automatically detects different typestyles from page to page which means you don't have to pre-sort pages when using the automatic sheet feeder. Its error rate is one error in every 150,000 characters. Average reading rate is between 120 and 150 characters per second (cps) although the maximum rate is 264 cps. In addition, it offers dual output to multiple word processors or message communications devices.

For those organisations preferring a desk-top machine, the Japanese Totec TO-5000B from Mitsui, costing from about £12,500, may be more appropriate. Again, available in the UK from GADC, it not only looks like a desktop copier, it acts like one. It can read, without manual switching, the intermixed fonts of unmodified golfball or daisy-wheel typewriters, with the power to handle up to 300 pages an hour.

Currently it can read 16 fonts, with others under development. Special typefaces are available to order. But up to six fonts can be fitted at any one time. Its multilingual capability does not include French, Spanish and Italian, says Kendata, because the accents are too microscopic. However, as with most OCR devices, it is possible to build in codes for accented characters which show up on the word processor screen, enabling the operator to key in the correct character manually.

A notable feature of this machine is its straight paper-path which eliminates paper jamming, a common fault with OCR machines. Although there are formatters for most word processors, GADC say they can usually develop interfaces for non-standard systems at short notice.

A competitor of that OCR machine is the American DEST Workless Workstation from Lexisystems, a wholly owned subsidiary of Formscan. It inputs a page in 25 seconds, which is about twenty times faster than the average typist. It reads two up to eight multilingual typestyles, greatly expanding the translator’s ability to read both in-house documents and those from outside sources. It will automatically format text to be compatible with the host system; it reads multiple typestyles on the same page – which means no dials or switches to set – and a 75-page paper tray feeds in face up and ejects face down, maintaining proper page sequence.

Prices for DEST Workless systems start at just over £6,000, but a multifont multilingual system would start at just over £8,000, with multilingual Courier 10 as standard and optional typefaces available in pack form from just over £1,500 for two extra fonts, £3,500 for a seven-font pack. Multilingual word-processing format processors cost just over £1,500, but it should be noted that these multilingual machines do not read proportional spacing.

However, it is the Kurzweil 4000 Intelligent Scanner, launched in the USA in 1984, which has created more interest than most in the field of character recognition. Indeed, since it was introduced, 200 systems have been installed worldwide, twelve of which are in the UK. It is not cheap at £41,000 from UK distributor Penta Systems, but then intelligence doesn't come cheap. It combines the high technology of artificial intelligence with optical scanning, whereby it learns new styles as it scans.

As you may realise from my previous comments, OCR is limited in the number and variety of type fonts and formats it can handle. Intelligent character-recognition technology, on the other hand, accepts the entire range of type styles and formats in everyday use, including proportionally spaced type as well as conventional typewritten copy.

Enhanced artificial intelligence capabilities enable the system to train itself to recognise a full array of bold and italic characters, engineering and mathematical symbols, foreign alphabets, and multiple fonts within a single document. It will also recognise word-processing and typesetting formats and font codes. When broken characters, smudged type or other problems make recognition uncertain, the system flags the operator for assistance, displaying an enlarged version of the character and asking for an identification decision.

With reasonably clean copy, the system's training process is completed in minutes, at which point production scanning begins. Other special features include the ability to identify and code font changes, scan from left or right, and distinguish such characters as left and right quotation marks, hyphens and em rules, and horizontal and vertical spacing.

Currently, its multilingual capability includes German, Dutch, Danish and Swedish, with French and Spanish due shortly. The system can also make use of an electronic tablet costing around £5,000 which enables portions of text to be separated from surrounding graphics or allows selective automatic data entry. This system could, therefore, be used for reading and inputting text from books or newspapers.

Three years ago, industry observers predicted the development of an OCR for the home computer market and, indeed, less than a year ago a British invention, the Omni-Reader, was launched at the incredible price of £399, answering the need for a low-cost method of entering hard copy into a microcomputer or word processor. It is somewhat pedantic in the typefaces it will read, but a number of people have voiced their satisfaction with its performance providing they keep to Courier 10 or 12, although Letter Gothic 12 and Prestige Elite 12 are also incorporated. Other typefaces, I understand, can be downloaded from a computer.

In fact, a prominent laser print bureau which has no less than five large laser printers costing around a quarter of a million pounds each was very happy using the Omni-Reader for typewritten texts received from clients, which has saved the company many hours of unnecessary re-keying.

It consists of two moving parts, the precision read head and the tracking guide or ruler. It is connected to the workstation in the same way as a modem, through an RS-232C serial interface. The hard copy is placed under the ruler and aligned and the information is scanned by the read head running backwards and forwards.

The drawback is that alignment of the guide over the text line has to be accurate to avoid triggering the misread 'beep', but on the other hand you could use the ruler edge to 'hide' underlining and avoid the misreads that other OCR devices are sometimes prone to. Being manual, it is not speedy but the manufacturers, Oberon, had planned further developments, among which was supposed to be a semi-automated version. Whether that will come about now is debatable. Sadly the company is currently in financial difficulties. But perhaps someone will step in and rescue them. There are still a number of units apparently available through dealers.

I hope this unfortunate circumstance will not damage the prospects of another British invention shown to prospective backers recently. It is called the Typereader, invented by Frank O'Gorman (a computer scientist who was recently a research fellow at Sussex University working on artificial intelligence vision projects) in collaboration with Southdata Ltd, a database management software company.

As the MD Peter Laurie has said, the objective is to make an OCR machine which will read misaligned, dirty letters in almost any typeface in any size, in roman, bold and italic variants; and read lines of text at random angles (within reasonable limits). This is not as easy as it might sound, but they believe they have solved the problems which have baffled many people by bringing advanced computer vision techniques to bear.

It will be an accessory which can be connected to any computer – from mainframe to home micro – and will look rather like a photocopier. The simplest version will probably only accept separate A4 sheets fed into a slot, but larger versions would be able to accept text in sheet or book form, laid on a glass plate.

The backing that Southdata is hoping for will produce a production machine combining facsimile input hardware and computer boards to run the analysis software in a single box. It is thought that there is a mass market available for a machine costing about £4,000 which could read A4 sheets of text in more or less any typeface, perhaps with manual feed, just like a normal fax or copying machine, and with automatic feeder options.

It is also thought that there is a market available for a more sophisticated system at round £25,000 which would comprise a stand-alone facsimile transmitter (the prototype uses the Muirhead Mufax 7600) or similar hardware and a separate 68000 microcomputer running the copyrighted software. Laurie says that there is no reason why the production machine should not be able to recognise foreign language typefaces and the software could be readily adapted for that purpose. The prototype shows that such a machine would read about fifty words a minute.

Facsimile and other image-scanning techniques are thought by many to be the way that OCR will go in future. Image scanning is often confused with OCR but does not, in itself, have any ability to recognise characters. Image and graphics scanners physically record the sequence of light and dark images on a page with no provision for editing or converting those images into computer code. To be capable of recognition so that the information can be processed, not just viewed, the image-scanning process needs to be complemented by artificial intelligence, as indicated above with the Typereader invention and the Kurzweil system.

Many people believed that optical disk technology would provide the answer but, again, such systems do not ‘understand’ the images they read, they only capture and store images. Also, the optical disks currently available are non-erasable. Therefore, captured images cannot be edited or amended except for deletion and optical manipulation such as enlargement and reduction of the captured image. What is more, such systems currently cost around a quarter of a million pounds.

Some people may be interested to know if there are OCR machines which can read handwritten documents and, indeed, there are. One of the leaders in this field is Scan Optics with their Easy Reader 1750: but such systems are intended for applications such as sales order entry, subscription fulfilment, motor vehicle registration, money order records and other primarily form-filling applications. They are not intended for intensive word- or text-processing applications.

However, just to keep you up to date on the subject, a brand new OCR system was launched in November 1985 at the COMPEC computer show, which is claimed to be the only system on the market able to read a mixture of alphanumeric handprint, typewriter output and computer print, including dot matrix, without the need to predefine which fonts will be used. From Computer Gesellschaft Konstanz, a Siemens subsidiary, and available in the UK from DRS Data & Research Services, the CSL 2610 will read handprinting, in block capitals, completed in pencil, ballpoint or felt tip pen, and in blue or black inks. Again, uniquely, or so they claim, it will automatically adjust to cope with handprint slopes of up to forty-five degrees.

Potential users include banks, travel companies, bureaux; any organisation, in fact, which handles a high volume of data entry tasks involving multiline documents completed by the public. The maximum sheet size it can handle, however, is A5.

While it may not have immediate application in the translator’s office it does illustrate that OCR research is just beginning to be taken more seriously by a larger number of manufacturers, even if it is only recently. Since DRS accepts systems commissioning projects, then maybe we can see adaptations for other fields.

It is somewhat surprising that the Japanese have not been faster off the mark in producing character-recognition equipment since they are the leaders in facsimile transmission and image-processing technologies. I do hear that NEC in Japan has been fairly active recently, although it doesn’t expect to be considering export to other countries until at least mid-1986.

NEC has produced, for example, an optical character reader that can read handwritten alphanumeric data input, graphics images and characters printed from terminals, either separately or directly from the printer, with no special conversion of control programs. Its price in Japan is around the £10,000 mark.

New high-performance hybrid equipment is expected to be introduced by Japanese manufacturers, that incorporates the capabilities and functions of several products. For example, a combination of a facsimile machine, which has input and output, image transmission and communications capabilities, and a personal computer which incorporates information processing and storage capabilities – quite an exciting prospect.

But there is a word of warning which comes from the IRD report mentioned at the beginning. Apparently lawyers are potentially prime users of OCR equipment, but the space age, high-tech designs do not complement the decor of their staid offices and libraries, apparently. So, says IRD, the solution is simple: suppliers should design the OCR equipment with leather or velvet upholstery (slightly worn, of course) and/or mahogany panelling.

Could it be a case of only accepting new technology providing it doesn't look like new technology? Is it really the outward designs which are holding up progress? Let us hope such attitudes do not dampen the enthusiasm of the developers and designers of the equipment which we all need if we are to realise the full benefits of word processing and computerisation.

USEFUL ADDRESSES

DRS Data & Research Services plc
Sunrise Parkway, Linford Wood, Milton Keynes MK14 6LR. Tel: (0908) 666088

General Audio and Data Communications Ltd
70/82 Akeman Street, Tring, Herts HP23 6AJ. Tel: (044282) 4011

Kendata Peripherals Ltd
Nutsey Lane, Totton, Southampton SO4 3NB. Tel: (0703) 869922

Lexisystems Ltd
Apex House, West End, Frome, Somerset BA11 3AS. Tel: (0373) 61446

Oberon International Ltd
2 Hall Road, Maylands Wood Estate, Hemel Hempstead, Herts HP2 7BH. Tel: (0442) 3803

Penta Systems (UK) Ltd
719 Banbury Avenue, Slough, Berks SL1 4LH. Tel: (0753) 29064

Scan-Optics Ltd
36 Sunbury Cross Centre, Sunbury, Middx TW16 7AZ. Tel: (09327) 88881

Siemens Ltd
Siemens House, Windmill Road, Sunbury-on-Thames, Middx TW16 7HS. Tel: (09327) 85691

Southdata Ltd
166 Portobello Road, London W11 2EB. Tel: (01) 727 7564

AUTHOR

Julie Harnett, Editorial Consultancy & Services, 10 Post Lane, Meadway, Twickenham, Middx TW2 6NZ, UK.