Optical Character

Mark Sealey looks at a pre-release version of an exciting new software development from Irlam.

Imagine the scenario: you have a document, a lengthy document with perhaps some unfamiliar spellings or terminology, repetitive wording and small print. Maybe you are even working with a document in a foreign language. You wish to alter it and reprint it or to use it in a desktop publishing (DTP) environment. Your heart sinks at the prospect of typing it all into a word processor or text editor.

Say you also own a scanner of some sort which is capable of scanning pages from the document and storing them as sprites. But to import them as sprites directly into your word processor would be nigh on impossible, and to drop them - grey smudges and all - into a DTP frame would be ungainly and unsightly.

To avoid hours spent tediously retyping the original (with errors), the ideal solution would be if the sprite could somehow be converted into raw text, as faithful to the original as possible, which could then be manipulated as if it had been entered or imported as ASCII text from scratch. Then you could spool it into the text editor or DTP package as cleanly as if it had been typed in from the keyboard.

At least two software houses are now in the process of developing packages that will do just that, one of which is previewed this month. A full review will follow soon.

Scanned image plus EluciData output

The technique of converting from what is essentially a graphic to text is called optical character recognition (OCR). Text imported using OCR still needs further corrective work to be performed on the text, but it is better than a complete retype. Irlam Instruments (already known and respected for their high quality scanning and similar hardware) are about to launch *EluciData*, an OCR package for the A3000, A5000 and Archimedes range. It will attempt to convert any mode 0 or mode 18 sprite that contains text into an ASCII file. It works in the Desktop and the pre-release version worked well in preliminary trials, but detailed examination will have to wait for the final product.

What is more, if you have bit-mapped images from other computers and/or full colour sprites, then you can convert them into two-colour sprites (which the OCR software requires) by means of the several utilities available to do this. A typical and popular one would be *ChangeFSI* from Acorn.

EluciData is fully Desktop compatible and appears to obey the rules of the Wimp environment. This is important, essential even, since it must be possible to deal with file import and output in an easy and fluent way.

The software sent for preview was supplied on a single disc, which required a registration process to be performed; then it could be copied to a hard disc or another floppy disc. It is a relatively easy package to get to grips with and also supports the Acorn Interactive Help system.

Before use, EludiData must have “seen” the *!Alphabets* folder provided. It is from here that data is obtained for use when comparing and decoding the graphic characters.

There is a series of options which can be set for the conversion process. The size and amount of cropping of the sprite, the fonts to be used, and so on can be set before the sprite is converted. There can be a graphical display of how accurate the conversion has been, and even (rather ambitiously) a spelling check afterwards. There is also the possibility of altering the threshold of error from the default of an 80% match.

It is also possible to select the font of the text in the scanned sprite, resulting in a generally more accurate rendering of the sprite. So, for example, you can tell the software that you are about to scan some text in Times Bold, and then the software knows what shapes the letters should be, increasing the chances of recognising the characters correctly. Alternatively you can choose a fast fonts option, which promotes faster processing.

To start conversion, all you need to do is drop the scanned sprite onto the EluciData icon bar icon, and you are away. There were - as there always will be with pre-release versions - one or two crashes and blips, but for something as complex as this, though, there were surprisingly few.

Once conversion has been attempted, a window offers the user the chance to correct suspected errors as detected by the software. As soon as this process is complete, the text is saved as a file for later use.

A RTIFICIALLY INTELLIGENT

OCR software needs to make intelligent guesses as to the identity of each character that it ‘sees’ in the sprite which is being converted. To do this it needs to have a typeface with which to compare each character. This is, for example, so that serifs (the tiny slabs at the foot and head of letters, as in the letter i for example) are not misinterpreted, nor the chunkiness of the Courier (Corpus) typewriter style mistaken.

Also, some method has to be found of producing a complete character set for any existing typeface as encountered in a sprite, and some means of producing OCR typefaces of your own. The advantage of this is that the next time you use a sprite with the same or similar typeface, EluciData can be told where to look for all its comparisons, regardless of their size.

The application *AlphaBite* that comes as part of the Irlam package deals with these problems, and effectively adds an alphabet training component to the scene.

CONCLUSION

EluciData performs an extremely complex task, and the fact that it works at all is very impressive. Hopefully a full review will appear in the near future to address the pros and cons of the system in operation, when it will be complete enough to allow criticism of performance and robustness.

Product	EluciData, OCR software
Supplier	Irlam Instruments Brunel Institute for Bio-Engineering, Brunel University, Uxbridge UB8 3PH.
Price	£159 + VAT (expected)