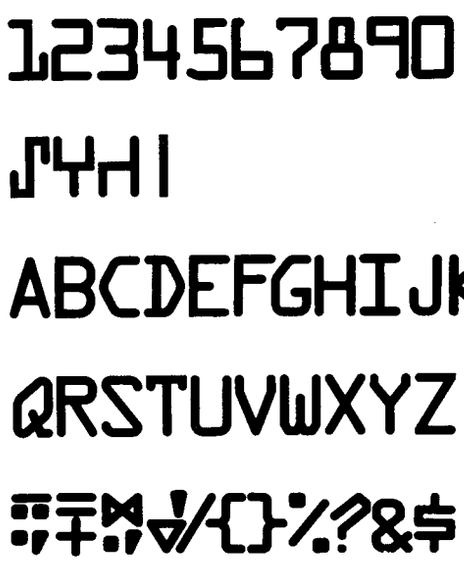


Figure 4: OCR-A

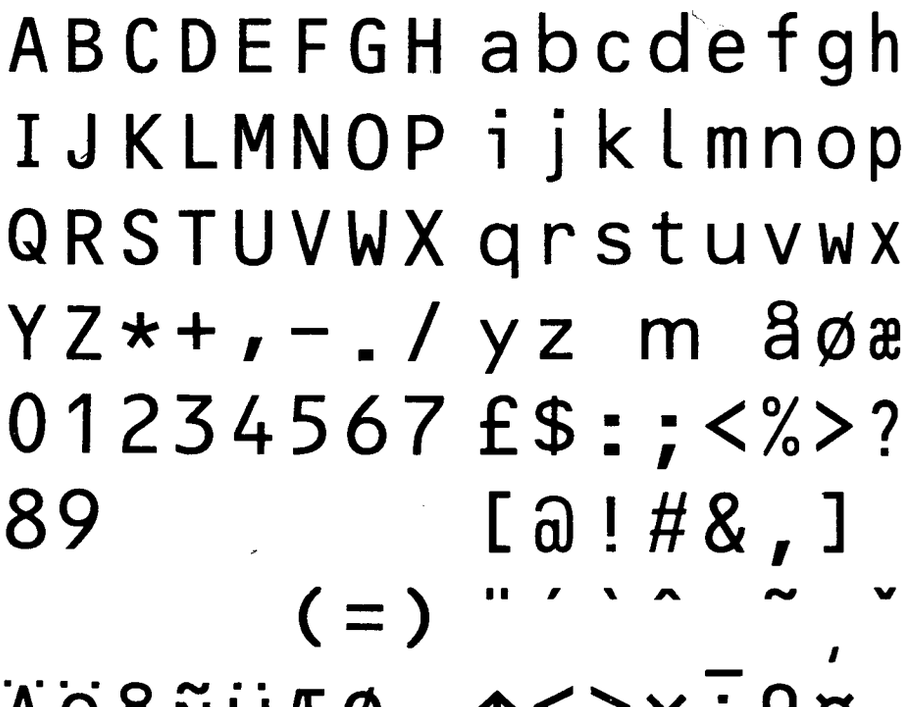


Figure 5: OCR-B

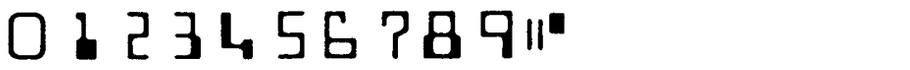


Figure 6: E13B.
(Developed by American Bankers' Association in 1958.)

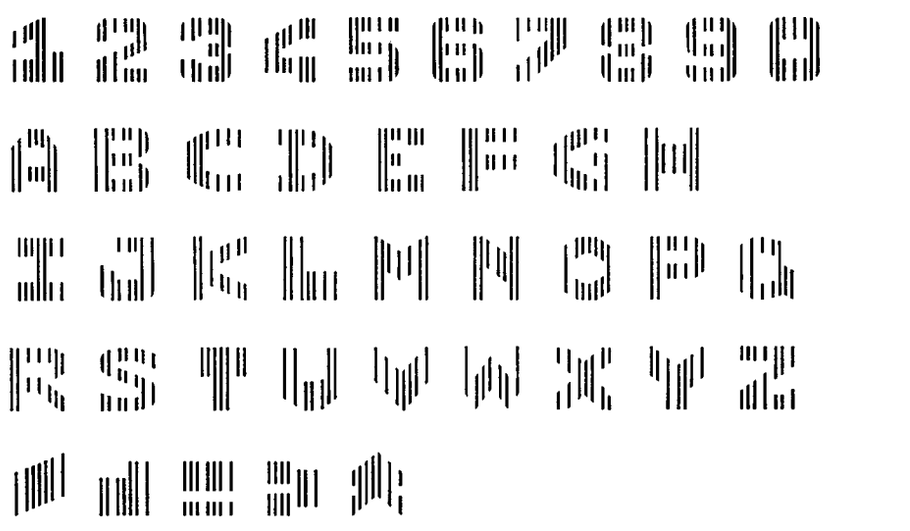


Figure 7: CMC7.
(Sponsored by European Computer Manufacturers Association, Geneva.)

A general purpose type fount suitable for use with optical reading equipment

G. G. Scarrott

Blocks illustrating this article are by Grout Engraving Co. Ltd.

The type fount to be described (shown opposite) was evolved by a technical committee set up by the European Computer Manufacturers Association in October 1961 and charged with the task of recommending standards for character recognition systems. The committee recommended two fount designs. One called class A was to satisfy the more immediate requirements. The other, which eventually became OCR–B, was intended to permit and encourage the widest possible use of optical character recognition by the use of character shapes which are as distinguishable as possible without undue sacrifice of their acceptability by the public as a general purpose type fount. Many of the later stages of the work were carried out in close collaboration with the International Standards Organisation and in 1966 OCR–B became part of an ISO recommendation (No.996) which was circulated to national standards organizations in June 1967.

1. *The technical situation*

It has long been recognized that one of the most inefficient parts of many data-processing systems is that concerned with the translation of printed information used in the human part of the system into the coded information used in the electronic part. Accordingly, efforts have been made in many laboratories to make Optical Character Reading equipment (OCR devices) which could carry out this task. Ideally such equipment should do the work at present undertaken by the keyboard and card punch operators, but faster, more accurately and at less cost.

Although several OCR systems have already come into field use they all fall short of this ideal and, consequently, OCR techniques are still undergoing rapid development. At present it is generally accepted by engineers working in the OCR field that it is quite easy to devise an OCR system with low cost and acceptable error rate, provided that three rather severe limitations are accepted: (a) the character repertoire must be restricted, often to numerals only; (b) the character shapes must be predetermined in detail; (c) the print quality must be good, e.g. that produced by an electric typewriter using a one-time ribbon.

Development efforts have been made in various laboratories to relax these limitations. Some engineers have sought to avoid the need for predetermined character shapes by developing multifount readers which can accept input documents with mixed founts. Others have tried to avoid the need for close control of print quality which is often, in present OCR systems, much tighter than is necessary for satisfactory human legibility and good appearance. Many laboratories have been working toward alpha-numeric readers with a full repertoire of recognized characters. The performance of most of these readers is still completely dependent, however, on print quality; substantially all rejects are due to print faults which do not seriously affect human legibility but prevent correct operation of the reading equipment. At least one reader has been demonstrated which can accept a restricted repertoire (numerals only) without requiring either a predetermined fount or particularly good print quality.

In short, engineering techniques are known for relaxing any two of limitations (a), (b), and (c), but there is no immediate solution in sight to the problem of relaxing all three.

2. *User's requirements for OCR equipment*

Because OCR equipment is not yet in very wide use the requirements have not yet been explicitly formulated by many users. It is reasonable, however, to presume that a prospective user will have the following questions in mind: (a) What will be the cost savings offered by the OCR system compared with alternative methods (usually card punch and keyboard operators or point of origin coding devices)? (b) Will the introduction of the system lead to any increase in the errors in the human part of the system? (c) Will it generate any opposition from the prospective user's customers, i.e. the general public?

An important component of the cost of an OCR system is the cost of dealing with errors and rejects. Since these must be corrected by highly skilled human effort, it follows that the higher the speed of the system the more important it is to achieve a low error and reject rate. Another matter of great importance to users is the cost

of printing. Not all users, for example, are prepared to use one-time ribbons on their typewriters when preparing source documents. Moreover, there are many applications involving turn-around documents generated by a high-speed printer in which some degree of smudging is unavoidable. Users will not be satisfied until the print quality requirements are not appreciably more onerous than those required for a satisfactory appearance and human legibility. Yet another factor of great importance to prospective users arises from the fact that, even after OCR equipment comes into wide use, the printing equipment used to produce the input documents will not be used exclusively for this purpose. Indeed, we can expect that new printing equipment will sometimes be put into use initially without any corresponding OCR equipment at all, but with the ultimate intention of using the equipment when it becomes available.

It follows from these considerations that printers and printed documents for use in an OCR system should also be suitable for general purpose use. Regarding fount design, it cannot be doubted that users would prefer to avoid any fount which leads to increased errors in the human clerical part of their data-processing system, even if such a fount were desirable from the OCR point of view. They would also prefer that a fount used in an OCR system should be completely clear to their customers, who may have little opportunity or inclination to learn the peculiarities of a special OCR fount. Ideally, most users would prefer to retain freedom to choose any fount they please if they could have this freedom without sacrificing other essential requirements of an OCR system.

3. *The need for a standard OCR fount*

The present state of development of OCR techniques is not adequate to meet the user's requirements, and this situation is expected to continue for several years. A possible compromise which would be acceptable to many users and is technically accessible to equipment manufacturers is to by-pass the technical problem posed by the existence of many founts by the administrative action of standardizing a fount for OCR, leaving

the other technical problems of accepting a wide repertoire of characters in normal commercial print quality to be solved by reading machine designers. The mere act of effective standardization of a fount for use in data-processing systems, irrespective of the details of the character shapes, would be of great value. Nevertheless, to do the standardization job properly, to facilitate achieving agreement on such a standard, and to encourage the use of the standard require that the standard shapes be made especially suitable for OCR systems and at least as satisfactory in other respects as those already available. Moreover, to encourage wide use of the standard, as many as possible of the user requirements for OCR systems must be satisfied by the standard OCR fount. This last requirement in particular implies that the fount must appear completely conventional to the uninitiated layman in many parts of the world. This does not mean that the fount must be designed solely by following tradition but rather that departures from tradition put in to improve the distinguishability or printability of the characters should not be immediately noticeable to ordinary non-expert users.

4. *Importance of standardized dimensions*

Type founts have evolved over a long period to serve as a code for communication. The interpretation of the code (reading) had to use techniques which can easily be learnt and used by an unaided human eye and brain. The naked eye is not well adapted to making precise measurements of a printed character; so it is hardly surprising that type founts have evolved in which each character is identified by a set of features which are taught at school and which we can recognize without making precise measurements. The dimensions, however, and even the dimensional ratios of a particular character in the many type founts in common use, vary widely since there has been no evolutionary pressure to standardize these parameters.

The development of optical reading machines has changed the situation in three ways: (a) Reading machines, unlike the human eye, can easily be designed to measure precisely a printed character and use the measurements in the recognition process.

Indeed, it is much more difficult to design a recognition system analogous to the human eye and brain than it is to design a system which makes use of the dimensions of the characters. (b) The speed of reading machines is so high that we must aim at error and reject rates far lower than were acceptable in a human communication system. This can be achieved much more readily if we design reading systems which make use of all the information conveyed by a printed character. The valuable dimensional information must, therefore, be given significance by the adoption of a completely defined standard in the information processing industry. (c) Optical reading systems are still in an early stage of development and are not yet in wide use in data-processing systems.

Since there is not yet, in this new field, a large investment in printers and reading equipment, it is still practicable and, indeed, essential that a fully defined standard be agreed. The standard must include specifications of print quality which must be satisfied to ensure adequate performance of the optical reading machine. The print quality standard may gradually fall into disuse as means are perfected for reducing errors due to poor print quality. The character shape standards, however, can be expected to have a long life in all applications where the lowest possible error rate is essential.

5. *Aims of the OCR–B design*

Evidently the OCR–B character shapes must satisfy three main criteria: (a) They must be suitable for use with a wide variety of printing devices. This is rather easy to satisfy by careful choice of standard sizes, avoiding the general use of serifs, and giving careful consideration to practical printing problems such as the tendency of openings to fill in. (b) The character shapes must be acceptable for use as a general purpose type fount. This criterion was satisfied by using the services of a skilled professional designer with wide knowledge of the traditions and practice of typography. (c) The character shapes must be suitable for use with OCR systems. This means that if an OCR system were designed with full knowledge of the character shapes of the OCR–B fount and if

full advantage were taken of that knowledge it must be possible to achieve lower error and reject rates and/or accept lower print quality than if any previously existing type fount had been used. This criterion is not easy to interpret since OCR reading techniques are still under active development, so that it is not possible to know for what reading system the character shapes should be adapted. It was therefore decided that no particular reading system should be taken into account, but instead the character shapes would be made, in an absolute sense, as distinguishable as possible within the limitations implied by (a) and (b) above. There would then be the maximum freedom for reading machine designers to choose an optimum compromise between cost and reading accuracy.

6. A measure of distinguishability

To maximize the distinguishability of a set of character shapes, it is first necessary to measure distinguishability; this in turn demands that a quantitative, relevant, and unambiguous meaning be assigned to the phrase 'distinguishability of two characters from one another'.

The interpretation chosen was based on three assumptions:
(a) The 'message' conveyed by a printed character is entirely associated with the shape of the centre lines of the strokes. The pattern formed by the centre lines of the strokes could be called a 'skeleton' character for brevity. It is a pattern of a class which could be termed degenerate since it has shape but zero area.
(b) The printing and scanning processes introduce a 'noise' mechanism which causes random distortions to the skeleton. This noise can be measured by the RMS distance σ between a point on the true skeleton and the corresponding point on the distorted skeleton. The noise mechanism can be visualized as that which would result if an attempt were made to write with a defined number of shots from a machine-gun. Each shot would suffer a random deflection in a random direction, so that the resultant skeleton would appear distorted. It is important to note that each attempt to write would cause different distortions, but that σ is still a useful and definable measure of the distorting

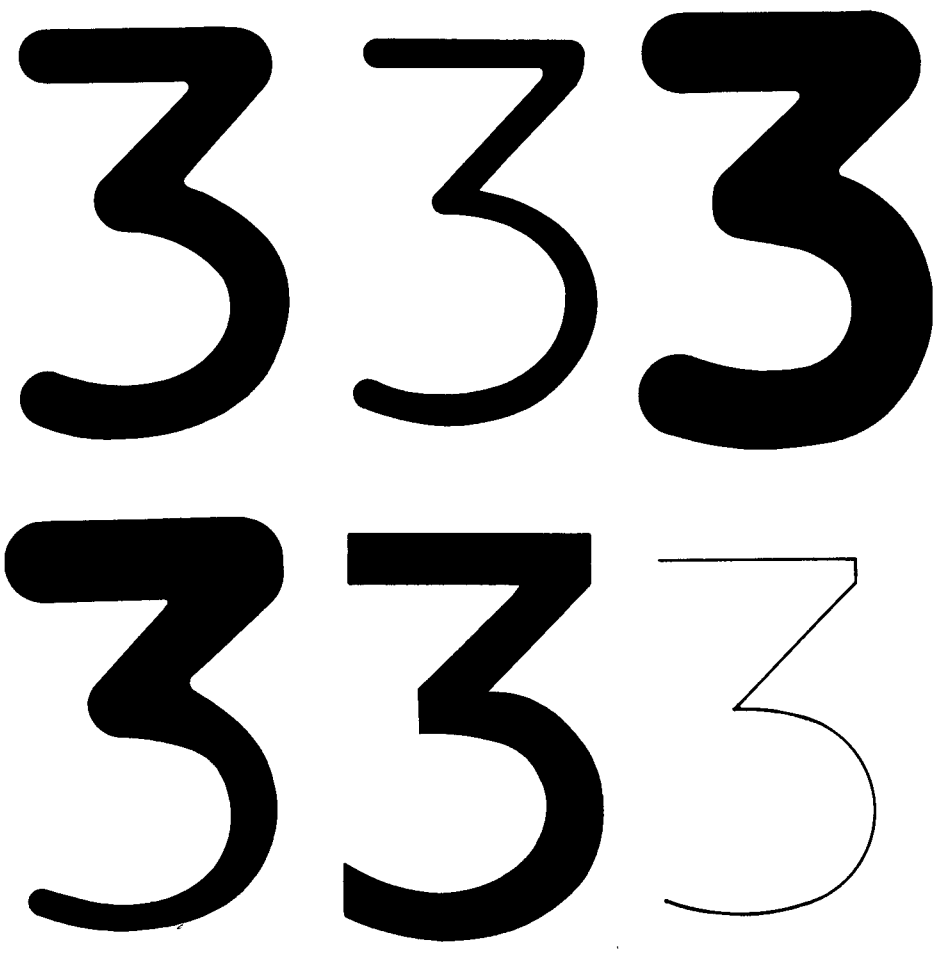


Fig.1. A set of different patterns which convey the same information.

mechanism. There are, of course, other printing noise mechanisms such as voids, spots, droop, etc., but these need not impair the inherent legibility of the message unless by chance they cause the skeleton to be misinterpreted and hence in effect distorted. Figures 1 and 2 illustrate these concepts.

Figure 1 shows a number of patterns all of which carry the same 'message', i.e. a figure 3. These patterns vary widely but all have in common the same skeleton shape, so that the skeleton alone evidently carries all the relevant information on the identity of the character depicted. Figure 2 shows a skeleton 3 together with a 'noisy' skeleton distorted by the mechanism posulated above.
(c) The distinguishability of two characters A and B is measured by the 'noise σ' which causes a standard probability that an A distorted by σ will be identical with a B distorted by σ. This measure of distinguishability has an arbitrary scaling factor depending on the choice of standard probability of confusion between A and B. However, this is of no importance since the purpose of the measure is to identify the most similar character pairs and this is not affected by an arbitrary scaling factor.

Of these assumptions (a) is certainly true and well known to many engineers engaged on the design of reading equipment; (b) and (c)

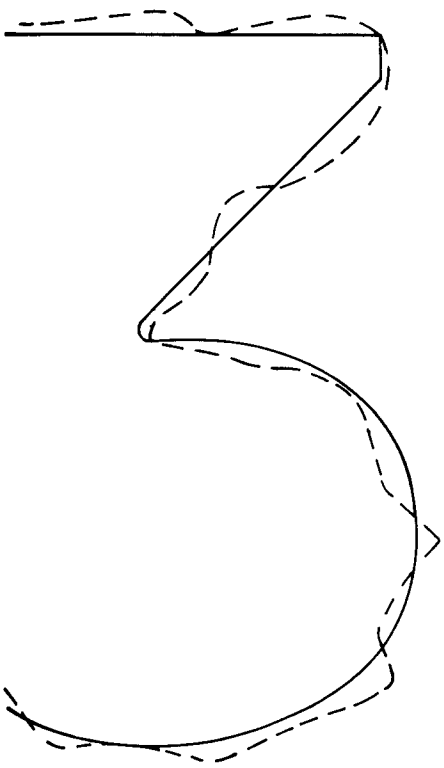
are somewhat artificial, but they accord with common sense since, for example, if two skeletons are very similar the proper measure of distinguishability implied by (b) and (c) is simply the RMS distance between the skeletons when they are superimposed on one another in the relative positions such that the RMS distance is a minimum. The definition of distinguishability in terms of assumptions (b) and (c) is, however, preferred since it can be more generally applied.

7. Representation of patterns in information space

To give a quantitative meaning to this definition demands a method of comparing degenerate patterns. Evidently a Hamming distance method is not directly applicable to degenerate patterns since the area common to two lines is zero, whether or not the two lines are of similar shape. However, a degenerate pattern can be converted into a related normal pattern (called the 'derived pattern') defined as the region swept out by a circle of radius x and whose centre follows the locus of the degenerate pattern (Fig.3).

The derived patterns can be compared by computing the minimum. Hamming distance between them and the result is a function of the parameter x which was introduced in the derivation process. As x is increased, the areas of the derived patterns increase but proportions of these areas which contribute to the Hamming distance between the patterns become smaller. The first effect is independent of the shapes of the patterns and it is, therefore, advantageous to eliminate it by normalizing the computed Hamming distance with respect to the geometric mean area of the derived patterns under consideration in order to concentrate attention on the second pattern dependent effect. The Hamming distance normalized in this way is a measure of the angle between the vectors representing the two derived patterns in information space. This angle cannot exceed $\pi/2$ (90°) and can be derived as follows:

Figure 4 illustrates the representation of two patterns P & Q by two points in information space. For this purpose, information



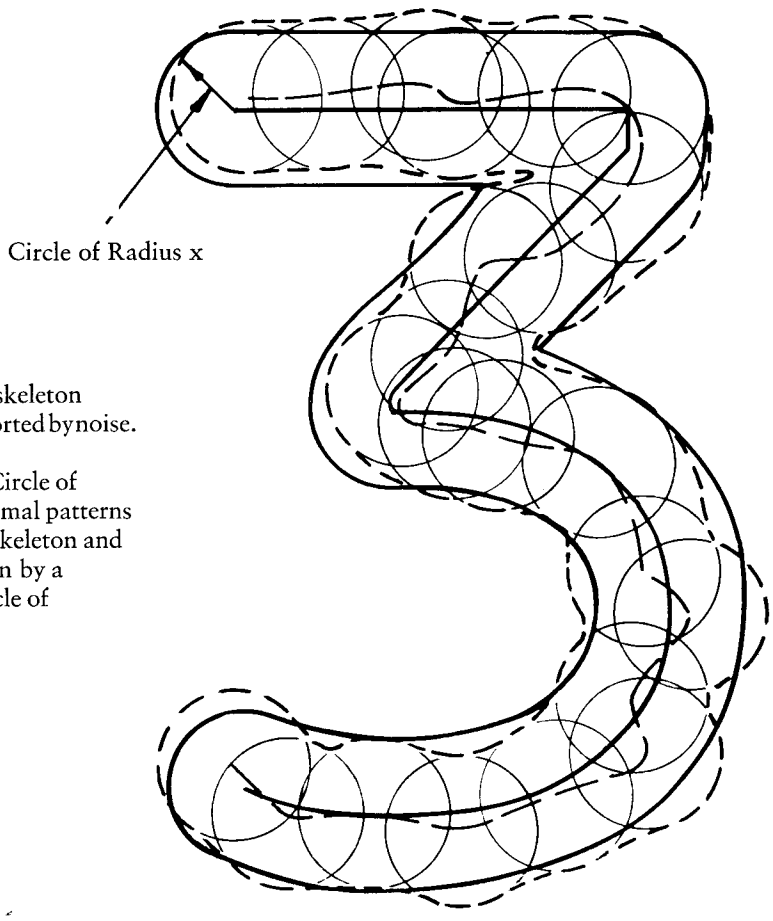


Fig.2. (left) A skeleton character distorted by noise.

Fig.3. (right) Circle of Radius x. Normal patterns derived from skeleton and 'noisy' skeleton by a generating circle of Radius x.

space is conceived as a hyper space with arbitrarily many orthogonal dimensions. A pattern such as P can be considered as a large number of elementary areas – some of which are black and some white. Each elementary area of the pattern is represented by one dimension in information space and along this dimension a vector of unit length represents a black cell and zero length a white cell. The complete pattern is therefore represented by the sum of a large number of vectors which are all mutually orthogonal to one another. Evidently the vector sum is a vector whose length is given by the square root of the number of unit vectors included since their summation law follows Pythagoras' theorem. Accordingly, Fig.4 shows the vector OP in information space of length, $\sqrt{S_P}$ where S_P is the area of the pattern P. The three points, O the origin representing blank paper, P representing pattern P, and Q representing pattern Q, are, of course, in hyper space but they collectively define a plane in it on which a triangle OPQ can be drawn which can be solved in the ordinary way to deduce that the distinction angle θ between OP and OQ is given by the relation

$$\text{Cos}\,\theta = \frac{S_{PQ}}{\sqrt{S_P S_Q}}$$ where S_{PQ} is the total area common to patterns P and Q.

This angle is a function of x which in general monotonically decreases from $\pi/2$ at $x=0$ to zero for large values of x. The curve

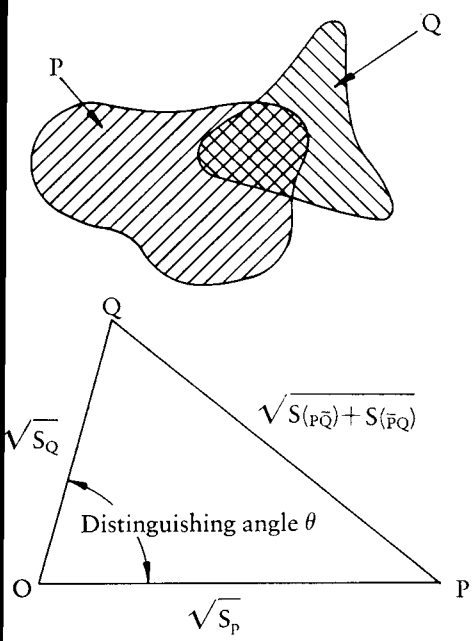


Fig.4. Vector representation of two patterns in information space Radius x of generating circle (expressed as fraction of skeleton character height).

relating the sine of the angle with x can be called the 'curve of merit' (COM) for the two degenerate patterns under consideration. It summarizes the differences between the two skeletons. The fact that at this stage of the argument a curve is necessary to summarize the distinguishability of two degenerate patterns is a direct consequence of the existence of two independent components of distinguishability, i.e. the distance between the patterns (when superimposed to give the best match) and the proportion of the total length of the skeleton pattern which provides the distinction.

If the degenerate patterns under consideration differ by a small distance over their entire length, the curve of merit converges rapidly to zero for small values of x. If on the other hand part of one of the skeleton patterns is very different from the other the curve of merit converges more slowly to zero for larger values of x.

This technique can be used to compare a skeleton character with its own 'noisy' version when distorted by the mechanism outlined in 8.2 and it is found that the effect of the distortion is to rotate the vector representing the derived pattern through an angle in information space which is dependent upon x. For most relevant values of x the noise angle has a narrow statistical distribution centred on a value proportional to $\sqrt{\sigma/x}$ but independent of the total length of the strokes in the character under consideration. If $x \to 0$ the noise angle tends to $\pi/2$ and not infinity as would be the case if the $\sqrt{\sigma/x}$ law applied down to $x = 0$.

Figure 5 depicts the average value of $\sin^2$ (noise angle) as a function of x/σ where x is the radius of the generating circle and σ is the RMS value of the deviation of the distorted 'noisy' skeleton from its undistorted position. Figure 6 shows the curves of merit for the worst character pairs on OCR–B, and for interest the worst pair in OCR–A (0 & 2). By combining the curve of merit with the noise angle law, a signal/noise ratio curve can be deduced. The S/N curve tends to unity if $x \to 0$ and tends to zero at large values of x. The curve rises to a peak at an intermediate value of x. The height of this peak is a measure in arbitrary units of the noise

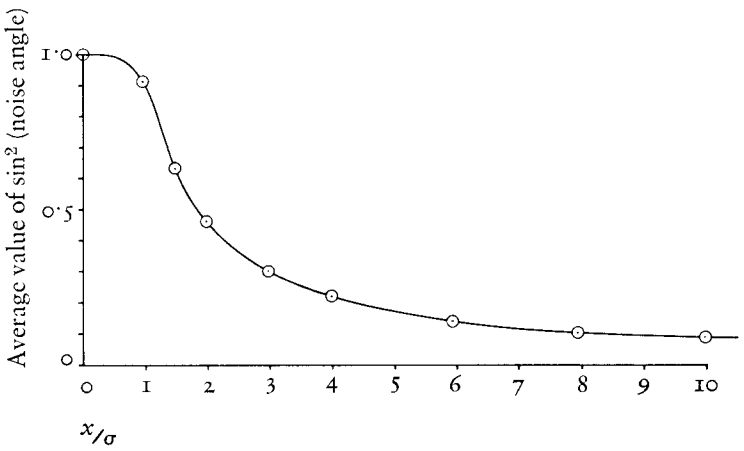


Fig.5.
x=radius of generating circle.
σ=standard deviation of 'noisy' skeleton from nominal skeleton pattern.

required to cause a standard probability that the two skeletons being compared will be confused and, therefore, gives the desired quantitative measure of distinguishability as defined in 4.3.

The value of x at which the best signal/noise ratio occurs depends on the character shapes under consideration. If they are very different, e.g. the worst pair of numerals, 3 & 5 in OCR–B, the optimum value of x is large, about 0·2 of the skeleton height, whereas if the shapes are less distinct, e.g. upper-case letters B and D, the optimum value of x is only about 0·12 of the character height. The optimum value of x deduced in this way is of considerable value in the design of recognition equipment.

8. *Application of the COM method in design of character shapes*
In the course of the character design work, it was always understood that the COM method, although a useful guide, should not be the sole criterion in character designs. In particular the result of the COM calculation cannot be applied to very small characters such as commas and full stops, since the noise calculation assumes that the character is composed of strokes which are long compared with their width. Moreover, to apply the COM analysis as defined in Section 2 above would require that the complete curve of merit (i.e. for a range of values of x) would have to be computed for all character pairs in all relative positions – a formidable task even with computer assistance. It was found

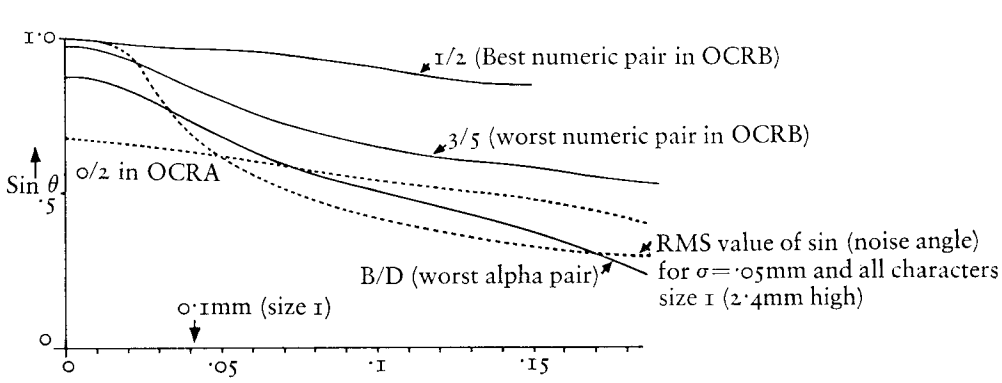


Fig.6. Curves of merit for best and worst character pairs.

Fig. 7

Nominal characters plotted on a 10×18 grid.
B E is worst alpha/alpha.
θ/O is worst alpha/number.
3/5 is worst number/number.
θ=zero.

(Nominal) Strokes 2 Cells Wide

Sin Distinction Angle	*Relevant Character Pairs*
0·59	B/E
0·60	B/D, F/R, I/T
0·61	
0·62	BH/, E/L, G/O, H/R
0·63	θ/O , E/F, F/T
0·64	I/L, O/Q
0·65	θ/C, O/G, 1/T
0·66	1/I, M/N
0·67	B/R, F/P
0·68	E/H
0·69	θ/Q, C/O, D/Q, O/U
0·70	B/C, B/F, B/S, C/G, E/R, I/J, L/U, P/R
0·71	O/G, D/P, M/U, P/T
0·72	θ/D, 3/5, 8/B, B/G, C/E, C/L, D/F, D/L, D/O, F/I, H/N

(Nominal +1) Strokes 3 Cells Wide

B D is worst alpha/alpha.
θ/O is worst alpha/number.
3/5 is worst number/number.
θ=zero.

Sin Distinction Angle	*Relevant Character Pairs*
0·50	B/D
0·51	θ/o
0·52	B/E, M/N
0·53	B/H, H/R
0·54	G/O, O/Q
0·55	θ/G, B/R
0·56	θ/C
0·57	8/B, B/S
0·58	E/F, F/P, H/N P/R
0·59	θ/D, D/O, D/Q, F/R, I/T, K/X, O/U
0·60	θ/Q, B/G
0·61	1/T, 8/R, C/O, E/H, E/L, E/R, N/R
0·62	θ/B, 2/Z, C/G, F/T, G/Q, N/W
0·63	7/Z, 8/H, B/C, C/D, E/G, G/S, I/L
0·64	3/5 , B/C, H/W

early in the work that the peak in the signal noise ratio usually occurred with x about 0·15 of the width of the skeleton character, so that most of the fount assessments were computed at this value of x. When used in this way the COM figure is effectively a normalized Hamming distance calculation with stroke width fixed at a value somewhat higher than occurs in normal printing.

To maximize the figure of merit computed in this way, a set of characters was proposed by the typographic designer. These were

described in terms suitable for computer analysis as a set of points on a matrix. The patterns were subjected to a stroke 'growing' process to the appropriate value of x and each character pair was compared in all possible positions to find the lowest COM figure. The COM figures for all pairs were printed out in order of demerit, so that the pair of characters which were most difficult to distinguish appear at the top.

The worst character pairs were then redesigned introducing subtle distortions to improve the distinguishability without causing difficulties with other pairs. The distortions were proposed by the typographic designer to ensure that the final result would be aesthetically acceptable. After several iterations of this process it became clear that the distinguishability of the fount was substantially better than that of previously existing founts, but that further improvement would be difficult. At that stage the fount design was frozen, apart from a few minor changes which were incorporated after the proposals had been widely circulated.

A computer print-out of the final COM figures is given in Fig.7. The print-out gives the values of sin (distinction angle) for the worst character pairs in order of demerit. Two complete sets of figures are given for two different values of x, the generating circle radius expressed as resultant stroke width.

9. *Examples of methods used to maximize machine legibility*
Since there are likely to be many applications of OCR in which numerical information only is handled, it was considered that the COM figure for the numeric subset should be as high as possible. The figure for the complete alpha-numeric set is inevitably somewhat lower since the repertoire is so much larger, but it was found possible to make the COM figure for the numeric subset substantially higher than for the complete repertoire without sacrifice of the figure for the complete repertoire.

The complete list of design adjustments made to achieve this is very long. Here are a few examples:
(1) The numerals are made the full printable height for each size

but the upper-case letters are made somewhat lower.
(2) Number zero, in addition to being higher than letter O is also made narrower. The difference is so marked that it is possible for the human reader to distinguish between the two symbols.
(3) The upper-case I has bold serifs which help to distinguish it from number 1 and letters T and L.
(4) The lower-case letters wth ascenders are made the full height, i.e. higher than the upper-case letters.
(5) The lower-case i has a serif to improve its distinction from lower-case j.

Innumerable other adjustments were applied to the character shapes and can best be seen by a close study of the standard drawings.

The OCR–B fount has been designed to be suitable for use with OCR systems. Since the whole of the design procedure was based on the assumption in section 6a above that the useful information conveyed by a character is entirely associated with its skeleton shape, the fount is certainly well suited for use with any reading system based upon this same assumption. This strategy in the design of reading systems facilitates the provision of defences against errors caused by poor print quality; reading systems already exist based on this concept and it is to be expected that many future developments will follow the same principle. There is little doubt that OCR–B will have important advantages for all reading systems of this kind. Any reading system designed to deal with conventional founts can accept OCR–B and at worst will suffer no disadvantage since OCR–B incorporates at least all the recognition features which can reliably be expected to occur with conventional founts. The maximization of the distinguishability in effect causes a great many additional recognition features to be included in the fount which would not necessarily occur in a 'natural' fount. It follows that feature recognition reading systems can be expected to be able to exploit the OCR–B fount to advantage.

There are some very simple reading systems which can operate

only with a special fount in which the characters are constructed from comparatively few straight line segments. Such systems will not be able to be used with OCR–B or indeed any fount of conventional appearance. An unusual but inescapable circumstance in the development and adoption of OCR–B as a standard was the fact that the standard had to be proposed and agreed before equipment for reading it could possibly be available. At the time of its introduction therefore, no hard proof of its suitability for OCR could be expected. The eventual proof that OCR–B is especially suitable for use with optical reading systems can only appear in practical experience after it has come into widespread use. When that occurs the pedantic argument on the measurement of distinguishability will be of academic interest only. These arguments were, however, an essential step in the development of OCR–B, and at present they are the chief justification for the belief by the engineers concerned in the work that OCR–B is indeed well suited for use with optical reading systems.

The work outlined here was a typical engineering job in that it had to be completed to a tight time scale in order to be useful. No doubt some of the detailed statistical arguments used in the course of the work lacked rigour. However, the final outcome obeys the dictates of common sense, and we already have some endorsement of the methods used from engineers who have built successful reading systems using reading methods which were devised long after the analysis methods used in the development of OCR–B had been decided.

The work described in this article was carried out by a group of engineers who constituted Technical Committee No.4 of the European Computer Manufacturers Association. It was a great pleasure to participate in the work and I would like particularly to acknowledge the great contributions made by M. Weill, who organized the fount design work, and Adrian Frutiger, the fount designer, who collaborated with the engineers to satisfy the recognition requirements without undue sacrifice of his artistic integrity.