An Experimental Study of Machine Recognition of Hand-Printed Numerals

RAIMO BAKIS, NOEL M. HERBST, and GEORGE NAGY, MEMBER, IEEE

Abstract—The recognition of hand-printed numerals is studied on a broad experimental basis within the constraints imposed by a raster scanner generating binary video patterns, a mixed measurement set, and a statistical decision function. A computer-controlled scanner is used to acquire the characters, to adjust the raster resolution and registration, and to monitor the black–white threshold of the quantizer. The dimensionality of the decision problem is reduced by a hybrid system of measurements.

In the measurement design, three types of measurements are generated: a set of "topological" measurements, a set of logical "n-tuples," both designed by hand, and a large set of n-tuples machine generated at random under special constraints. The final set of 100 measurements is selected automatically by a programmed algorithm that attempts to minimize the maximum expected error rate between every character pair. Computer simulation experiments show the effectiveness of the selection procedure, the contribution of the different types of measurements, the effect of the number of measurements selected on recognition, and the desirability of size and shear normalization.

The final system is tested on four data sets printed under different degrees of control on the writers. Each data set consists of approximately 10 000 characters. For this comparison, a first-order maximum likelihood function with weights quantized to 100 levels is used. Error versus reject curves are given on several combinations of training and test sets.

INTRODUCTION

THE STEADY accumulation of recognition methods, techniques, and algorithms has brought the commercial reading machine for hand-printed numerals to the threshold of realizability.[1] Nevertheless, papers which quote error and reject rates on hand-printed material are still in the minority, and performance figures on other than greenhouse data sets remain rare. This lack of published reports on large data sets probably reflects the fact that most of the research to date has been directed at exploring new methods instead of aiming at the best possible overall performance. In an attempt to obtain results significant from an engineering viewpoint, we worked with data sets larger by an order of magnitude than customary; even these sets may not be large enough.

The various components of our recognition system were tested and modified on the basis of experiments performed on a 7000 character data set to be referred to as "Backroom I." For a final performance test, we reserved another 30 000 characters of diverse origins, described in detail in the Appendix. The decision experiment on this data set was run but once in order to guard against "implicit inference" from the test set. All too often the scrupulous separation of data into "analysis" and "test" sets is observed only in the derivation of the hyperplane coefficients. Yet, the selection of other design parameters and alternative configurations on the basis of comparisons of the test data is conducive to highly misleading expectations.

Previous work on machine recognition of hand-printed numerals is summarized in the literature.[1]–[4] We have been influenced most by Doyle's[5] "topological" measurements and Kamentsky and Liu's[6] n-tuple schemes. The n-tuples were modified by the introduction of local rather than global shifts. Many of the ideas for preprocessing are the result of previous work on multifont character recognition as described by Liu[7] and Liu and Shelton.[8] The linear decision schemes we used are well known.[9] Highleyman[10] was one of the first to apply them to hand-printed numerals.

We have avoided three methods often invoked to improve recognition. On-line scanning by means of a light pen or captive stylus does, indeed, yield better signal-to-noise ratio, but its use must surely be restricted to special applications. Individual training of each writer by means of labeled training sets also does not seem to be widely acceptable. The third artifice involves dot and line constraints;[11] the difficulty of introducing such restrictions in the field can hardly be appreciated by those who have not tried it.[12]

Another fundamentally different approach, taken by Greanias *et al.*[4] in hardware, and Kuhl[13] in software, is the technique of curve following, i.e., tracing the inner and outer edges of each line forming the character. Although the features derived are quite powerful, and may be designed intuitively, line following itself is troubled by tics, and broken, noisy, and misshapen characters. Our work was principally motivated by the development of the IBM multifont recognition system.[6]–[8],[14] This system uses a program-controlled flying spot scanner, a set of registration-invariant n-tuple measurements, and a digitally implemented hyperplane decision. The n-tuples have shown surprising capability for line and feature detection.

In the character acquisition phase, each sample numeral was converted into a 25 × 32 binary matrix by means of a cathode-ray tube scanner. Under the supervision of an IBM 1401 computer, the scanning system controls document changing, character localization, separation of adjacent characters, noise suppression, threshold and line

Manuscript received November 7, 1967.
The authors are with the IBM Watson Research Center, Yorktown Heights, N. Y.
[1] Since the preparation of this manuscript, one machine, the IBM 1287, has been announced.

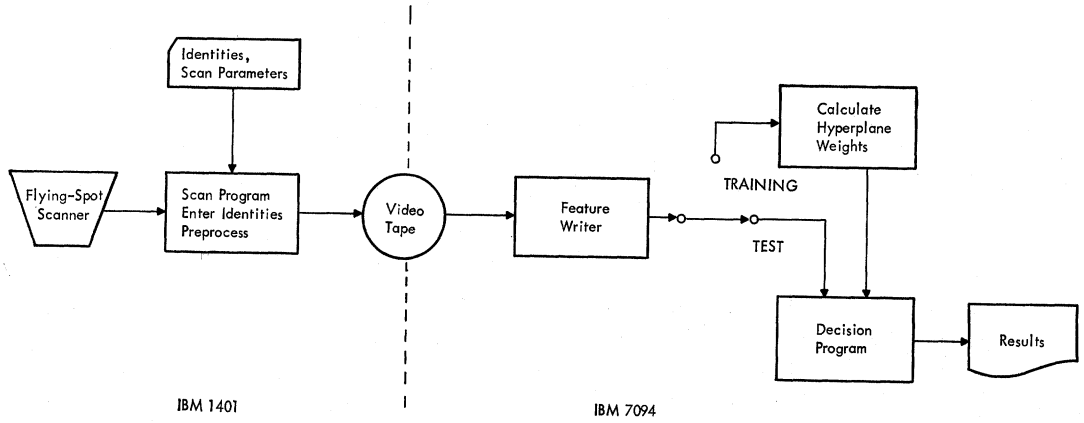


Fig. 1. Flowchart of the recognition system.

width, and size normalization. The volume of data processed requires that manual intervention be held to a minimum.

The remaining portions of the recognition cycle, shown in Fig. 1, are carried out on the IBM 7094. Before the final categorization, the dimensionality of the data, viewed as vectors in binary n space, is reduced from 800 (25 × 32) to about 100 by the extraction of features, or measurements.[2] These measurements represent the best of a composite set generated partly by machine, partly by hand. The selection rule is completely automatic; while the underlying assumptions are only approximately true, experimental results testify to its efficacy. The 100-bit feature vectors serve as input to several piecewise-linear hyperlane categorizers, whose performance is then compared to one another.

Experimental System

The opaque scanner is a light-tight box containing the document and cathode-ray tube (CRT). A lens assembly focuses the light from the phosphorescent spot on the face of the CRT onto the document. The intensity of the reflected light is measured by monitoring the sum of the cathode currents in eight photomultipliers (PMTs) facing the document. The PMT current is then quantized in both time and amplitude to provide a binary input to the remainder of the system.

Three additional PMTs directed at the screen are used in conjunction with subtraction circuitry to compensate for phosphor irregularities in the CRT. Pincushion correction and dynamic electromagnetic focusing help maintain linearity over the 2400 × 2000 grid which constitutes the 10- by 8-inch working area of the scanner. Additional details about the scanner may be obtained elsewhere.[14]

A special purpose digital-to-analog interface unit accepts "macro" commands from an IBM 1401 computer and translates them into deflection and threshold voltages. The approximate vertical and horizontal location of the start of each vertical sweep, the length of the sweep, the

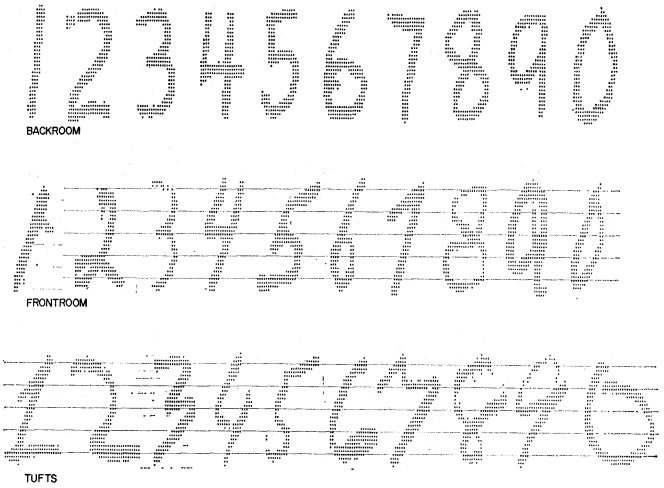


Fig. 2. Scanned samples from the three different data sets after size normalization and quantization. The backroom data was the most constrained, while the Tufts set was written without any control. The aspect ratio is distorted by the printer.

horizontal increment between sweeps, the size of the spot, and the video threshold level for differentiating between "black" and "white" are controlled by the IBM 1401. Each sweep, irrespective of its length, is quantized into 32 bits to represent one vertical column of a character. The vertical resolution of the system is thus inversely proportional to the sweep length. Fig. 2 illustrates the quantized video image of scanned hand-printed characters.

During "search" mode scanning, the horizontal resolution is decreased (i.e., spacing between scans is increased) and the spot zig-zags back and forth across the page until it encounters a black area. After a preliminary check of the number and configuration of black bits to reject stray dots, the system switches into the "center" mode. The entire line is centered approximately on the basis of the mean vertical cross section of the characters. Then the machine alternates between "individual center," "threshold," and "normalize" modes until the parameters are adjusted to produce a video pattern about 29 bits high with a white row above and below it, and an average line width of about 3 bits. For some characters, even ten iterations are insufficient to arrive at a satisfactory setting.

[2] The terms are used interchangeably.

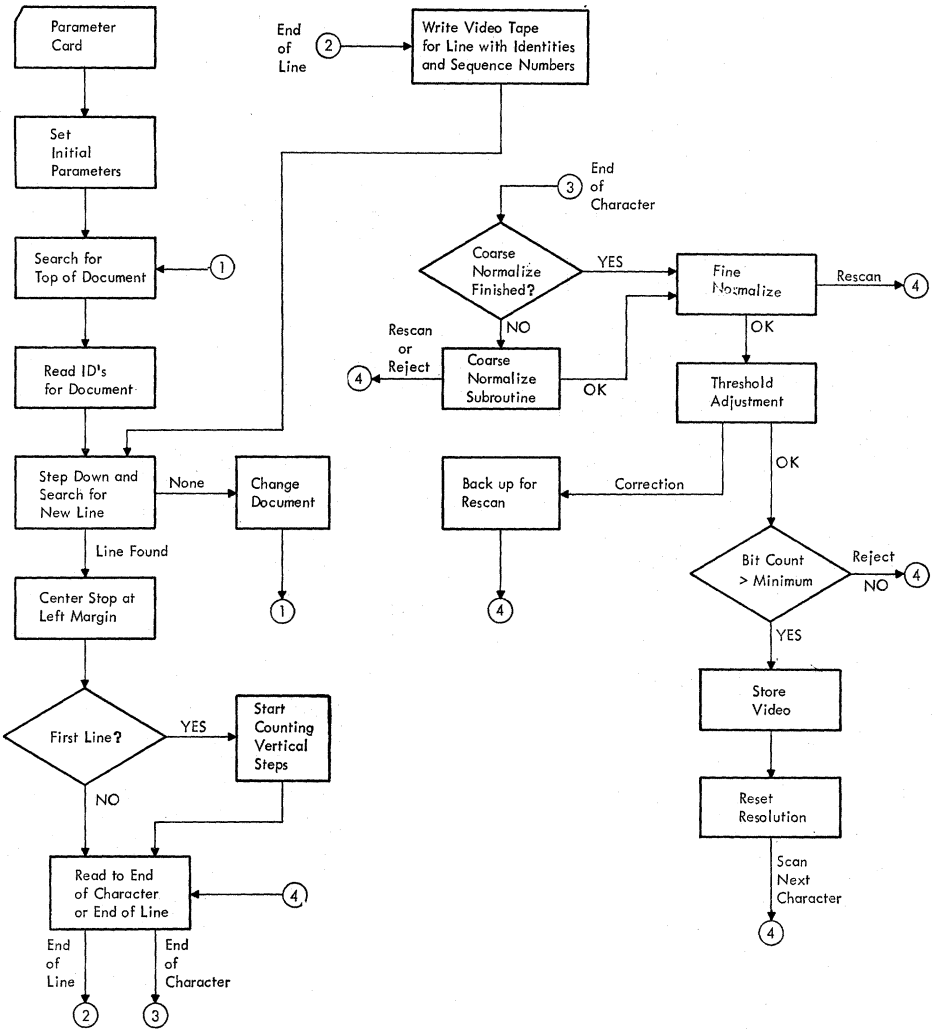


Fig. 3(a). An overall flowchart for the scanner program.

A flow chart for the scan program is shown in Fig. 3(a). Different document formats, margin bars, fiducial marks, background, and other details may be met by modifying the program flow and including suitable logical tests. The ease with which these changes can be made is a major advantage of a program-controlled scanner.

Two size normalization techniques were used. The coarse technique of Fig. 3(b) attempts to predict the final position and resolution settings to achieve the desired height. A histogram of the character heights obtained by this method is shown in Fig. 4. Accordingly, a fine normalization was added, shown in Fig. 3(c), which begins with the coarse estimate and adjusts the registers one step at a time until there are two adjacent bits in the top and bottom rows of the desired character matrix. The improvement achieved is also shown in Fig. 4. The large number of rescans necessary with this method slowed the processing time to about 5 seconds per character.

Adjacent characters are separated by looking for a "white" path between them. Only paths slanting from left-bottom to right-top are permitted in order to retain the tops of disconnected fives. The threshold was adjusted by monitoring the average line width as described elsewhere.[8]

A method of shear normalization was also investigated by simulation on the IBM 7094. Each horizontal row in the character was shifted by an amount proportional to its height, the constant of proportionality being chosen to minimize the inclination of the principal axis having the smallest moment. Some sample characters illustrating the action of the algorithm are shown in Fig. 5.

Measurement Design

The specification of measurements which will simultaneously yield significant information about several categories is certainly the most difficult part of the character recognition problem. Not only do we not know how to design a good measurement set, we are not even sure we will recognize a good one when we see it. Nevertheless, it is easier to discard worthless measurements than to synthesize good ones; for this reason, most feature extraction schemes reported in the literature operate by selection.

The role assigned to the computer may consist only of testing hand-designed measurements,[4],[5],[11],[13] or the

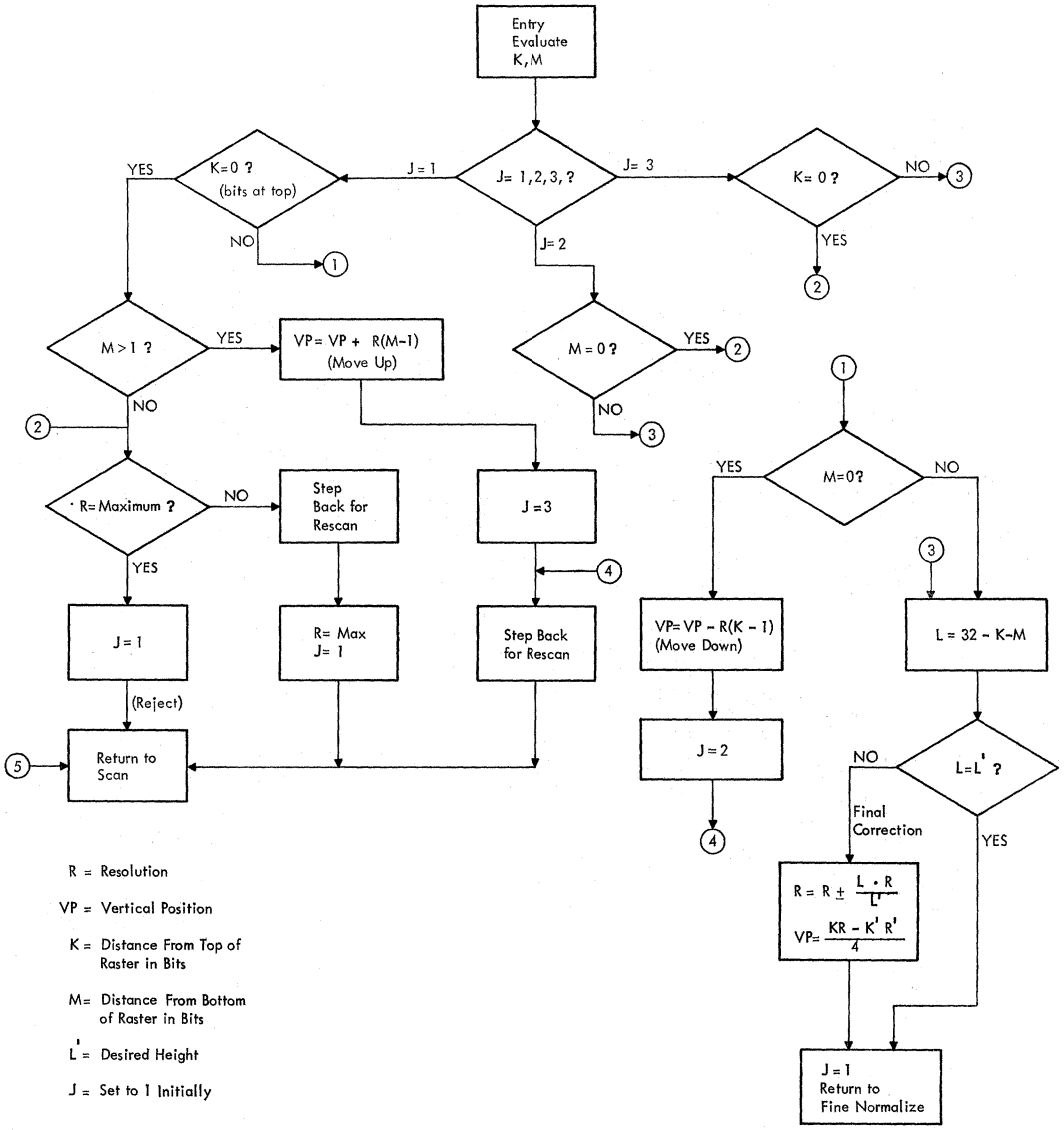


Fig. 3(b). Flowchart for the coarse normalization subroutine, which attempts to set the size and resolution predictively.

computer may perform the selection on the basis of some more or less readily computed success criterion,[6],[17] or it may even take part in generating the measurements under some previously specified constraints.[15],[16] Our computer has been helpful in all three ways.

Of the many types of measurements which could be applied to raster-scanned data (stroke and sequence data not available), we restricted our attention to "topological" measurements (for lack of a better name), and to "n-tuples." Threshold gates, of which n-tuples form a subclass, were deliberately avoided in the absence of any systematic design procedure.

Measurement Selection

The automatic selection of a subset of measurements from a large pool to optimize the error rate (with a specified decision procedure) must cope with certain difficulties. The large number of possible subsets of a pool of reasonable size makes exhaustive computation of the statistics of each subset impossible. Calculation of actual error rates on a valid character sample is also impractical. Taking the statistical dependencies of the measurements into account also broadens the computations. (A rigorous treatment of measurement selection, taking dependencies into account, can be found in Estes.[17] Another method has recently been developed by Chow.[18]) It is thus tempting and practical to use some readily calculated figure of merit, defined for any number of measurements and related to the discriminating power of the whole set. An information measure has been used successfully by Kamentsky and Liu,[6] but the computing time becomes lengthy for large numbers of measurements. In a two-class problem with linear decision and uncorrelated measurements, another measure, loosely related to the error rate, may be derived as follows.

Let the two classes be 1 and 2. The ith measurement, x_i, is known to have mean $\mu_i^{(j)}$ and variance $\sigma_i^{(j)}$ (j = 1, 2) within class j. The discriminant is

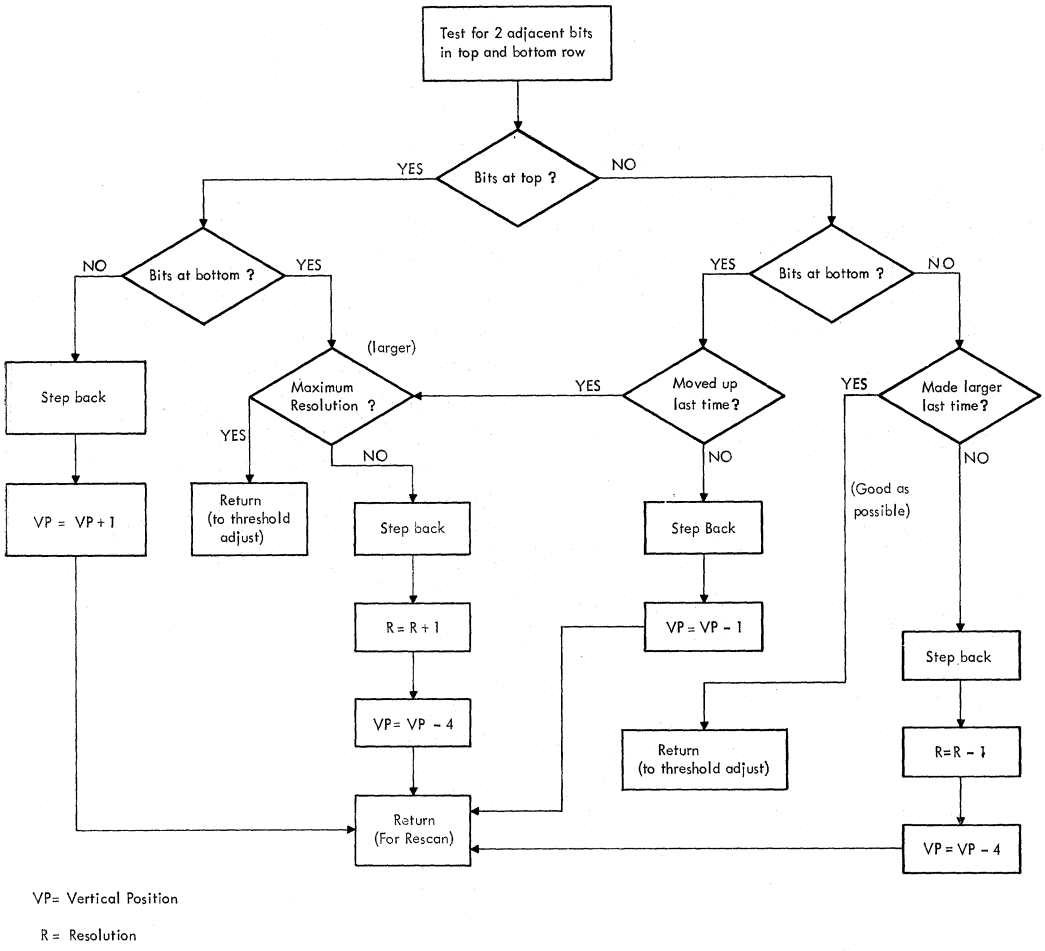


Fig. 3(c). Flowchart for the fine normalization subroutine, which rescans the character until the desired height, as measured between the topmost and bottommost horizontally adjacent black bits, is achieved.

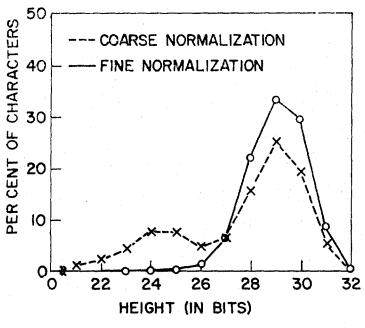


Fig. 4 Distribution of character height using coarse normalization alone and together with fine normalization.

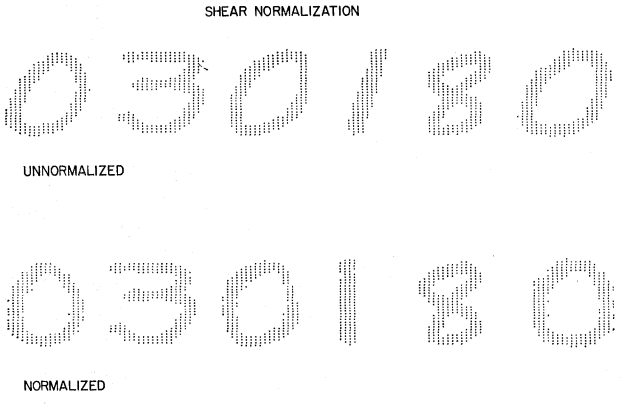


Fig. 5. Quantized version of test characters with and without shear normalization.

$$d = \sum_{i=1}^{n} a_i x_i. \tag{1}$$

Under the assumption that the measurements are uncorrelated, we have

$$M_j = \sum_i a_i \mu_i^{(j)} \quad j = 1, 2 \tag{2}$$

and

$$V_j = \sum_i a_i^2 \sigma_i^{(j)} \tag{3}$$

where M_j and V_j represent the mean and variance of d for class j. As in discriminant analysis,[19] we wish to choose the coefficients a_i such that the difference between the means $M_1 - M_2$ is as large as possible, while keeping the variances to some prescribed value, say

$$V_1 + V_2 = 1. \tag{4}$$

Then using Lagrange multiplier λ, we must solve the set of equations

$$\frac{\partial(M_1 - M_2)}{\partial a_i} + \lambda \frac{\partial(V_1 + V_2)}{\partial a_i} = 0 \quad i = 1, \cdots, n \tag{5}$$

where λ is chosen to satisfy (4). Evaluating the derivatives from (2) and (3), we find

$$\mu_i^{(1)} - \mu_i^{(2)} + 2\lambda a_i(\sigma_i^{(1)} + \sigma_i^{(2)}) = 0 \tag{6}$$

and solving for a_i,

$$a_i = \frac{1}{2\lambda} \frac{\mu_i^{(1)} - \mu_i^{(2)}}{\sigma_i^{(1)} + \sigma_i^{(2)}}. \tag{7}$$

But

$$\begin{aligned} V_1 + V_2 &= \sum_i a_i^2(\sigma_i^{(1)} + \sigma_i^{(2)}) \\ &= \frac{1}{4\lambda^2} \sum_i \frac{(\mu_i^{(1)} - \mu_i^{(2)})^2}{\sigma_i^{(1)} + \sigma_i^{(2)}} \\ &= \frac{1}{4\lambda^2} \sum f_i \end{aligned} \tag{8}$$

where

$$f_i \equiv \frac{(\mu_i^{(1)} - \mu_i^{(2)})^2}{\sigma_i^{(1)} + \sigma_i^{(2)}} \tag{9}$$

so that from (4)

$$\lambda = \frac{1}{2} (\sum_i f_i)^{1/2}. \tag{10}$$

Similarly,

$$\begin{aligned} M_1 - M_2 &= \sum_i a_i(\mu_i^{(1)} - \mu_i^{(2)}) \\ &= \frac{1}{2\lambda} \sum \frac{(\mu_i^{(1)} - \mu_i^{(2)})^2}{\sigma_i^{(1)} + \sigma_i^{(2)}} \\ &= \frac{1}{2\lambda} \sum f_i \end{aligned} \tag{11}$$

so that from (10) and (4),

$$M_1 - M_2 = (\sum_i f_i)^{1/2}.$$

We now define a "figure of merit" F^* for the set of measurements:

$$F^* \equiv \frac{(M_1 - M_2)^2}{V_1 + V_2}.$$

Then from (8) and (11),

$$F^* = \sum_i f_i. \tag{12}$$

Thus F^* is the total figure of merit for all the measurements, defined in terms of the means and variances along the normal to the best separating plane. If the measurements are independent, then F^* may be obtained also by summing the individual f_i. If the independence assumption is violated, F^* is less than Σf_i. We use $F = \Sigma f_i$ as the measured figure of merit.

Using the optimum threshold to equalize the probabilities of both types of errors between the two distributions

$$a_0 = -\frac{M_1 V_2^{1/2} + M_2 V_1^{1/2}}{V_1^{1/2} + V_2^{1/2}} \tag{13}$$

it can easily be shown by Chebyshev's inequality that the total probability of misclassification P_E is

$$P_E \leq \frac{1}{F}\left[1 + \frac{2\sqrt{V_1 V_2}}{(V_1 + V_2)}\right] \leq \frac{2}{F}. \tag{14}$$

When both distributions are symmetric, the bound may be divided by 2. Unfortunately, use of this bound in a precise way is impossible. The effects of measurement dependencies usually make F^* far lower than Σf_i, while the Chebyshev bound is usually a large overestimate. Nonetheless, since these two factors tend to cancel, one may hope that $2/F$ is a rough estimate of the order of magnitude of the error rate.

This figure of merit is used to select subsets of measurements from a larger pool in a multiclass problem as follows. Each measurement of the pool is evaluated on the analysis set of characters and f_i estimated for every class pair. The total F for each pair is then calculated by summing the individual f_i. Then the measurement contributing the smallest f_i to the character pair having the smallest original F is deleted. The procedure is then repeated until only the desired number of measurements remain.

The computation could be modified to add measurements to a set; Estes[17] has shown that under certain circumstances the two procedures of sequential selection and sequential reduction lead to equivalent results.

Computer Generated n-Tuples

Our first attempts at generating the n-tuples at random produced a very low yield of acceptable measurements.

(For a given character pair, the f_i of the pseudorandomly generated measurements is normally distributed. The variance of this distribution is as an indication of the "yield" of a set of constraints since the higher the variance, the higher the fraction of measurements above any desired threshold.) The restriction of the n-tuples to specific regions of the character (zones) resulted in a higher yield, but also in an increase in the number of free parameters. The zones were eventually fixed at the entire character, one of its halves, or one of its quadrants. The best results were obtained with the following set of "line-seeking" constraints. A black point was placed at the matrix center. The next black point was chosen from a two-dimensional spherical Gaussian population with a mean distance of 3 from the center. Three black points were then chosen by proceeding a random distance down the line defined by the previous two points, and then adding a perturbation from another spherical Gaussian generator. Three white points were then chosen from another two-dimensional distribution. A set of 1800 measurements was generated in this manner.

Topological Measurements

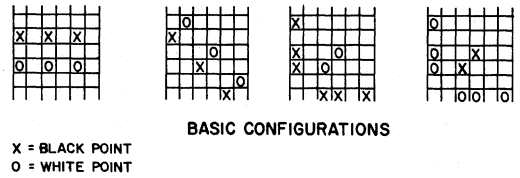
This family of measurements consisted of a set of subroutines designed intuitively (but with the aid of computer-generated statistics to optimize parameters) to detect certain prominent characteristics of the characters, such as symmetries and number of line segments. Most were chosen from Doyle[5] because they lend themselves to hardware implementation with a raster scanner. Additional measurements of this type may be readily designed in this manner to discriminate among the worst confusion pairs. Such programmable measurements make it possible to introduce characteristics that are evident to the eye.

Eight of the original pool of seventeen measurements were selected for the final set.

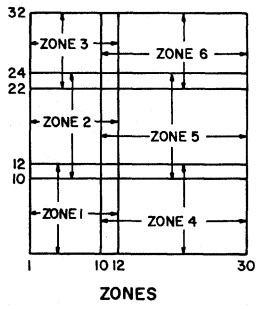
Five measurements turn on when the maximum number of segments (a string of 3 bits or more) encountered by a vertical (or horizontal) slice passed through the entire character exceeds some threshold. The measurements which test for one to three vertical and one or two horizontal segments have significant information.

Two measurements involve the counting of bits to test for symmetries. Thresholds are set by computing the sample distributions on the analysis set and calculating the best separating value. One measurement, designed for the character pair 5/3, turns on if more than 80 percent of the black bits occur within the leftmost 80 percent of the character width. The other measurement, for 6/8, turns on if more than 45 percent of the black bits occur within 40 percent of the character height from the bottom.

The last measurement is used to detect bays (CAV LEFT in Doyle[5]). Successive scans from left to right are "ORed" together. The number of segments is counted and if it decreases within z percent of the character width, the measurement is set on ($z = 30$ was eventually chosen).



(a)



(b)

Fig. 6. (a) Four basic measurement configurations. (b) Six overlapping zones governing the acceptable position of the lowest rightmost bit of each n-tuple.

Hand-Designed n-Tuples

The four basic configurations illustrated in Fig. 6(a) are rotated in four cardinal directions and applied to the characters in the six zones shown in Fig. 6(b). Thus, whenever one of the 16 n-tuples fits a character anywhere in a specified zone, the corresponding feature bit is set on. This accounts for 96 bits; four more similar "zoned n-tuples" round the number out to 100.

The measurements are designed to respond, both singly and in conjunction with one another, to lines, line ends, and sharp bends of various orientations and in various portions of the character. The final configurations and the locations of the zone boundaries were reached after about twenty 1401 runs on 1000 characters from the Backroom I set. The information used to improve the measurements consisted of the probability matrix specifying the frequency of occurrence of each measurement on each class and printouts of the digitalized versions of erroneously identified characters.

The n-input AND gates and the required character-shifting mechanism were readily implementable on the experimental system: both zones and measurements are specified by means of a convenient plugboard. Later, the measurements were also programmed on the IBM 7094 for participation in the final measurement selection.

Decision Methods

Both decision procedures tested are of the statistical hyperplane variety. It is not our intention to enter the controversy between the advocates of sequential tree-type decisions and of parallel-probabilistic methods; the latter simply seemed more suitable for our proposed machine organization, and the automatic methods for generating the coefficients were already at our disposal.

Hyperplane categorization requires, by definition, the comparison of expressions of the following variety:

$$\alpha_j \cdot x = \sum_{i=1}^{m} a_{ij}x_i + a_{oj} \tag{15}$$

where x_i is the outcome of a measurement on an unknown pattern and a_{ij} is a coefficient determined from analysis data.

Bayes' Method

The application of the maximum likelihood ratio for categorization is well known.[19] With the assumption of independence between the inputs, the coefficients are computed as follows:

$$a_{ij} = \log \frac{p_{ij}}{1 - p_{ij}} \tag{16}$$

where p_{ij} is the probability of occurrence of measurement i on class j. The character class giving the maximum sum in (15) is chosen, provided that its supremacy is not too closely contested.

Some improvement in recognition may be obtained by taking into account the correlations between measurements by means of iterative procedures.[20] The ultimate performance is similar to that obtained by adaptive perceptron-like algorithms. These more time-consuming methods were not attempted.

Anderson–Bahadur Method

This procedure[21] requires a decision plane between every pair of classes, and so becomes uneconomical for larger alphabets. For the numerals, however, the 45 planes represent only a moderate increase over the strictly linear methods, and with 100 measurements they can be readily accommodated in core storage.

The normal to the plane separating the jth class from the kth class is calculated by

$$b_{jk} = (aC_j + dC_k)^{-1}\delta_{jk} \tag{17}$$

where C_i is the covariance matrix of the ith class, a and d are scalars, and δ_{jk} is the difference of the mean vectors of the jth and kth classes.

If there is a disagreement between several planes in the classification of an unknown sample, the sample is rejected. Although this intuitive reject procedure does not really minimize any risk function based on uniform penalties for errors, and uniform, though different, penalties for rejects, more rational reject criteria require elaborate computation.

The Anderson–Bahadur procedure is the statistical counterpart of Highleyman's linear decision functions. An experimental comparison would be required to determine which is superior in a problem, such as this one, where neither linear separability nor multivariate normality obtains.

Experiments and Results

The experiments about to be described were designed with a twofold purpose: to provide a guideline in the selection of useful features and to establish performance levels on the data sets available.

Some 150 separate computer runs, totaling about 50 hours of IBM 7094 time, were required (in addition to the preparation and testing of the programs and the scanning of the data). Fig. 7 is a flow chart showing the relationship of the various programs.

Size of Measurement Set

From a pool composed of 96 hand-designed, and 1800 computer-designed features, the best N were selected by means of the figure-of-merit criterion. The selection took place on the basis of the first 1000 characters of the 10 517 character (Backroom I) data set. N was varied from 50 to 200, and a Bayes' decision was implemented on the full set. The recognition results, with no rejects allowed (forced decision), are shown in Fig. 8.

The eventual rise in error rate with the number of measurements has been observed by other investigators and may be attributed both to the suboptimal estimation of the categorizer parameters on the basis of an analysis sample of finite size, and to violation of the independence assumptions. These phenomena are discussed fully by Allais[25] and Chow.[18]

Generalization Capability of Measurement Selection Scheme

The initial measurement pool contained 100 hand-designed and 1800 machine-designed n-tuples, and 17 topological measurements. Two sets were selected using the figure of merit in order to determine, in spite of the nonhomogeneity and relatively small size of the data sets, whether the measurement selection scheme is too tightly data-dependent. Table I shows that the particular set of characters used to select the measurements has little influence on the recognition results (1000 characters were a sufficient sample for the figure-of-merit selection algorithm).

Figure of Merit versus Random Selection

While the restrictiveness of the underlying assumptions allow no claims of optimality on behalf of the figure of merit, we can at least compare it to one often-used scheme, random selection.[16],[17],[22],[23] Thus a third set of measurements was chosen at random from the pool of 1917 measurements. The three measurement sets were then tested using Bayes' decision. The forced decision error rate is shown in Table II for both methods of selection. The performance is about 6 to 1 in favor of the figure-of-merit selection.

Generalization Capability of Weight Selection Algorithm

To investigate how closely the weights selected by the maximum likelihood program were tailored to the particular character set on which they were designed, Bayes'

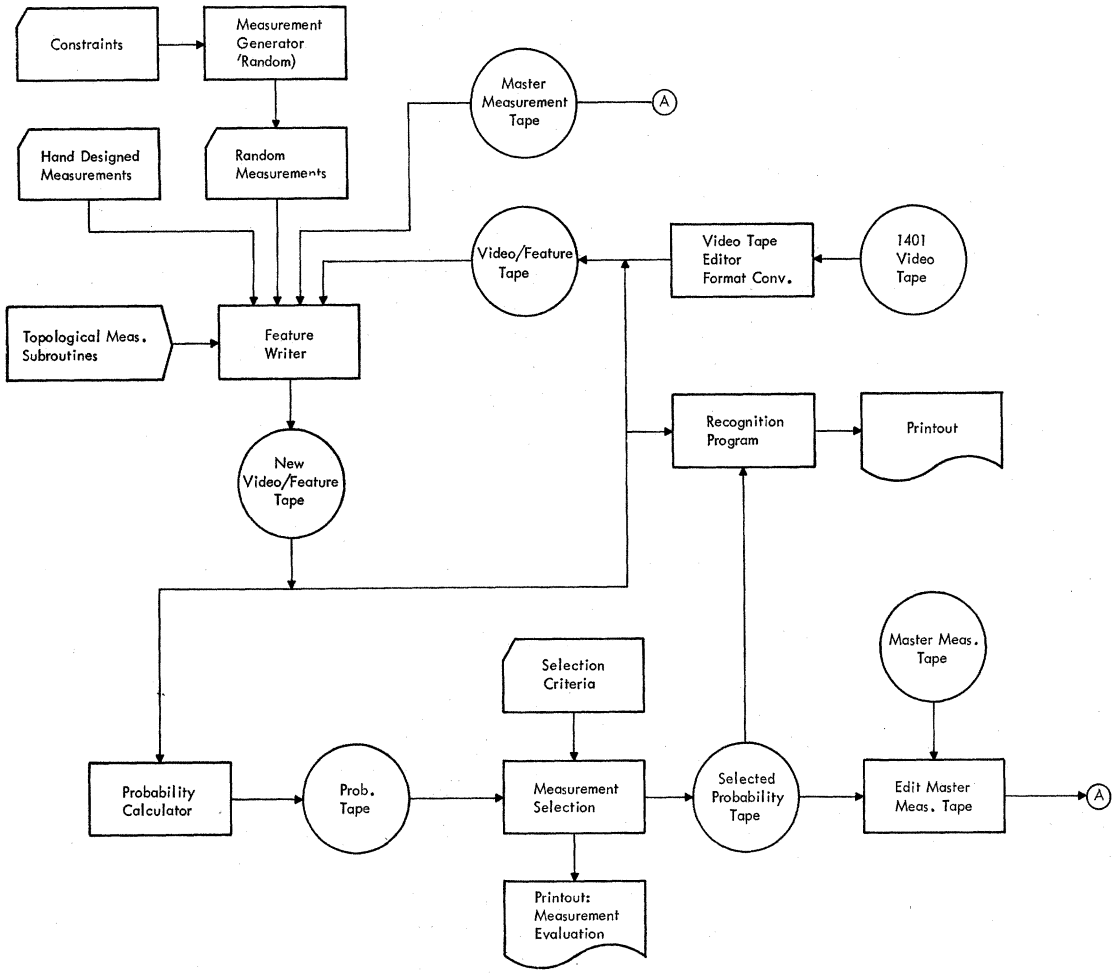


Fig. 7. Overall flowchart for measurement design and recognition.

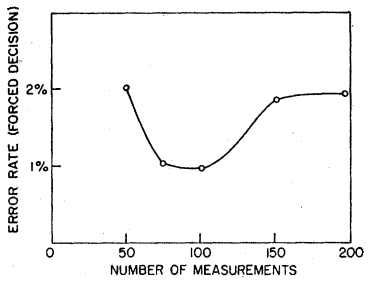


Fig. 8. Error rate on Backroom I data as a function of the number of measurements selected.

analysis and test runs were conducted on the next two 3000-character sets of the Backroom I tape. The two were then reversed to nullify the effect of any nonhomogeneity. The results are shown in Table III. Although the second character set is of better quality than the first, by averaging the error rates on "new" samples and also on "training" samples, one can see clearly that there is a twofold decrease in performance to be expected on going from training set to new data. In a statistical weight design algorithm of this type, the number of "training" characters is considered sufficient only if recognition performance on "test" and "training" sets is substantially the same.

Thus, while even as small a sample as 1000 characters seems sufficient to yield a good "general purpose" measurement set for this data, a much larger sample than 3000 characters is needed for the more finely tuned weight calculating algorithms.

Measurement Types

Of what benefit is the merging of these different measurement types? To answer this question, we selected sets of 100 measurements from several combinations of the three types and ran training and recognition runs on 7000 characters. The results appear in Table IV. The final set, consisting of 56 randomly generated, 36 hand-designed,

TABLE I
GENERALIZATION CAPABILITY OF FIGURE-OF-MERIT MEASUREMENT SELECTION

Measurement Selection	Bayes' Weights Computed on	Error Rate on	Error Rate (Forced Decision)
#1–#1000	#1–#1000	#1–#1000	0.8%
#1000–#2000	#1–#1000	#1–#1000	0.6%
#1–#1000	#1000–#2000	#1000–#2000	0.5%
#1000–#2000	#1000–#2000	#1000–#2000	0.6%

# Denotes the serial number of the character on the Backroom I tape.

TABLE II
FIGURE OF MERIT VERSUS RANDOM SELECTION

Method of Selection	Bayes' Weights Computed on	Error Rate on	Error Rate (Forced Decision)
Figure of merit, selected on #1–#1000	#1–#2000	#1–#2000	0.6%
Figure of merit, selected on #1000–#2000	#1–#2000	#1–#2000	0.6%
Random	#1–#2000	#1–#2000	3.7%

# Denotes the serial number of the character on the Backroom I tape.

TABLE III
GENERALIZATION CAPABILITY OF BAYES' DECISION ALGORITHM

Measurement Selection	Training Sample	Test Sample	Error Rate (Forced Decision)
#1–#1000	#1000–#4000	#1000–#4000	0.47%
#1–#1000	#4000–#7000	#4000–#7000	0.15%
		Average	0.31%
#1–#1000	#1000–#4000	#4000–#7000	0.68%
#1–#1000	#4000–#7000	#1000–#4000	0.65%
		Average	0.66%

# Denotes the serial number of the character on the Backroom I tape.

and 8 topological measurements, yields the lowest error rate on the Backroom I tape.

Preprocessing

These observations, while not directly related to measurement design and decision methods, are reported in order to emphasize the role of simple preprocessing techniques.

The importance of adequate normalization was confirmed on examination of the printouts of misidentified characters of the design set. Frequent failure of the normalization routine was shown by the histogram of character height distributions (Fig. 4). Removal of 3000 characters less than 27 bits from the analysis set reduced the forced decision error rate from 1.3 to 0.6 percent with the hand-designed measurement set. The other experiments described up to this point were run on this "doctored" set, but the normalization routine was improved in time for the experiments reported further on.

Shear normalization was programmed too late for inclusion in the complete series of experiments. Its promise was shown by the reduction of the error rate on the size-selected analysis data from 0.61 to 0.42 percent with the hand-designed measurements. The zoning of the n-tuples had to be changed to make effective use of shear normalization. While this is a simple procedure with the hand-designed measurements, other priorities on computer utilization prohibited recycling through the automatic measurement design and selection routines to include shear normalization in the subsequent experiments.

Decision Experiments

A set of weights for the Bayes' decision was estimated from each data set, and the decision carried out on all four data sets. The reject-error curves for these runs are shown in Figs. 9 through 12. As expected, recognition performance on each data set is best when its own statistics are used. On the other hand, the relative degree of care taken in the printing tended to swamp the other considerations. For example, the uncontrolled Tufts data gave a forced decision error rate of 8.25 percent on its own references, while the Backroom I set gave an error rate of 0.33 percent on its own references.

TABLE IV
COMPARISON OF VARIOUS TYPES OF MEASUREMENTS

Measurement Pool	Measurement Selection	Bayes' Weight Computed on	Error Rate on	Error Rate (Forced Decision)
R_{1800}	#1–#1000	#1–#7000	#1–#7000	1.20%
H_{100}	#1–#1000	#1–#7000	#1–#7000	0.61%
$R_{1800}H_{100}$	#1–#1000	#1–#7000	#1–#7000	0.43%
$R_{1800}H_{100}T_{17}$	#1–#1000	#1–#7000	#1–#7000	0.42%

# Denotes the serial number of the character on the Backroom I tape.
The subscript refers to the number of measurements of that type available.
R is randomly generated n-tuples, H intelligently designed n-tuples, T topological measurements.

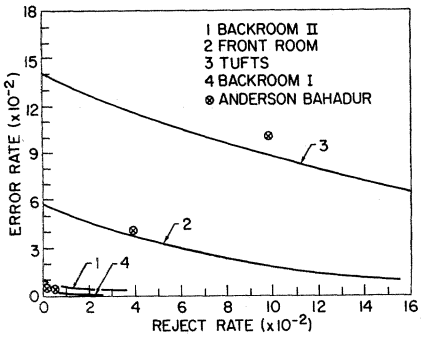


Fig. 9. Reject error curves using Bayes' decision with references estimated on the Backroom I data with Anderson–Bahadur results shown.

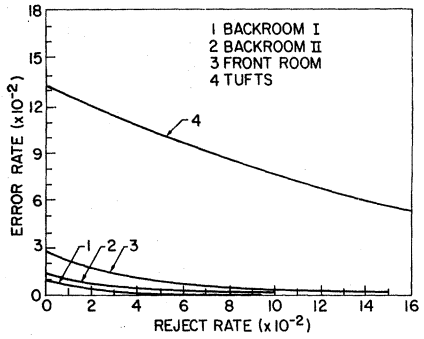


Fig. 11. Reject error curves using Bayes' decision with references estimated on frontroom data.

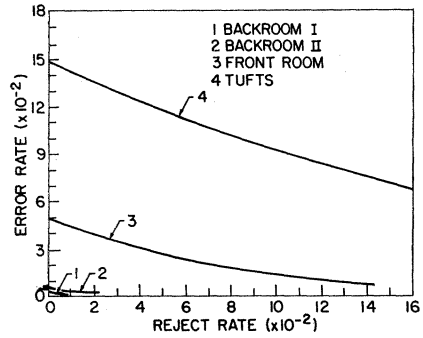


Fig. 10. Reject error curves using Bayes' decision with references estimated on Backroom II data.

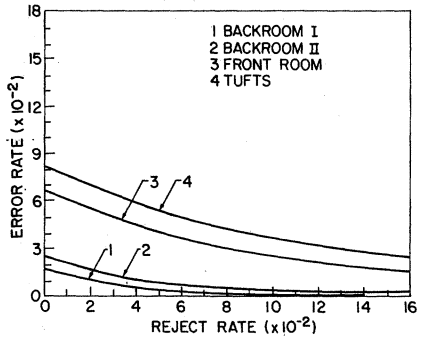


Fig. 12. Reject error curves using Bayes' decision with references estimated on Tufts University data.

Typical error patterns are shown in Fig. 13. Many of these errors are due to failure of the normalization because of stray bits, to misshapen characters, or lack of measurements to sense some peculiar stylistic variation. However, it must be conceded that many well-formed characters are also misrecognized, as shown in Fig. 14.

A set of 45 planes was designed by the Anderson–Bahadur method on the 7000-character Backroom I subset. The design took approximately 15 hours on the IBM 7094. (The program, which can accommodate up to 108 measurements, operates by a relaxation method and was developed by Casey.[24]) The results are shown in Table V and for purposes of comparison on Fig. 9. In all cases, the performance was very close to that of the maximum likelihood decision. This indicates that either the independence assumption was indeed satisfied or that there was an insufficient number of design samples. Because of the large amount of computer time required and our policy of one decision run, the planes were not redesigned on a larger data set.

This data was previously used by Greanias *et al.*[4] on their curve follower. Recognition results were 8-percent reject plus error on the Tufts data, they were 0.67-percent reject with 0.137-percent error on the Backroom I data.[3]

[3] Personal communication.

Concluding Remarks

The design and testing of a complete system for recognition of hand-printed numerals has been described. Using raster-scanned binary video, a mixed measurement set designed partially by machine, and linear decisions, it has given recognition rates from 86 to 99.7 percent (forced decision) depending on the quality of the data set and the appropriateness of the particular set of references used. The dominant factor affecting performance is the degree of care taken in the printing.

A novel measurement selection algorithm was used yielding a subset of measurements considerably better than one selected at random. It was found that 1000 characters were a sufficiently large sample for the selection procedure, and that combination of all three kinds of measurements resulted in recognition performance superior to that obtained by one kind alone, illustrating that the information extracted by each type is, to some extent, additive.

Height and shear normalization for each character decreased the error rate. This is certainly related to the size and position sensitivity of the zoned n-tuple measurements. But this type of invariance can be obtained by a program-controlled scanner with relatively straightforward logic and seems worth having in any event.

The Bayes' decision assuming independence gave better results than the Anderson–Bahadur method, although the

Fig. 13. Printouts of typical errors and rejects. IDENT. stands for the intended identity; DEC. for the computer decision.

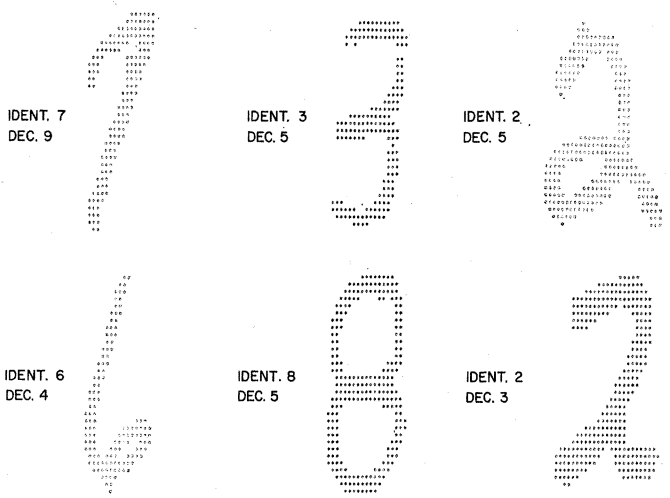


Fig. 14. Printouts of well-formed characters that are misidentified.

latter used more weights and more design time. The Bayes' procedure uses the probabilities for each measurement within each class, which may be estimated with fewer samples than the covariance matrix required in the Anderson–Bahadur method. The number of samples

TABLE V
ANDERSON–BAHADUR DECISION WEIGHTS OBTAINED ON BACKROOM I

Data Set	Errors	Rejects
Backroom I	24/6918 = 0.347%	11/6918 = 0.159%
Backroom II	65/11999 = 0.542%	47/11999 = 0.392%
Frontroom	492/11996 = 4.10%	469/11996 = 3.91%
Tufts	675/6746 = 10.01%	727/6746 = 10.78%

used for design appeared insufficient even for the Bayes' procedure. This illustrates how the choice of decision method must be influenced by the number of design samples available. Hand printing is ordinarily of such variability, that it may be the case that the large design samples required will make more complex procedures impractical. Among the promising techniques that should be investigated are clustering of the design set by machine, and self-adaptive procedures which continually update the decision weights. The latter seems attractive where the system has knowledge of the writer or of the document length.

A method for generation of random n-tuples using "line-seeking" constraints was developed. Existing meth-

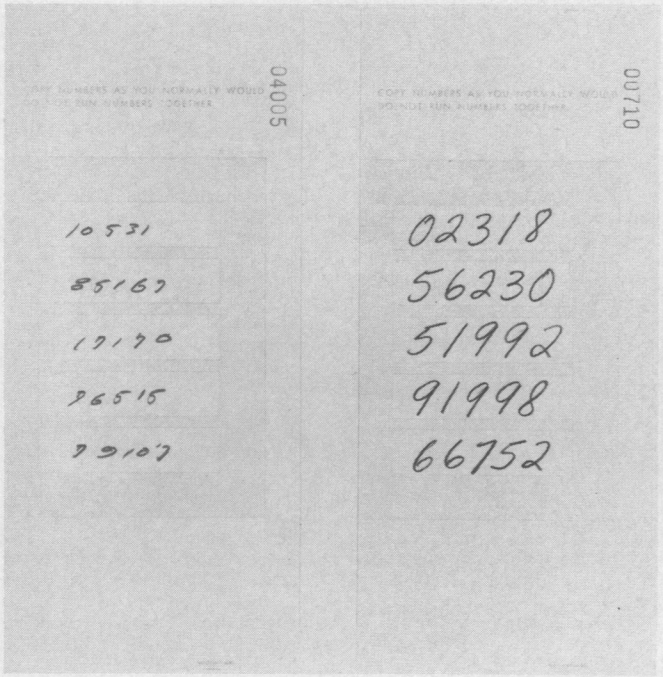


Fig. 15. Original data cards from Tufts University data set. The background is light blue which appears white to the scanner.

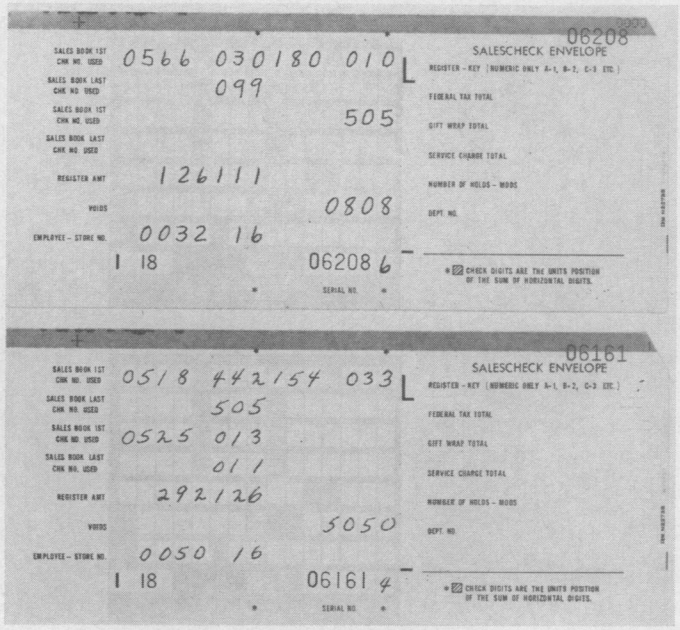


Fig. 16. Samples of the department store backroom data. Context information and check digits were not used in the experiments.

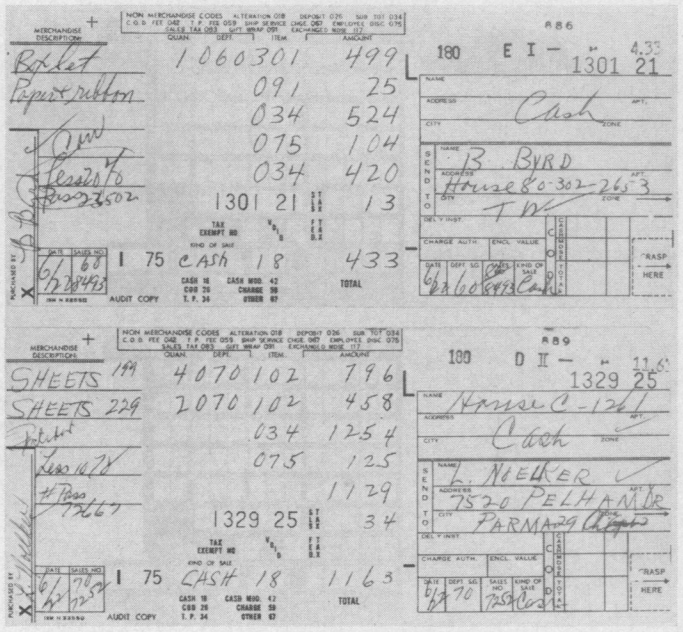


Fig. 17. Samples of the department store frontroom data printed by busy sales clerks.

ods were not applicable because they rely on the presence of stable points. It is interesting that this family of constraints gave significantly better results than any others tried. In a sense, we found our adaptive measurement design procedure trying to design curve-following measurements. It seems reasonable that the heuristic curve-following features do indeed extract the significant information from hand printing, and that an n-tuple machine, with enough zones and majority statements, could successfully accomplish the same feat. The advantages of automatic design for larger alphabets and a multiplicity of operating environment are obvious. Thus it would seem that the course to follow is to combine curve-following type measurements (derived either directly or from a raster scan) with automatic feature selection and parallel decision logic.

Appendix

Data Sets

Recognition results on hand-printed material depend markedly on the source of the data. In order to obtain a representative cross section, experiments were performed on data generated in widely varying circumstances over a period of about four years.

Tufts University Data Set

The 7000 characters in this data set were obtained from 120 college students, 52 high school students, 29 sales clerks, and 22 "statistical clerks and miscellaneous" at the Institute of Psychological Research at Tufts University, Medford, Mass., in 1962. Details of the data collection are discussed in Crook and Kellogg.[12] This set was printed freely, without any stylistic specifications. Samples of the original characters are shown in Fig. 15.

Department Store Backroom Data Set

The 20 000 numerals in this set are a subset of a larger sample produced in the course of routine operations by four inventory clerks in an Ohio department store. Characters were to be formed according to the following rules.

1) Gaps (as in top of 5) are not permitted.
2) Bays (as in 2, 3, 5) are required to be open.

3) Loops (as in 6, 8, 9, 0) are required to be well rounded and closed.
4) Lines cannot cross over materially where they close a figure (at top of 8 and 0).
5) Fancy strokes and extra-long tails are to be avoided.
6) Numerals are to be well proportioned, with proper balance between upper and lower portions.

The preprinted IBM cards on which the numerals were written were fed into an experimental character-recognition machine installed on the premises. Daily feedback was thus provided to maintain a relatively high level of neatness. Examples of this material are shown in Fig. 16. The background outlining the number fields is blue, which appears as white in the scanner.

Department Store Frontroom Data Set

About 40 sales clerks from 6 departments of the store contributed to this set of about 10 000 characters. Although each of the clerks had been subjected to a brief lecture outlining the desired print quality, subsequent supervision and feedback was lax. The clerks were often under time pressure in making out the preprinted sales slips, thus erasures, overprints, and crossed-out characters were common, despite instructions to begin a new card rather than make corrections. Typical samples of this set are shown in Fig. 17.

The identities of all the characters were keypunched and stored on magnetic tape to be copied later onto the scanned binary versions of the patterns. These identities were used to keep tally of errors and rejects. Whenever the number of characters obtained from a document did not match the number of identities stored on tape, a flag was set to indicate a scanning failure. About 15 percent of all the characters were thus rejected. These could have been recovered by manual labeling, but as the rejected frames were caused principally by random stray bits from the margin, this was not deemed necessary.

Acknowledgment

The authors are indebted to their colleagues in the Systems Science Group at the IBM Watson Research Center, Yorktown Heights, N. Y., and in particular to Miss M. Miller who wrote the scanning programs and to C. Marr and A. Sebastiano who operated the scanners. They also are obliged to E. Greanias and R. Norman, IBM ASD Mohansic, who furnished the data described in the Appendix.

References

[1] M. E. Stevens, "Automatic character recognition—a state of the art report," National Bureau of Standards, Washington, D.C., Tech. Note 112, May 1961.
[2] J. Munson, "Recognition of handprinted FORTRAN coding sheets," *Proc. 1966 IEEE Pattern Recognition Symp.* Washington, D.C., Thompson Book Publishing Co., April 1968.
[3] L. A. Kamentsky, "The simulation of three machines which read rows of handwritten arabic numbers," *IRE Trans. Electronic Computers*, vol. EC-10, pp. 489–501, September 1961.
[4] E. C. Greanias *et al.*, "The recognition of handwitten numerals by contour analysis," *IBM J. Res. and Develop.*, vol. 7, pp. 14–21, January 1963.
[5] W. Doyle, "Recognition of sloppy, hand-printed characters," *Proc. Western Joint Computer Conference*, pp. 133–142, May 1960.
[6] L. A. Kamentsky and C. N. Liu, "Computer-automated design of multifont print recognition logic," *IBM J. Res. and Develop.*, vol. 7, pp. 2–13, January 1963.
[7] C. N. Liu, "A programmed algorithm for designing multifont character recognition logics," *IEEE Trans. Electronic Computers*, vol. EC-13, pp. 586–593, October 1964.
[8] C. N. Liu and G. L. Shelton, Jr., "An experimental investigation of a mixed-font print recognition system," *IEEE Trans. Electronic Computers*, vol. EC-15, pp. 916–925, December 1966.
[9] C. K. Chow, "An optimum character recognition system using decision functions," *IRE Trans. Electronic Computers*, vol. EC-6, pp. 247–254, December 1957.
[10] W. H. Highleyman, "Linear decision functions, with application to pattern recognition," *Proc. IRE*, vol. 50, pp. 1501–1514, June 1962.
[11] J. L. Masterson and R. S. Hirsch, "Machine recognition of constrained handwritten arabic numbers," *IRE Trans. Human Factors in Electronics*, vol. HFE-3, pp. 62–65, September 1962.
[12] M. N. Crook and D. S. Kellogg, "Experimental factors for a handwritten numeral reader," *IBM J. Res. and Develop.*, vol. 7, pp. 76–79, January 1963.
[13] F. Kuhl, "Classification and recognition of hand-printed characters," *1963 IEEE Internat'l Conv. Rec.*, vol. 11, pt. 4, pp. 75–93, March 1963.
[14] R. J. Potter, "An optical character scanner," *SPIE J.*, vol. 2, p. 75, 1964.
[15] L. A. Kamentsky and C. N. Liu, "A theoretical and experimental study of a model for pattern recognition," in *Computer and Information Sciences*. Washington, D.C.: Spartan Books, 1964, pp. 194–218.
[16] W. W. Bledesloe and I. Browning, "Pattern recognition and reading by machine," *Proc. Eastern Joint Computer Conf.*, pp. 225–233, 1959.
[17] S. E. Estes, "Measurement selection for linear discriminants used in pattern classification," IBM Research Rept. RJ-331, April 1965.
[18] C. K. Chow, "A class of nonlinear recognition procedures," *IEEE Internat'l Conv. Rec.*, vol. 14, pt. 6, pp. 40–50, March 1966.
[19] S. J. Wilks, *Mathematical Statistics*. New York: Wiley, 1962, p. 574.
[20] G. Nagy and G. L. Shelton, Jr., "Self-corrective character recognition system," *IEEE Trans. Information Theory*, vol. IT-12, pp. 215–222, April 1966.
[21] T. W. Anderson and R. Bahadur, "Classification into two multivariate normal distributions with different covariance matrices," *Ann. Math. Stat.*, vol. 33, pp. 420–431, 1962.
[22] O. G. Selfridge, "Pandemonium: a paradigm for learning," *Proc. Symp. on Mechanization of Thought Processes*. London, England: H.M.S.O., 1959.
[23] G. Palmieri and R. Sanna, "Automatic probabilistic programmer analyzer for pattern recognition," *Estratto Rivista Methodos*, vol. 12, pp. 1–26, January 1960.
[24] R. G. Casey, "Linear reduction of dimensionality in pattern recognition," IBM Research Rept. RC-1431, March 1965.
[25] D. C. Allais, "The selection of measurements for prediction," Systems Theory Laboratory, Stanford University, Stanford, Calif. Tech. Rept. 6103-9, November 1964.