A DIGITAL SIMULATION OF AN AIDED ADAPTIVE
CHARACTER READING MACHINE

Paul Baran
Mathematics Division
The RAND Corporation

Gerald Estrin*

P-1989

May 17, 1960

*University of California at Los Angeles,
California

Reproduced by

The RAND Corporation • Santa Monica • California

The views expressed in this paper are not necessarily those of the Corporation

26

SUMMARY

A pattern recognition system utilizing an initial man-machine learning phase was simulated on an IBM 709 computer. Transformations on a deformed set of 48 samples of each of ten numerals were used to form separation filters, while a second set of 480 similarly varied numerals served as the "unknown" characters which were examined.

Measured probability density distributions of the inked areas of all characters are established, and a weighted stencil or filter is created to distinguish each character relative to the possible set of characters.

The numerical experiment demonstrated the extent to which the actual value of the best "score of match" between the unknown and each character in the set provided confidence in recognition. Whenever the best score is too low, it is possible to call for more complex processes to aid recognition permitting the construction of recognition systems of greater accuracy than the basic reading mechanism.

INTRODUCTION

A digital simulation on a general purpose computer of a proposed recognition scheme was undertaken to determine the feasibility of a new approach to machine recognition of a set of characters encountered in regular printed material, such as in foreign language publications, where a large variety of type faces is encountered.*

In the design of such mechanisms, the important effectiveness criteria include the ability to recognize characters deformed by imperfect printing devices.

The philosophy adopted was one which established recognition criteria based upon the prior scanning of a significant sample of the text to be read. During the statistics gathering phase, the characters are identified by a human operator. The data gathered from the scanner and operator is then used to effect a transformation on unknown characters so that the recognition algorithm may be computationally simpler. This separation between the learning phase and the recognition phase permits the definition of a hierarchy of transformations. The simulation described here deals with a very simple model utilizing elementary sub-areas and disregards conditional probabilities in the plane of the character.

*Part of this work was performed in partial fulfillment of the requirements of the M.S. degree in Engineering at U.C.L.A. and is described in the companion P-1990.

The digital simulation was performed on the IBM 709 at the Western Data Processing Center at U.C.L.A., and we express our appreciation to the staff of that organization. We further thank Dr. Yates M. Hill and the International Business Machines Corporation for supplying us with the raw data cards used in their experiments on automatic recognition.*

These sample characters were originally printed on a single IBM 407 business machine at various phases of the life cycle of the printing ribbon, ranging from a brand new over-inked (sample 00) ribbon to the same ribbon at 150 percent of normal life. These latter samples represent a much greater deterioration of printing than is encountered in conventional printed matter.

* E. C. Greanias, C. J. Hoppel, M. Kloomok, and J. S. Osborne, "The Design of a Logic for the Recognition of Printed Characters by Simulation," Convention on Digital Computer Techniques, April, 1956, A Supplement to the Proceedings of the Institution of Electrical Engineers, Vol. 102, Paper No. 2149.

THE EXPERIMENTAL PROGRAM

The parameters of the recognition simulation are established by the constraint of original character cards supplied by IBM:

10 characters each scanned into
30 vertical scans, or rows,
32 columns and
96 samples

Of the 96 samples of each of the ten separate characters, one-half were used for generating a matrix of probability distribution frequencies associated with each cell, while the other 48 samples were retained as a source of unknown characters to be analyzed.

Each of the 30 IBM cards per character represented one scan or row with 32 column positions.

Thus, there were a total of 960 cells per character requiring 30 cards each for the $10 \times 96 = 960$ sample characters, or a total of 28,800 cards to be handled. As indicated above, these were broken into two decks; the first, a probability frequency distribution function generating deck (F-deck) and, the second, the cards for the 48 samples comprising the "unknown" characters (U-deck).

The following illustrations, Figures 1 and 2, display print-outs of some of these representative samples. Note the range of character density between the low numbered samples and the higher numbered samples. The edge data was caused by adjacent characters in the text.

Figure 1 Sample Characters

Figure 2 Sample Characters

EXPERIMENTAL RESULTS, RUN NUMBER 1

Run No. 1 used all forty-eight samples of the F-deck to form the frequency distributions, and all cells of the unknown characters were examined.

Figures 4 to 12 indicate the probability density distribution for the occurrence of ink spots for all examined characters. Only the 0.1, 0.5, and 0.9 contour lines are shown. Examination of Figure 5 shows, for example, no portions of the character "1" would be expected to have ink spots for over .9 of the samples of character "0" examined. The "cell weight factor" of Figure 3 is a number associated with a particular cell of any character and is derived for the complete set of 480 "F" samples. It may be viewed as an expected value of the occurrence of a black cell at a particular coordinate location in the matrix. Hence, there are certain areas of the character which, although always present, still may convey little differentiating information because such areas would also probably be found in other characters as well. The cell weight factor is the a priori probability density distribution for the occurrence of an ink spot for any unknown character.

Figure 4 shows the frequency distribution for the character "0." Examination of the contour diagrams show how extremely poor the distribution is, i.e., there is only one cell having as many as 90 percent of the samples of character "0" having a black spot in such a cell. Even the area between the 50 percent contour lines is very small. Hence, a badly

CELL WEIGHT FACTOR -BLACK

Figure 3

CHARACTER 0

PROBABILITY OF AN INK MARK
OCCURRING IN A TEST SAMPLE
OF CHARACTER "0" :

Figure 4

CHARACTER 1

Figure 5

CHARACTER 2

Figure 6

CHARACTER 3

Figure 7

CHARACTER 4

Figure 8

CHARACTER 5

Figure 9

CHARACTER 6

Figure 10

CHARACTER 7

Figure 11

CHARACTER 8

Figure 12

CHARACTER 9

Figure 13

F-1989
12

"smeared" distribution existed. It should be noted that the edges of the characters have values of high cell occupancy of an apparent non-random nature. The character samples provided were not developed on a random basis with regard to adjacent character contacts. Character "2," for example, always had a "1" to the left and a "3" to the right of itself. The cards provided were not wholly suitable for the purpose contemplated in the sense that such strongly correlated contextual information would not usually be present.

Another limitation of the samples used was the high degree of misalignment of the characters in the sample. There is little misalignment for most of the samples with low sample numbers, but in the higher numbered samples some characters were many rows and columns displaced as indicated by the frequency distribution of the characters and the wide spread of the fringes.

Thus, having considered the limitations of the samples of characters used, the results of the actual computer runs will now be described.

Out of a total of 480 unknown samples, only 45 were incorrectly read in the first run without any shifting, to give an average correct reading rate of 91 percent. For the first half of the samples, up to sample number 47, only 13 characters were read incorrectly out of a total of 240, or a raw reading rate of 94.5 percent.

The number of errors increased with character deterioration as the sample numbers increased.

Figure 15 shows the scores of match as a function of character deterioration for character "5." It should be noted that the three incorrectly read samples had low scores of match as might be expected.

Figure 16 shows a table with the distribution of the number of samples read correctly vs. the scores of match for Run No. 1. Also shown is the distribution for those characters that were incorrectly read. The difference in these two distributions provides a device for error removal. As is shown in the companion paper* high error rates--higher than those found in this examination--are tolerable in certain applications provided an indicator is available as to the certainty of the guess.

Figure 17 shows the probability of a correct estimate vs. the value of the score of match and shows that if a character had a score of match of about 16,000 the probability of it being incorrectly read would be low.

If each score of match bar on the graph is assumed as a separate sample space, we may estimate the probability of a sample being correctly read by virtue of the magnitude of the score of match. Thus, for samples having a score of match between 7750 and 8250, the probability of a correct estimate is only 0.43 since, of the 7 samples having such a score, only

* P. Baran, and G. Estrin, "An Aided Adaptive Character Reader for Machine Translation of Language," The RAND Corporation, Paper P-1990, May, 1960.

Scores of Match as a Function of Character Deterioration (Sample Number)
 for Character Number 5
 Run #1, F-Distribution Composed of 48 Samples, All Cells Examined

- - Indicates Character was Correctly Read
- x - Indicates Character was Incorrectly Read

Figure 15

Figure 16

Figure 17

3 were read correctly. In a similar manner, the probability of a correct estimate can be computed for the entire range of scores of match, and such data is shown on Figure 17. Note how the curve starts at a low value and very rapidly rises to approach a probability of correct estimate of 1.0 asymptotically.

Examination of Figures 16 and 17 reveals that the great bulk of the characters fall in the domain where the probability of correct estimates is extremely close to 1.0. Thus, a powerful means is available to the designer of the character reading machine to exploit the scores of match to separate the doubtful characters meriting more detailed examination.

Another indicator that can be used to appraise the confidence of the estimate is the value of the first differential score and is shown in Figures 18 and 19. Here the first differential score, i.e., the score of the character with rank 10 minus the score of the character with rank 9 is plotted for a number of samples. The distribution of the occurrence of the first differential score is plotted. Again, the heavy dark line corresponds to the correctly read characters, the dotted line to characters incorrectly read, and the dash, dot, dot line to characters for which rank 8 was assigned.

No significant improvement is obvious using the differential scores. Further, the differential score computation requires subtraction of the scores of match and so is less attractive than original score of match value estimator.

Figure 18

Figure 19

EXPERIMENTAL RESULTS, RUN NUMBER 2

In Run No. 2, the size of the sample forming the frequency distribution was reduced from 48 samples to 24. The results are of interest even though generally what one might expect. The total number of samples read incorrectly increased by a large amount. Sixty-seven out of a total of 480 samples were incorrectly read. Not only was the number of errors much greater in samples 49-99, but even samples 1 through 47, in which the characters were relatively well formed and dense, showed an increase in errors. In this latter group, there were 18 errors in Run No. 2 compared to 13 in Run No. 1. Although the 24 samples of Run No. 2 represented more closely the denser samples used in the frequency distribution, this did not improve the recognition process. Better results were obtained when the sample distribution contained the entire range of variation in samples likely to be encountered in the reading process rather than only relatively complete and dense samples.

EXPERIMENTAL RESULTS, RUN NUMBER 3

Run number 3 was designed to determine the changed performance when the examination is restricted to the central portion of the matrix. The intent was to remove material on the edges of the character samples. It is to be remembered that the flying spot scanner used to prepare the samples always saw character "3," for example, between character "2" and character "4" since the original IBM sample cards were not

prepared on a random basis. In a deck prepared on a random basis, one would not expect side information to be coherent. In Run No. 3, in which only 352 cells out of the total 960 cells were used in the comparison, the number read incorrectly increased to about double the number wrong in Run No. 1. Thus, the coherence of the edge noise materially aided recognition in Run No. 1. All 480 samples were included in the frequency distribution for Run No. 3. The number of samples read incorrectly was 99, approximately double the number of errors found in Run No. 1.

It should be emphasized that such a relationship exists to the degree found here because of the constant contact of the original character samples adjacent to one another at the time of preparation of the original frequency distributions.

OPTICAL SEPARATION FILTERS

The mathematical function used in the character recognition algorithm may be written in the form of a summation of weights for each sub area of the character analyzing matrix field. It is possible to create an optical separation stencil for each character where light is projected through each stencil consecutively. The value of total reflected light from the text is then exactly proportional to the score of match.

Figure 20 shows the shape and transmittance values for such a stencil designed to separate character "0." This stencil looks very little like the probability density distribution of ink spots for character "0," Figure 4. Reference to the

Figure 20

probability density distribution for character "O" shows that only a single spot had a 90 percent probability of occupancy, yet in the filter stencils these areas had little weight in the separation determination of the character. This difference may be explained by reference to Figure 3 which shows the probability density distribution for any character taken at random. Here the probability of each spot in the 90 percent contour of Figure 4 is also within the 50 percent contour of Figure 3. Thus, it is apparent that a filter to separate a particular character must be designed to separate that single character relative to all others. Figure 20 is such a filter and has been computed from the data of Figures 3 and 4.

Four things are to be noted:

- 1) The separation filter does not greatly resemble the character being separated.
- 2) The light output is normalized relative to the collection of possible characters.
- 3) The critical areas of the stencils are small compared to the 30 X 32 quantizing matrix used. Thus, an even finer cell resolution and better positional alignment would appear useful.
- 4) Language bigram and trigram contact frequencies may be utilized automatically by increasing the size of the analyzing matrix.