AN ITERATIVE TECHNIQUE OF SELECTING OPTIMAL MATRIX CHARACTER SETS FOR HANDPRINT RECOGNITION AND COMPUTER OUTPUT SYSTEMS

Chyi Shiau

A Thesis
in
The Department
of
Computer Science

Presented in Partial Fulfillment of the Requirements for
the Degree of Master of Computer Science at
Concordia University
Montreal, Quebec, Canada

March, 1978

ABSTRACT

CHYI SHIAU

AN ITERATIVE TECHNIQUE OF SELECTING OPTIMAL MATRIX CHARACTER SETS FOR HANDPRINT RECOGNITION AND COMPUTER OUTPUT SYSTEMS

Due to the facts that current optical readers have limited capability in the recognition of handprinted characters and the wide variety of dot-matrix alphanumeric designs offered by display manufacturers, an iterative method was developed to produce an optimum set of distinctive handprint 29 X 39 models for reliable optical character recognition and the most legible set of 5 X 7 matrix models for computer output systems.

A total of 90 different handprint and 121 dot-matrix models were investigated in this thesis. The models were compiled from an extensive survey of over 30 different handwriting and computer output systems. Eight quantitative measurements were used in the process of successive elimination of undesirable models.

ACKNOWLEDGEMENTS

I am greatly indebted to Professor C. Y. Suen for suggesting the topic of this thesis, for allowing me a generous amount of his time and for his many constructive comments during the preparation of this thesis. I also wish to thank Mr. V. McLeod and other staff members of the Computer Center for assistance in this project.

This project is partially supported by the National Research Council of Canada and the Department of Education of Quebec. The financial support of Concordia University is gratefully acknowledged.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Continuing advances in electronic technology and the steady reduction in equipment costs have resulted in the proliferation of man-machine communication devices. There are many commercial devices which can automatically recognize machine-printed characters at high speed with low error rates, not to mention there are also many computer output systems such as CRT's, thermo-printers and high-speed printers which are widely used. Due to the advantage of eliminating the expensive and error-prone process of data transcription, considerable interest has been developed recently in the automatic recognition of handwriting. Although some OCR machine can read multi-font and mixed font characters, most of them have very limited capabilities in the reading of handprinted characters. On the other hand, the problem of legibility of dot-matrix characters arises with the increasing usage of computer output systems. This thesis describes an iterative method aimed at establishing an optimum set of hand-print models so that a high recognition rate can be achieved using current OCR equipment. It is also the purpose of this thesis to determine the most legible set of dot-matrix characters for computer output systems.

Chapter 1 describes the procedures of the iterative method and each of the eight quantitative measurements used in the process. Various distance and information functions, including some of those employed to develop the OCR-A font [16], are discussed.

Chapter 2 provides a brief introduction to the need of reliable recognition of handprint characters. Each of the four elimination rules and the 12 iterations used in the process of eliminating undesirable models from a set of alphanumeric handprint models is discussed. The original set of 90 29 x 39 alphanumeric and the final optimum set of 36 models are presented.

Chapter 3 presents a brief discussion on the legibility problem and the need of an optimum dot-matrix character set for computer output systems. The four elimination rules and 12 iterations used in obtaining the most legible set of dot-matrix characters are discussed. Both the original set of 121 alphanumeric and the final set of 36 5 x 7 dot-matrix models are presented.

CHAPTER 1 ITERATIVE PROCESS

1.1 INTRODUCTION

In this chapter we will discuss the iterative technique used in obtaining the most distinct set of 5 x 7 dot-matrix characters among a set of 121 alphanumeric characters for the computer output systems (Chapter 3) and finding an optimum set of 29 x 39 matrix characters among a set of 90 handprinted characters for reliable optical character recognition (Chapter 2).

Eight different quantitative measurements were made when applying the iterative process to eliminate undesirable models (characters and models will be used synonymously in this thesis). These measurements were obtained from the average values of the following functions:

1) Similarity Function,
2) Hamming Distance,
3) Linear Correlation Function,
4) Cross Correlation Function,
5) Information Content,
6) Entropy,
7) Nearest Neighbor Distance-1,
8) Nearest Neighbor Distance-2.

By using the above functions, measurements were made between each character and the remaining characters in the set. These measurements were summed and averaged. The averages (or sums) of different models of the same symbol (e.g., the symbol "B" has 3 models B1, B2, and B3 in the original set of 90 29X 39 handprint characters as shown in Figure 1.) were compared. Those models rated as "undesirable" by all (100% rule) eight quantitative measurements were eliminated from the corpus. Measurements were computed again for the remaining set of characters. This "pruning" process continued until the 100% rule could not be applied any longer. Then the 87.5% rule (for a majority of 7 against 1) was used and the elimination process continued until this rule could not be applied any longer. Successive rules included the 75% (for a majority of 6 against 2) and 62.5% (for a majority of 5 against 3) rules. The remaining set of characters became smaller and smaller as the iterative process continued. The last 36 alphanumeric models were considered as the most distinct set of matrix characters.

1.2 NOTATIONS

Since the same eight quantitative measurements and iterative process were used for two different sets of data of different sizes (a set of 121 5 X 7 matrix models and a set of 90 29 X 39 matrix models), for simplicity, the following notations would be used throughout the remaining sections of this chapter:

S : The set of all character models being investigated in each stage of the iterative process. The total number of models in the character set S can be divided into 36 different symbols (26 character symbols from A to Z and 10 numeric symbols from 0 to 9).

C, D : Any two models in the character set S. Each of the character models in the set S is coded into an m × n matrix of m columns and n rows.

C_{ij} : The i-th column and j-th row cell of the character model C. $C_{ij} = 1$ if the ij-th cell is "occupied" and $C_{ij} = 0$ if it is not "occupied".

N_C : The number of matrix cells occupied by the model C.

N : The number of models in the character set S.

S(C) : The subset of character set S without the model C.

S(C,*) : The subset of S without all models belonging to the same symbol indicated by the model C.

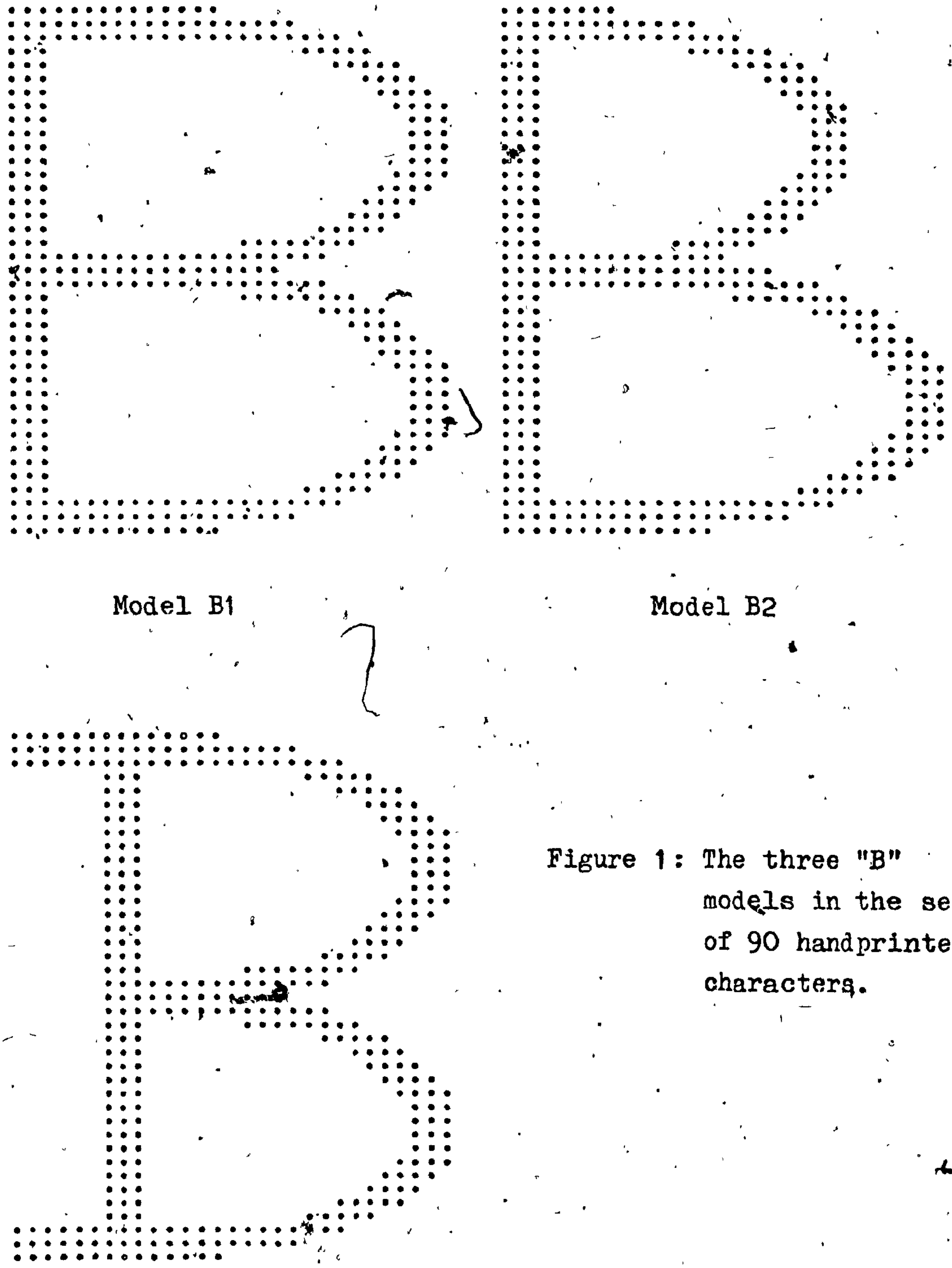
N(C,*) : The number of character models in the subset S(C,*).

T : T is an m × n density matrix, where T_{ij} indicates the number of times the ij-th cell being occupied by all models of the character set S.

P : P is an m × n probability matrix, where

$$P_{ij} = T_{ij} / \left(\sum_{j=1}^{n} \sum_{i=1}^{m} T_{ij} \right).$$

I : I is an m × n matrix, where $I_{ij} = -\log_2 P_{ij}$.



Model B1

Model B2

Model B3

Figure 1: The three "B" models in the set of 90 handprinted characters.

1.3 SIMILARITY FUNCTION

The Similarity Function A(C,D) measures the number of matrix cells common to both character models C and D. It can be expressed by the equation

$$A(C,D) = \sum_{i=1}^{m} \sum_{j=1}^{n} (C_{ij} .A. D_{ij}) \qquad \text{where}$$

$$C_{ij} .A. D_{ij} = \begin{cases} 1 & \text{, if the ij-th cell is occupied by both models C and D,} \\ 0 & \text{, otherwise.} \end{cases}$$

The smaller the value of A(C,D) is, the less the "common area" is shared by character models C and D and therefore the less the degree of "similarity" between models C and D.

The Similarity Function can be used to obtain two measurements which, in turn, can be used to indicate the desirability among models of the same symbol in the set S.

(a) By using Similarity Function, we can calculate the total sum and average of the measurements between each character model C and the remaining models in the set S (i.e., models in the subset S(C)):

$$SUM(AND,C) = \sum_{D \in S(C)} A(C,D) \qquad \text{and}$$

$$AVE(AND,C) = SUM(AND,C) / (N - 1) .$$

A larger value of SUM(AND,C) (and hence AVE(AND,C)) indicates that the total common area shared by the model C and all the remaining models in the character set S is larger and therefore less desirable. Conversely, a smaller value of SUM(AND,C) (and hence AVE(AND,C)) indicates that the common area shared by the model C and all the remaining models in the character set is smaller and therefore more distinctive and desirable. Consequently, among models of the same symbol, the one with the smallest value of AVE(AND,C) is preferred.

(b) By using Similarity Function, we can calculate the total sum and average of the measurements between each character model C and the remaining models of different symbols in the set S (i.e., models in the subset S(C,*)):

$$\mathrm{SUM(AND,C,*)} = \sum_{D \in S(C,*)} A(C,D) \qquad \text{and}$$

$$\mathrm{AVE(AND,C,*)} = \mathrm{SUM(AND,C,*)}\ /\ N(C,*) \quad .$$

Similarly, a smaller value of SUM(AND,C,*) (and hence AVE(AND, C,*)) indicates that the common area shared by the model C and all character models of different symbols is smaller. Therefore the smaller the value of SUM(AND,C,*) (and hence AVE(AND,C,*)) the more distinctive and desirable the model C is against all models of different symbols. Consequently, among models of the same character symbol, the one with the smallest value of AVE(AND,C,*) is preferred.

Both measurements AVE(AND,C) and AVE(AND,C,*) could be used for evaluation of models in our iterative process. Since consistent results were obtained from these two measurements, only the measurement AVE(AND,C,*) was used in the iterative process.

Table 1 illustrates the results of the operations AVE(AND, C) and AVE(AND,C,*) to the three character models B1, B2, and B3 of the same symbol "B" (in Figure 1) during the first iterative process of 90 29 X 39 character models from which we concluded that the character model B3 is the most (and B1 is the least) desirable model among the three.

		Measurements	
		AVE(AND,)	AVE(AND, ,*)
Models	B1	157.55	154.52
	B2	135.72	133.16
	B3	118.64	115.95

Table 1: The measurements of the three "B" 29 X 39 character models in the first iteration.

1.4 HAMMING DISTANCE

A very simple error-detecting and error-correcting encoding method has been devised by Hamming [2] . This method is based upon the concept of "distance" between code words and can be used to detect the "desirability" of character models.

The Hamming Distance X(C,D) measures the number of uncommon matrix cells between any two models C and D in the character set S. It can be expressed by the equation

$$X(C,D) = \sum_{i=1}^{m} \sum_{j=1}^{n} (C_{ij} \text{ .XOR. } D_{ij}) \qquad \text{where}$$

$$C_{ij} \text{ .XOR. } D_{ij} = \begin{cases} 1 & \text{, if models C and D are unequal at the ij-th cell,} \\ 0 & \text{, otherwise.} \end{cases}$$

Obviously, the larger the value of X(C,D) is, the greater the "difference" is between the two models C and D.

As is in the case of the Similarity Function, the Hamming Distance can be applied to obtain two measurements which, in turn, can be used in the evaluation of the "desirability" of character models of the same symbol:

(a) By applying Hamming Distance, the total sum and average of the measurements between each character model C and the remaining models in the set S (i.e., models in the subset S(C)) are given by

$$\text{SUM(XOR,C)} = \sum_{D \in S(C)} X(C,D) \qquad \text{and}$$

$$\text{AVE(XOR,C)} = \text{SUM(XOR,C)} \;/\; (N - 1) \;.$$

A smaller value of SUM(XOR,C) (and hence AVE(XOR,C)) indicates that the total common area shared by the model C and all remaining character models in the set S is larger and hence is less desirable. Conversely, a larger value of SUM(XOR,C) (and hence AVE(XOR,C)) indicates that the common area shared by the model C and all remaining character models in the set is smaller and therefore the model C is more "distinct" from the remaining models in the set. Consequently, among models of the same character symbol, the one with the largest value of AVE(XOR,C) is preferred.

(b) By applying Hamming Distance, the total sum and average of the measurements between each character model C and the remaining models of different symbols in the set S (i.e., models in the subset S(C,*)) are given by

$$\mathrm{SUM}(\mathrm{XOR},C,*) = \sum_{D \in S(C,*)} X(C,D) \quad \text{and}$$

$$\mathrm{AVE}(\mathrm{XOR},C,*) = \mathrm{SUM}(\mathrm{XOR},C,*) \;/\; N(C,*) \quad .$$

Similarly, a larger value of SUM(XOR,C,*) (and hence AVE(XOR, C,*)) indicates that the common area shared by the model C and all character models in the set S(C,*) is smaller. Therefore, a larger value of AVE(XOR,C,*) indicates that the model C is more "distinct" against all models of different symbols and thus more "desirable". Consequently, among models of the same character symbol, the one with the largest value of AVE(XOR,C,*) is preferred.

Since both measurements AVE(XOR,C) and AVE(XOR,C,*) produced consistent results, here again, only measurement AVE(XOR,C,*) was used in the iterative process.

Here again, Table 2 presents the results after applying the operations AVE(XOR,C) and AVE(XOR,C,*) to the three models B1, B2, and B3 (in Figure 1) during the first iterative process from which we concluded that the character model B3 is the most (and B1 is the least) desirable model among the three.

	Measurements		
	AVE(XOR,)		AVE(XOR, ,*)
Models	B1	348.45	352.52
	B2	367.39	370.23
	B3	408.47	411.64

Table 2: The Hamming Distance measurements of the three "B" character models in the first iteration.

1.5 LINEAR CORRELATION FUNCTION

The Similarity Function A(C,D) discussed in section 1.3 measures the number of matrix cells "common" to both character models C and D. The smaller the value of A(C,D) is, the more C is distinct from D. The function A(C,D) can be modified to obtain a new measurement LA(C,D) by taking into consideration of the various degrees of misalignment and stroke width variation of models. The measurement LA(C,D), called linear correlation function, is obtained by dividing the Similarity Function A(C,D) by the arithmetic mean of the number of cells "occupied" by models C and D, i.e.

$$
\begin{aligned}
LA(C,D) &= A(C,D) \; / \; (\; (N_C + N_D) \; / \; 2 \;) \\
&= 2(\; A(C,D) \; / \; (N_C + N_D) \;) \quad ;
\end{aligned}
$$

where N_C and N_D are the numbers of matrix cells occupied by models C and D respectively. The measurement LA(C,D) gives us a normalized area correlation between the pair of models C and D. It is clear that LA(C,C) = 1 for any model C.

By using the Linear Correlation Function LA(C,D), a magnified sum of the measurements between each character model C and all models in the set S is given by

$$SUM(LA,C) = \left(\sum_{D \in S} LA(C,D) \right) \cdot N$$

where N is the number of models in the character set S. Thus, the average of this sum is

$$AVE(LA,C) = SUM(LA,C) \; / \; N$$

$$= \sum_{D \in S} LA(C,D) \quad = \sum_{D \in S} 2(\; A(C,D) \; / \; (N_C + N_D) \;).$$

Obviously, a smaller value of AVE(LA,C) indicates that the normalized common area shared by the model C and all character models in the entire set S is smaller. Consequently, among models of the same character symbol, the one with the smallest value of AVE(LA,C) is preferred. For instance, as shown in Table 3, the character model B3 is the most (and B1 is the least) desirable model among the three "B" models.

	AVE(LA,)
B1	47.24
B2	42.40
B3	36.89

Table 3: The Linear Correlation measurements of the 3 "B" models.

1.6 CROSS CORRELATION FUNCTION

By taking into consideration the degree of misalignment and stroke width variation of character models, another measurement CA(C,D) can be obtained by normalizing the Similarity Function A(C,D). The measurement CA(C,D), called cross correlation function, is obtained by taking the square of the quotient of the Similarity Function A(C,D) and the algebraic mean of the number of cells occupied by models C and D. That is,

$$CA(C,D) = \left(A(C,D) \;/\; (N_C \cdot N_D)^{1/2} \right)^2$$

$$= \left(A(C,D) \right)^2 / (N_C \cdot N_D) \quad .$$

The measurement CA(C,D) gives us a normalized area correlation between the pair of models C and D. It is clear that CA(C,C)=1 for any model C.

By using the Cross Correlation Function CA(C,D), a magnified sum of the measurements between each character model C and all models in the set S is given by

$$SUM(CA,C) = \left(\sum_{D \in S} CA(C,D) \right) \cdot N$$

where N is the number of models in the character set S. Thus, the average of this sum is given by

$$AVE(CA,C) = SUM(CA,C) \;/\; N$$

$$= \sum_{D \in S} CA(C,D) \quad = \sum_{D \in S} \left(\left(A(C,D) \right)^2 / N_C \cdot N_D \right) \quad .$$

Similarly, a smaller value of AVE(CA,C) indicates that the normalized common area shared by the model C and all character models in the entire set S is smaller. Consequently, a model with smaller value of AVE(CA,C) indicates that it is more "distinct" against all models in the character set S. Therefore, among models of the same character symbol, the one with the smallest value of AVE(CA,C) is preferred. For instance, as shown in Table 4, by applying the measurement AVE(CA,C) to the three models B1, B2, and B3 of character "B" (in Figure 1) during the first iteration we concluded that the character model B3 is the most (and B1 is the least) desirable model among the three.

	AVE(CA,)
B1	26.42
B2	20.96
B3	16.36

Table 4: The Cross Correlation measurements of the three "B" models in the first iteration.

1.7 INFORMATION CONTENT AND ENTROPY

Ingels [5] defined that "the information content of a message symbol is the negative of the logarithm of the probability that this symbol will be emitted from the source." He explained that if we were to assign numbers relating to the information carried by statements, it would be natural to assign a small number to common everyday statements such as "Hello", "How are you?" and a large number to unusual items such as the announcement of the bombing of the Pearl Harbor. Thus we would want to denote information content of a statement as a function that is decreasing in numerical value as the probability of occurrence is increasing [17].

The information measure is a logarithmic function that depends upon the uncertainty, or probability of occurrence, associated with the message symbol. Thus if a particular message symbol s_{ij} were to occur with probability p_{ij}, we would say that the information content associated with this symbol is defined as

$$I(s_{ij}) = -\log_2 p_{ij} \quad \text{bits} .$$

In the determination of the optimum sets of 5 × 7 matrix characters for computer output systems and of 29 × 39 handprint matrix characters for reliable optical recognition, the concept of information content was used to obtain the following two measurements:

Consider a set S of N character models is being investigated. Each model is coded into an m × n matrix of m columns and n rows. In each stage of the iterative process, an m × n density matrix T is obtained by counting the number of times the ij-th cell being occupied by all N character models. Thus the density matrix T indicates the distribution of the entire set of character models currently being investigated. Next, an m×n probability matrix P is obtained by calculating the probability of occurrence for each cell of the density matrix T. That is,

$$P_{ij} = T_{ij} / (\sum_{i=1}^{m} \sum_{j=1}^{n} T_{ij})$$

for each ij-th cell of P and P_{ij} indicates the probability of "occupancy" of the ij-th cell. Finally, an m×n information matrix I is obtained by

$$I_{ij} = - \log_2 P_{ij}$$

for each ij-th cell and I_{ij} indicates the information content of the ij-th cell of the entire set of character models.

For each model C in the set S, the following two measurements are then obtained:

1.7.1 INFORMATION CONTENT MEASUREMENT

$$INF(C) = \sum_{i=1}^{m} \sum_{j=1}^{n} (I_{ij} \cdot C_{ij}) .$$

This is a measure of the information carried by the cells of the model C based on the distribution of the entire set of character models under investigation. In view of the above discussion, it is clear that among models of the same symbol, the one with the smallest value of INF(C) is preferred.

1.7.2 ENTROPY MEASUREMENT

From the above discussion and our knowledge of logarithms we should suspect that the information measure is additive, i.e., if we have two items we would expect the average information of the combination to be the sum of the information of the individual items, weighted in accordance with their probability of occurrence. In general, if s_1 and s_2 occur with probability p_1 and p_2 respectively, we have as the individual information content $-\log_2 p_1$ and $-\log_2 p_2$. The average information of the combination would then be

$$H(s_1, s_2) = p_1(-\log_2 p_1) + p_2(-\log_2 p_2) .$$

The function H is termed the average information content (or entropy) of the source consisting of items s_1 and s_2 . In view of the above discussion, we then obtain the entropy measurement

$$\text{ENT}(C) = \sum_{i=1}^{m} \sum_{j=1}^{n} (P_{ij} \cdot I_{ij} \cdot C_{ij})$$

for each model C in the character set S. Thus, the entropy measurement ENT(C) is the weighted average of the information content per cell of the character model C. Therefore, among models of the same symbol, the one with the smallest value of ENT(C) is preferred. For instance, as shown in Table 5, by applying the measurement INF(C) and ENT(C) to the three models

B1, B2, and B3 of character "B" (in Figure 1) during the first iteration we concluded that the model B2 is the most (and B1 is the least) desirable model among the three with respect to the Information Content measurement INF(C) and model B3 is the most (and B1 is the least) desirable model with respect to the Entropy measurement ENT(C).

	INF( )
B1	3660.56
B2	3502.25
B3	3660.45

5 (a)

	ENT( )
B1	5.347
B2	4.641
B3	4.151

5 (b)

Table 5 (a): The Information Content measurements of the three "B" models in the first iteration.

(b): The Entropy measurements of the three "B" models in the first iteration.

1.8 NEAREST NEIGHBOR DISTANCE

In this section, two new distance measurements will be discussed. These two measurements could be used to distinguish the "desirability" of different character models.

Consider two $m \times n$ character models C and D. For each ij-th cell of the model C, a "nearest cell distance" $d(C_{ij},D)$ can be obtained by comparing the nearest distance between the cell C_{ij} and cells occupied by model D:

$$d(C_{ij},D) = \begin{cases} 0 \quad , \text{ if } C_{ij} = 0 \;, \\ \min\limits_{\substack{1 \leq k \leq m \\ 1 \leq l \leq n}} \left\{ (k-i)^2+(l-j)^2 \;\middle|\; D_{kl} \neq 0 \right\}, \quad \text{if } C_{ij} \neq 0, \end{cases}$$

Obviously, a larger value of the nearest cell distance $d(C_{ij},D)$ indicates a larger distance between the cell C_{ij} and the "nearest cell" occupied by the model D. It is clear that $d(C_{ij},D) \neq d(D_{ij},C)$ in general. Two distance measurements between models C and D can be obtained by the following formulae:

$$MID1(C,D) = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} \left(d(C_{ij},D) \right)^{1/2} \right) / N_C \; + $$

$$\left(\sum_{i=1}^{m} \sum_{j=1}^{n} \left(d(D_{ij},C) \right)^{1/2} \right) / N_D$$

and

$$MID2(C,D) = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} d(C_{ij},D)/N_C + \sum_{i=1}^{m} \sum_{j=1}^{n} d(D_{ij},C)/N_D \right)^{1/2}$$

where N_C and N_D are the number of cells occupied by the models C and D respectively. Clearly, both MID1(C,D) and MID2(C,D) are distance measurements, and MID1(C,D) = MID1(D,C) and MID2(C,D) = MID2(D,C).

The measurement MID1(C,D), called the Nearest Neighbor Distance-1, is obtained by adding the normalized sum of the square roots of the nearest cell distance $d(C_{ij},D)$ for all ij-th cells of C and the normalized sum of the square roots of the nearest cell distance $d(D_{ij},C)$ for all ij-th cells of D. Therefore, a larger value of MID1(C,D) indicates that the pair C and D is more distinctive. Consequently, among models of the same symbol, the one with the largest value of MID1(C,D) is preferred.

The measurement MID2(C,D), called the Nearest Neighbor Distance-2, is obtained by adding the normalized sum of $d(C_{ij},D)$ for all ij-th cells of C and the normalized sum of $d(D_{ij},C)$ for all ij-th cells of D and then taking the square root of the result. Similarly, a larger value of MID2(C,D) indicates the pair C and D is more distinctive. Consequently, among models of the same symbol, the one with the largest value of MID2(C,D) is preferred.

For example, consider the three 5 X 7 character models I1, I2, and L1 in Figure 2. It is not difficult to calculate the "nearest cell" distances between the pair I1 and L1. The results are shown in Table 6. The nearest neighbor distances between the pair I1 and L1, and between the pair I2 and L1 are presented in Table 7. Since the two nearest neighbor distances between the pair I1 and L1 are greater than the corresponding distances between the pair I2 and L1, thus the pair I1 and L1 is more distinguishable than the pair I2 and L1. Therefore the pair I1 and L1 is considered more desirable between the two pairs.

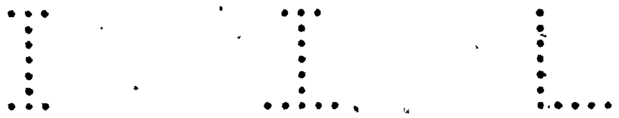


Figure 2: The three 5×7 character models

0	1	4	9	0
0	0	4	0	0
0	0	4	0	0
0	0	4	0	0
0	0	4	0	0
0	0	1	0	0
0	0	0	0	0

6(a)

1	0	0	0	0
2	0	0	0	0
4	0	0	0	0
4	0	0	0	0
4	0	0	0	0
2	0	0	0	0
1	0	0	0	1

6 (b)

Table 6 (a): The value in the ij-th cell of the matrix indicates the nearest cell distance of the cell $I1_{ij}$ with respect to the model L1 ,

(b): The value in the ij-th cell of the matrix indicates the nearest cell distance of the cell $L1_{ij}$ with respect to the model I1 .

		I1	I2
L1	MID1	2.4	2.0
	MID2	2.1	2.0

Table 7: The nearest neighbor distances between I1,L1 and between I2 and L1 .

The total sum and average of the nearest neighbor distances between each character model C and the remaining models of different symbols are given by

$$\mathrm{SUM}(\mathrm{MID1},C,*) = \sum_{D \in S(C,*)} \mathrm{MID1}(C,D) \quad ,$$

$$\mathrm{AVE}(\mathrm{MID1},C,*) = \mathrm{SUM}(\mathrm{MID1},C,*) \;/\; N(C,*) \quad ,$$

$$\mathrm{SUM}(\mathrm{MID2},C,*) = \sum_{D \in S(C,*)} \mathrm{MID2}(C,D) \qquad \text{and}$$

$$\mathrm{AVE}(\mathrm{MID2},C,*) = \mathrm{SUM}(\mathrm{MID2},C,*) \;/\; N(C,*) \quad .$$

Clearly, a larger value of AVE(MID1,C,*) (or AVE(MID2,C,*)) indicates that the model is more "distinct" against the models of different symbols. Consequently, among models of the same symbol, the one with the largest value of AVE(MID1,C,*) (or AVE(MID2,C,*)) is preferred. For instance, as shown in Table 8, applying the two Nearest Neighbor Distance measurements to the three models B1, B2, and B3 of character "B" (in Figure 1) during the first iteration we concluded that model B3 is the most (and B1 is the least) desirable model among the three with respect to the measurement AVE(MID1,C,*) and model B2 is the most (and B1 is the least) desirable model with respect to the measurement AVE(MID2,C,*) .

	AVE(MID1, ,*)
B1	4.32
B2	4.75
B3	4.99

8 (a)

	AVE(MID2, ,*)
B1	5.10
B2	5.33
B3	5.25

8 (b)

Table 8 (a): The results obtained by applying the measurement AVE(MID1, ,*) to the three "B" models during the first iteration.

(b): The results obtained by applying the measurement AVE(MID2, ,*) to the three "B" models during the first iteration.

1.9 CONCLUSION

During each stage of the iterative process, eight different quantitative measurements AVE(AND,C,*), AVE(XOR,C,*), AVE(LA,C), AVE(CA,C), INF(C), ENT(C), AVE(MID1,C,*), and AVE(MID2,C,*) were made between each character model C and the remaining models in the character set (or the remaining models of different symbols). The measurements of different models of the same symbol were compared among themselves. The models rated as "undesirable" by all (100% rule) eight measurements were eliminated from the corpus. Measurements were computed again for the remaining set of models. This "pruning" process continued until the 100% rule could not be applied any longer. Then the successive 87.5%, 75%, and 62.5% rules were used in the same manner.

The last 36 alphanumeric models were considered as the most distinct set of matrix characters. For instance, as shown in Table 9, after the first iteration for the set of 90 29 X 39 character models, model B1 was eliminated from the three models of character "B" in Figure 1.

Measurements	Model B1	Model B2	Model B3
AVE(AND, ,*)	154.52	133.16	115.95
AVE(XOR, ,*)	352.52	370.23	411.64
AVE(LA,)	47.24	42.40	36.89
AVE(CA,)	26.42	20.96	16.36
INF( )	3660.56	3502.25	3660.45
ENT( )	5.347	4.641	4.151
AVE(MID1,,*)	4.32	4.75	4.99
AVE(MID2,,*)	5.10	5.33	5.25

Table 9: The eight quantitative measurements of the three "B" models.

CHAPTER 2 RELIABLE RECOGNITION OF HANDPRINT DATA

2.1 INTRODUCTION

During the last two decades, optical character recognition (OCR), an essentially new field in technology, has come into being and has been nurtured intensively. The principal impetus to the development of OCR has been given by the need to cope with an enormous flood of paper generated by an expanding technological society. When the numbers of bank checks, commercial forms, government records, credit-card imprints, and pieces of mail to be sorted and accounted reach several billion each week, man requires the help of his machines. Other pressures arise from the desire to index and retrieve literature references automatically, to have machine translation between languages, and to provide sensory prostheses for the blind. Finally there is the purely intellectual challenge to find ways to make machines replicate the functions of humans.

Character recognition systems generally fall into two categories. First, there are those that recognize only machine printed characters, and second, there are those that handle data printed (or written) by hand. The first class is conceptually fairly simple, since the data to be recognized is essentially invariant. Thus, a simple template matching scheme can be the basis for reading machine prints. In fact, such systems have reached commercial practicality and are currently in use.

In recent years, considerable interest has been developed in the automatic recognition of handwriting, not only because it is a very challenging problem, but also because it has many practical applications, in particular, the advantage of eliminating the expensive and error-prone process of data transcription.

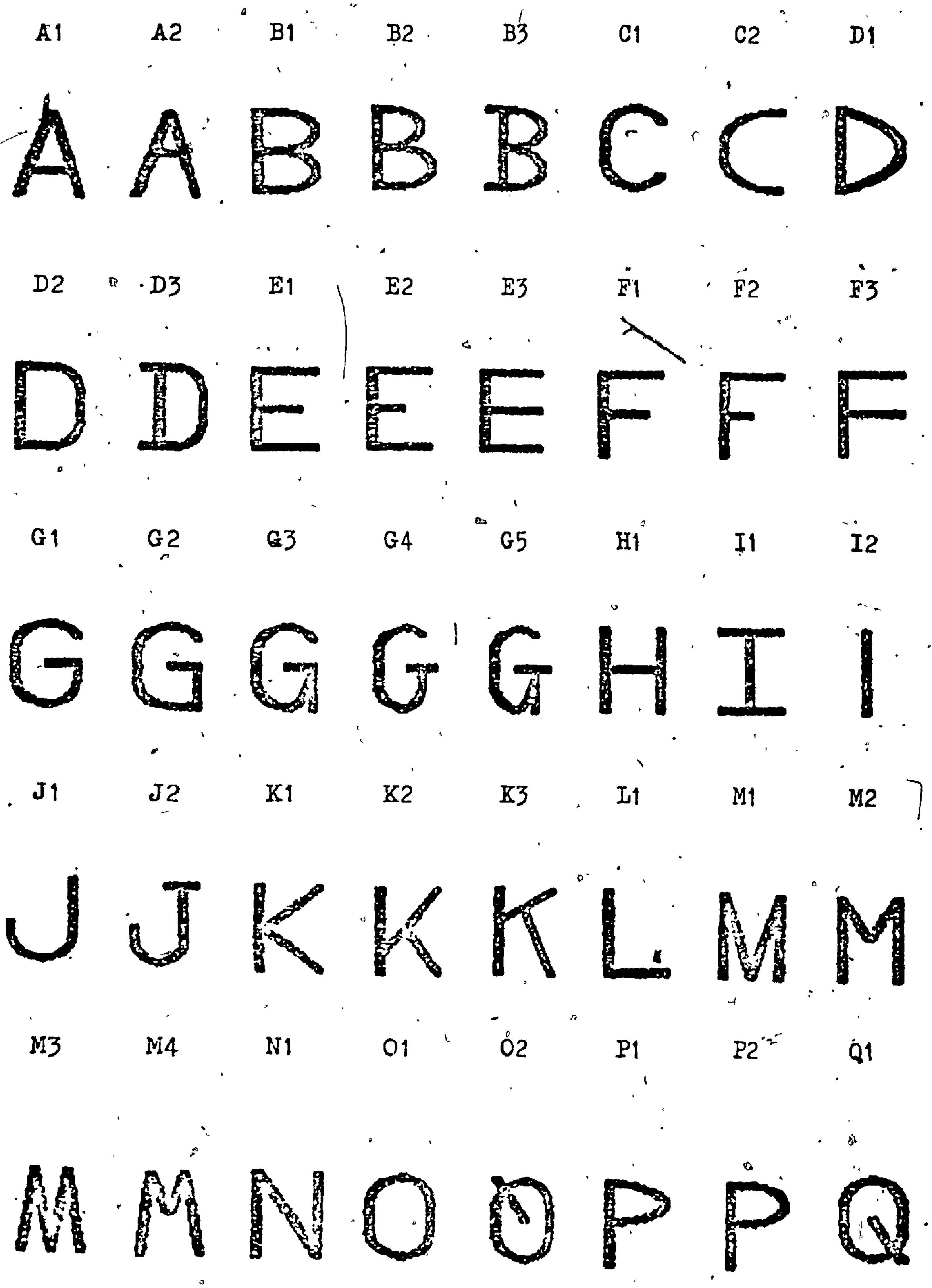
Although some OCR machines can read multi-font and mixed font characters, most of them have very limited capabilities in the reading of handprinted characters. Since each writer has his/her own style of writing, there is little doubt that handprints are less tractable than multi-font characters. The difficulty of handprint recognition was illustrated very clearly by Neisser and Weene [8] who showed that the human recognition rate of a selected sample of about 650 relatively unconstrained handprint characters was only about 95.9% correct. In another experiment conducted by Suen [12] , 30 subjects were instructed to handprint as quickly and carefully as possible for recognition by another group of 30 subjects. A more encouraging recognition rate (> 98% correct) was achieved. This indicates that simple instruction can improve the quality of characters significantly. This fact, plus the sensitivity of OCR to print degradation and document mutilation (Suen [13]), indicates that in order to attain a reliable recognition rate (< 0.1% error) for use in commercial applications, some kind of constraints must be imposed on the writer when entering the data. One simple solution to achieve a high recognition rate by using current OCR equipment is to develop a standard such that the person entering the data can write in a style as close as possible to the given standard (Suen [11]).

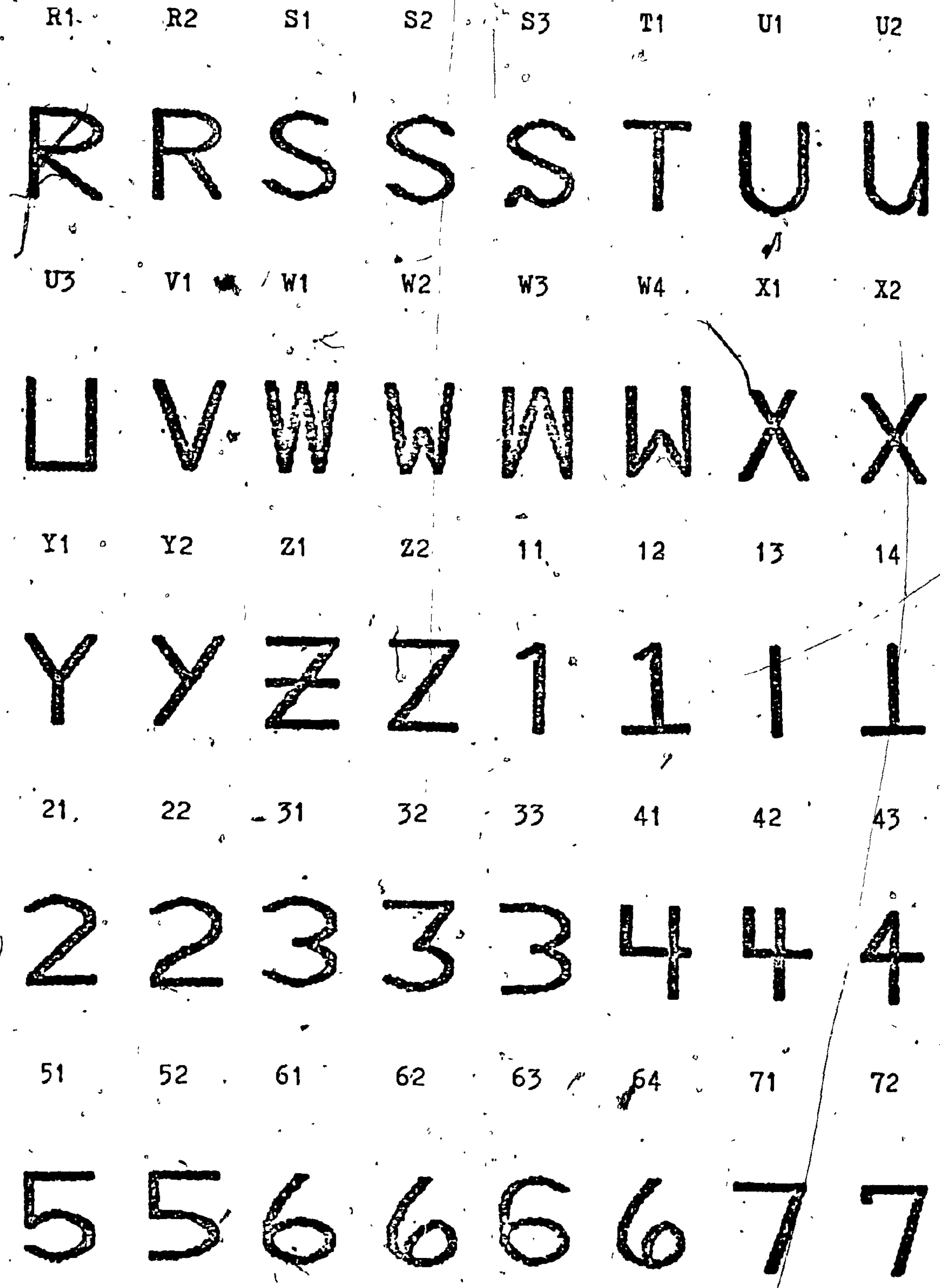
In view of the above, an iterative method was developed to produce an optimum set of alphanumeric models so that a high recognition rate could be achieved using current OCR equipments.

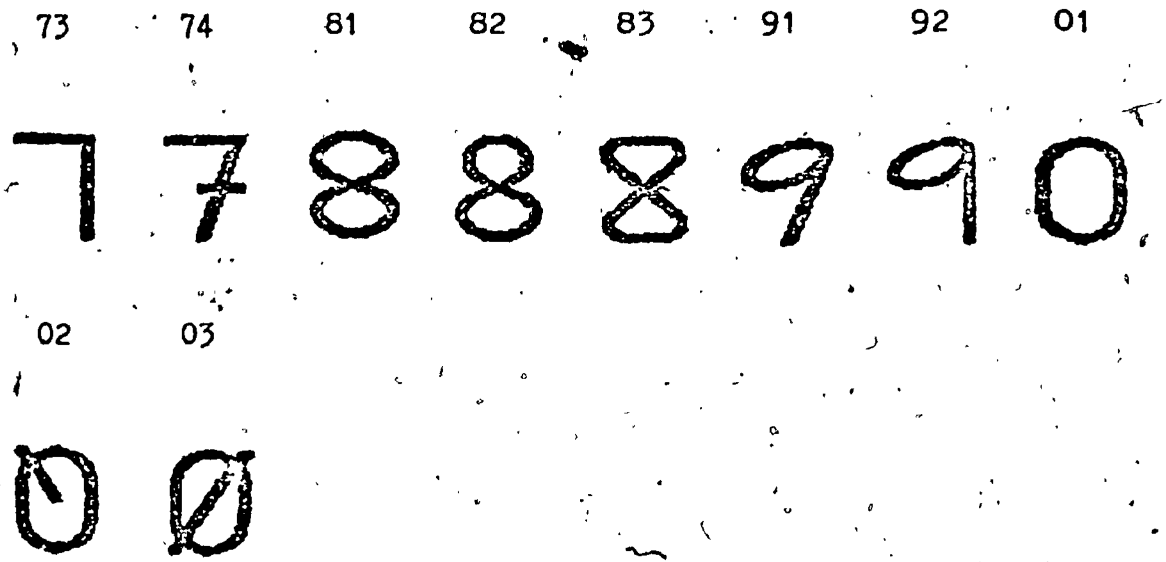
2.2 DATA COLLECTION

In the preparation of the optimum set of handprint alphanumeric models for reliable recognition, an extensive review of the handwriting systems taught in elementary schools and in OCR applications was conducted by Suen [10, 12]. In this investigation, more than 30 different handwriting systems were studied and the ANSI handprint standard was also examined. As a result, 90 handprint alphanumeric models (as shown in Figure 3) were chosen. Each model was coded into a 29 x 39 matrix. Based on the concept of majority rule, a FORTRAN program was written to calculate the eight quantitative measurements of each model. Models of the same character symbol were ranked and those undesirable ones were eliminated by an iterative process.

Figure 3: The 90 29 × 39 handprint models.




R1
R2
S1
S2
S3
T1
U1
U2
R R S S S T U U
U3
V1
W1
W2
W3
W4
X1
X2
U V W W W W X X
Y1
Y2
Z1
Z2
11
12
13
14
Y Y Z Z 1 1 1 1
21
22
31
32
33
41
42
43
2 2 3 3 3 4 4 4
51
52
61
62
63
64
71
72
5 5 6 6 6 6 7 7



2.3 ELIMINATION PROCEDURES

A total of 90 handprinted alphanumeric models (as shown in Figure 3) were chosen and used in an iterative process. In each stage of the iterative process, eight quantitative measurements were made and compared among different models of the same symbol. Those models rated as "undesirable" were eliminated by 100%, 87.5%, 75%, and 62.5% elimination rules accordingly.

2.3.1 THE 100% ELIMINATION RULE

The 100% elimination rule was used in the first two iterations. In each of these two iterations, the eight quantitative measurements of different models of the same symbol were compared among themselves and those models rated as undesirable by all measurements were eliminated from the set.

A total of 11 alphanumeric models were eliminated after the first iteration. They are designated as models B1, C1, O1, O2, P2, R2, S2, W3, X2, 63, and 81 in Figure 3. For instance, as shown in Table 9, the model B1 was eliminated after the first iteration because it was rated as undesirable against the model B2 by all eight measurements. Due to a very large amount of computing time used and the fact that the pair of character model O1 and numeric model 01 and the pair of character model O2 and numeric model 02 are exactly the same, we decided to drop character models O1 and O2 after the first iteration and use one of the numeric models 01 and 02 as character "O" later.

After the second iteration, six more alphanumeric models were eliminated from the remaining set of 79 models. They are designated as models D2, G1, G2, G3, M1, and 41 in Figure 3. For instance, as shown in Table 10, the first three models of character "G" were eliminated because the models G1 and G2 were rated as undesirable against the model G4, and model G3 was rated as undesirable against model G5 by all eight measurements. Thus there were only 73 alphanumeric models remaining after the application of the 100% elimination rule.

Measurements	Model G1	Model G2	Model G3	Model G4	Model G5
AVE(AND, ,*)	109.85	115.28	116.62	90.18	100.26
AVE(XOR, ,*)	335.09	349.23	363.55	349.45	372.28
AVE(LA,)	41.61	41.72	41.06	36.27	37.05
AVE(CA,)	21.23	20.87	20.39	16.06	16.30
INF( )	2651.87	2909.16	3104.55	2483.97	2921.29
ENT( )	3.977	4.186	4.256	3.318	3.693
AVE(MID1,,*)	5.14	5.22	5.15	5.62	5.34
AVE(MID2,,*)	5.61	5.74	5.64	5.83	5.64

Table 10: Measurements of the five "G" models in the second iteration.

2.3.2 THE 87.5% ELIMINATION RULE

The 87.5% elimination rule was used in the third to seventh iterations. In each of these iterations, the eight quantitative measurements of different models of the same symbol were compared among themselves and those models rated as undesirable by a majority of 7 against 1 were eliminated from the set.

A total of 6 alphanumeric models were eliminated from the remaining set of 73 models after the third iteration. They are designated as models E3, F3, 12, 14, 74, and O3 in Figure 3. For instance, as shown in Table 11, the third model E3 of the character "E" was eliminated because it was rated as undesirable against the model E2 by a majority of 7 to 1. Since the shape of model E1 is "desirable" we decided to keep it for another iteration even though it could be eliminated in this iteration.

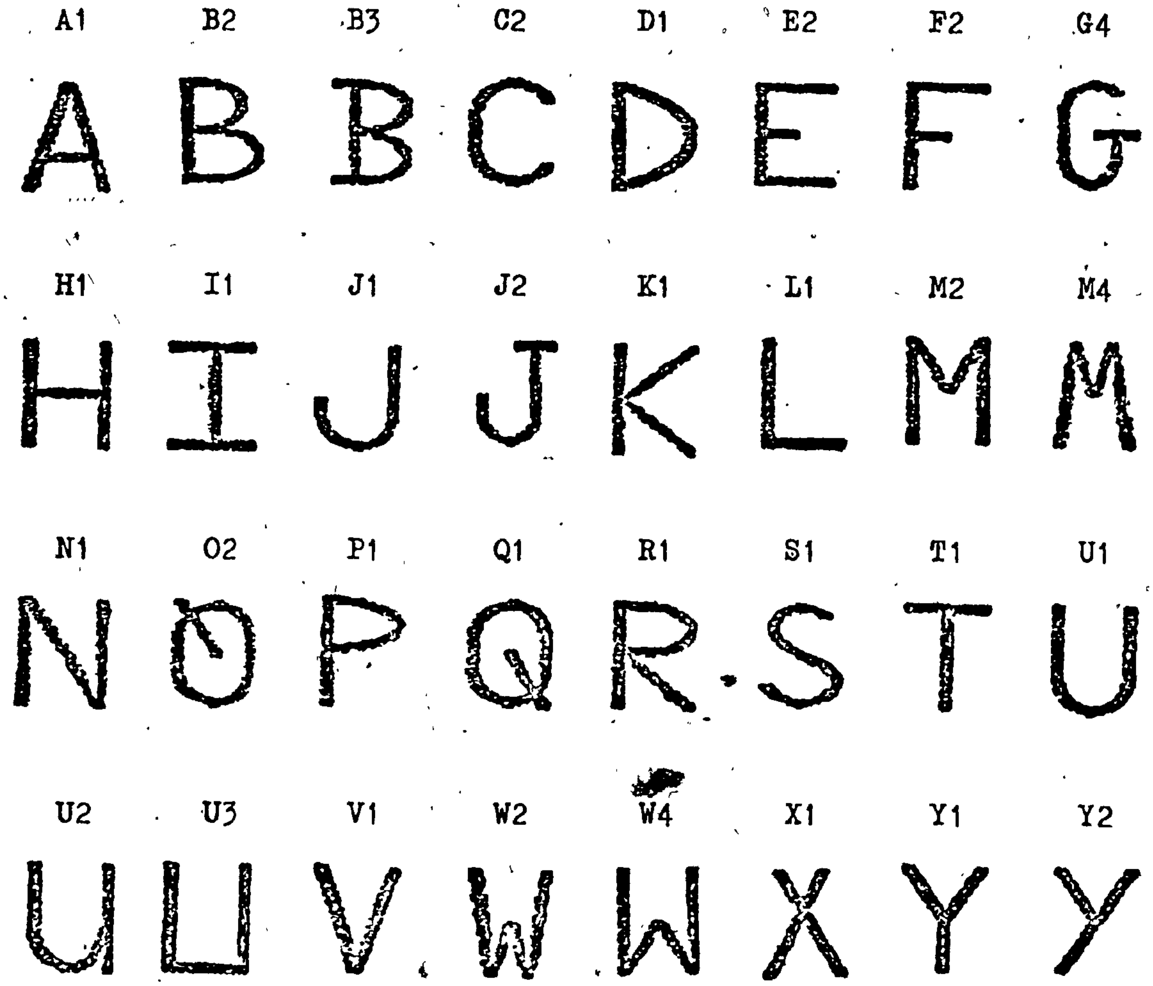
Measurements	Model E1	Model E2	Model E3
AVE(AND, ,*)	128.41	121.97	134.70
AVE(XOR, ,*)	338.84	333.73	350.27
AVE(LA,)	44.88	43.99	45.15
AVE(CA,)	24.60	23.81	24.86
INF( )	3076.36	2904.39	3317.42
ENT( )	4.699	4.436	4.902
AVE(MID1,,*)	5.55	6.20	5.26
AVE(MID2,,*)	6.20	6.90	5.94

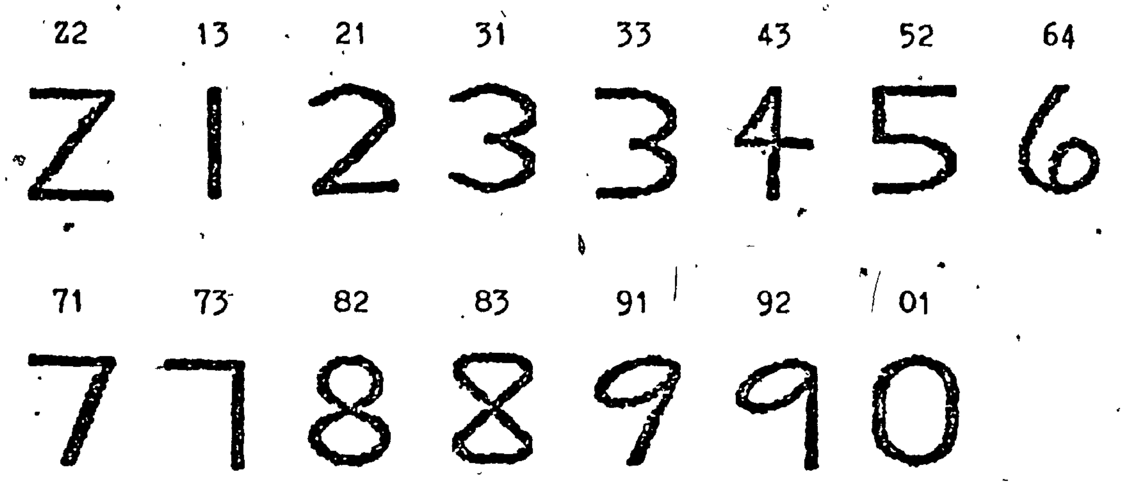
Table 11: Measurements of the three "E" models in the third iteration.

After the fourth iteration, 7 more alphanumeric models were eliminated from the remaining set of 67 models by the 87.5% elimination rule. They are designated as models A2, E1, F1, I2, K2, K3, and 22 in Figure 3. From the remaining two models 01 and 02 of the numeral "0" we decided to designate the numeric model 02 as character "0" and the model 01 as the number "zero". The choice was decided upon the desire to use the standard symbol 01 for the number "zero".

After the successive fifth, sixth and seventh iterations, additional seven (G5, M3, W1, 11, 32, 51, 72), five (D3, S3, Z1, 42, 62) and one (61) alphanumeric models were eliminated by the 87.5% elimination rule. Thus a total of 26 models were eliminated in this stage. The remaining 47 models are shown in Figure 4.

Figure 4: The remaining 47 alphanumeric models after the 87.5% elimination rule.





2.3.3 THE 75% ELIMINATION RULE

The 75% elimination rule was used in the eighth and nineth iterations. In each of these two iterations, the eight quantitative measurements of different models of the same symbol were compared among themselves and those models rated as undesirable by a majority of 6 against 2 were eliminated from the set.

A total of 4 alphanumeric models (three models B2, U3, 83,after the eighth and one model 92 after the nineth iterations) were eliminated by the 75% rule. The numeric model 92, as shown in Table 12, could not be eliminated by the 75% rule in the eighth iteration (a majority of 5 against 3 between models 91 and 92), but it was eliminated by the 75% rule in the nineth iteration. This fact indicates that the measurements of a model are affected by the models eliminated. Therefore the iterative technique is effective.

Measurements	Model 91	Model 92
AVE(AND, ,*)	61.16	77.27
AVE(XOR, ,*)	394.49	370.27
AVE(LA,)	26.08	31.16
AVE(CA,)	8.99	12.73
INF( )	2521.85	2517.17
ENT( )	2.445	2.961
AVE(MID1,,*)	7.06	8.37
AVE(MID2,,*)	6.97	8.91

12 (a)

Measurements	Model 91	Model 92
AVE(AND, ,*)	60.14	75.07
AVE(XOR, ,*)	391.12	369.26
AVE(LA,)	26.14	30.81
AVE(CR,)	9.13	12.63
INF( )	2515.87	2518.88
ENT( )	2.471	2.956
AVE(MID1,,*)	7.09	8.46
AVE(MID2,,*)	7.00	8.94

12 (b)

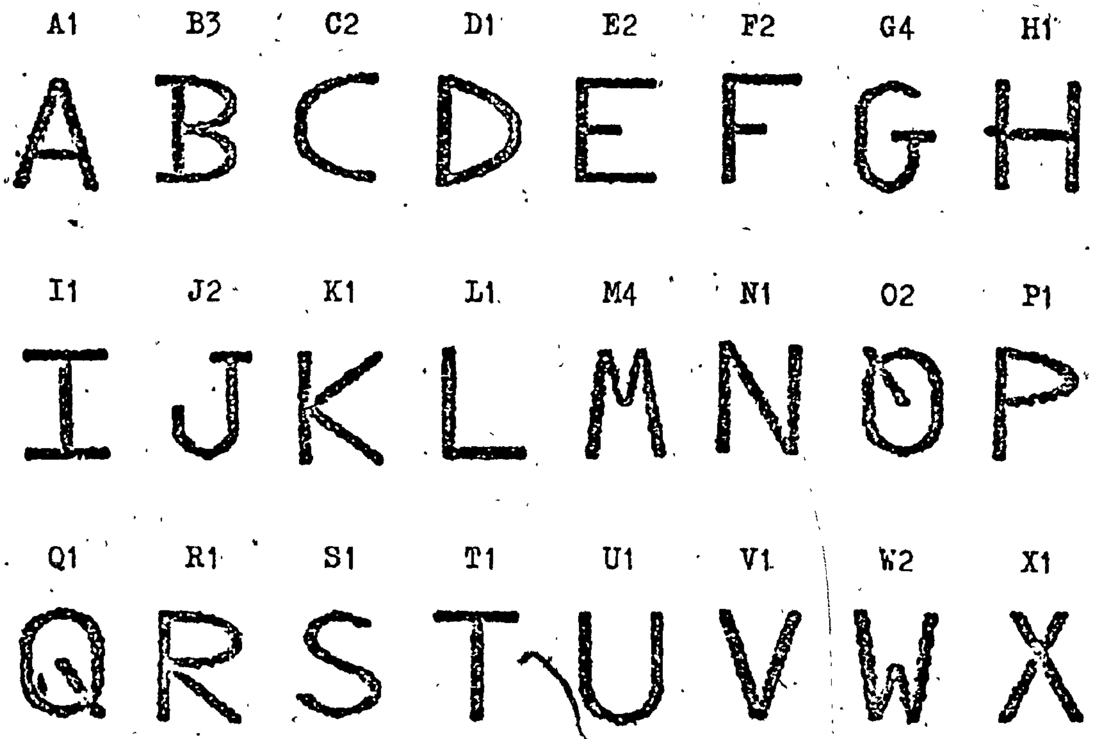
Table 12 (a): Measurements of the two "9" models after the eighth iteration.
(b): Measurements of the two "9" models after the ninth iteration.

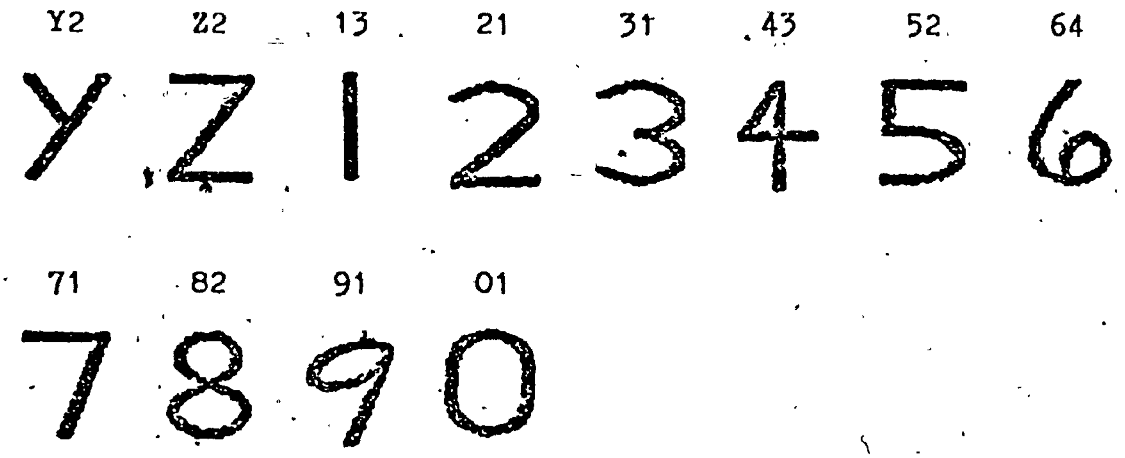
2.3.4 THE 62.5% ELIMINATION RULE & RESULTS

The 62.5% elimination rule was used in the tenth to twelfth iterations. In each of those iterations, the eight quantitative measurements of different models of the same symbol were compared among themselves and those models rated as undesirable by a majority of 5 against 3 were eliminated from the set.

A total of 7 alphanumeric models (4 models J1, W4, 33, 73 after the tenth, one model Y1 after the eleventh, and 2 models M2, U2 after the twelfth iterations) were eliminated from the remaining set of 43 models during those stages. The remaining 36 models, as shown in Figure 5, are considered to be the most distinct handprint characters. Although these 7 alphanumeric models could be eliminated altogether by the 62.5% rule after the tenth iteration, we decided to do it by two more iterations so that the desirability of the models could be examined more closely.

Figure 5: The 36 most distinct handprint models





For instance, as shown in Table 13, the desirability of models Y2 against Y1 was 5 to 3 in the tenth iteration, it increased to 7 against 1 in the eleventh iteration. This indicates that the desirability of model Y2 has improved after the elimination of models J1, W4, 33 and 73 in the tenth iteration.

Measurements	Model Y1	Model Y2
AVE(AND, ,*)	43.49	43.54
AVE(XOR, ,*)	363.78	369.68
AVE(LA,)	22.39	21.58
AVE(CA,)	8.37	7.89
INF( )	1868.85	1939.65
ENT( )	1.808	1.822
AVE(MID1,,*)	8.99	9.09
AVE(MID2,,*)	8.23	8.68

13 (a)

Measurements	Model Y1	Model Y2
AVE(AND, ,*)	45.43	44.00
AVE(XOR, ,*)	362.22	371.08
AVE(LA,)	23.47	22.07
AVE(CA,)	9.07	8.31
INF( )	1852.40	1935.13
ENT( )	1.878	1.845
AVE(MID1,,*)	8.63	8.70
AVE(MID2,,*)	7.99	8.35

13 (b)

Table 13 (a): Measurements of the two "Y" models in the tenth iteration.
(b): Measurements of the two "Y" models in the eleventh iteration.

2.4 REMARKS

In sections 2.3.3 and 2.3.4 we have seen cases that the desirability of a model increases as the number of the models in the set decreases. This, naturally, leads us to wonder whether the iterative process is necessary. We might ask whether the same 36 most distinct characters could be obtained, in the first iteration, by the same majority rules. Unfortunately, as shown in Table 14, this is not the case. Comparing the two models 82 and 83 of the number "8", the desirability of model 82 against model 83 was 3 to 5 in the first iteration, 6 to 2 in the third iteration, 4 to 4 in the seventh and 6 to 2 in the eight iterations. Thus if the majority rule were used in the first iteration, then the model 83 would be chosen instead of model 82. This indicates that the iterative process is necessary and it should be used as many times as possible to obtain better results.

At first sight, one might think that AVE(AND, ,*) and AVE(LA,) are linearly dependent. However, AVE(LA,) varies with the number of points occupied by each model, thus it is not linearly dependent on AVE(AND, ,*) as demonstrated in Table 14 (a). Similarly we can also say about other measurements.

Measurements	First Iteration		Third Iteration	
	Model 82	Model 83	Model 82	Model 83
AVE(AND, ,*)	103.05	111.86	96.35	109.21
AVE(XOR, ,*)	387.44	421.80	393.42	419.70
AVE(LA,)	35.67	35.20	33.56	34.85
AVE(CA,)	15.48	14.02	13.81	14.02
INF( )	3145.34	3711.47	3162.37	3697.59
ENT( )	3.663	3.973	3.528	4.008
AVE(MID1,,*)	4.77	4.84	5.07	5.02
AVE(MID2,,*)	5.07	5.26	5.33	5.43

14 (a)

Measurements	Seventh Iteration		Eighth Iteration	
	Model 82	Model 83	Model 82	Model 83
AVE(AND, ,*)	94.43	105.02	92.67	105.20
AVE(XOR, ,*)	392.57	423.39	396.42	423.36
AVE(LA,)	33.92	34.53	33.34	34.60
AVE(CA,)	14.53	13.99	14.03	14.07
INF( )	3149.41	3694.11	3160.76	3693.20
ENT( )	3.580	4.003	3.526	4.008
AVE(MID1,,*)	5.19	5.21	5.25	5.23
AVE(MID2,,*)	5.52	5.65	5.56	5.66

14 (b)

Table 14 (a): Measurements of the two "8" models in the first and third iterations.

(b): Measurements of the two "8" models in the seventh and eighth iterations.

CHAPTER 3 OPTIMUM MATRIX CHARACTER SET FOR COMPUTER OUTPUT SYSTEMS

3.1 INTRODUCTION

In recent years, matrix characters, especially the 5×7 font, have widespread use in computer output systems such as CRT's, thermo-printers and high-speed printers (McLaughlin [7]). This is due to the increasing usage of digital devices, the steady cost reduction in computer output systems, the flexibility of point and graphical presentation and the ready availability of matrix character generators in IC chips. As matrix characters become more widely used, their legibility become imperative. Legible characters do not only provide better reading, they also play a vital role in applications such as aerospace operations where penalties for common misreading errors could be catastrophic.

The "dot" matrix characters used in the computer output systems are quite different in appearance from their conventional "stroke" counterparts in that stroke characters are composed of continuous line segments. It has been recognized for some time that certain characteristics of stroke alphanumerics affect their relative legibility. Much research has been undertaken to ascertain which stroke font is the most legible under certain conditions (Cornog and Rose [1]). However, it has not been satisfactorily demonstrated that the conclusions from stroke font research are directly transferable to dot-matrix fonts.

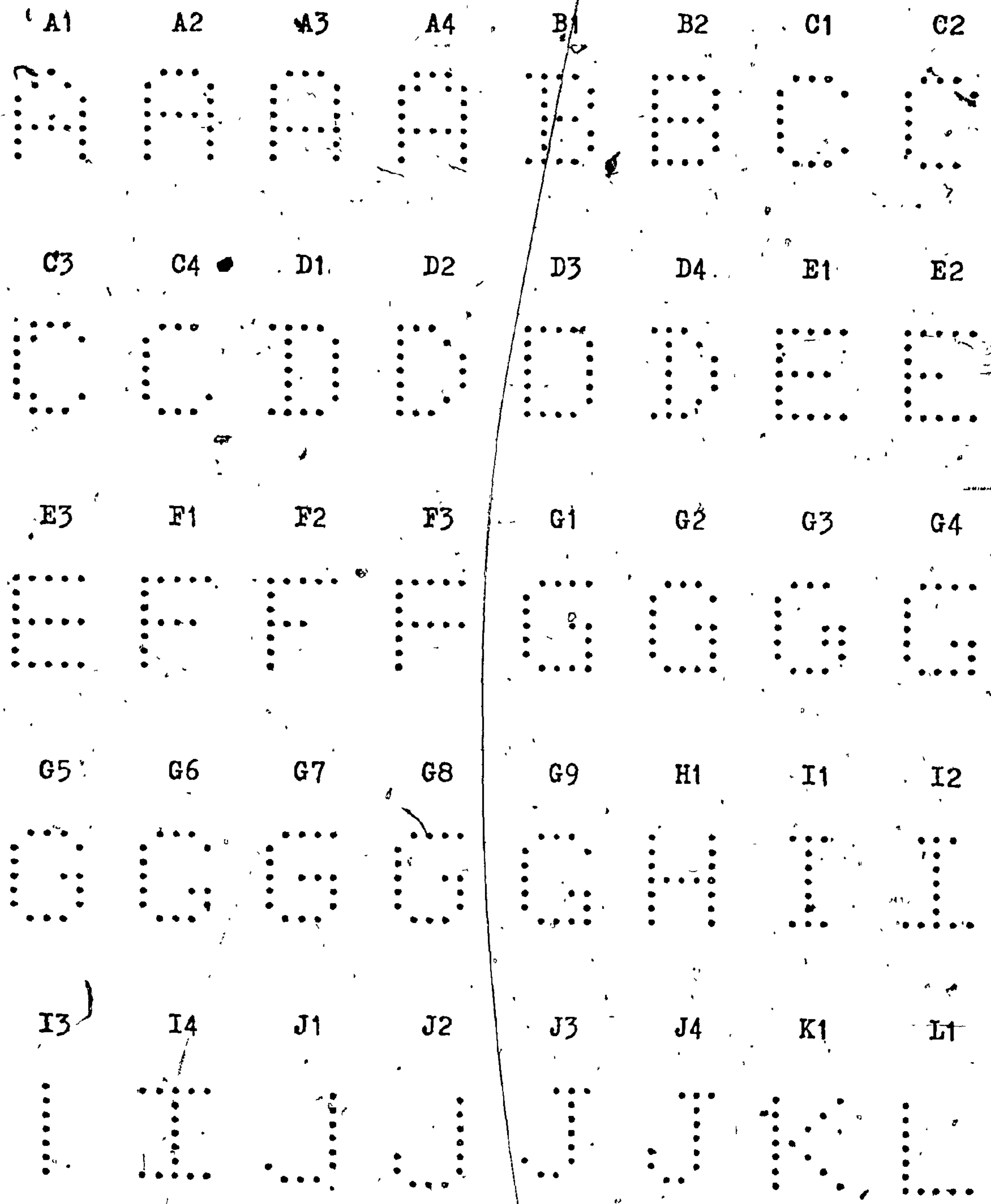
In view of the wide variety of dot-matrix alphanumeric designs used or offered by broadcasting companies and display manufacturers, many experiments with human subjects have been conducted by several researchers ([2, 4, 6, 9, 15]) to investigate the legibility of those designs. Unfortunately, not a great deal of effort has been spent in the development of evaluation methods which could be used to measure the distinguishability (and hence the legibility) of those dot-matrix designs (Suen et al, [14]).

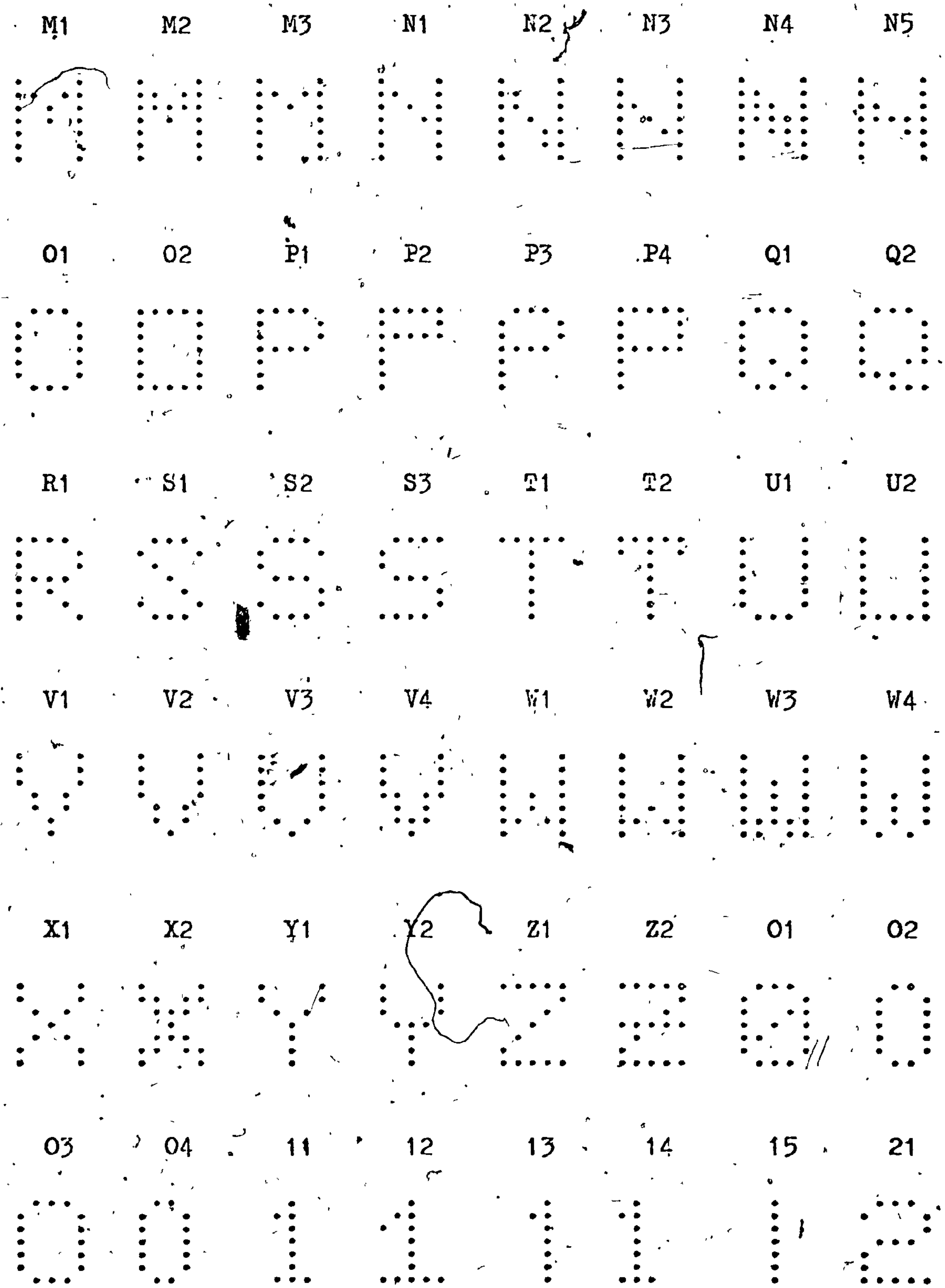
In view of the above, an evaluation technique for a set of 121 5 × 7 alphanumeric models was investigated in this research. In this investigation, the eight quantitative measurements discussed in chapter 1 were used in an iterative process to determine the most distinct set of 5 × 7 matrix characters for computer output systems.

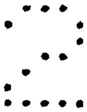
3.2 DATA COLLECTION

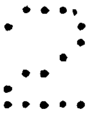
In the determination of the optimum set of 5 × 7 dot-matrix characters for easy distinction, an extensive review of various computer systems was conducted. In this investigation, more than 30 different systems (including those models developed by Huddleston, 1971) were examined. They comprised models generated by leading manufacturers of ROM matrices, matrix printers, CRT's and a variety of computer terminals. Altogether, 121 different models of the 36 alphanumeric matrix symbols were assembled. They comprised 78 models of letters and 43 models of numerals. Thus, on the average, a symbol was represented by 3.36 different shapes. These 121 alphanumeric models are presented in Figure 6.

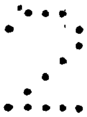
Figure 6: The 121 5 X 7 matrix models.

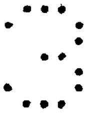


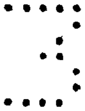

M1
M2
M3
N1
N2
N3
N4
N5
O1
O2
P1
P2
P3
P4
Q1
Q2
R1
S1
S2
S3
T1
T2
U1
U2
V1
V2
V3
V4
W1
W2
W3
W4
X1
X2
Y1
Y2
Z1
Z2
01
02
03
04
11
12
13
14
15
21

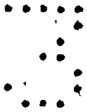
22


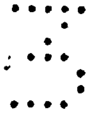
23


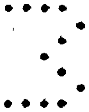
24


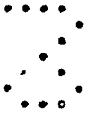
31


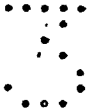
32


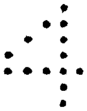
33


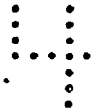
34


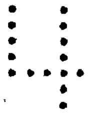
35


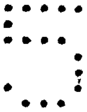
36


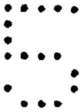
37


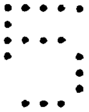
41


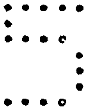
42


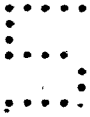
43


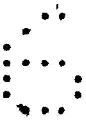
51


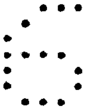
52


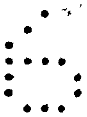
53


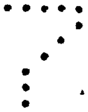
54


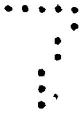
55


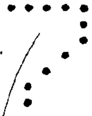
61


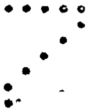
62


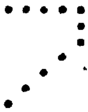
63


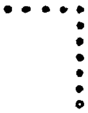
71


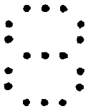
72


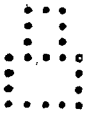
73


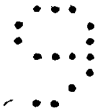
74


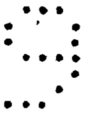
75


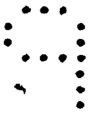
76


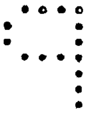
81


82


91


92


93


94


3.3 ELIMINATION PROCEDURES

A total of 121 alphanumeric 5 × 7 matrix models (as shown in Figure 6) were compiled and used in an iterative process. In each stage of the iterative process, eight quantitative measurements were made and compared among different models of the same symbol. Those models rated as "undesirable" were eliminated by 100%, 87.5%, 75%, and 62.5% elimination rules accordingly.

3.3.1 THE 100% ELIMINATION RULE

The 100% elimination rule was used in the first three iterations. In each of these iterations, the eight quantitative measurements of different models of the same symbol were compared among themselves and those models rated as undesirable by all measurements were eliminated from the set.

A total of 36 alphanumeric models were eliminated after the first iteration. They were designated as models A2, A3, A4, C1, D1, D3, G1, G2, G5, G7, G8, J2, J4, M1, P1, P4, S2, S3, T2, W1, Y2, 11, 12, 14, 21, 22, 23, 31, 32, 33, 61, 62, 91, 92, 94, and 03 in Figure 6. For instance, as shown in Table 15, the model C1 was eliminated from the set of 4 character "C" models because it was rated as "undesirable" against model C2 by all eight measurements; but the remaining 3 models of "C" could not be reduced further by the 100% elimination rule.

Measurements	Model C1	Model C2	Model C3	Model C4
AVE(AND, ,*)	8.08	7.53	8.39	7.45
AVE(XOR, ,*)	12.03	12.13	13.40	12.28
AVE(LA,)	69.98	67.80	67.89	66.99
AVE(CA,)	45.25	42.82	41.74	42.51
INF( )	59.61	54.83	72.72	55.02
ENT( )	2.487	2.313	2.627	2.297
AVE(MID1,,*)	1.16	1.25	1.16	1.27
AVE(MID2,,*)	1.30	1.39	1.28	1.39

Table 15: Measurements of the 4 character "C" models in the first iteration.

Five more character models I1, I2, I4, O2, V3 and one numeric model 01 could be eliminated after the first iteration. The model O2 of character "O" and the model 01 of numeral "0" were kept for a few more iterations due to the common difficulty in distinguishing the character "O" from the numeral "0". We decided to keep the character models I1, I2, I4 and V3 for one or two more iterations because their shapes were considered to be quite pleasing and also because we would like to use as many iterations as possible in obtaining the most distinct set of alphanumeric models.

A total of 4 alphanumeric models were eliminated from the remaining set of 85 models after the second iteration. They were designated as models I4, V2, V3 and 43 in Figure 6. Note that the models V2 and 43 were eliminated in the second iteration but could not be eliminated in the first iteration by the same 100% rule. This clearly indicates the effectiveness of the iterative process.

After the third iteration we discovered that no more alphanumeric models (except models I1, I2, 02 and 01) could be eliminated by the 100% rule. Thus it was time to decide the fate of the models I1 and I2. Up to that point, there were three models I1, I2, I3 of the character "I" and two models 13, 15 of the numeral "1" remaining in the set of 81 models. If we eliminated models I1 and I2 by using the 100% rule, then in the end we had to use model I3 for the character "I" and therefore model 13 for the numeral "1" due to the similarity between models I3 and 15. Obviously, they were not desirable because their similar shapes and the difficulty in distinguishing them. Since the "desirability" between the remaining two models 13 and 15 of the numeral "1" was 4 against 4 in each of the first three iterations (as shown in Table 16), thus the obvious solution to the problem was to eliminate the model 13 from the two "1" models and model I3 from the three "I" models. Therefore there were only 79 alphanumeric models remaining at the end of the third iteration.

Measurements	1st iteration		2nd iteration		3rd iteration	
	Model 13 ⋮	Model 15 ⋮	Model 13 ⋮	Model 15 ⋮	Model 13 ⋮	Model 15 ⋮
AVE(AND, ,*)	2.60	2.53	2.60	2.52	2.57	2.52
AVE(XOR, ,*)	18.15	17.28	18.02	17.19	18.15	17.25
AVE(LA,)	27.64	29.89	23.39	24.24	23.12	24.30
AVE(CA,)	9.30	12.13	8.33	9.93	8.09	9.81
INF( )	46.70	40.02	45.82	39.80	46.18	40.01
ENT( )	0.906	0.868	0.934	0.864	0.920	0.859
AVE(MID1,,*)	2.12	2.16	2.14	2.16	2.15	2.16
AVE(MID2,,*)	1.89	1.84	1.90	1.84	1.90	1.83

Table 16: Measurements of the two "1" models in the first three iterations.

3.3.2 THE 87.5% ELIMINATION RULE

The 87.5% elimination rule was used in the fourth to seventh iterations. In each of these iterations, the eight quantitative measurements of different models of the same symbol were compared among themselves and those models rated as undesirable by a majority of 7 against 1 were eliminated from the set. A total of 14 alphanumeric models were eliminated from the remaining set of 79 models after the fourth iteration. They were designated as models C2, E3, F3, G9, N2, N4, N5, W3, 34, 51, 52, 55, 73 and O1 in Figure 6. For instance, as shown in Table 17, the model C2 was eliminated from the set of character "C" models because it was rated as undesirable against the model C4 by a majority of 7 to 1.

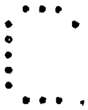
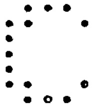
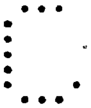
Measurements	Model C2 	Model C3 	Model C4 
AVE(AND, ,*)	7.42	8.29	7.37
AVE(XOR, ,*)	12.54	13.80	12.64
AVE(LA,)	55.23	55.41	54.72
AVE(CA,)	34.61	33.70	34.50
INF( )	55.22	73.07	55.34
ENT( )	2.277	2.592	2.266
AVE(MID1,,*)	1.26	1.18	1.28
AVE(MID2,,*)	1.39	1.28	1.39

Table 17: Measurements of the remaining three "C" models in the fourth iteration.

After the successive fifth, sixth and seventh iterations, seven (B2, D2, E1, F1, M2, V4, O4), six (C3, I2, U2, W4, 36, 53) and four (O2, Z2, 42, 72) alphanumeric models were eliminated successively by the 87.5% rule. Thus a total of 17 models were eliminated during these stages. The remaining 48 models could not be reduced further by the 87.5% rule.

3.3.3 THE 75% ELIMINATION RULE

The 75% elimination rule was used after the eighth iteration. In the eighth iteration, the eight quantitative measurements of different models of the same symbol were compared among themselves and those models rated as undesirable by a majority of 6 against 2 were eliminated from the set. Only two alphanumeric models N3 and 37 were eliminated. For instance, as shown in Table 18, the model N3 was eliminated from the remaining 2 models of character "N" because it was rated as undesirable against the model N1 by a majority of 6 to 2. The remaining 46 alphanumeric models could not be reduced further by the 75% rule. They are presented in Figure 7.

Measurements	Model N1	Model N3
AVE(AND, ,*)	7.78	8.04
AVE(XOR, ,*)	15.74	15.22
AVE(LA,)	50.49	52.16
AVE(CA,)	29.79	31.15
INF( )	84.86	83.61
ENT( )	2.731	2.802
AVE(MID1,,*)	1.32	1.31
AVE(MID2,,*)	1.36	1.37

Table 18: Measurements of the remaining two "N" models in the eighth iteration.

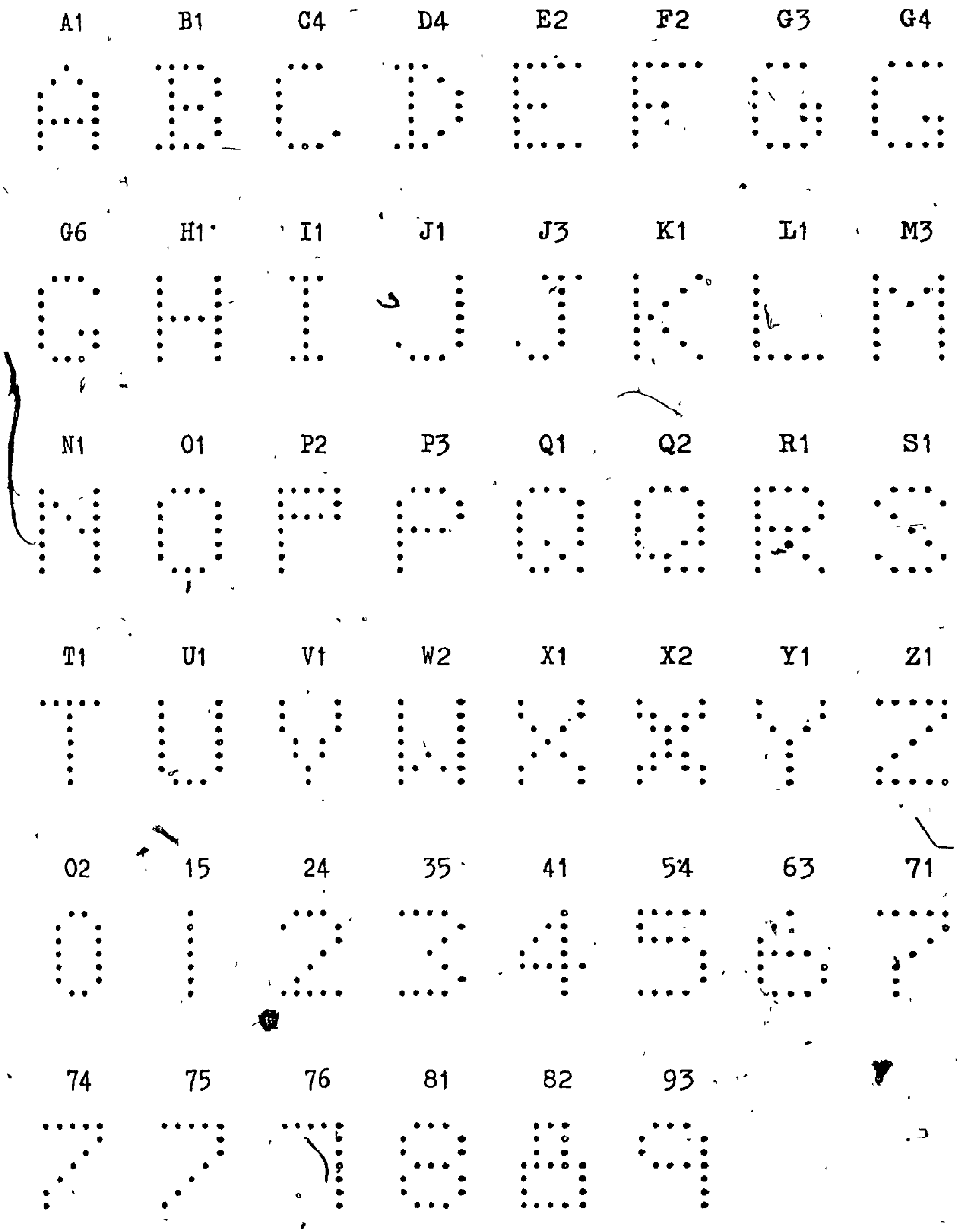


Figure 7: The remaining 46 alphanumeric models after the 75% elimination rule.

3.3.4 THE 62.5% ELIMINATION RULE & RESULTS

The 62.5% elimination rule was used in the ninth to the twelfth iterations. In each of these iterations, the eight quantitative measurements of different models of the same symbol were compared among themselves and those models rated as undesirable by a majority of 5 against 3 were eliminated from the set.

After the successive ninth, tenth, eleventh and twelfth iterations, four (J1, P2, X2, 76), three (G6, 74, 81), one (75) and two (G4, Q1) alphanumeric models were eliminated successfully by the 62.5% rule. Thus a total of 10 models were eliminated during those stages. The remaining 36 alphanumeric models , as shown in Figure 8, are considered to be the most distinct set of models. Although those 10 models could be eliminated altogether by the 62.5% rule after the ninth iteration, we decided to do it by few more iterations so that the desirability of the models could be examined more closely. For instance, as shown in Table 19, the desirability of models 71 against 75 was 5 to 3 in the ninth iteration, it decreased to the rate of 4 against 4 in the tenth iteration and then increased to the rate of 5 against 3 in the eleventh iteration.

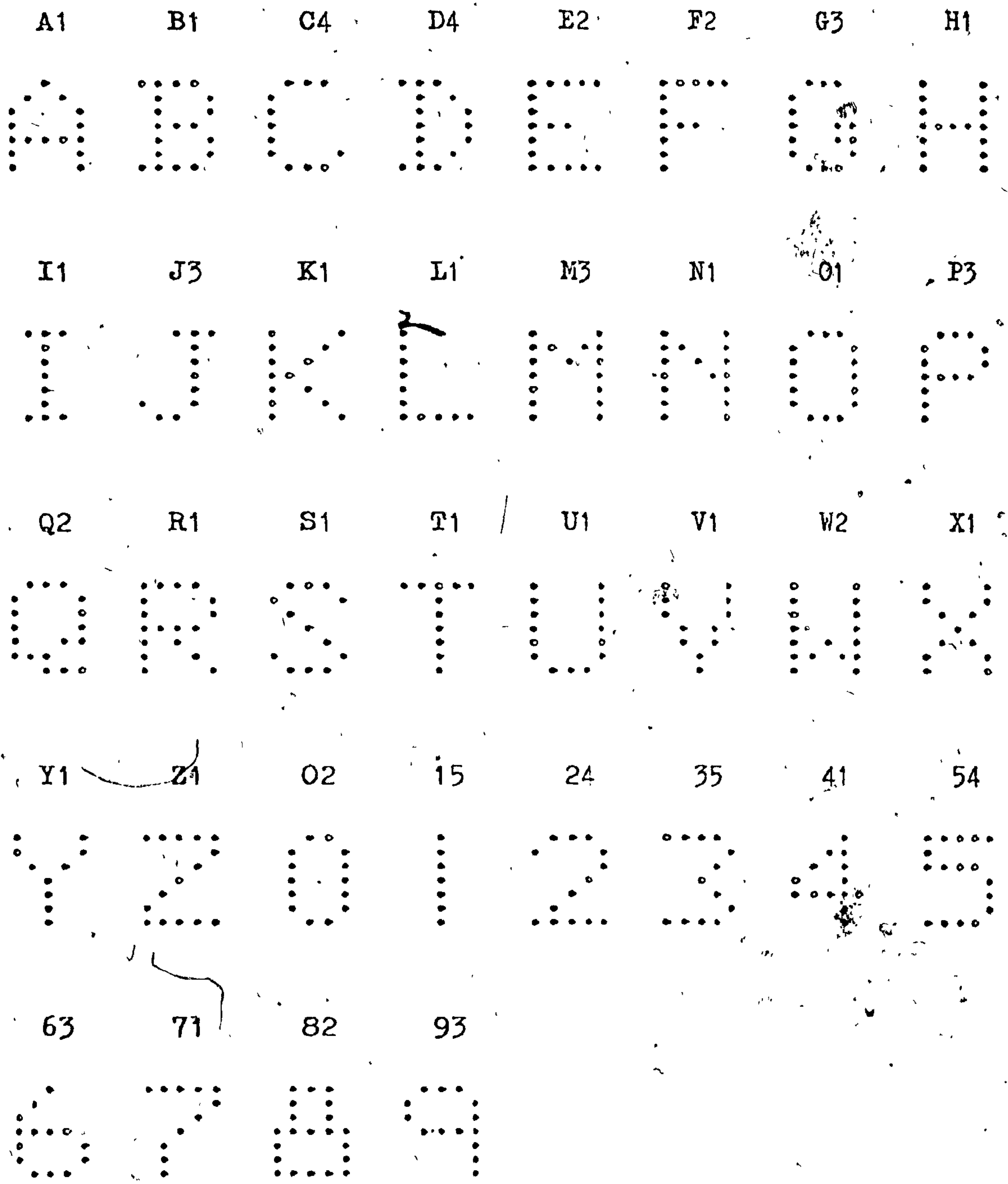


Figure 8: The most distinct set of 36
5 x 7 alphanumeric models.

Measurements	9th iteration		10th iteration		11th iteration	
	Model 71	Model 75	Model 71	Model 75	Model 71	Model 75
AVE(AND, ,*)	5.02	5.19	4.97	5.15	4.95	5.16
AVE(XOR, ,*)	15.64	15.31	15.72	15.36	15.70	15.27
AVE(LA,)	42.17	42.87	41.86	42.39	40.90	42.14
AVE(CA,)	21.14	21.27	21.03	20.87	20.05	20.69
INF( )	55.14	54.51	55.41	54.70	55.65	54.60
ENT( )	1.771	1.809	1.755	1.793	1.732	1.792
AVE(MID1,,*)	1.60	1.63	1.60	1.63	1.60	1.63
AVE(MID2,,*)	1.56	1.59	1.55	1.58	1.55	1.59

Table. 19: Measurements of the two "7" models.

3.4 DISCUSSION

The last 36 alphanumeric models resulting from the iterative process in this study are considered to be the most distinct models among the original set of 121 models. Upon inspection the shapes of the last 36 models, we realize that model O1 of character "O" and model O2 of numeral "O" are not very distinctive. This is due to the fact that the former was selected from a set of two character "O" models O1 and O2 while the latter was selected among a set of four numeral "O" models O1, O2, O3 and O4 (as shown in Figure 9). Since the iterative process was applied to models of the same symbol, we might be able to obtain a better result if we had treated the models of character "O" and numeral "O" as one symbol at the beginning of the iterative process.

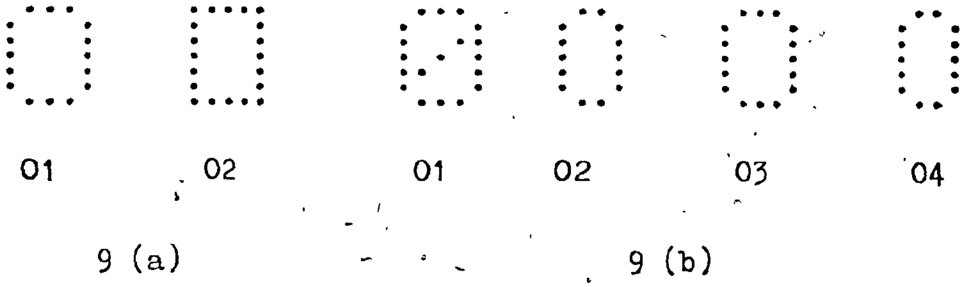


Figure 9 (a): The two models of character "O" ,
(b): The four models of numeral "0" .

Recently, an experiment using human subjects was conducted by Suen and Strobel [15] to determine the most legible set of matrix characters for computer-man communications. A set of 122 stimuli was used in the experiment. Their set of stimuli contains one less model of character "I" and one more model each of character "Y" and numeral "9" than the alphanumeric set used in our study. The results show that there are some discrepancies between the optimal set selected in their experiment and the optimum set selected by our investigation. The discrepancies can be attributed to the fact that humans are not well adapted to making precise measurements of a displayed character. Thus it is hardly surprising that the stimuli used in their experiment are identified by subjects based on features which are taught at school and which can be recognized without making precise measurements. The discrepancies could also be attributed to the procedure used in their experiment where the stimuli were presented in random order. Thus the subjects could not compare the desirability among models of the same symbol. We believe that a better assessment could be obtained by investigating models of the two optimum sets in another experiment.

BIBLIOGRAPHY

1. Cornog, D. Y. and Rose, F. C., "Legibility of Alphanumeric Characters and Other Symbols, II", A Reference Handbook, Washington, D. C., U. S. Government Printing Office, 1967.

2. Hamming, R. W., "Error Detecting and Error Correcting Codes", Bell System Technical Journal, 29, pp. 147-150, 1950.

3. Harmon, L. D., "Automatic Recognition of Print and Script", Proc. IEEE , Vol. 60, pp. 1165-1176, 1972.

4. Huddleston, H. F., "An Evaluation of Alphanumerics For A 5 × 7 Matrix Display", Proceedings, Conference on Displays, IEE Pub. No. 80, pp. 145-147, Sept. 1971.

5. Ingels, F. M., "Information & Coding Theory", Intext Educational Publishers, College Division of Intext, International Textbook Company, 1971.

6. Maddox, M. E., Burnette, J. T., and Gutmann, J. C., "Font Comparisons for 5 × 7 Dot Matrix Characters", Human Factors, 19(1), pp. 89-93, 1977.

7. McLaughlin, R. A., "Alphanumeric Display Terminal Survey", Datamation, pp. 71-92, Nov. 1973.

8. Neisser, U., and Weene, P., "A Note on Human Recognition of Handprinted Characters", Information and Control, Vol. 3, pp. 191-196, 1960.

9. Shurtleff, D. A., "Studies of Display Symbol Legibility: XXII. The Relative Legibility of Four Symbol Sets Made With a Five by Seven Dot Matrix", MITRE Technical Report, Dec. 1969.

10. Suen, C. Y., "Factors Affecting the Recognition of Handprinted Characters", Proc. International Conference on Cybernetics and Society, pp. 174-175, Nov. 1973.

11. Suen, C. Y., "Optical Character Recognition - The State of the Art", Canadian Datasystems, Vol. 6, pp. 40-44, May 1974.

12. Suen, C. Y., "Human Factors in Character Recognition", Proc. International Conference on Systems, Man and Cybernetics, pp. 253-258, Oct. 1974.

13. Suen, C. Y., Shiau, C., Shinghal, R., Kwan, C. C., "Reliable Recognition of Handprint Data", Proc. Joint Workshop on Pattern Recognition and Artificial Intelligence, pp. 98-102, June 1976.

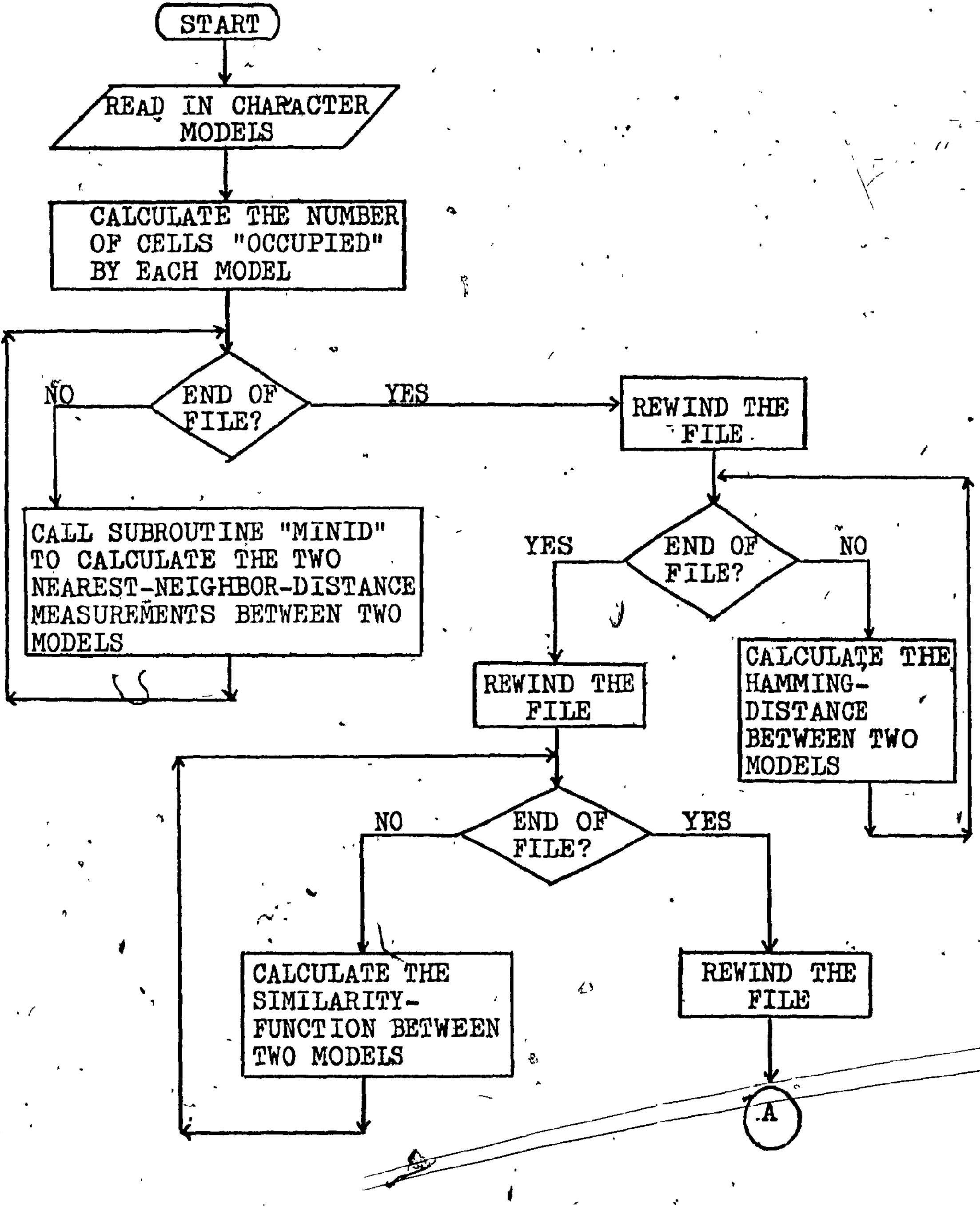
14. Suen, C. Y. and Shiau, C., "Optimum Matrix Character Set for Computer Output Systems", SID International Symposium Digest of Technical Papers, pp. 52-53, April 1977

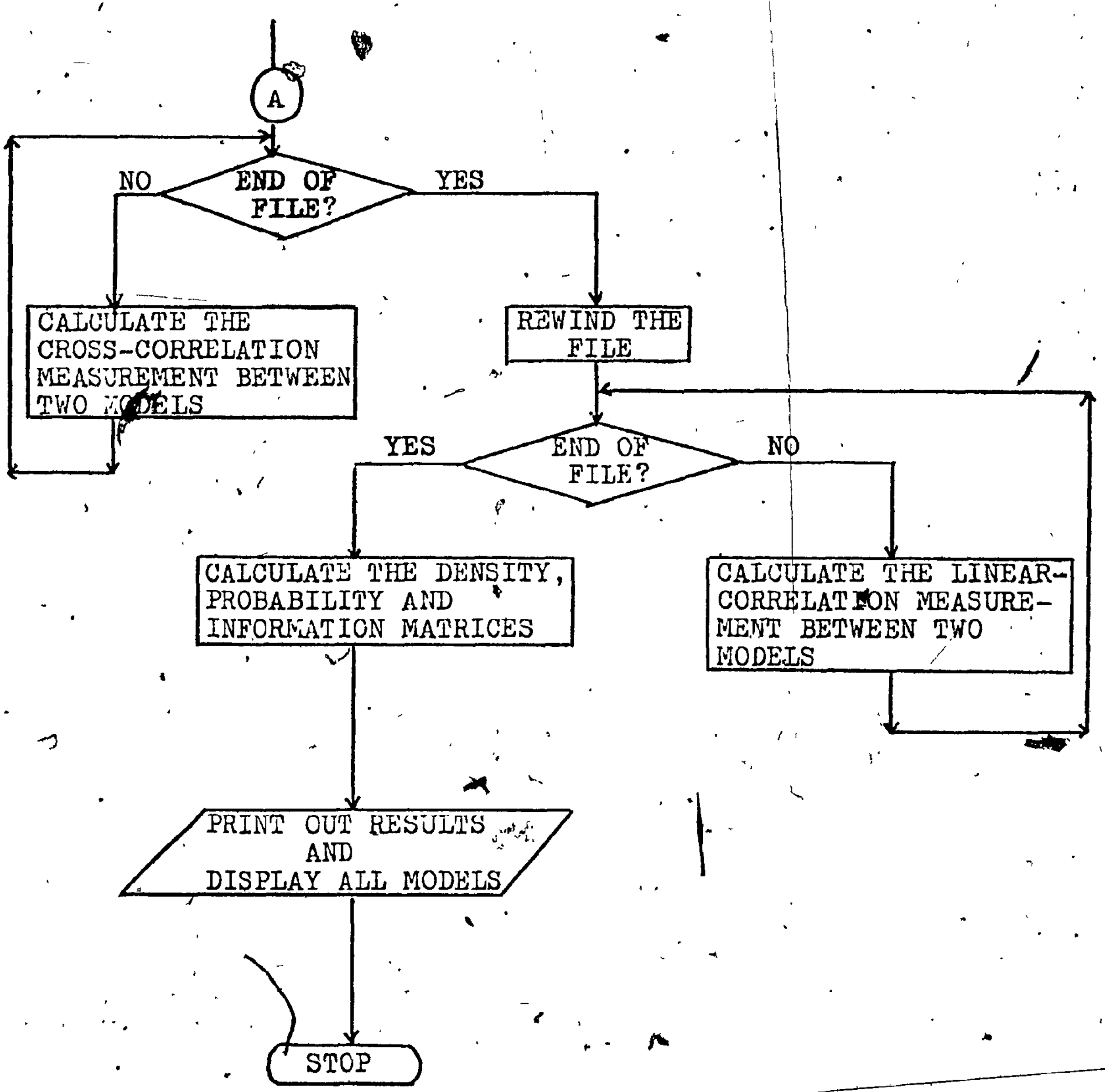
15. Suen, C. Y. and Strobel, M. G., "Selection of the Most Legible Matrix Characters in Computer-Man Communications", Proc. 5th Man-Computer Communications Conference, pp. 45-52, May 1977.

16. Trickett, T., "The Design of A Standard Type Font for Optical Character Recognition", Honeywell Computer Journal, pp. 3-11, Winter 1969.

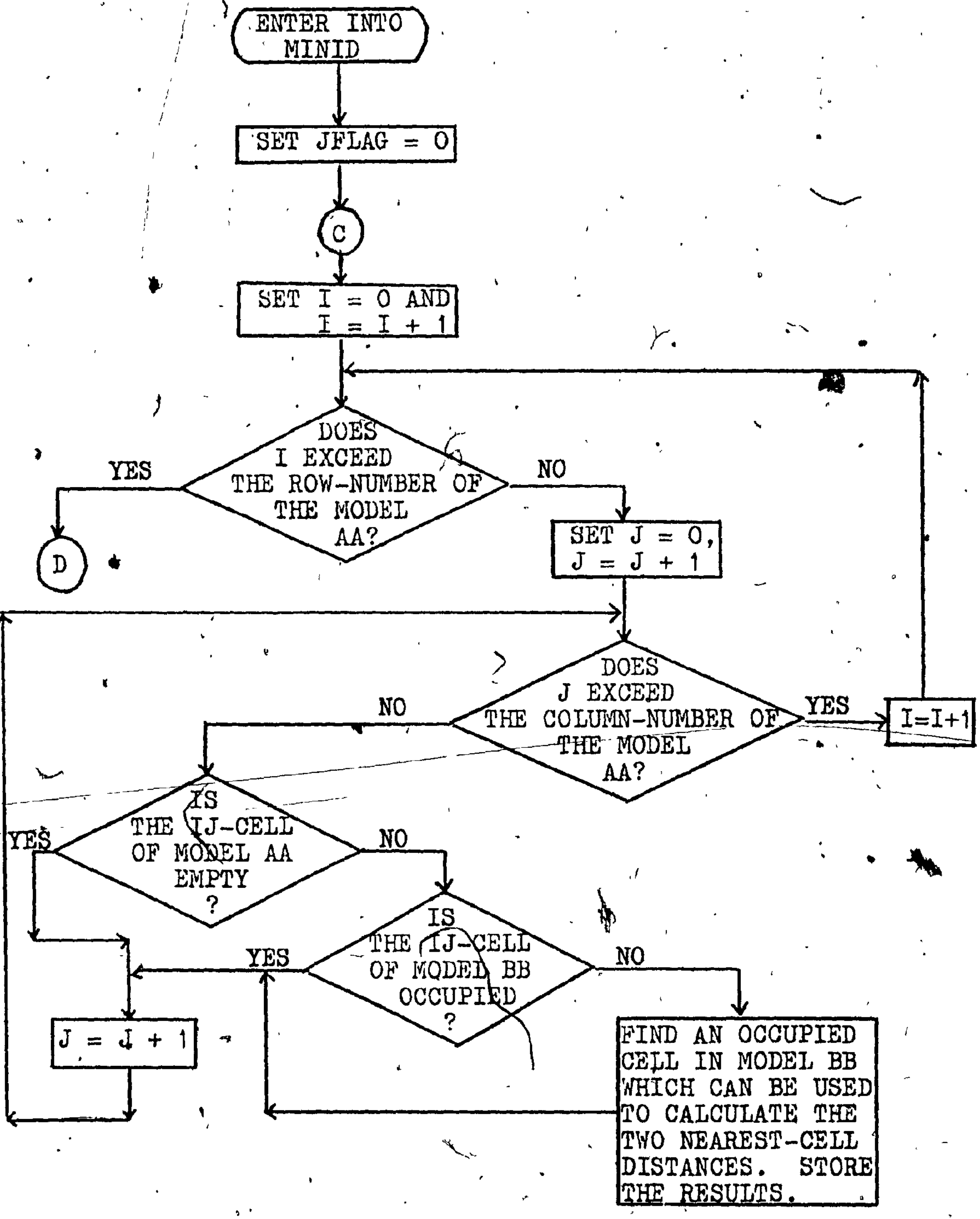
17. Shannon, C. E., "Prediction and Entropy of Printed English", Bell System Tech. J., Vol. 30, pp. 50-64, 1951.

APPENDIX: PROGRAM FLOWCHART

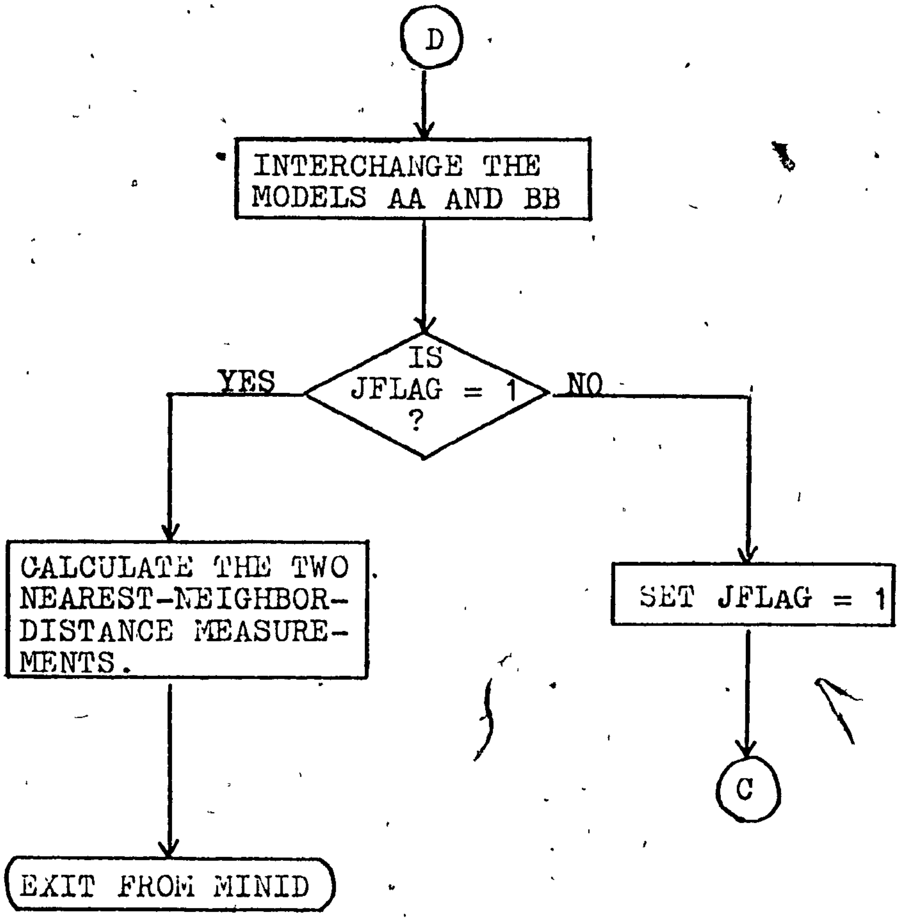

START
READ IN CHARACTER MODELS
CALCULATE THE NUMBER OF CELLS "OCCUPIED" BY EACH MODEL
END OF FILE?
NO
YES
CALL SUBROUTINE "MINID" TO CALCULATE THE TWO NEAREST-NEIGHBOR-DISTANCE MEASUREMENTS BETWEEN TWO MODELS
REWIND THE FILE
END OF FILE?
YES
NO
REWIND THE FILE
CALCULATE THE HAMMING-DISTANCE BETWEEN TWO MODELS
END OF FILE?
NO
YES
CALCULATE THE SIMILARITY-FUNCTION BETWEEN TWO MODELS
REWIND THE FILE
A


A
END OF FILE?
NO
YES
CALCULATE THE CROSS-CORRELATION MEASUREMENT BETWEEN TWO MODELS
REWIND THE FILE
END OF FILE?
YES
NO
CALCULATE THE DENSITY, PROBABILITY AND INFORMATION MATRICES
CALCULATE THE LINEAR-CORRELATION MEASUREMENT BETWEEN TWO MODELS
PRINT OUT RESULTS AND DISPLAY ALL MODELS
STOP

SUBROUTINE MINID


ENTER INTO MINID
SET JFLAG = 0
C
SET I = 0 AND I = I + 1
DOES I EXCEED THE ROW-NUMBER OF THE MODEL AA?
YES
NO
D
SET J = 0, J = J + 1
DOES J EXCEED THE COLUMN-NUMBER OF THE MODEL AA?
NO
YES
I=I+1
IS THE IJ-CELL OF MODEL AA EMPTY ?
YES
NO
IS THE IJ-CELL OF MODEL BB OCCUPIED ?
YES
NO
J = J + 1
FIND AN OCCUPIED CELL IN MODEL BB WHICH CAN BE USED TO CALCULATE THE TWO NEAREST-CELL DISTANCES. STORE THE RESULTS.

SUBROUTINE MINID


D
INTERCHANGE THE MODELS AA AND BB
IS JFLAG = 1 ?
YES
NO
CALCULATE THE TWO NEAREST-NEIGHBOR-DISTANCE MEASUREMENTS.
SET JFLAG = 1
C
EXIT FROM MINID