The Perspectives for Practical Optical Character Recognition

M. Nadler

Research on optical character recognition was begun at the Cie des Machines BULL in 1960. At that time, you will recall, there were no practical machines on the market, but the state of the literature was such that many people thought that OCR was "just around the corner." Two main approaches were widely publicized—one based on stylised characters, where the authors thought that the fact of stylising would make the problem of recognition logic design simpler, the other oriented towards "multi-font" recognition.

At the start we rejected the goal of reading stylised characters, for a number of reasons. In the first place, we felt that the effort would be wasted because the inevitable progress to be expected in the understanding of the problem would make it possible to read conventional characters as cheaply as stylised ones, and therefore the eventual clients would naturally prefer to buy machines reading characters which appear more natural than the grotesque designs proposed for stylised characters. This opinion was reinforced by the idea that "natural looking" characters are such because of a certain "adaptation," or evolutionary process since Gutenburg which has resulted in matching the characters to the human recognition process, whatever that may be. Thus, if we could somehow arrive at an "optimal" recognition logic, this logic would find highly legible, "natural looking" characters at least as easy to read as stylised ones.

Another aspect of the problem, which we expected to solve by the same means, is that of print quality. The majority of readers available today require quality approaching that of typography itself. In general, one-time ribbons must be used. The few exceptions are either restricted

to a single stylised, even coded font, or else are very expensive indeed (1/2 to 1 million dollars).

Thus the goal we set ourselves was to find the basic principles for the construction of relatively inexpensive universal optical character readers.

The Problem

Seen in the abstract, a graphic sign, a character, is printed or written according to an ideal figure which is composed of straight lines, broken lines, or curves, which form nodes at their junction points. Their size is of no significance in the first approximation, as long as the relative dimensions of their composants are respected.

When we approach the reality of the sign, examining it if necessary under microscope, we see it as a surface bounded by contours between which we find, in the abstract, the lines which compose the ideal character.

Taking a further step towards the concrete, we are forced to observe that the contours containing the ideal figure are generally distorted by "noise," due to ribbon structure, the character of the paper, inking variations, and so on. Not only the contours, but the inked surface itself is "noisy." In the last analysis, the character appears as a cloud of discrete points more or less densely packed into the nominal area of the intended sign, with a certain "overflow" outside that area, in the form of spots.

The principal function of a reading machine should therefore consist in the transformation of the initial data measured on the character to be read into a surface which is homogeneous and sharply delimited by its contours. The character can then be described on the basis of measures performed on these idealised contours.

The description of a sign based on the structure and form of the strokes composing it, or the contours of these strokes, is something that many of the workers in the field have attempted. It is a rather difficult problem; the difficulty is indicated by the fact that while many have attempted it, the vast majority have rejected this approach in favor of mask-matching methods, whether these methods are purely logical, purely statistical ("correlation methods"), or a combination. The essential difficulty consists in the following: on the one hand, it may happen that certain very significant parts of a sign can have dimensions

which are very small compared to the rest of the sign, or at least signs in the same character set (an example is the tail on the comma in any nonstylized font); at the same time, the alterations in the contours due to the noise can have dimensions of the same order of magnitude as these significant parts.

This makes it impossible to adopt a single set of variables which can simultaneously neglect (or filter) the disturbances due to noise, and at the same time take into account the smallest significant details. The solution which we have found to this problem, described in broad outlines in the next section, enables us in fact to distinguish the full stop, or point, from the comma, as well as all other such "difficult" distinctions.

The Technical Principles of Our Solution

We have just seen that the recognition problem breaks down in a natural way into two main subproblems: the transformation of the signal scanned on the paper into an idealized uniform bounded surface, and the logical description of this bounded surface. This corresponds to the subproblems usually distinguished in character-recognition literature under the designations: "a set of measurements"; "a decision process."

Since we are looking for the boundary between the area which "belongs" to the character and the "outside," it appeared natural to us to seek a measurement system which would give us this boundary directly, rather than to attempt to deduce it from the usual type of OCR measurement—the generation of a rectangular matrix containing merely information on positions presumed "black" and positions presumed "white." The solution which we found has been described in detail for the specialist;[1] we summarize it here by stating that we are detecting the contours directly in terms of their local orientation in the scanning matrix. We distinguish only eight directions, in jumps of 45°; thus, in place of simple yes-no information, at each point we decide if we are observing an element of the contour, and its direction.

This approach is extremely powerful in that a very economical calculation gives us more useful information than in the usual scanning head. The simple fact of calculating these directions, or unit vectors as they are called in the mathematical theory of our machine, enables us to

[1] Nadler, M., "An Analog-digital Character Recognition System," *IEEE Trans.* EC-12, 1963, 814.

filter out practically all of the small-scale noise, while retaining most of the small-scale detail. This can be seen on computer printouts of actually scanned characters which we have available; a discussion of these is beyond the scope of the present paper.

An interesting sidelight on this detection of oriented contrast boundaries is that present knowledge of the visual cortex in the higher mammals indicates that this is one of that organ's main functions; it occurs at approximately the fourth or fifth neurophysiological logic level removed from the retina.[2]

From this field of vectors we can then generate the idealized forms mentioned in the preceding section. Voids and spots have been removed, microscopic variations in the edges of the contours have been straightened out, variations in ink density eliminated, and small-scale detail (and medium- and large-scale noise) reproduced.

Our actual recognition, or decision logic, avoids the difficulty of trying to recognize interesting details while neglecting accidental variations of the same or larger dimensions by the use of three levels of description, or analysis/synthesis. The three levels are analysed simultaneously, and then the resulting information is used to synthesize an ideal, or logical, description of the character in an hierarchic manner. Each level uses a set of variables proper to its particular function in the overall process.

At the first and most global level, the principal structural elements of the sign are determined. Thus the "O" is defined as a simple loop, the "4" as a loop with a tail descending, the "d" as a loop with a tail ascending, etc. It is quite clear that at this level we cannot do more than to regroup the various signs into classes, within each of which the structural descriptions are equivalent; thus "B,8," "O,D," etc. For high-quality print, such as electric typewriters with onetime ribbons, these classes will correspond to our intuitive decomposition into classes. For degraded print, or hand-lettered "block" letters and numerals, the noise will cause the classes to be enlarged. Thus, if the tail on the "4" gets lost, as happens all too often, it enters into the "O,D" class.

A second level of description will enable us to distinguish within each class distinguished at the first level the differences defined by the general form of each of the contours composing the elements of the ideal figure. Thus, in the class "O,D,4," we shall find that at the left there is an arc in

[2] Hubel, D., "The Visual Cortex of the Brain," *Scientific American*, 1963, 209, (11), 54.

the "O," a vertical straight line terminated by two right angles in the "D," and an oblique slanting upwards to the right in the "4."

However this is still insufficient. In certain fonts—particularly the sans-serifs, and with degraded printing—it may not always be possible at the second level to obtain a clear-cut decision, say between the "O" and the "D." Furthermore, particularly in the punctuation marks, the second level breaks down completely. This is the case, for example, with the "." and the "," (not to speak of the " ' "). This problem is treated at the third level, which analyzes the complete fine detail in the elements of the sign to be recognized. It is clear that if the entire burden of analysis had been assigned to this level, its complexity and volume would render it prohibitive in price, if not simply impossible of realization. Since its role is restricted to the distinctions to be made after the first two levels have terminated their work, on the contrary, it becomes a very interesting little gadget. In passing, we remark that at this level we apply an automatic "learning" algorithm which permits a digital computer to design the logic on the basic of actual characters actually read by the machine. This does not mean that our client will require the computer; the learning is done once and for all for each font to be read on a sample large enough to be representative of the range of print qualities to be covered, and then the logic is reproduced in the usual way for the machines supplied to the clientele.

A Modular Solution

The principles which we have just described result in a modular approach. The modules which will be available are: the transport and reading head, the structural code generator, the general form analyser, the detail recognizer.

The Modules and Their Roles. At present we are oriented towards business edp applications. Therefore the transports which are under consideration are stub handlers, document (block) handlers, tally roll handlers, and page scanners. It is this latter which would seem to be of principal interest. The role of the page scanner is simply to pass successive pages of a typescript (or other printed document) under the optical head in such a way that the necessary information arrives at the recognition logic. We are thinking in terms of a machine which will accept the standard business formats, and therefore your standard ms. sizes, with no restrictions on layout, other than those generally imposed

on authors by publishing houses (double spacing of lines, and so on). Whether a page reader for business edp can be used for direct input of manuscripts in your work is a question of functional specifications, which can be tailored to your needs, probably in the accompanying software.

If, which is certainly not your case, the alphabet to be read contains only signs whose structural codes are sufficiently differentiated, the structural code generator would be enough. This is the case, for example, in document readers where only numeric information is required to be read.

If, as may be proposed to you by some suppliers, you could restrict your work to caps only on the input side, with certain circumlocutions for special signs and exaggerated punctuation marks such as appear in otherwise normal looking special OCR fonts, you would have to add the general form analyser.

With the addition of the final module, the detail recognizer, all is changed. In this case we can recognize the entire typewriter keyboard, and this in any number of prescribed fonts. Thus we could supply a machine which would read, say, the ten most widely used fonts, upper and lower case, numerals, signs, punctuation marks, and all.

In passing, we mention a configuration which is probably beyond your possible requirements, but may be required for special computer applications such as machine translation (also reputed in 1960 to be "just around the corner"); we can deliver the machine hooked up to the computer with the learning program, able to learn any new font that may be presented to it.