A STUDY OF CHARACTER RECOGNITION USING GEOMETRIC MOMENTS UNDER CONDITIONS OF SIMPLE AND NON-SIMPLE LOSS

N. D. Tucker

A Thesis Submitted for the Degree of PhD
at the
University of St Andrews

1974

A STUDY OF CHARACTER RECOGNITION USING GEOMETRIC MOMENTS UNDER CONDITIONS OF SIMPLE AND NON-SIMPLE LOSS.

A thesis presented by

N. D. Tucker B.Sc.,

to the

University of St. Andrews

in application for the Degree
of Doctor of Philosophy.

ABSTRACT.

The theory of Loss Functions is a fundamental part of Statistical Decision Theory and of Pattern Recognition. However it is a subject which few have studied in detail. This thesis is an attempt to develop a simple character recognition process in which losses may be implemented when and where necessary.

After a brief account of the history of Loss Functions and an introduction to elementary Decision Theory, some examples have been constructed to demonstrate how various decision boundaries approximate to the optimal boundary and what increase in loss would be associated with these sub-optimal boundaries. The results show that the Euclidean and Hamming distance discriminants can be sufficiently close approximations that the decision process may be legitimately simplified by the use of these linear boundaries.

Geometric moments were adopted for the computer simulation of the recognition process because each moment is closely related to the symmetry and structure of a character, unlike many other features. The theory of Moments is discussed, in particular their geometrical properties. A brief description of the programs used in the simulation follows.

Two different data sets were investigated, the first being hand-drawn capitals and the second machine-scanned lower case type script. This latter set was in the form of a message, which presented interesting programming problems in itself. The results from the application of different discriminants to these sets under conditions of simple loss are analysed and the recognition efficiencies are found to vary between about 30% and 99% depending on the number of moments being used and the type of discriminant.

Next certain theoretical problems are studied. The relations between the rejection rate, the error rate and the rejection threshold are discussed both theoretically and practically. Also an attempt is made to predict theoretically the variation of efficiency with the number of moments used in the discrimination. This hypothesis is then tested on the data already calculated and shown to be true within reasonable limits. A discussion of moment ordering by defining their resolving powers is undertaken and it seems likely that the moments normally used unordered are among the most satisfactory.

Finally, some time is devoted towards methods of improving recognition efficiency. Information content is discussed along with the possibilities inherent in the use of digraph and trigraph probabilities. A breakdown of the errors in the recognition system adopted here is presented along with suggestions to improve the technique. The execution time of the different decision mechanisms is then inspected and a refined 2-stage method is produced. Lastly the various methods by which a decision mechanism might be improved are united under a common loss matrix, formed by a product of matrices each of which represents a particular facet of the recognition problem.

ACKNOWLEDGMENTS.

I would like to thank Mr. D.C.Stark for assisting with some of the more abstruse theoretical problems and Mr. N.R. Paterson for correcting some of the programming errors and for writing the OS Subroutine to convert the tape data into Fortran format. I am also indebted to IBM for supplying the magnetic tape data.

Finally, I would like to thank Deirdre, my wife, who has patiently laboured through the typing of this thesis, and Mr. F.C.Evans for supervising this work and giving a guiding hand when most needed.

The work was supported by the Science Research Council to whom I would also like to express my gratitude.

CAREER.

The author began his University career in 1966 when he won the Russell Scholarship to St. Andrews University. After studying Physics and Mathematics, he graduated with a First Class Honours Degree in Physics in 1970.

In October of the same year, he matriculated as a Research Student in the Physics Department of St. Andrews University and then pursued his research into Pattern Recognition leading to this thesis which is presented for the Degree of Doctor of Philosophy.

From 1970 to 1973, the author was supported by a Science Research Council Grant.

Declaration:

I hereby certify that I am the sole composer of this thesis, that it is a record of my own work and that it has not been previously presented for a higher degree.

This research was carried out in the Physical Science Laboratory of St. Salvator's College, in the University of St. Andrews, under the supervision of Mr. F.C.Evans.

Certificate:

I certify that N.D.Tucker B.Sc., has spent nine terms at research work in the Physical Science Laboratory of St. Salvator's College in the University of St. Andrews under my direction, that he has fulfilled the conditions of Ordinance No.16 (St. Andrews) and that he is qualified to submit the accompanying thesis in application for the Degree of Doctor of Philosophy,

Research Supervisor.

CONTENTS.

INTRODUCTION		1
CHAPTER 1	THE DEVELOPMENT OF THE LOSS FUNCTION CONCEPT.	
1.1	The Historical Development of Loss Functions.	3
1.2	The Development of Definition.	6
1.3	The Definition of Expected Loss, Error and Risk.	8
1.4	Additive and Non-additive Losses.	9
1.5	The Bayes Solution.	11
1.6	The Relation between Rejection and Loss.	12
CHAPTER 2	THE OPTIMAL DECISION BOUNDARY.	
2.1	The Selection of the Decision Maker.	14
2.2	An Elementary Approach to the Problem of Simplification.	19
2.3	A Comparison of the Different Systems of Discrimination by Example.	21
2.4	The Degree of Approximation between the Bayes Rule and the Equiprobability Boundary.	22
2.5	The Loss Associated with Different Boundaries.	25
2.6	The Geometric Form of the Optimal Boundary.	27
2.7	The Piece-wise Linear Boundary.	28
CHAPTER 3	GEOMETRICAL MOMENTS IN DECISION MAKING.	
3.1	Introduction.	29
3.2	The Theory of Geometric Moments.	30

3.3 The Geometrical Properties of Moments. 32

3.4 The Decision Mechanism. 36

3.5 The Calculation of the Moments. 38

3.6 Program Description. 41

3.7 The Efficiency of the Decision Mechanism. 44

CHAPTER 4 THE APPLICATION OF THE MOMENTS DATA TO CERTAIN THEORETICAL PROBLEMS

4.1 Introduction. 50

4.2 The Relation between the Rejection Rate and the Threshold Function. 51

4.3 The Relation between the Error Rate and the Rejection Rate. 52

4.4 The Variation of Efficiency with Number of Moments used. 58

4.5 The Experimental Validity of the Hypothesis for Efficiency Variation. 61

4.6 The Resolving Power of Moments. 65

CHAPTER 5 THE IMPROVEMENT OF RECOGNITION EFFICIENCY AND THE MINIMIZATION OF LOSS

5.1 Recognition with Minimum Expected Loss. 68

5.2 The Forms of Error in a System. 70

5.3 The Relation between Execution Time and Efficiency. 77

5.4 Forms of Loss Matrix 78

CONCLUSION 82

APPENDICES

1 Power and Size in the Neyman-Pearson Theory. 86

2 A priori Probabilities and Information Content of Alphabetic Characters. 88

3 A Normality Theorem. 90

4 The Hand-Drawn Letter Set. 91

5 The Computer Tape Message. 100

6 Examples of Comparative I and L in the Tape Message. 102

7 Computer Programs used in Character Recognition. 103

8 Calculation of the Expected Overlap of $(A_n \cap g_{n+1})$. 110

9 Proof of a Combination Relation. 111

10 Straight Line Fitting Equations. 111

REFERENCES 112

LIST OF FIGURES.

2.1.1 Sketch of Variation of Efficiency with Simplicity 17

2.3.1 Example of Euclidean Boundary 17

2.4.1 Example of Bayes Rule Boundary with Different a priori Probabilities 24

2.5.1 Example of Difference between Equiprobability Boundary and Euclidean Boundary 24

3.3.1 Graph of Variation of Moment Intensity with Angle for M_{21} 35

3.5.1 Demonstration of Method of Character Digitization 39

3.5.2 Graph of Variation of Efficiency with Number of Moments used, with Different Decision Mechanisms 40

3.7.1 Variation of Exponent Distance for a well-resolved Letter 45

3.7.2 Variation of Exponent Distance for a poorly-resolved Letter 45

3.7.3 The Effect of Stepwise Variation of Letter Shape on the Modified Hamming Distance 48

4.3.1 The Theoretical Relation between Error Rate and Rejection Rate 53

4.3.2 The Experimental Relation between Error Rate and Rejection Rate 54

4.3.3 Linearity Test for Theoretical Relation between E and R 55

4.3.4 Test for Theoretical Relation between Rejection Threshold and Rejection Rate 57

4.4.1 Set Diagram for Calculating the Total Amount of

Information Extracted by n Moments of Equal Information Capacity 59

4.5.1 The Experimental Variation of Error Rate with Moment Number 62

5.2.1 Variation of Efficiency with Different Rejection Thresholds 72

5.2.2 Comparison between I and L in Tape Test Set 72

5.2.3 Confusion Matrix for Tape Test Set without Rejection Threshold 75

5.2.4 Confusion Matrix for Tape Test Set with Rejection Threshold 76

5.3.1 Variation of Execution Time with Recognition Efficiency 79

INTRODUCTION

Frequent reference has been made in the literature to certain loss functions and their applications to Decision Theory. In most cases, however, little regard has been paid to whether the approximations made during the course of a theoretical analysis are valid or whether more sophisticated loss functions ought to be implemented.

In the present work, the author has attempted to outline some of the knowledge already accumulated and to suggest new lines of approach for our understanding of loss functions. To do this it has proved necessary to reconsider the definitions, so that loss functions may be utilised in decision problems as opposed to the normal practice of simplifying them to abstraction. A certain amount of work has been completed on the use of Geometrical Moments in Character Recognition. This has provided a stable basis from which to develop the computer simulations which have revealed the relative efficiencies of a variety of methods available, including the effect of the introduction of an elementary rejection threshold. Finally a number of allied problems have been investigated in order to answer some of the questions that arose during the main body of the research.

The thesis falls naturally into three parts, one of which has been further divided to improve clarity of display. Also a number of appendices have been added in order to remove definitions and lengthy tables from the body of the work. A few of the more interesting computer programs have been included, as well as the hand drawn letter set and a few examples of the tape data set. Some of the results on Geometric Moments are soon to be published in a paper by F.C. Evans and the author entitled: 'The Development of a Two-Step Strategy for Character Recognition by Geometrical Moments.'

CHAPTER 1

THE DEVELOPMENT OF THE LOSS FUNCTION CONCEPT.

Before any detailed discussion of loss functions can be undertaken, a summary of past work must be made. Also a certain amount of elementary Decision Theory has to be included in a thesis of this kind to provide a basis from which the ideas suggested herein may develop with some continuity. It must be expected, however, that any review of general Decision Theory will fall short of the accumulation of knowledge that has been continuing over the last thirty years. Hence only the most relevant topics will be covered in this chapter.

1.1 The Historical Development of Loss Functions.

Much of our understanding of loss functions has developed within the sphere of Statistical Decision Theory. This has lead to a wealth of knowledge about the theoretical aspects of loss functions and their importance in relation to Decision Theory. The application of this theory to fields unknown at the time of its development has led to a deeper understanding in many channels. However, it seems that because of the intractability of loss functions to analysis and because of their dependence upon the particular problem being studied, they have often been neglected in the transformation from theory to practice.

In 1936, the Neyman-Pearson Theory of Hypothesis Testing [1,2] was published and it was here that the first notions of statistical risk were formulated. Two types of risk were considered which they called the Power and Size of the Critical Region - that region of observation space throughout which the Hypothesis was rejected (for definitions of Power and Size and their relationship to Statistical Decision Theory, see Appendix 1). This theory was generalized in 1947 when a sequential method for testing Hypotheses was developed by Abraham Wald,[3] the pioneer of much Decision Theory, which removed the restriction that the experiment was to be carried out in a single stage. However, certain assumptions had to be imposed:

(i) Each stage of the experiment consists of a single observation,

(ii) The chance variable X_i is observed in the ith. stage.

Wald [4] explains that there is no loss of generality in the first restriction if it is assumed that the cost of experimentation depends on the total number of observations but not on the number of stages in which the experiment is carried out. The second restriction, he continues, is more serious since it does not leave freedom of choice for the selection of the chance variable to be observed at any stage of the experiment. In the special case when the N chance variables are independently and identically distributed, there is no loss of generality in the second restriction either.

The concept of a loss function was further developed by Wald who considered that part of the problem of choosing a particular decision function involved stating the relative degree of preference given to the various elements of the terminal decision space when the true distribution of the random variable is known; the cost of experimentation was also important in determining the decision function. The degree of preference given to the various elements of the terminal decision space can be expressed by a non-negative function called a weight function which is defined over the whole decision space and the whole range of possible probability distributions. Also the cost of experimentation may depend upon the chance variables selected for observation, on the actual observed values obtained and on the stages in which the experiment is carried out. This then gives a clear comparison between the 'Weight Function' and the 'Cost Function', which may be considered collectively as the loss function in the problem of Pattern Recognition.

In 1951, T.W.Anderson[5] published the first of his papers on Multivariate Analysis, wherein he gave a clear description and summary of the properties of the loss function, and the way in which it is related to the calculation of optimum strategies. This work is discussed in a later chapter. After this paper, the study of statistical utility, of which loss is only the negative, developed properly both in its role in statistical decision theory and axiomatically in its own right. This study is summarised in several books, e.g. Luce and Raiffa[6] and De Groot[7]. Further development in the statistical field has been carried out by Raiffa and Schlaifer[8] who discuss a theory of non-additive utility in part of their book.

The use of loss functions in Pattern Recognition is widely mentioned, not least in Highleyman's [9] excellent article on 'Linear Decision Functions with applications to Pattern Recognition.' and in his doctoral thesis[10]. In general, however, it is found that loss functions have been regarded as 'simple' and hence could be ignored for lack of evidence to the contrary. The validity of this assumption is analysed in later sections.

1.2 The Development of Definition.

Throughout the growth of Decision Theory, there has been little agreement as to the precise meaning to be attached to such words as 'Loss Function' and 'Cost Function'. In this thesis these words will be defined, and used in different contexts.

The problem of interchangeability arises in the field of statistics since loss and cost are different words for the same phenomenon; that is, the statistician, wishing to solve a problem of point estimation, might talk of the 'cost' or 'penalty' imposed for not achieving correct identification (see for example Sage and Melsa[11]). If the true value of a parameter to be identified were θ and a value $\hat{\theta}$ were assumed, then a suitable cost function might be $\kappa(\theta - \hat{\theta})^2$. Other functions besides the quadratic form can be used in different circumstances (see Hays and Winkler[12]). However, this type of loss is formulated purely to allow the point estimation to be carried out. Certainly this is decision making under uncertainty, but of a different form from that which we shall be mainly studying.

In Pattern Recognition, the problem reduces to deciding which of a number of choices is correct. The existence of a correct choice within the set of those available is generally assumed; otherwise rejection occurs. Here again the loss is the penalty imposed for not achieving correct recognition, but it can be formulated independently of the parameters θ and $\hat{\theta}$. Instead loss becomes tied to whether or not the recognition has occurred correctly. Suppose measurements are made to determine the state of a system which could be in any one of n possible states $(\theta_1, \theta_2, \ldots, \theta_n)$, and as a result of these measurements, a choice θ is made. Then, because of our understanding of the nature of loss, we can say that if the recognition is correct, the loss is zero,

$$L(\theta|\theta_i)=0 \quad \text{if} \quad \theta = \theta_i \qquad 1.2.1.$$

where θ_i is the actual state of the system. Furthermore, all misrecognitions have losses associated with them that are greater than zero,

$$L(\theta|\theta_j)>0 \quad \text{for all} \quad j \neq i \qquad 1.2.2.$$

This formulation is the same as that described by Anderson[5] and Fukunaga[13]. It is intuitively clear, and will become mathematically clear below, that an absolute definition of misrecognition loss is unnecessary and only the ratio between different losses is of importance. Hence there is no loss of generality in defining an upper bound to $L(\theta|\theta_j)$ of unity,

$$\max\,[L(\theta|\theta_j)] = 1 \qquad 1.2.3.$$

There is another customary inclusion to the definition and that is one of rejection loss. If the sample measurement is rejected when the sample actually belongs to θ_γ, then the loss associated with that rejection is $L(0|\theta_\gamma)$, the rejection loss and

$$0 \leqslant L(0|\theta_\gamma) \leqslant 1 \qquad 1.2.4.$$

are the bounds imposed on such a rejection loss.

In future this notation will be simplified so that θ_γ will be represented by γ etc., for example the rejection loss $L(0|\theta_\gamma)$ will be written as $L(0|\gamma)$. The formulation of detailed loss functions from the above definitions is dealt with in the last chapter.

1.3 The Definition of Expected Loss, Error and Risk.

Before going further into the theory of loss functions, it is necessary to define the concepts of loss, error and risk. These elementary definitions serve as an introduction to Decision Statistics.

Let $\pi_1, \pi_2, \ldots, \pi_m$ be m populations with probability density functions $p_1(x_1, \ldots, x_p), \ldots, p_m(x_1, \ldots, x_p)$ respectively.* We wish to divide the space of observations into m mutually exclusive and exhaustive regions $R_1, \ldots, R_m$. If an observation falls into R_g, we shall say that it comes from π_g. Let the loss due to an observation, actually belonging to π_g, being classified as if from π_h, be $L(h|g)$. Then the probability of this misclassification is

$$p(h|g,R) = \int_{R_h} p_g(x_1, \ldots, x_p)\, dx_1 \ldots dx_p \qquad 1.3.1.$$

* after Anderson[5]

If the observation is from π_g, the expected risk is

$$r(g, R) = \sum_{h=1}^{m} L(h|g)\, p(h|g, R) \qquad 1.3.2.$$

Suppose we have a priori probabilities of the populations $q_1, q_2, \ldots, q_m$ then the expected loss is

$$\tilde{L} = \sum_{i=1}^{m} q_i\, r(i, R) = \sum_{i=1}^{m} q_i \left\{ \sum_{j=1}^{m} L(j|i)\, p(j|i, R) \right\} \qquad 1.3.3.$$

If we do not know the a priori probabilities, we cannot define an unconditional expected loss for a classification procedure. Then we consider the maximum of the risk $r(g, R)$ over all values of g and the decision problem becomes the choice of $R_1, \ldots, R_m$ which minimizes this maximum expected loss.* Fortunately in Character Recognition we nearly always know the a priori probabilities so that the Bayes strategy can be used.

To summarise, the average loss,

$$\tilde{L} = \sum_{i=1}^{m} \sum_{j=1}^{m} q_i\, L(j|i)\, p(j|i, R) \qquad 1.3.4.$$

The error rate, i.e. the average rate of misclassification is

$$\tilde{E} = \sum_{i=1}^{m} \sum_{j=1}^{m} q_i\, p(j|i, R) \qquad 1.3.5.$$

and the average risk

$$\tilde{R} = \sum_{i=1}^{m} r(i, R) = \sum_{i=1}^{m} \sum_{j=1}^{m} L(j|i)\, p(j|i, R) \qquad 1.3.6.$$

As is immediately apparent, if the losses are equal, $L(i|j) = L$ for all i and j then the difference between the error $\tilde{E}$ and the average loss $\tilde{L}$ becomes only an academic one; similarly if the a priori probabilities are considered equal, then the difference between the loss $\tilde{L}$ and the risk $\tilde{R}$ becomes also academic.

1.4 Additive and Non-additive Losses.

In statistical decision problems, there are times when a sampling cost has to be associated with the observation of a random variable X. This cost has to be considered when the risk of using any decision function based on X is calculated. This cost is particularly important where a choice between different random variables has to be made or whether it is better not to include a particular observation.

Once the value of the costs in appropriate units has been assigned, the expected cost of observation may be calculated. The total risk is defined to be the sum of the risk of using the decision function and the expected cost of observation. Normally all work in statistical decision theory uses this additive form for the total risk (De Groot's[7] words), but Raiffa and Schlaifer[8] discuss this assumption in some detail. They point out that even in problems in which the cost cannot be dimectly related to monetary cost, the decomposition of the risk into additive terms can occur when;

(i) the consequence of each action, (i.e. the performance of an experiment resulting in a particular consequence) and the cost of taking an action when the prevailing state is specified, is measureable in some common unit such that the total consequence is the sum of the two partial consequences,

(ii) the cost of this common unit is linear over the entire range of consequences involved in the given problem.

This common unit can, in many cases, be found even though the two actions have no common factor. This can be accomplished by an estimation on the part of the decision maker of how much of one action is equivalent to the other. Raiffa and Schlaifer quote the example of a scientist, wishing to estimate some physical constant θ, who may feel that whatever the error $(a-\theta)$ of his estimate a may be, he would be willing to make 10k more observations if he could be sure that by so doing, he would reduce $(a-\theta)^2$ by k units. If so, the cost of the consequence of a given action can be measured by the sum of the actual number of observations plus $10(a-\theta)^2$. The reference continues in some detail upon the theme of additive and non-additive losses and the author recommends it as a succinct account of the subject.

1.5 The Bayes Solution.

If R is the rule of classification, then this implies the division of a p-dimensional observation space into two regions R_1 and R_2. Also, if the observation is drawn from set i, the probability of correct classification $p(i|i,R)$ is the probability of the sample point falling in R_1, and the probability of misclassification $p(j|i,R)$ is the probability of the sample point falling in R_2. If we assign the sample point to the region which gives the lower expected loss (as defined above), then

$$R_1 : L(j|i)\, q_i\, p_i(x_1,\ldots,x_p) > L(i|j)\, q_j\, p_j(x_1,\ldots,x_p) \qquad 1.5.1.$$

and

$$R_2 : L(j|i)\, q_i\, p_i(x_1,\ldots,x_p) < L(i|j)\, q_j\, p_j(x_1,\ldots,x_p) \qquad 1.5.2.$$

We could also write

$$R_1 : \qquad \frac{p_i(x_1,\ldots,x_p)}{p_j(x_1,\ldots,x_p)} > \frac{L(i|j)\, q_j}{L(j|i)\, q_i} \qquad 1.5.3.$$

$$R_2 : \qquad \frac{p_i(x_1,\ldots,x_p)}{p_j(x_1,\ldots,x_p)} < \frac{L(i|j)\, q_j}{L(j|i)\, q_i} \qquad 1.5.4.$$

This is the so-called 'Bayes solution' and shows the mathematical reasoning behind the intuitive approach of believing that loss functions are only important in ratio-form and so are strictly relative.

Note: A distinction is made between two types of error occurring in 2-state systems.

(i) When a measurement belonging to the first category is classified by the decision maker as belonging to the second.

(ii) When a measurement belonging to the second category is classified as belonging to the first. These errors, although less definitive in a multi-category system, are called respectively errors of the first and second kinds.

1.6 The Relation between Rejection and Loss.

The above simple method of classification is a comparison of the conditional probabilities that a given measurement belongs to a given class. This conditional probability

$$\beta(m|i) = \frac{q_i p(x|i)}{\sum_{k=1}^{p} q_k p(x|k)} \qquad 1.6.1.$$

can be shown to lead to an optimal strategy. Chow[14] showed that the error rate can be minimized for a given rejection rate by utilizing the following strategy:

(i) Select the class for which the conditional probability is greatest.

(ii) Reject the measurement if this conditional probability is below a given probability threshold.

Highleyman[15] later showed that this same strategy led to the minimization of the expected loss given a constant loss function of the form

$L_{ii}=0$ for all classes i

$L_{ij}=1$ for all classes $i \neq j$ 1.6.2.

$L_{0j}=\beta$ for the rejection loss when the actual class was j.

This is perhaps the simplest useful form of loss function and will be discussed in some applications later.

CHAPTER 2

THE OPTIMAL DECISION BOUNDARY.

In this section, an attempt is made to extend the previously-discussed preparatory concepts to the field of Decision Theory. It is a chapter containing ideas and suggestions put forward with the aim, not of producing a comprehensive treatise on the subject of decision boundaries, but of stimulating criticism and further thought in the hope that the understanding of these concepts is increased in some measure.

The chapter opens with an ordered comparison of several discriminants in common usage and continues with attempts to calculate, theoretically or by example, certain decision boundaries and the losses associated with them. Examples have been used fairly widely to illustrate the theoretical mechanisms as and where it seemed important to emphasize the extent of the approximation dealt with in a typical problem.

2.1 The Selection of the Decision Maker.

The choice between the different decision discriminants available rests on the twin criteria of theory and expediency. However what is optimal in theory is rarely best in practice. Hence the gulf between the two has widened to such an extent as to allow well over a dozen discriminants into the applications field, many with the flimsiest theoretical backing, but of use because they work in the particular cases to which they have been applied.

Suppose we have a set of readings which make up the n components of the pattern vector $\underline{x} = (x_1, x_2, \ldots, x_n)$ and we are asked to compare these readings with two standard normal sets with means $\underline{\mu}_1 = (\mu_{11}, \mu_{12}, \ldots, \mu_{1n})$ and $\underline{\mu}_2 = (\mu_{21}, \mu_{22}, \ldots, \mu_{2n})$, and covariance matrices Σ_1 and Σ_2 respectively. If we are asked to compare the value $\underline{x}$ with $\underline{\mu}_1$ and $\underline{\mu}_2$ with no other demands to be satisfied, it is clearly best to calculate the probability that the reading $\underline{x}$ belongs firstly to $\underline{\mu}_1$, $p(x|\mu_1)$ and secondly to $\underline{\mu}_2$, $p(x|\mu_2)$. We can then assign $\underline{x}$ to the distribution for which $p(x|\mu_i)$, $i = 1, 2$, has a higher value. This can obviously be extended to an arbitrary number of normal sets. However this calculation, for practical purposes, has two difficulties:

(i) it is long and complicated,

(ii) it assumes a complete knowledge of all the distributions involved.

These problems have been recognized for a long time and much work has been done to avoid these difficulties. However, allowing for these drawbacks, this is clearly the best calculation that can be undertaken.

Suppose, on the other hand, we are asked to compare $\underline{x}$ with $\underline{\mu}_1$ and $\underline{\mu}_2$, but we do not know the covariances and we wish to use the simplest possible calculating procedure. The difficulties increase immediately because the concept of simplest procedures is obviously intimately related to the efficiency of the recognition technique. We might expect the efficiency to vary in much the same way as is shown in figure 2.1.1. As the simplicity of the decision mechanism increases, the efficiency tends to zero; as the simplicity decreases, the efficiency tends to 100%. Points representing recognition procedures that lie below the curve are less useful than those lying on the curve, so that the curve is a measure of all those procedures that have the highest efficiency for a given simplicity of execution.

A simple technique commonly used is the calculation of the Euclidean distance. For each of the distributions

$$d(\underline{x}) = \sqrt{(\underline{x}-\underline{\mu}).(\underline{x}-\underline{\mu})} \qquad 2.1.1.$$

and the distribution which yields the lowest value of $d(\underline{x})$ is chosen. It is not difficult to show that the decision boundary, the plane of which represents the discontinuity of choice, is a flat surface. Consider the boundary surface between $n_1(\underline{\mu}_1, \Sigma_1)$ and $n_2(\underline{\mu}_2, \Sigma_2)$. This is defined by

$$d_1(\underline{x}) = d_2(\underline{x})$$

or

$$(\underline{x}-\underline{\mu}_1).(\underline{x}-\underline{\mu}_1) = (\underline{x}-\underline{\mu}_2).(\underline{x}-\underline{\mu}_2)$$

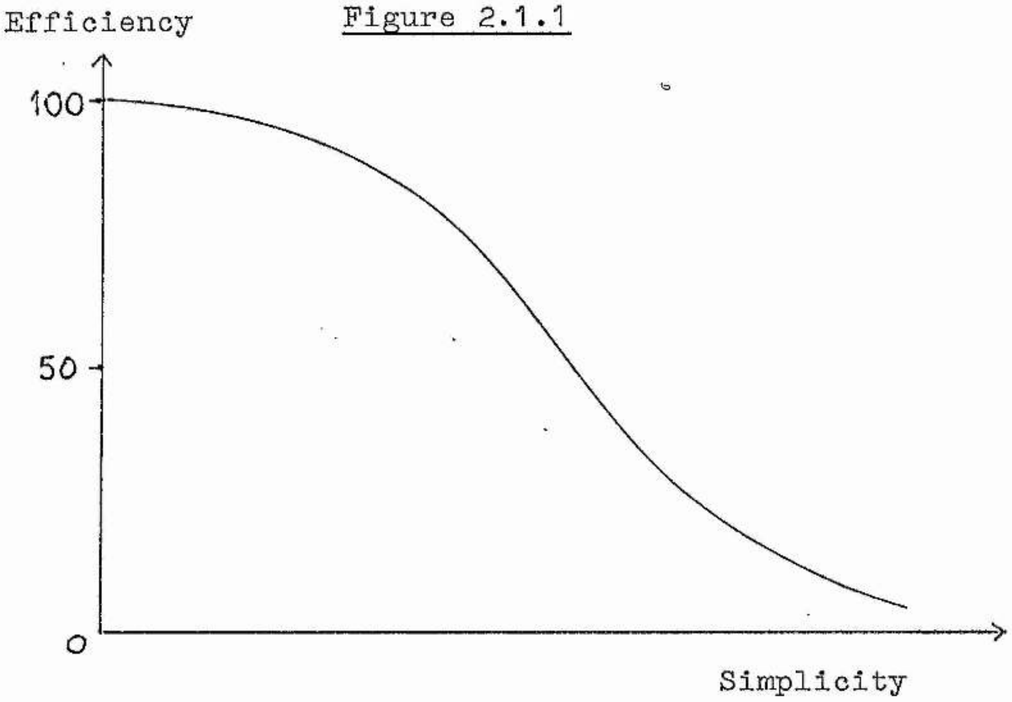
which contracts to

$$\underline{x}.(\underline{\mu}_2-\underline{\mu}_1) = \tfrac{1}{2}(\underline{\mu}_2.\underline{\mu}_2-\underline{\mu}_1.\underline{\mu}_1) \qquad 2.1.2.$$

This is the linear equation of a plane, with a perpendicular vector $(\underline{\mu}_2-\underline{\mu}_1)$, the vectorial line joining the means of the two distributions.

A calculation of this sort is very simple to undertake. The sign of the function $(\underline{a}\cdot\underline{x}-c)$ where $\underline{a}=(\underline{\mu}_2-\underline{\mu}_1)$ and $c=\frac{1}{2}(\mu_2^2-\mu_1^2)$ gives the side of the plane on which $\underline{x}$ lies, hence the calculation of this linear function determines to which region $\underline{x}$ will be assigned.

Figure 2.1.1



Sketch of Variation of Efficiency with Simplicity of decision mechanism.

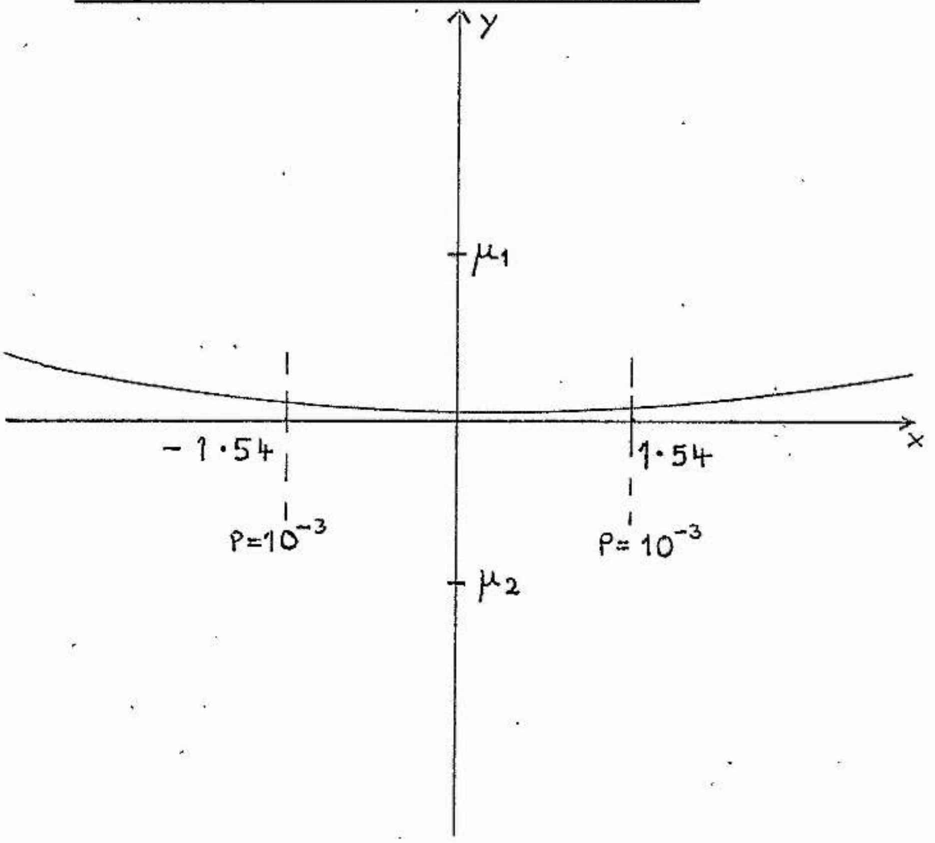


Figure 2.3.1

Example of Euclidean Boundary.

A similar discriminant is the Hamming distance

$$d(\underline{x}) = |\underline{x} - \underline{\mu}| \qquad 2.1.3.$$

Here again the calculation is simple, though slightly different, and the decision boundary consists of a set of planes, the main one of which is the same as the Euclidean boundary plane. The main plane is so designated because it cuts the line joining the two means $\underline{\mu}_1$ and $\underline{\mu}_2$ between these means, and so discriminates those points, or values of $\underline{x}$, that are most difficult to separate.

A systematic degradation of decision boundaries can be carried out from the probability function mentioned first to the Euclidean distance. The decision boundary between two normal multivariate probability distributions $n(\underline{\mu}_1, \Sigma_1)$and $n(\underline{\mu}_2, \Sigma_2)$ is given by

$$p(\underline{x}|\underline{\mu}_1) = p(\underline{x}|\underline{\mu}_2) \qquad 2.1.4.$$

Given that for a normal distribution

$$p(\underline{x}|\underline{\mu}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} \exp -\tfrac{1}{2}\left[(\underline{x}-\underline{\mu})' \Sigma^{-1} (\underline{x}-\underline{\mu})\right]$$

Then taking Logs., equation 2.1.4. simplifies to

$$\ln|\Sigma_1| + (\underline{x}-\underline{\mu}_1)'\Sigma_1^{-1}(\underline{x}-\underline{\mu}_1) = \ln|\Sigma_2| + (\underline{x}-\underline{\mu}_2)'\Sigma_2^{-1}(\underline{x}-\underline{\mu}_2) \qquad 2.1.5.$$

This has a quadratic form, and in two dimensions is a conic. If the assumption of equal covariance matrices is made, equation 2.1.5. reduces to the well-known discriminant function

$$g(\underline{x}) = -2\underline{x}'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) + \underline{\mu}_1'\Sigma^{-1}\underline{\mu}_1 - \underline{\mu}_2'\Sigma^{-1}\underline{\mu}_2 \qquad 2.1.6.$$

This approach is treated in numerous text books and does not need any further comment (see for instance S.S.Viglione paper[16]). If, on the other hand, the less sweeping approximation is made that $|\Sigma_1| = |\Sigma_2|$ or that $\ln|\Sigma_1^{-1}| << (\underline{x}-\underline{\mu}_1)'\Sigma_1^{-1}(\underline{x}-\underline{\mu}_1)$

for all $\underline{x}$ within the normal field of variation, and similarly for

$$\ln|\Sigma_2^{-1}| << (\underline{x}-\underline{\mu}_2)'\Sigma_2^{-1}(\underline{x}-\underline{\mu}_2)$$

Then the decision boundary becomes

$$(\underline{x}-\underline{\mu}_1)'\Sigma_1^{-1}(\underline{x}-\underline{\mu}_1) = (\underline{x}-\underline{\mu}_2)'\Sigma_2^{-1}(\underline{x}-\underline{\mu}_2) \qquad 2.1.7.$$

This is another quadratic form whose boundary has a slight displacement in space from that of equation 2.1.5.

If the assumption is now made that all the x_i's are independent, that is that Σ_1 and Σ_2 are diagonal matrices, then the decision boundary simplifies to a modified form of the Euclidean boundary in that it contains the standard deviations. If the diagonal elements of Σ are labelled Σ_j, then the simplified boundary is

$$\underset{\text{over } j}{\mathcal{S}} (\underline{x}-\underline{\mu}_1)_j^2\, \Sigma_{1j}^{-1} = \underset{\text{over } j}{\mathcal{S}} (\underline{x}-\underline{\mu}_2)_j^2\, \Sigma_{2j}^{-1} \qquad 2.1.8.$$

where $\mathcal{S}$ denotes the summation $j = 1, n$.

The last simplifying step is to assume all the Σ_j's are equal or unity and the boundary becomes the Euclidean boundary (equation 2.1.2.). It is clear from this process of degradation that the most simple boundary is a result of some drastic assumptions, the main one being that the two normal distributions have covariant matrices that are diagonal with equal elements.

2.2. An Elementary Approach to the Problem of Simplification.

During the stages of simplification in paragraph 2.1, it is theoretically often possible to state an assumption that leads to a simplifying step as:

$$\text{If } x << \alpha \text{ , then } \mathrm{I} \rightarrow \mathrm{II} \qquad 2.2.1.$$

where x is some positive variable, and a is a positive threshold linked to the more complex of the two stages (labelled I), and I and II are the two decision makers.

If the assumptions can be formulated in this manner, the efficiency of discrimination can be linked by a factor $f(x/a)$

$$\text{eff.}_{II} = \text{eff.}_{I} \cdot f(x/a) \qquad 2.2.2.$$

where $f(x/a)$ has the following properties:

(i) $1 > f(x/a) > 0$

(ii) $f(x/a) \to 1$ as $x/a \to 0$

(iii) $f(x/a) \to 0$ as $x/a \to \infty$

An elementary function of this form is

$$f(x/a) = \exp(-K.x/a) \qquad 2.2.3.$$

The parameter K is inserted to adjust the rate of fall-off of efficiency as the approximation becomes worse. However the form of the equation 2.2.3. certainly obeys conditions (i) and (iii) for $K > 0$. For $K(x/a)$ small, equation 2.2.3. may be expanded in a Taylor series

$$f(x/a) = 1 - K(x/a) + K^2/_2(x^2/a^2) - (O)(x^3/a^3)$$

$$\sim 1 - K(x/a) \qquad 2.2.4.$$

that is, the drop in efficiency becomes proportional to the degree of approximation involved in the simplifying step.

This step can be cast into a more rigorous form by the use of information theory (see for example ref. 17 p. 445). But at present it is enough to note this relationship and observe in later work whether this approach obeys the harsh criterion of working in practice.

2.3 A Comparison of the Different Systems of Discrimination by Example.

Suppose we take the two-dimensional case as an example, with

$$\underline{\mu}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \qquad \underline{\mu}_2 = \begin{pmatrix} 0 \\ -1 \end{pmatrix} \qquad \underline{x} = \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\Sigma_1^{-1} = \begin{pmatrix} 4 & 0 \\ 0 & 5 \end{pmatrix} \qquad \Sigma_2^{-1} = \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix}$$

Then the equi-probability decision boundary is

$$\sqrt{20}\exp -1/2\,(\underline{X}-\underline{\mu}_1)'\,\Sigma_1^{-1}(\underline{X}-\underline{\mu}_1) = \sqrt{12}\exp -1/2\,(\underline{X}-\underline{\mu}_2)'\,\Sigma_2^{-1}(\underline{X}-\underline{\mu}_2) \qquad 2.3.1.$$

We must expect that, since the diagonal elements in the matrices are fairly large with respect to unity, the distributions are mostly concentrated within circles radius 1, centres $\underline{\mu}_1$ and $\underline{\mu}_2$, so that the decision boundary will occur at a value of

$$p(X|\mu_1) = p(X|\mu_2)$$

that is fairly small. Taking Logs. of equation 2.3.1 and simplifying we obtain

$$x^2+(y-9)^2 = r^2 \quad \text{where } r^2 = 80+\ln 5/3 \qquad 2.3.2.$$

hence the probability boundary is a circle radius 8.97, centre $\binom{0}{9}$. In this case the Euclidean boundary is given by

$$x^2+(y-1)^2 = x^2+(y+1)^2$$

or y = 0 - the x-axis (see figure 2.3.1). These two boundaries are obviously very similar throughout the important region around the origin, and in this example, an approximation to the Euclidean distance discriminant would not seriously impair the efficiency of recognition, yet would increase the calculation speed.

From the above example, it becomes apparent that it is important to know what loss in efficiency is to be expected when choosing a simplifying procedure. The significance of this will be studied in paragraph 2.5. Before leaving this example, it might be noted that where the probability decision boundary crosses the y-axis, at y = 0.03, the probability $p(x|\mu)$ is about 0.067, whereas at y = 0 (the intersect formed by the Euclidean boundary and the y-axis), $p(x|\mu_1) = 0.0585$ and $p(x|\mu_2) = 0.0748$. The relevance of these figures is discussed in the latter section of paragraph 2.4.

2.4 The Degree of Approximation of the Bayes Rule to the Equi-probability Boundary.

As shown elsewhere, the Bayes decision boundary, which minimizes the expected loss, is of the form:

$$\frac{p_i(x_1,\ldots,x_p)}{p_j(x_1,\ldots,x_p)} = \frac{L(i|j)\,q_j}{L(j|i)\,q_i} \qquad 2.4.1.$$

The equi-probability boundary (equation 2.1.4) is clearly a special case of equation 2.4.1, where

$$L(i|j)\,q_j = L(j|i)\,q_i \qquad 2.4.2.$$

It is vitally important to consider the justification of this approximation. If it is valid, exact values of L and q do not have to be calculated. Alternatively, 'simple' loss functions can be instituted of the form

$$L(i|j) = 1 \quad \text{if } i \neq j \quad ; \; L(i|j) = 0 \quad \text{if } i = j \qquad 2.4.3.$$

which reduces equation 2.4.1 to one involving a priori probabilities only.

The calculation of this boundary, the Bayes boundary, simplifies from equation 2.4.1 to

$$\ln p_1 - \ln p_2 = \ln\{L(i|j)\,q_j\} - \ln\{L(j|i)\,q_i\} \qquad 2.4.4.$$

To justify the approximation of equating the right hand side to zero, we might limit the deviation of p_1 from p_2 at the boundary to 10% i.e

$0.9\,p_1 < p_2 < 1.1p_1$ at the boundary. 2.4.5.

Then ln p_1/p_2 = ln1.1 = 0.0953. In the work set out in the next chapter, the values of the exponents are in general very much greater than 10^3, so that 0.095 in 10^3 is small enough to warrant the approximation in at least that particular case. But if sample points very close to the decision boundary were given some special importance, then such a minute shift in the boundary radically alters the decision structure.

The size of the modification involved may be seen from the last example. If $a = q_1 L(1|2)$ and $b = q_2 L(2|1)$, then the Bayesian boundary is a circle with radius

$r = 80 + \ln(5a^2/3b^2)$ 2.4.6.

rather than the previous value

$r = 80 + \ln(5/3)$ 2.4.7.

In general, in Character Recognition, the maximum ratio $a : b$ might be 100:1. Then the circle radius could vary as

$r = 8.94 \pm 0.53$

Now in this case, a brief look at the figure below will show that this has made an important difference to the decision boundary. The intersect at the y-axis could vary between 0.56 and -0.50, a full 50% of the difference between the means (see figure 2.4.1).

It is suggested, then, that for so simple a case as this, the Euclidean distance measure is completely adequate to resolve the two sets, and that if losses could be associated with misrecognition, the Bayes decision boundary would be so

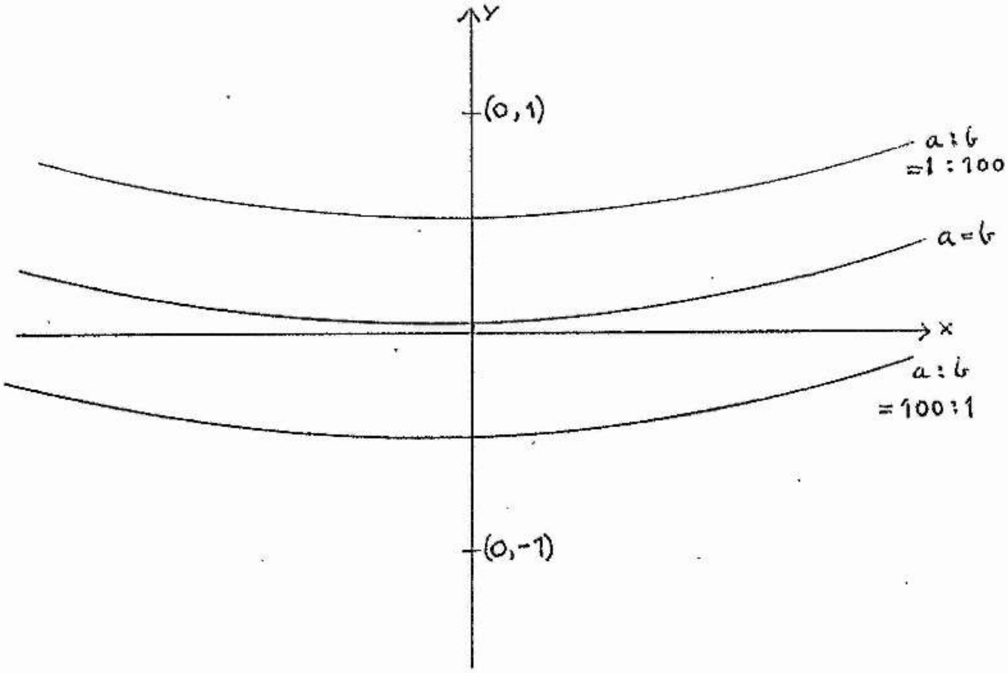


Figure 2.4.1

Example of Bayes rule boundary with different a priori probabilities.

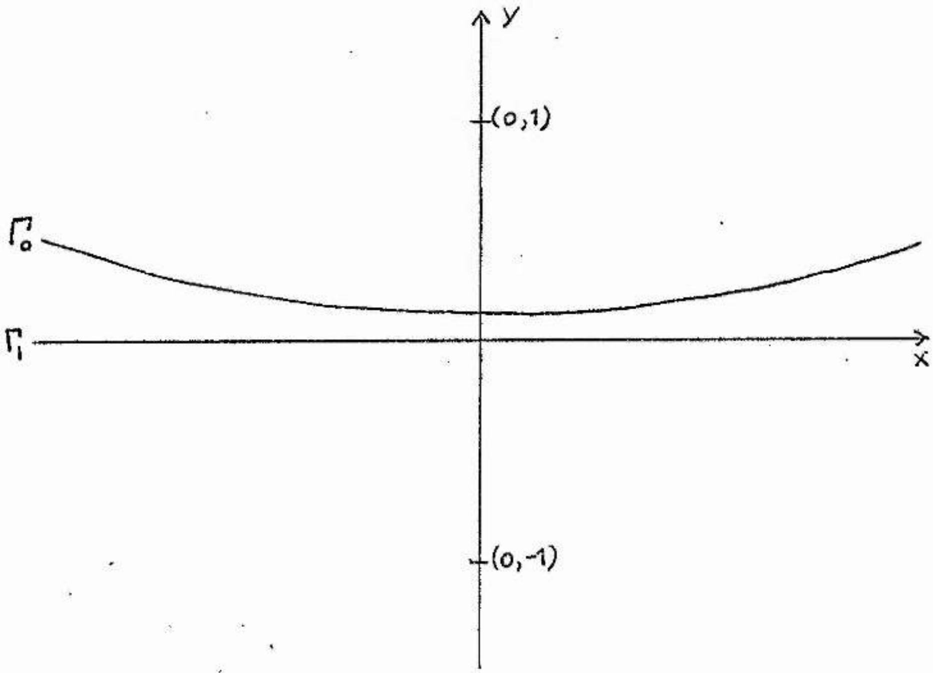


Figure 2.5.1

Example of difference between equiprobability boundary and Euclidean boundary.

affected that an adaptation to the Euclidean boundary could be made that would include some measure of the loss functions and the a priori probabilities.

2.5 The Loss associated with Different Boundaries.

Calculations of the expected loss in two and higher dimensions for optimal boundaries and two multivariate distributions have been successfully attempted. But what it is important to clarify is how close the loss associated with other decision boundaries approaches the minimum expected loss. In the example cited here, it is possible to calculate the expected loss for the Euclidean boundary.

If the equiprobability boundary is given by Γ_0 and the Euclidean boundary by Γ_1 (see figure 2.5.1), then the minimum expected loss

$$\tilde{L} = \iint_{-\infty}^{\Gamma_0} p(x|\mu_1)\,dS + \iint_{\Gamma_0}^{\infty} p(x|\mu_2)\,dS \qquad 2.5.1.$$

and the expected loss associated with the Euclidean boundary

$$\tilde{L}_{\text{subopt}} = \iint_{-\infty}^{\Gamma_1} p(x|\mu_1)\,dS + \iint_{\Gamma_1}^{\infty} p(x|\mu_2)\,dS \qquad 2.5.2.$$

If the first part of equation 2.5.2. is called L^1_{subopt} and the second part L^2_{subopt} , we have

$$L^1_{\text{subopt}} = \int_{-\infty}^{y=0} \int_{-\infty}^{\infty} \frac{a\sqrt{20}}{2\pi} \exp - \frac{1}{2}\left(4x^2 + 5(y-1)^2\right) dx\,dy \qquad 2.5.3.$$

which is easily evaluated as the terms in x and y are separable to give

$$L^1_{\text{subopt}} = 0{\cdot}159\,a$$

Similarly

$$L^{2}_{subopt} = \int_{y=0}^{\infty} \int_{-\infty}^{\infty} \frac{b\sqrt{12}}{2\pi} \exp -\tfrac{1}{2}\left(3x^2 + 4(y+1)^2\right) dx\,dy \qquad 2.5.4.$$

$$= 0.159\,b$$

Hence

$$\tilde{L}_{subopt} = 0.159\,(a+b)$$

If the integral is terminated at the $P = 10^{-3}$ level i.e. at $x = \pm 1.54$, then the area of integration becomes $-1.54 < x < 1.54$, $-\infty < y < +\infty$, and the value of $\tilde{L}_{subopt}$ is reduced to 0.876 of its extended value.

To compare these values with the optimal expected loss, this second term must be calculated. Fukunaga and Krile[18 or 19,20] give an accurate method for this calculation or an estimate may be made as follows:

the area Δ bounded by the Euclidean boundary y = 0, the optimal boundary Γ_0 and the lines $x = \pm 1.54$, is about 0.2. The probability density on the optimal boundary varies between 0.001 and 0.0675 and hence has an average of 0.034. At (x,y) = (0,0), the value of P_1 is 0.0585 and P_2 is 0.0748; hence an average over the whole area might be

P_1 average = 0.032

P_2 average = 0.040

$$\therefore\ \tilde{L}(\delta) = L_{subopt} + 0.032\,\Delta - 0.040\,\Delta \quad \text{where } \Delta = 0.2$$

$$= L_{subopt} - 0.0016$$

$$= 0.316,$$

a reduction of about ½% (evaluated for a and b unity).

Hence there is in this example only a marginal improvement in using the optimal boundary as opposed to the Euclidean boundary.

2.6 The Geometric Form of the Optimal Boundary.

As previously stated and as is well known, the form of the Bayesian decision boundary can be written

$$(\underline{x}-\underline{\mu}_2)' V_2^{-1} (\underline{x}-\underline{\mu}_2) - (\underline{x}-\underline{\mu}_1)' V_1^{-1} (\underline{x}-\underline{\mu}_1) = c \qquad 2.6.1.$$

where

$$c = 2 \ln \left\{ \frac{L(1|2)\, q_2}{L(2|1)\, q_1} \right\} + \ln \frac{|V_2|}{|V_1|} \qquad 2.6.2.$$

This may, with a little manipulation, be converted into a perfect square of the form

$$(\underline{x}-\underline{d})'(V_2^{-1} - V_1^{-1})(\underline{x}-\underline{d}) = c' \qquad 2.6.3.$$

where

$$\underline{d} = (V_2 - V_1)(V_2^{-1}\underline{\mu}_2 - V_1^{-1}\underline{\mu}_1)$$

and

$$c' = c - \left\{ \underline{\mu}_2' (V_1^{-1} - V_2^{-1}) \underline{\mu}_1 - \underline{\mu}_1' (V_1^{-1} - V_2^{-1}) \underline{\mu}_2 \right\} + \left\{ \underline{\mu}_2' V_2^{-1} V_1 V_2^{-1} \underline{\mu}_2 - \underline{\mu}_1' V_1^{-1} V_2 V_1^{-1} \underline{\mu}_1 \right\}$$

Equation 2.6.3 is a general quadratic form in n-dimensional space, and a conic in two dimensions. What is of interest is that as c varies, the quadratic form remains confocal and hence the family of equations of the form 2.6.3 given by the variation of c is locally parallel. Thus if $L(1|2) \to L'(1|2)$ and $L(2|1) \to L'(2|1)$ (see Appendix 3) then

$$c' = c + 2 \ln \left\{ \frac{L'(2|1)}{L'(1|2)} \right\} - 2 \ln \left\{ \frac{L(2|1)}{L(1|2)} \right\} \qquad 2.6.4.$$

or

$$c' = c + 2 \ln \left\{ \frac{L'(2|1)}{L(2|1)} \right\} - 2 \ln \left\{ \frac{L'(1|2)}{L(1|2)} \right\} \qquad 2.6.5.$$

Thus the distançe shifted, $(c' - c)$, is proportional to $\ln\left(\frac{L+\Delta L}{L}\right) \sim \frac{\Delta L}{L}$ where $(L+\Delta L)$ is either one of the new L''s.

2.7 The Piece-wise Linear Boundary.

Owing to the computational difficulties implicit in the quadratic form of the normal probability distribution, it is important to investigate alternative forms or approximations to such non-linear boundaries. Without resorting to the elementary ones already discussed, squared Euclidean etc., a good case can be made for piece-wise linear boundaries (see for instance Helstrom[43] or J.R.Ullmann[44]).

A recent article by Chhikara and Odell[45] suggests that computer algorithms or calculations involving normal probability integrals may be unnecessary in real Pattern Recognition. They are concerned with the use of discriminant analysis of complex images with many resolution elements. By generating a set of 'normed' exponential densities, of which the normal distribution is a special case, a good argument develops that the assumption of normality, which never was adopted unconditionally, can be lifted and the new form of decision boundary confidently expected to perform as well in practice as the old one, but with substantially faster implementation time.

CHAPTER 3

GEOMETRICAL MOMENTS IN DECISION MAKING.

3.1 Introduction.

Having summarised the theory of decision strategies and loss functions, a study of their effect upon the efficiency of different decision mechanisms has to be conducted. But before this can be done, a working model needs to be adopted by which to test the theoretical reasoning. To arrive at such a model, the different features that have been used before in Character Recognition were considered:-

(i) Spatial features [21, 22]

(ii) Selected n-tuples [23]

(iii) Random n-tuples [24]

(iv) Spatial Fourier transforms [25, 26]

(v) Temporal Fourier transforms [27]

(vi) Statistical Moments [28, 29]

(vii) Geometrical Moments [30]

(viii) Whole Character recognition [31, 32, 33]

As a system of features was required that would generate in a relatively simple manner sets of numbers that were linked preferably to the geometric structure of the characters, it was decided that Geometrical Moments should be used. A model could be then established based on these features, with the aim of improving as much as possible the recognition efficiency and other variables associated with the minimization of loss. Franz Alt's work[30] provided a good basis from which to start. A letter written by Minsky[34] criticises this approach as being by no means the only method for producing invariants. However the method is simple and hence has that much to commend it.

Ming-Kuei Hu, who has performed much research into Statistical/Geometric Moments mentions an optical method for Moment calculation[35] developed by E.Kletsky. A mask was formed with an optical density that varied according to the power of the Moment. Then, when the focussed image was transmitted through the mask onto a photoelectric cell, the resulting current was proportional to the Moment for that particular image. This method seems to be a very quick and inexpensive way of calculating the Moments. Unfortunately it was found that the accuracy was limited to around 1% and this, combined with the usual alignment difficulties, was enough for the project to be abandoned. However, the fact that such an approach is possible probably warrants further study, especially since Moments seem to be so rewarding in recognition efficiency.

3.2 The Theory of Geometric Moments.

Any pattern can be quantized into a matrix of numbers, each number referring to the greyness level at that coordinate point. The pattern can then be regarded as a set S of coordinate positions and greyness levels:

$$S = \{x, y, f(x,y)\}$$

Normally the greyness $f(x,y)$ is taken as an integer value, and in the subset considered here, $f(x,y)$ is 0 or 1, an elementary black/white system. The Moment M_{ij} is defined by the relation

$$M_{ij} = \sum_{\text{over } S} f(x,y)x^i y^j \qquad 3.2.1.$$

where the sum over S is taken over all black points on the matrix in the special case being considered. Hence

$$M_{ij} = \sum_{\substack{\text{over all} \\ \text{black points}}} x^i y^j \qquad 3.2.2.$$

Alt[30] extended the definition by successive transformations to normalize the following variables:

(i) Position of the Character

(ii) Size (number of black bits)

(iii) x and y spread

and (iv) Slant;

the total transformation being

$$M_{ij} = \sum_{\text{over } S} X^i Y^j \qquad 3.2.3.$$

where $X = [(x-\bar{x})/\sigma_x - \rho(y-\bar{y})/\sigma_y]/\sqrt{1-\rho^2}$

and $Y = (y-\bar{y})/\sigma_y$

σ_x = standard deviation of the character about the x-axis

σ_y = standard deviation of the character about the y-axis

ρ = the correlation coefficient

$$= \left\{\sum_{\text{over } S} (x-\bar{x})(y-\bar{y})/\sigma_x\sigma_y\right\} \Big/ \sum_{\text{over } S} (x-\bar{x})^2/\sigma^2_x$$

Note: the normalizing procedure does not necessarily have to be carried out on each coordinate point before the Moments are calculated. They may be calculated unnormalized first and then the appropriate algebraic manipulation done on the final numbers.

Fifteen Moments were calculated per character, that is, all those up to $i+j = 5$, in all 21, less those used in the normalizing procedure (see table 1).

i+j						
0	M^*_{00}					
1	M^*_{01}	M^*_{10}				
2	M^*_{02}	M^*_{11}	M^*_{20}			
3	M_{03}	M_{12}	M_{21}	M_{30}		
4	M_{04}	M_{13}	M_{22}	M_{31}	M_{40}	
5	M_{05}	M_{14}	M_{23}	M_{32}	M_{41}	M_{50}

* Moments used in normalization.

TABLE 1

3.3 The Geometrical Properties of Moments.

The Moment M_{ij} reveals information about two properties of a character,

(i) The presence of an axis of symmetry,

(ii) The spread of a character about a given axis.

Consider letters that have been orientated for zero slant. Then, for Moments with either i or j odd, for instance M_{30}, if a character possesses symmetry about the y-axis (in this case), then the Moment contribution from those parts of the character, for which the x-coordinate is positive, is equal but opposite to the Moment contribution of the parts for which the x-coordinate is negative. Hence the Moment M_{30} for an A would be small, or zero if the A possessed complete symmetry about the y-axis.

Secondly, the spread of the character about a given axis may be revealed as follows:-

Consider the Moment

$$M_{ij} = \sum_{S} x^i y^j$$

and substitute for x and y the polar coordinates

$$x = r \cos\theta \qquad y = r \sin\theta \qquad 3.3.1.$$

respectively. This is taking the centre of gravity of the character at the origin.

Then $$M_{ij} = \sum_{\text{over } S} r^{i+j} \cos^i\theta \sin^j\theta \qquad 3.3.2.$$

In order to determine the values of θ for which M_{ij} is a maximum or minimum, equation 3.3.2 is differentiated with respect to θ.

$$\frac{\partial M_{ij}}{\partial\theta} = \sum_{\text{over } S} r^{i+j}(j \cos^{i+1}\theta \sin^{j-1}\theta - i \cos^{i-1}\theta \sin^{j+1}\theta)$$

$$= \sum_{\text{over } S} r^{i+j}(j \cos^2\theta - i \sin^2\theta) \sin^{j-1}\theta \cos^{i-1}\theta \qquad 3.3.3.$$

$$\frac{\partial M_{ij}}{\partial\theta} = 0 \quad \text{when } \sin\theta = 0, \quad \text{when } \cos\theta = 0,$$

$$\text{or when } \tan\theta = \pm\sqrt{j/i}.$$

For further calculation, specific cases are best calculated individually.

E.g. for M_{21}, $i = 2$ and $j = 1$.

Also

$$\frac{\partial M_{21}}{\partial\theta} = \sum_{\text{over } S} r^3 \cos\theta (\cos^2\theta - 2\sin^2\theta) \qquad 3.3.4.$$

Now $\frac{\partial M}{\partial \theta}21 = 0$ when $\cos\theta = 0$ or when $\tan\theta = \pm 1/\sqrt{2}$
In order to find for which values of θ M_{21} is a maximum or minimum, we calculate the sign of

$$\frac{\partial^2 M_{21}}{\partial \theta^2} = \sum_{\text{over } S} -r^3 \sin\theta (7 \cos^2\theta + 2 \sin^2\theta) \qquad 3.3.5.$$

$$\begin{array}{llllll}
\text{So for } \cos\theta = 0, & \frac{\partial^2 M_{21}}{\partial \theta^2} & > 0 & \theta = (2n+1)\pi & M_{21} & \text{min.} \\
 & & < 0 & \theta = 2n\pi & M_{21} & \text{max.} \\
\tan\theta = +1/\sqrt{2}, & \frac{\partial^2 M_{21}}{\partial \theta^2} & > 0 & \pi < \theta < 3\pi/2 & M_{21} & \text{min.} \\
 & & < 0 & 0 < \theta < \pi/2 & M_{21} & \text{max.} \\
\tan\theta = -1/\sqrt{2}, & \frac{\partial^2 M_{21}}{\partial \theta^2} & > 0 & \pi/2 < \theta < \pi & M_{21} & \text{min.} \\
 & & < 0 & 3\pi/2 < \theta < 2\pi & M_{21} & \text{max.}
\end{array}$$

This is shown more clearly in the graph (figure 3.3.1). Hence if a character is concentrated about the axis

$$\theta = \pm \tan^{-1} 1/\sqrt{2},$$

and has parts lying in the quadrants $0 < \theta < \pi/2$, $3\pi/2 < \theta < 2\pi$, such as the letter K, then M_{21} will be large and positive. If, on the other hand, the letter was concentrated in the other two quadrants, M_{21} would be large and negative.

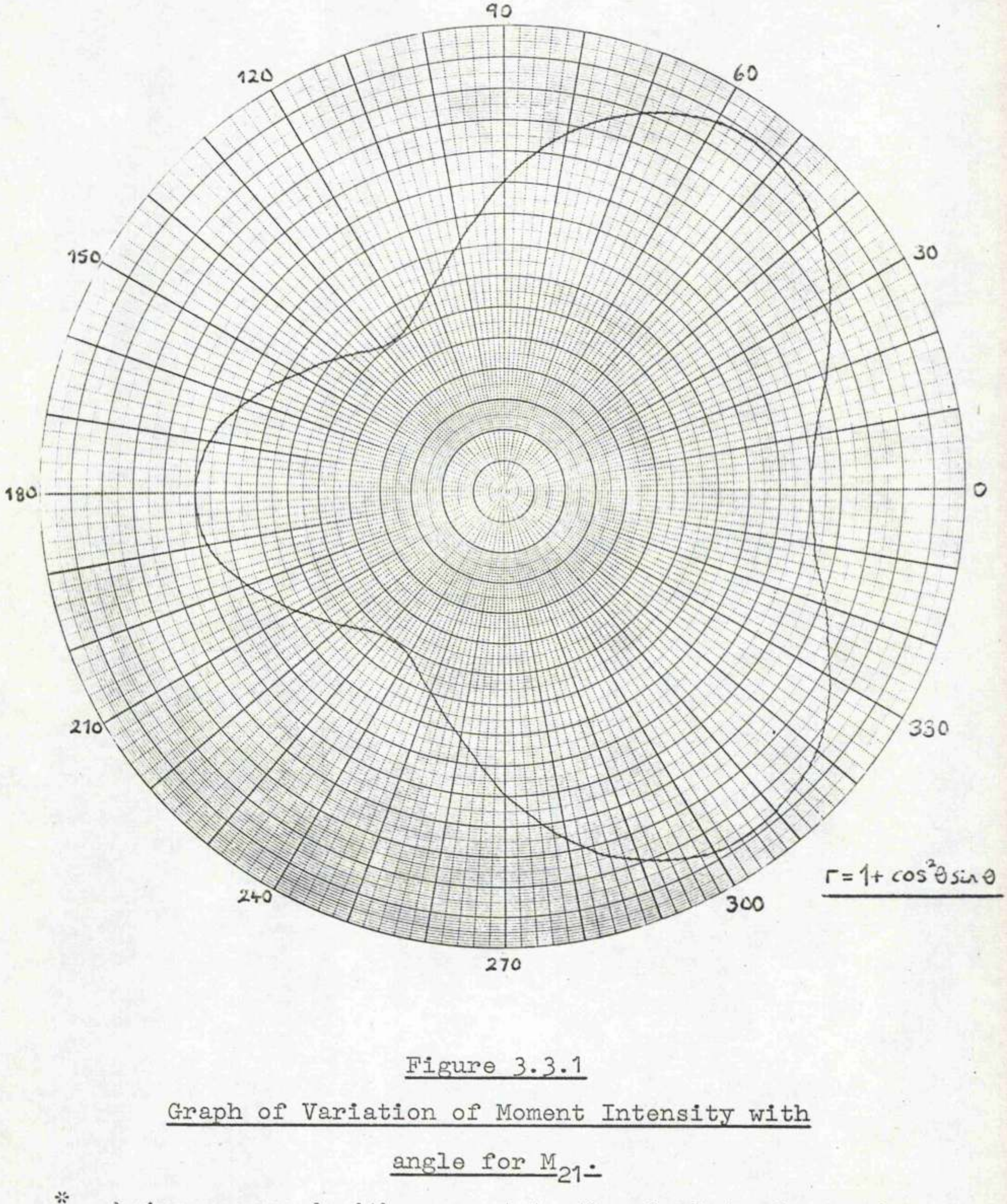


Figure 3.3.1

Graph of Variation of Moment Intensity with angle for M_{21}.

* being measured with respect to the absolute frame of reference of the character already normalized for zero slant.

3.4 The Decision Mechanism.

A) Moments. Alt found experimentally that ten of the moments worked better than the others, namely:

M_{13}, M_{12}, M_{21}, M_{30}, M_{04}, M_{31}, M_{40}, M_{14}, M_{03} and M_{50}.
It was these that he used in a Decision Tree form of recognizer.

However, we attempted to use all the moments equally at the beginning of our research and to devise a resolution technique later to assess the individual value of each moment.

B) Discriminants. In relation to the discriminants discussed in paragraph 2.1, we used five main methods:

(i) The squared Euclidean distance

$$\sum_{i=1}^{n} (X_i - \mu_i)^2,$$

where the subscript i refers to one of the n moments; X_i to the calculated ith moment of the unknown character and μ_i to the ith moment of the mean of the distribution with which X is being compared.

(ii) The Hamming distance

$$\sum_{i=1}^{n} |X_i - \mu_i|$$

(iii) The modified squared Euclidean distance

$$\sum_{i=1}^{n} (X_i - \mu_i)^2/\sigma_i^2,$$

σ_i being the standard deviation of the distribution for which μ_i is the mean.

(iv) The modified Hamming distance

$$\sum_{i=1}^{n} |X_i - \mu_i|/\sigma_i.$$

(v) The exponent distance

$$(\underline{X} - \underline{\mu})' V^{-1} (\underline{X} - \underline{\mu}),$$

V being the variance/covariance matrix of which the σ_i^2 are the diagonal elements, $\underline{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$ and $\underline{X} = (X_1, X_2, \ldots, X_n)$.

In the case of moments, recognition efficiency may be expected to be low for (i) and (ii) since each individual moment may be of a different order of magnitude e.g. M_{50} contains terms in X^5, whereas M_{30} contains only terms in X^3, so that M_{50} will be far larger than M_{30}.

Few discussions in print mention any need to divide by the deviations since, if they are comparable, the normal Hamming distance method is quite adequate. However, when using Geometric Moments, it is clear that such a simplification should not be undertaken (see table 2). The difference in the efficiency of the Hamming and modified Hamming methods reveals the experimental backing for this argument (see table 3 in the following section).

	M_{13}	M_{12}	M_{21}	M_{04}	M_{31}	M_{40}	M_{14}
μ_A	-11.8	12.4	150	-21.4	-12.5	-3.72	-9.38
σ_A	4.48	3.92	32.4	7.84	5.98	2.90	4.30

TABLE 2

Comparison of μ and σ for letter A of the Tape Data.

3.5 The Calculation of the Moments.

Moment data for the decision mechanisms were obtained from a set of hand-drawn capital letters, containing eight widely varying types to each alphabetic character. The quantization of each character was carried out by drawing it as a series of points on a 20 x 20 matrix (see figure 3.5.1). The moments were then calculated and normalized on an IBM 360 computer. From the eight types, the means and standard deviations were found for each alphabetic letter. Using the decision techniques outlined above, the learning set was then tested against itself, so that the resolution and efficiency of the methods could be measured (see figure 3.5.2). Letters were then constructed on the computer's remote access units (RAX Terminals) as a test set and the Modified Hamming Distance technique used for recognition. Finally the same methods were applied to an IBM tape of some 4000 characters in the form of a message. The first 500 characters were used as a learning set and the whole 4000 letters as a test set.

Number of moments used	3	6	9	12	15
Hamming distance efficiency %	50.9	70.2	63.5	75.0	77.9
Modified Hamming efficiency %	49.5	70.2	85.6	88.0	87.0

TABLE 3

Construction of the Data Set: For each letter of the alphabet, eight types of letter were hand-drawn onto 20 x 20 squared graph paper. These were encoded into five cards per character in the following way :-

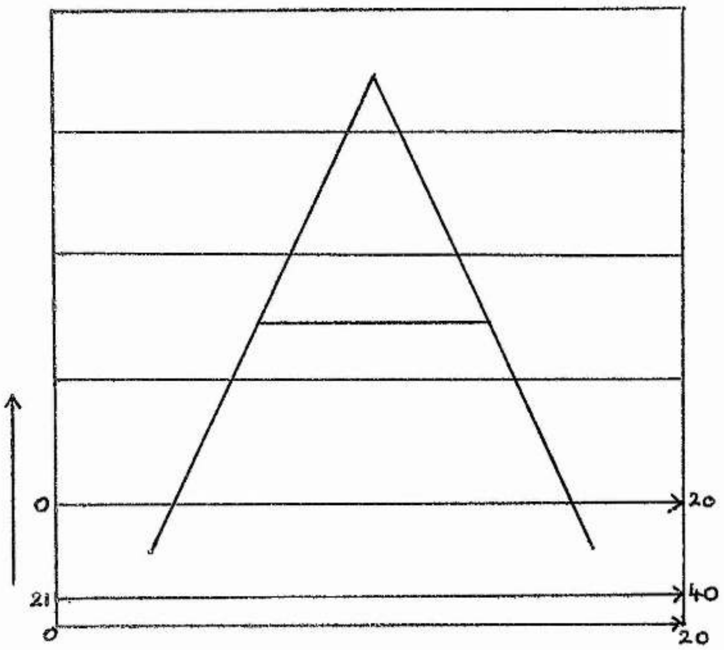


Figure 3.5.1.

Each 80-column computer card was assigned one of the oblong areas shown in figure 3.5.1. Starting at the bottom left-hand corner and working from left to right up the drawing, 1's were printed in the appropriate column if part of the letter crossed that particular square. In practice, x's were used to draw the character rather than continuous lines, so that there was no confusion as to how much of a line there needed to be before a 1 was punched. This method was used as opposed to one in which the actual Cartesian Coordinates of each point are measured and recorded. This was because the number of x's, used on average, took up fewer recorded columns, and the way in which the letters were drawn made this technique simpler to implement. Furthermore the method gave a fixed number of cards per character.

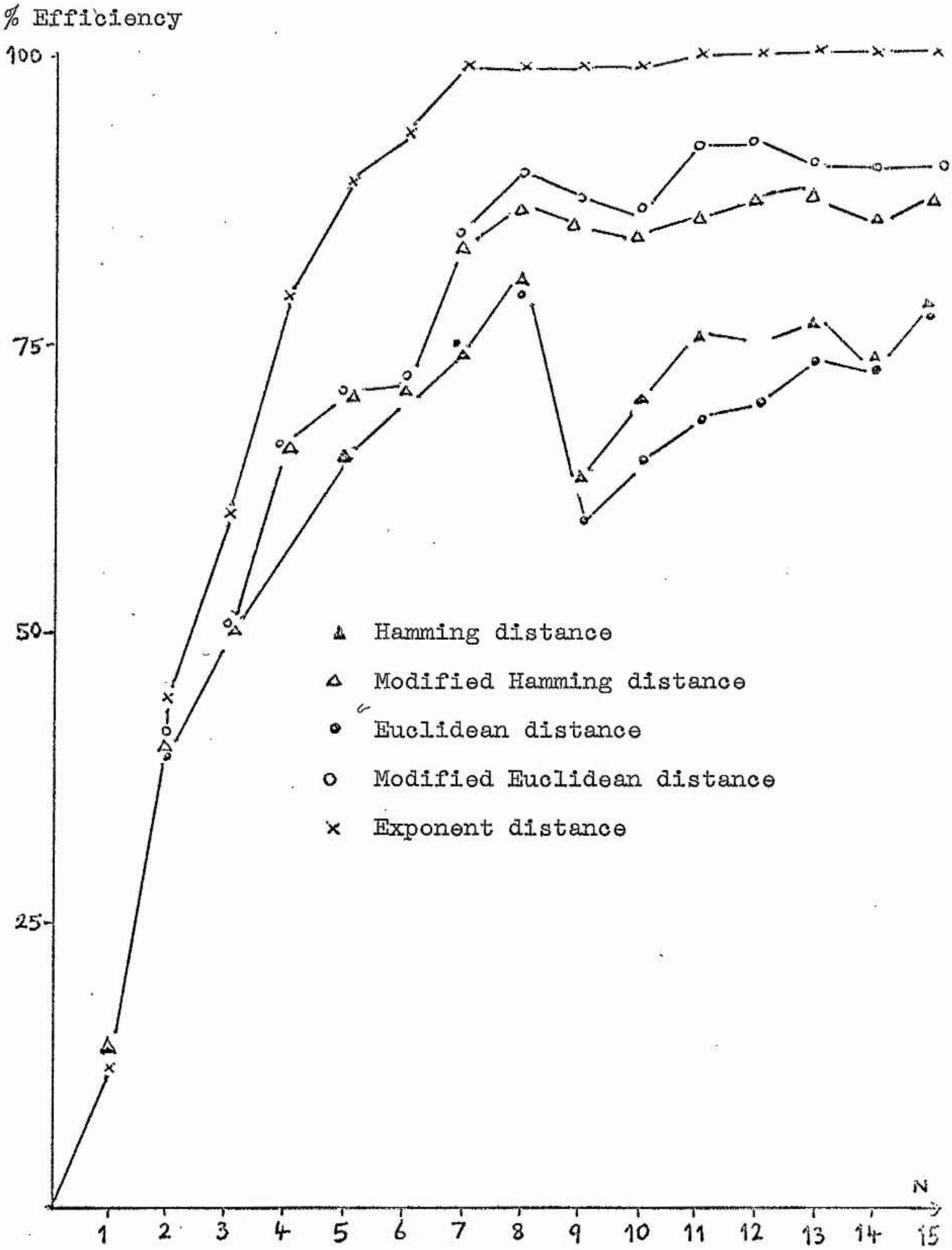


Figure 3.5.2

Graph of Variation of Efficiency with Number of Moments used, with different decision mechanisms.

The different types of letter in this hand-drawn letter set were all capitals, but varied as much as possible within the limited grid spacing. Little variation in size and position on the grid was written into the set as these variables are accounted for in the normalization procedure in the program. The entire hand-drawn letter set is displayed in Appendix 4 for reference.

3.6 Program Description.

It was decided as general practice, to calculate the fifteen moments of each character and store them in some tangible form, either on cards, as in the case of the hand-drawn letter set, or on magnetic tape, as in the much longer IBM message. This meant that during most tests, the moments did not need to be recalculated for each run. The principle exception to this was the calculation of execution time with various decision mechanisms and different numbers of moments (see Section 5.3).

The first program shows the technique used on the analysis of the tape characters with subroutine CHANGE and subroutine MOMENT. Subroutine CHANGE was especially written in Assembler language to convert the rather unusual coding of the tape letter scan into a conventional matrix. The IBM scanner read each character in two halves and scanned in vertical lines, feeding the digitized information straight onto the magnetic tape in the form of half-words of 4 bytes each. This needed to be converted to a matrix of logical elements, representing a black section by True (1) and a white section by False (0). This accomplished, subroutine MOMENT was called to calculate the moment invariants. The entire process from scanning to the generation of moment numbers could have been accomplished electronically using integrated circuit multipliers at a considerable saving in expenditure and calculation time, but for the purposes of the simulation and having an IBM 360 available, the adopted system was the most flexible one available.

Having generated and stored the moments, together with a code to identify to which character the moments belong, a program was developed to calculate the means and standard deviations of each character set assuming a normal distribution about that mean. With the small size of sample sets used, it may be presumed that the normality assumption is acceptable. Again the tape message presented certain difficulties as the letter sequence had to be ordered alphabetically, naturally, and the punctuation marks removed, since in general these were of too limited a set. Also capital letters had to be noted and three letters omitted entirely because in the learning set of the first 500 tape characters, they either appeared only once or not at all. The presence of capitals caused the most trouble, as it was not until after many of the experiments had been performed that a proper listing of the message with spaces and punctuation (see Appendix 5) was obtained. It was only then that capitals could be identified and noted. In general these letters were recognized incorrectly by any decision mechanism, and these letter errors have been omitted from the efficiency figures except where stated.

Finally, once both the character moments and the letter set moment means and standard deviations were available, recognition tests could be carried out on the different mechanisms referred to elsewhere. The only decision method that caused trouble was the exponent mechanism because not only the standard deviations but the entire covariance matrix needs to be calculated for each letter set. This matrix then has to be inverted. Because the moments are generally numerically large, overflow and underflow problems were encountered when the matrix determinants were being calculated. This difficulty, however, was bypassed by inserting a suitable power of ten before and after inversion, before to allow the computer to invert the matrix, and after to restore the matrix elements to their proper form. Even then it was found that the subroutine supplied for matrix inversion consistently generated the wrong sign for the determinant, so that calculations of $(\underline{X} - \underline{\mu})' V^{-1} (\underline{X} - \underline{\mu})$ would be correct except for a spurious minus sign (this expression must be positive for a positive definite symmetric matrix V). That this was the only fault of the subroutine was proved by calculating the product of the inverted and non-inverted matrix and for each letter set this was found to generate the identity matrix as expected. This slight anomaly was easily corrected, once found, by taking the modulus of the scalar terms $(\underline{X} - \underline{\mu})' V^{-1} (\underline{X} - \underline{\mu})$. This restored the recognition efficiency from chance to over the 90% level.

The last program listed in the appendix calculates not only the recognition efficiency with different rejection thresholds, but also under no-rejection conditions. This was used extensively as described in Chapter 4.

3.7 The Efficiency of the Decision Mechanism.

(A) Learning set on the learning set: Figure 3.5.2 shows the variation of efficiency of the different decision mechanisms as the number of moments involved in the calculation changes. The exponent method performs better than the other mechanisms as it contains the cross-correlation factors between different moments. Although this method is undoubtedly more complicated than the others, it is of interest as it shows the sort of improvement that can be gained by using the entire covariance matrix. Both the modified methods are more efficient than their unmodified forms, but only the exponent method efficiency increases monotonically with the number of moments used. This suggests that the cross-correlations play an important part in reducing the error rate. The peak performance occurs at 7 or 8 moments. The precise reason why the efficiency tails with the addition of more moments to the discriminant suggests some other process, such as the introduction of noise, that increases with moment number until it becomes a more important factor to the change in recognition efficiency than the sheer increase of the number of moments.

Figure 3.7.1 shows the resolution of the exponent method applied to a typical letter J. It identifies the letter as the only possible character within six orders of magnitude. The modified Hamming and Euclidean methods also identify it successfully, but with a greatly reduced recognition threshold.

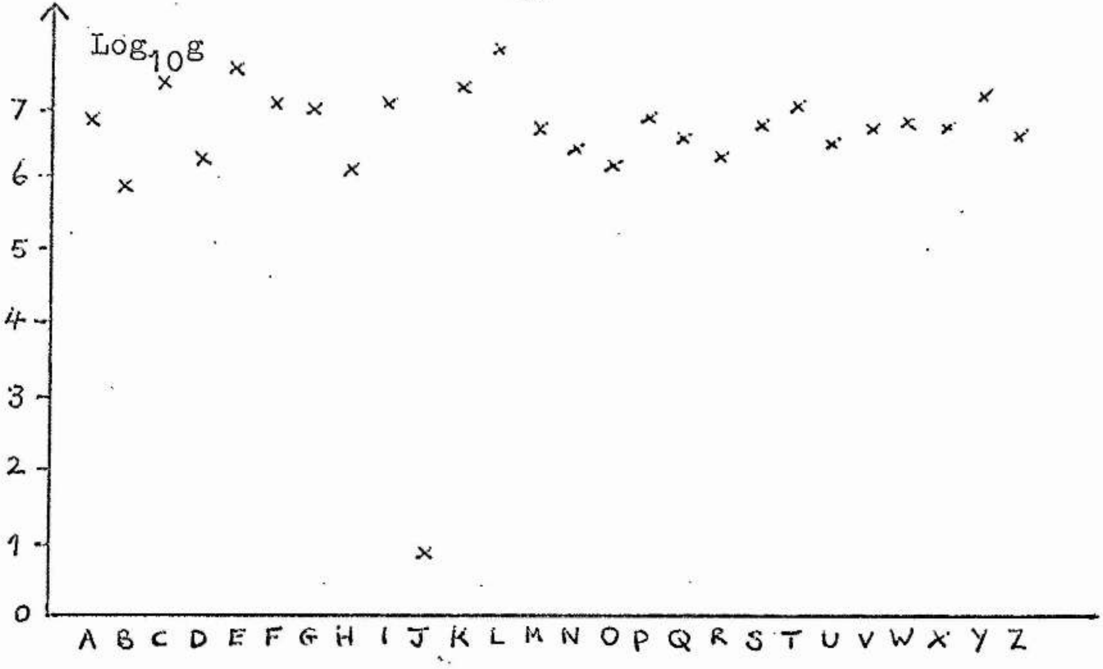


Figure 3.7.1

Recognition of the Letter J.

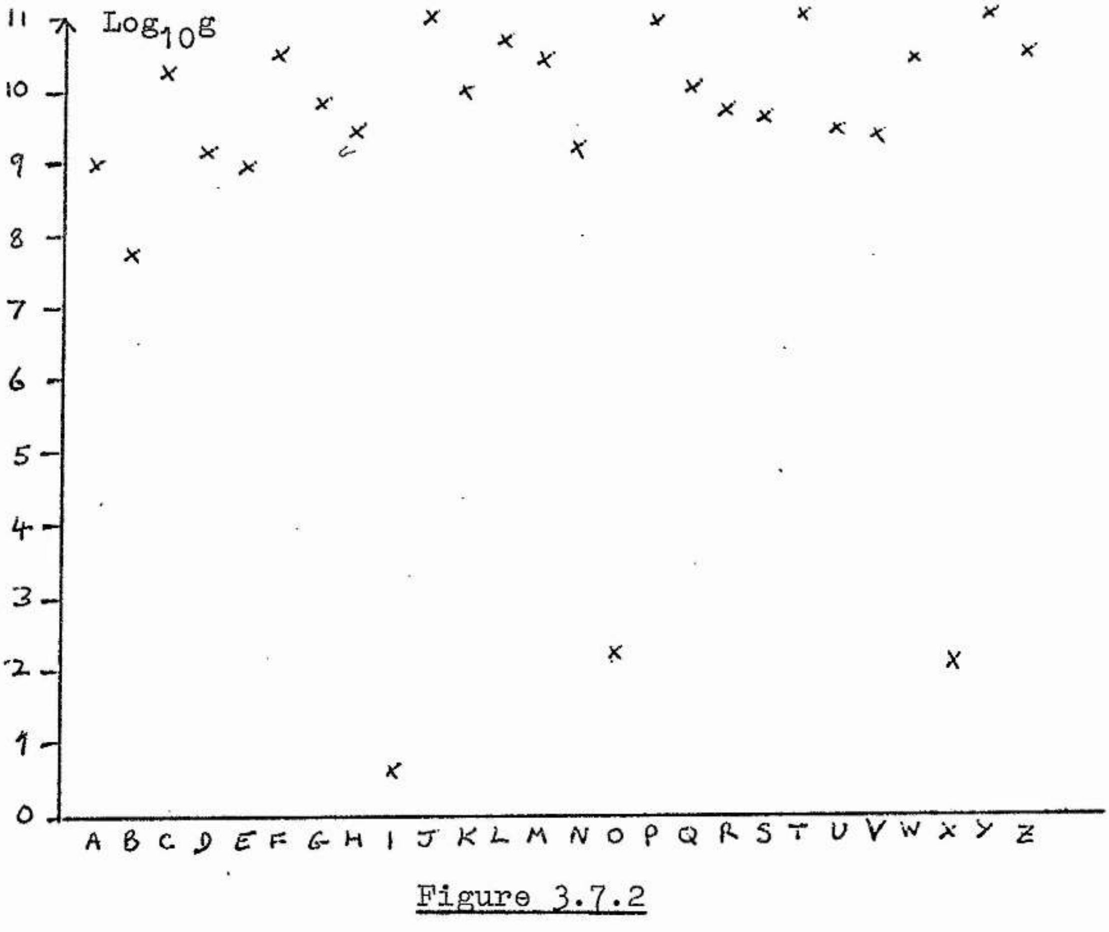


Figure 3.7.2

Recognition of the Letter I.

Figure 3.7.2 shows the resolution of the exponent method applied to the far more difficult letter I, on which the other methods fail. O and X both score fairly low, but the I is still adequately resolved. The reason why O and X are regarded by the computer as being similar to I is because the letter widths are normalized to a standard size. Hence a very narrow letter is expanded into a solid block which then has Geometric Moments very similar to those of O and X. This breadth normalization can be suppressed but was left in for the purposes of this investigation.

(B) Test set on the learning set: Of the 52 hand-drawn letters typed onto the RAX terminal screens, 36 were correctly recognised, an efficiency of 69.4%. There were two of each alphabetic character and the modified Hamming distance using 15 moments was the method chosen. This was judged to be quite a good result although the small size of test set prevents any important conclusions from being drawn.

The use of the IBM tape proved more difficult, as it was in the form of a message, hence the letters had to be categorized by the computer. Also the letter scan had to be translated into a matrix form using a special subroutine. However, once these difficulties had been overcome, the results were very rewarding. In the first 500 characters on the tape there were 384 letters in a 23 letter alphabet (Q,X,Z did not appear). These were processed using a similar moment generating program to the previous one and the means and standard deviations of the letter sets calculated. The combination of the learning set and the decision mechanisms resulted in the efficiencies shown in Table 4.

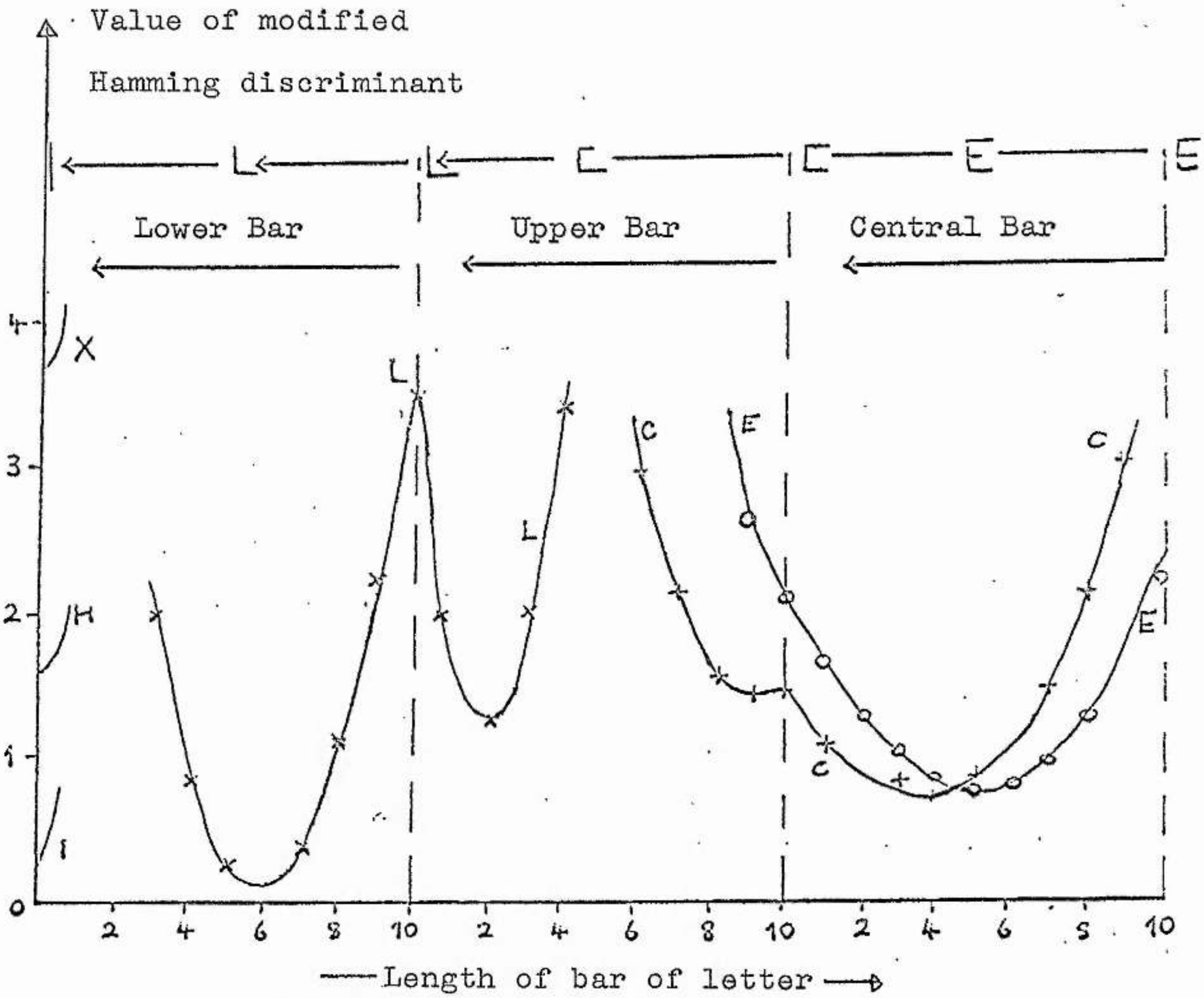
	First 500	Next 2000	First 500	Next 2000
Eff. of Hamming	81.88	85.34	69.04	63.82
Eff. of Euclidean	81.42	83.88	67.43	63.98
Eff. of Mod. Hamming	95.64	91.13	95.64	92.32
Eff. of Mod. Euclidean	96.33	92.75	95.87	92.86
	using 15 moments		using 8 moments	

TABLE 4

The first 500 characters constitute the learning set, whereas the next 2000 constitute the test set. Due to a misunderstanding, it was not until later that it was discovered that the tape message contained 4000 characters, and not 2500, so later results are referred to a larger test set.

The general standard of the efficiencies reflects the quality of the type characters. A breakdown of the forms of error that occurred in this analysis revealed that the largest source of error was a few recurring mistakes such as O for S rather than a general scattering of misrecognitions. Again no attempt was made to apply a threshold of rejection, although this will be considered later.

Finally, figure 3.7.3 shows the results of a test to demonstrate the precise change-over points of the recognition system as a letter was transformed stepwise into other letters. This was done by a) reducing the central arm of the letter E towards the C shape; b) then reducing the upper arm towards the L shape; c) and lastly reducing the lower arm towards the



× Discriminant value for choice L.

+ Discriminant value for choice C.

o Discriminant value for choice E.

Figure 3.7.3

Stepwise Variation of Letter Shape.

I. The effect on the character recognized is shown, the values of the different discriminants being plotted against the length of arm. The cross-over points are clearly defined and also regions are apparent where no character is recognizable with that discriminant at less than two standard deviations.

The best recognition system is one that combines speed of operation with high efficiency. The Modified Hamming method using 8 moments was 87% efficient and fast. However the exponent method was far more efficient but much slower. It is suggested, therefore, that a two layer machine be used,

(i) With the Modified Hamming method to find the three lowest discriminants, or all those lying below a given threshold, say two standard deviations from the mean,

(ii) then using the exponent method to decide between those three characters.

Since only a few characters would be involved in such a calculation, the slower method could be used to advantage. This 2-step process yields 100% recognition in the hand-drawn learning set discussed above. This promising result deserves further study and is the subject of a paper soon to be printed.

CHAPTER 4

THE APPLICATION OF MOMENTS DATA TO CERTAIN THEORETICAL PROBLEMS.

Reference has been made to a number of related variables in Character Recognition; these are efficiency of recognition, error rate, rejection threshold and others. In this chapter an attempt will be made to show the relations that exist between these variables, both at the theoretical and at the experimental level. This latter is accomplished using the moment techniques developed in the last chapter. Finally individual attention is paid to certain isolated problems that have occurred in the use of moments.

4.1 Introduction.

In a recognition strategy involving thresholds, such as that mentioned in paragraph 1.6, a certain amount of theoretical work can reveal the relations that exist between those variables most commonly used:

(i) the error rate E which is the probability that a character will be wrongly recognized, after the rejection of those possibilities that lie beyond the rejection threshold has occurred,

(ii) the rejection rate R which is the probability that a sample character will be rejected given a particular rejection threshold,

(iii) the probabilistic rejection threshold $(1 - \alpha)$ which is that value of the probability of a sample character belonging to a given character set below which rejection occurs,

and (iv) the rejection threshold t which is a simple function of that geometrical distance, between the sample character vector $\underline{X}$ and the mean of the given character set $\underline{\mu}$, beyond which rejection occurs.

In 1957, C.K.Chow published the first [36] of a number of articles [36,37,14] discussing 'the functional relationship of recognition error and rejection trade-off'. He pointed out that part of the optimum recognition rule which minimizes the error rate E for a specific rejection rate R is to reject the pattern if the maximum of the a posteriori probabilities is less than the threshold $(1 - \alpha)$, $(0 \leq \alpha \leq 1)$. Hence the constant α provides a control over the E - R trade-off. He showed that the two are related by the integral

$$E = \int_{R(\alpha)}^{R(0)} \alpha(R)\,dr \qquad 4.1.1.$$

which, although surprising, is a very useful way of measuring the actual recognition errors which can be otherwise unidentifiable.

4.2 The Relation between the Rejection Rate and the Threshold Function.

Suppose the threshold function t is defined as

$$t^2 = (\underline{W} - \underline{\mu})' \, V^{-1} \, (\underline{W} - \underline{\mu}) \qquad 4.2.1.$$

for a given normal distribution, $n(\mu, V)$ and sample vector $\underline{W}$, such that rejection occurs for all values of the exponent distance greater than the scalar t^2. Then the rejection rate may be defined to be

$$R = k \cdot \int_t^{\infty} e^{-q^2/2}\, dq \qquad 4.2.2.$$

where k is a normalizing constant chosen such that the rejection rate tends to unity as t tends to zero. As is well known, $k = \sqrt{2/\pi}$, and the integral can be written in terms of the error function, erf(x) where

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-q^2} dq \qquad 4.2.3.$$

$$\text{and} \quad \operatorname{erfc}(x) = 1 - \operatorname{erf}(x) \qquad 4.2.4.$$

$$\text{Then} \quad R = \operatorname{erfc}(t/\sqrt{2}) \qquad 4.2.5.$$

This is the theoretical relation between the rejection rate and the threshold function.

4.3 The Relation between Error Rate and Rejection Rate.

As can be seen from equation 4.1.1, the differential ratio dE/dR is given by

$$-\frac{dE}{dR} = 1 - \alpha \qquad 4.3.1.$$

(see figures 4.3.1 and 4.3.2).

If α is hypothesised as being

$$\alpha = AR^n \qquad 4.3.2.$$

where A and n are constants, then equation 4.3.1 is integrable.

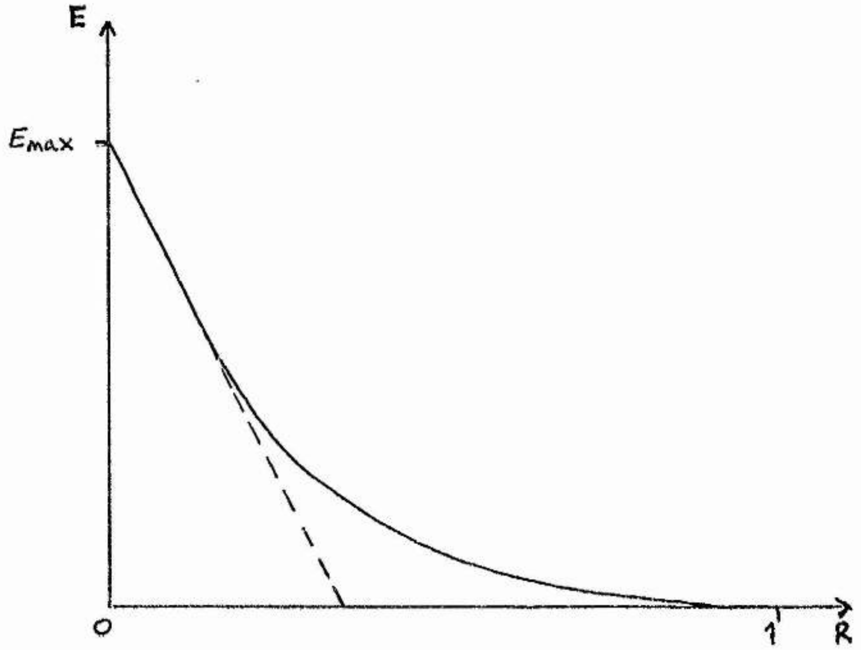


Figure 4.3.1.

Knowing that R = 0 when E = E_{max}, the result of integration becomes

$$E + R - E_{max} = A/(n + 1) .\ R^{n + 1} \qquad 4.3.3.$$

which may be tested graphically (see figure 4.3.3). Straight lines were obtained from the data provided, indicating the validity of the hypothesis.

This hypothesis can be put on a more substantial level by a little mathematics. The differential form of equation 4.2.2 is

$$\frac{dR}{dt} = - \alpha \qquad 4.3.4.$$

since α and t have been defined so that $\alpha = ke^{-t^2/2}$.

Or $$- \alpha = \frac{dR}{d\alpha} . \frac{d\alpha}{dt}$$

Hence $$\frac{dR}{d\alpha} = - \alpha \ \frac{dt}{d\alpha} \qquad 4.3.5.$$

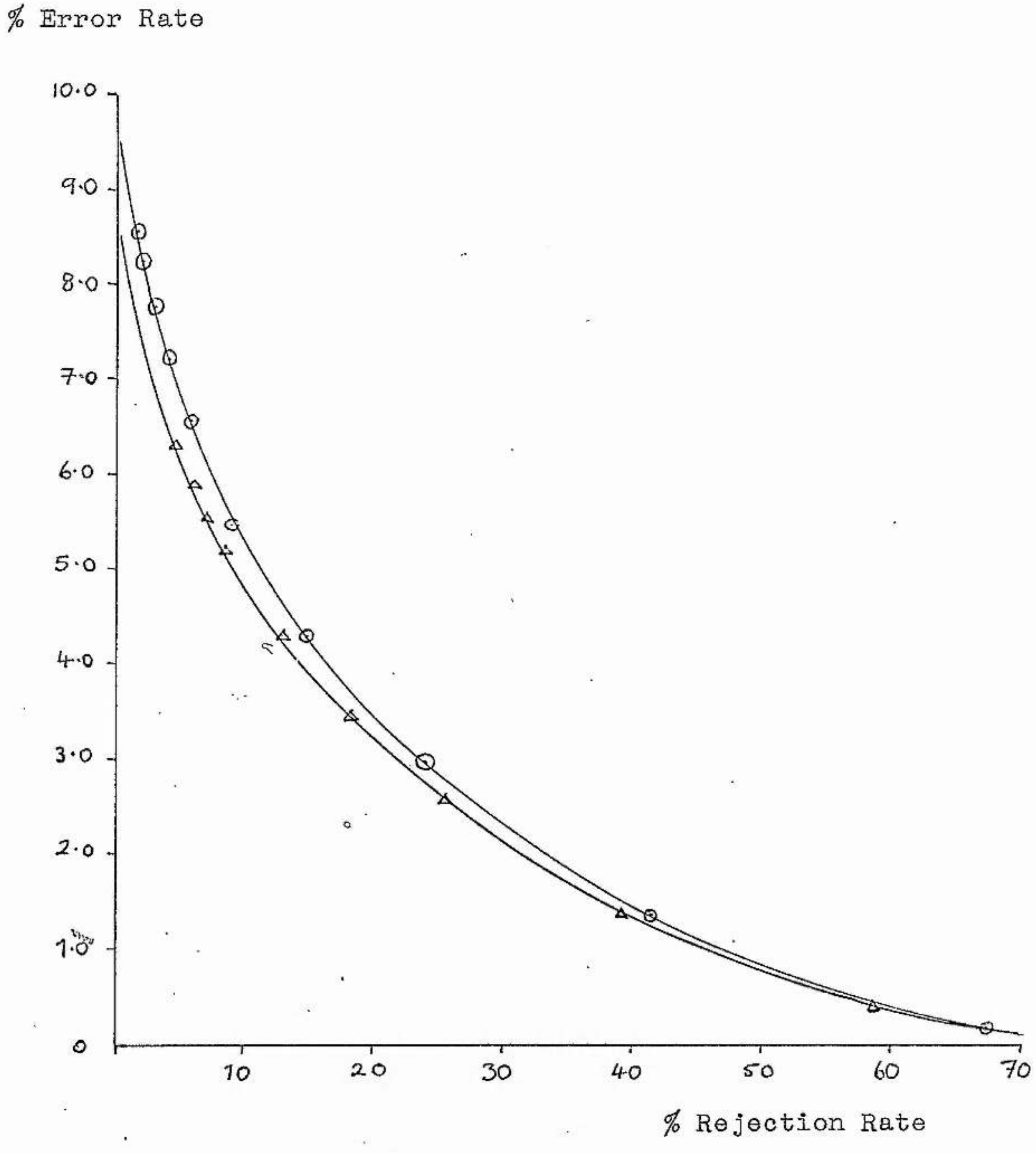


Figure 4.3.2

Graph of Variation of Error Rate with Rejection Rate.

⊙ Modified Hamming distance measure on 4000 letters, 8 moments.

△ Modified Euclidean distance measure on 4000 letters, 8 moments.

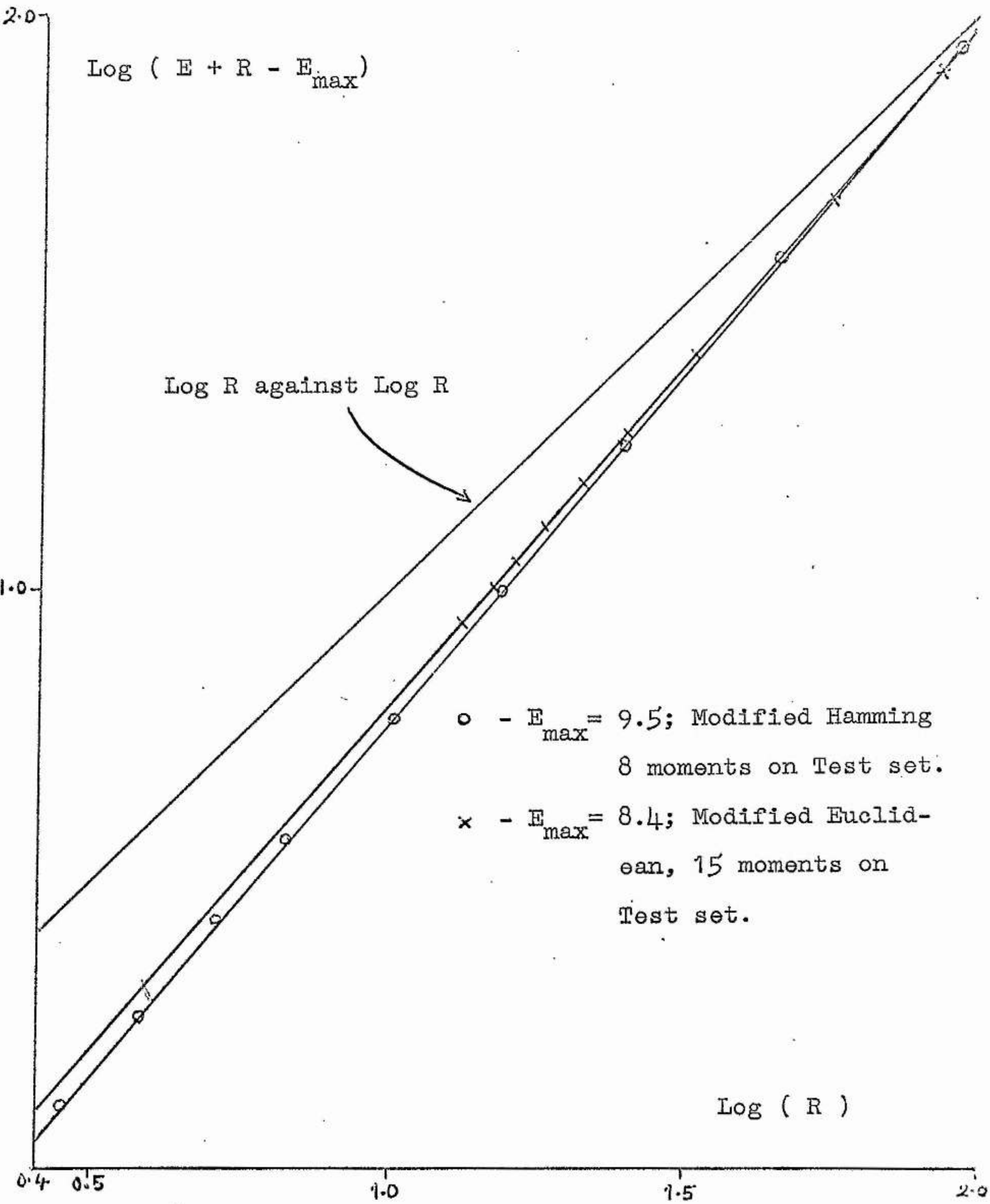


Figure 4.3.3

Graph of Log (E + R - E_{max}) against Log R.

But from equation 4.3.1.

$$-\frac{dE}{dR} = 1 - \alpha$$

So $$\frac{dE}{d\alpha} = \frac{dE}{dR}\cdot\frac{dR}{d\alpha} = +\alpha(1-\alpha)\,\frac{dt}{d\alpha} \qquad 4.3.6.$$

Hence, from equation 4.3.5

$$[R] = \int -\alpha\,dt \qquad 4.3.7a$$

and from equation 4.3.6

$$[E] = \int \alpha(1-\alpha)dt$$

$$= -[R] - \int \alpha^2 dt \qquad 4.3.7b$$

When $t = 0$ $R = 1$, $E = 0$ and

when $t = \infty$ $R = 0$, $E = E_{max}$ and

when $t = t$ $R = R_1, E = E_1$

$$\therefore \quad [E]_{E_1}^{E_{max}} = -[R]_{R_1}^{0} - \int_{t_1}^{\infty}\alpha^2 dt \qquad 4.3.8.$$

or $$E_{max} - E_1 - R_1 = -\int_{t_1}^{\infty}\alpha^2 dt$$

or $$(E + R - E_{max}) = \int_{t}^{\infty}\alpha^2 dt \qquad 4.3.9.$$

omitting the subscripts.

Now $$\int_t^{\infty}\alpha^2 dt = \int_t^{\infty}\frac{4}{2\pi}e^{-t^2}dt = \frac{1}{\sqrt{\pi}}\,\mathrm{erfc}(t)$$

and $R = \mathrm{erfc}(t/\sqrt{2})$ from equation 4.2.5.

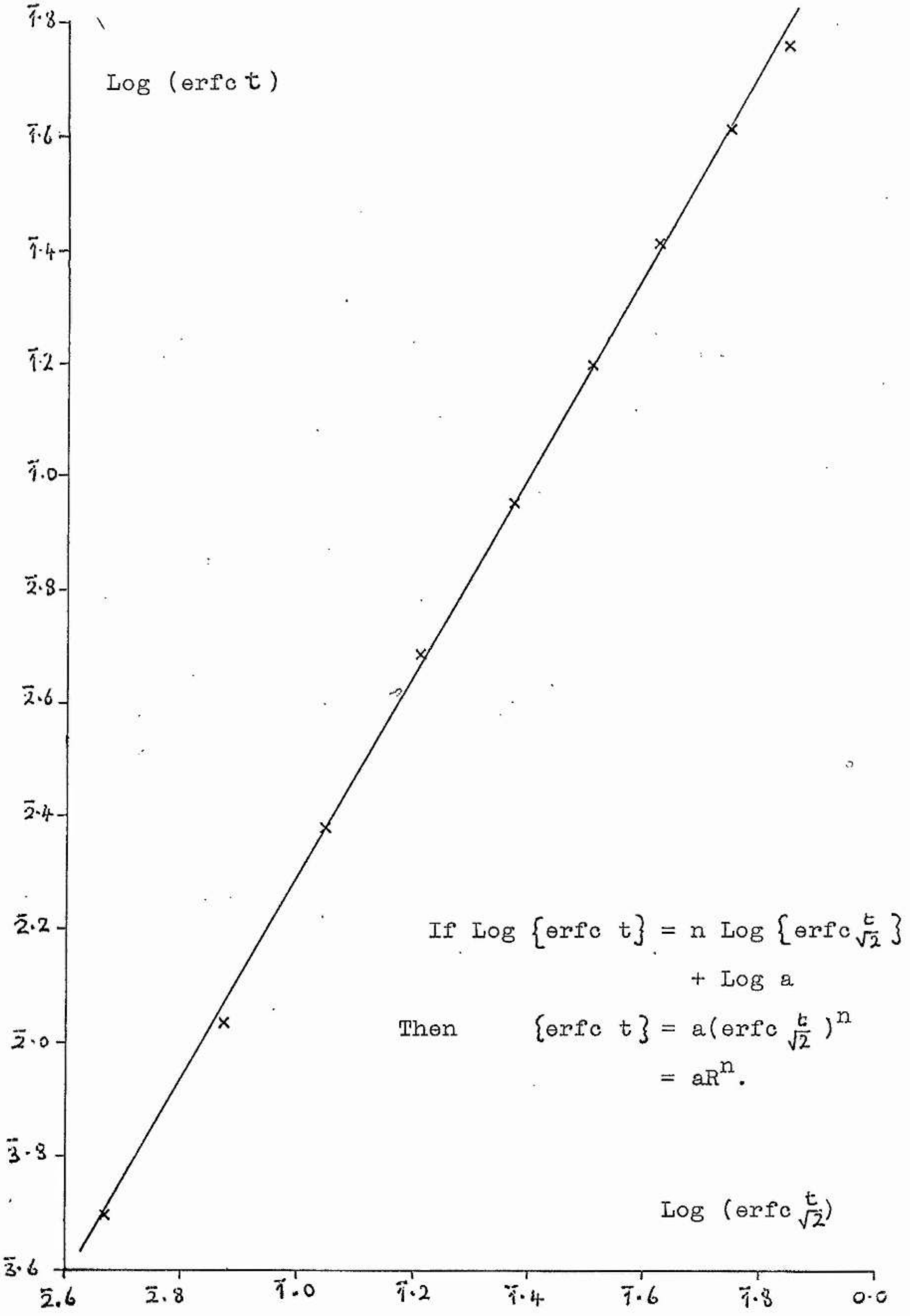


Figure 4.3.4

Test for theoretical relation between Rejection Threshold and Rejection Rate.

It so happens that AR^n is a fair approximation to $1/\sqrt{\pi}\cdot\mathrm{erfc}(t)$ within the range in which we are interested i.e. $0.4 \leq t \leq 2.4$ (see figure 4.3.4), and so the hypothesis has some strong theoretical backing.

4.4 The Variation of Efficiency with Number of Moments Used.

The graphs of efficiency variation using different decision mechanisms (see figure 3.5.2) seem to require some theoretical explanation. Sampling and information theories being both highly sophisticated,[17] it seems reasonable to attempt applying them to this problem of efficiency variation. In fact Bowman and McVey [38] suggest a similar approach as an ad hoc hypothesis.

Suppose we have a block of data out of which samples are drawn, measured and replaced (see figure 4.4.1). How much information is probably withdrawn after n draws? Suppose an equal amount of information is drawn every time, containing m bits of data. The set containing all the information withdrawn at the nth draw is A_n. Then if g_{n+1} is the amount of data withdrawn at the (n + 1)th. draw, then

$$A_{n+1} = A_n \cup g_{n+1} \qquad 4.4.1.$$

If $\mu(A_n)$ is the volume of data in A_n, it follows from equation 4.4.1 that

$$\begin{aligned}\mu(A_{n+1}) &= \mu(A_n \cup g_{n+1})\\ &= \mu(A_n) + \mu(g_{n+1}) - \mu(A_n \cap g_{n+1}) \qquad 4.4.2.\end{aligned}$$

If the expectation of $\mu(A_n)$ averaged over n is x_n, it follows that

$$x_{n+1} = x_n + m - \langle\langle \ldots \langle A_n \cap g_{n+1} \rangle_1 \cdots \rangle_n \rangle_{n+1} \qquad 4.4.3.$$

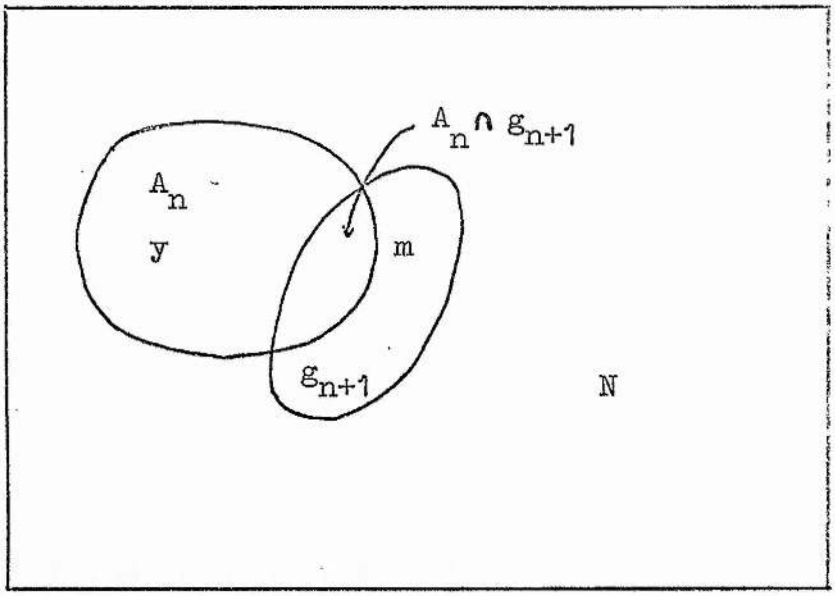


Figure 4.4.1

Set Diagram for Calculation of $A_n \cap g_{n+1}$.

N	Total Number of Bits of Data.
A_n	Information withdrawn after nth. Draw.
y	Average Volume of Data in A_n.
g_{n+1}	Information withdrawn at (n+1)th. Draw.
m	Volume of Data in g_{n+1}.

The last term of equation 4.4.3 is the sum over all intersections of A_n and g_{n+1}:

$$S = \frac{\sum_{r=0}^{m} r.{}^{y}C_r.{}^{N-y}C_{m-r}}{{}^{N}C_m} \qquad 4.4.4.$$

(see appendix 8), where $y = x_n$; N = total number of bits of data in the block. Now, remembering that

$$\sum_{r=0}^{m} {}^{y}C_r.{}^{N-y}C_{m-r} = {}^{N}C_m \qquad 4.4.5.$$

(see appendix 9). We can write r as y - (y - r) in equation 4.4.4 to obtain

$$S = \frac{\sum_{r=0}^{m} y.\ {}^{y}C_r.{}^{N-y}C_{m-r}}{{}^{N}C_m} - \frac{\sum_{r=0}^{m} (y - r).{}^{y}C_r.{}^{N-y}C_{m-r}}{{}^{N}C_m}$$

$$4.4.6.$$

$$S = y - \frac{\sum_{r=0}^{m} y.\ {}^{y-1}C_r.{}^{[(N-1) - (y-1)]}C_{m-r}}{{}^{N}C_m}$$

Thus $S = y.m/N$ 4.4.7.

Hence equation 4.4.3 simplifies to

$$x_{n+1} = x_n + m - x_n.m/N$$

$$x_{n+1} = x_n(1 - m/N) + m \qquad 4.4.8.$$

Using the method of induction, we know:

$x_0 = 0$; $x_1 = m$

and if we write $x_{n+1} = Ax_n + B$ where $A = 1 - m/N$, $B = m$

Then

$$x_{n+1} = A^{n+1}x_0 + B \sum_{r=0}^{n} A^r \qquad 4.4.9.$$

$$x_{n+1} = m. \frac{(1 - A^{n+1})}{(1 - A)} = m. \frac{1 - (1 - m/N)^{n+1}}{m/N}$$

Thus $x_{n+1} = N(1 - (1 - m/N)^n)$ 4.4.10.

4.5 The Experimental Validity of the Hypothesis of Efficiency Variation.

Having outlined the hypothesis that the volume of information gained on the (n+1)th. draw is $(1 - a^n)$ times the total amount of information available in Section 4.4, it is a short step to equating the efficiency of a decision process with that fraction of information extracted from the total available; that is, that the efficiency is

$$Q = (1 - a^n).\ 100\% \qquad 4.5.1.$$

where a is some constant equal to 1- that fraction of the total information available contained in one moment. The assumption is made here, it must be remembered, that each moment contains an equal amount of information.

Straight line graphs of Log(100 - Q) against n will test the hypothesis. The tables 5 and 6 below show the results of this analysis (see also figure 4.5.1).

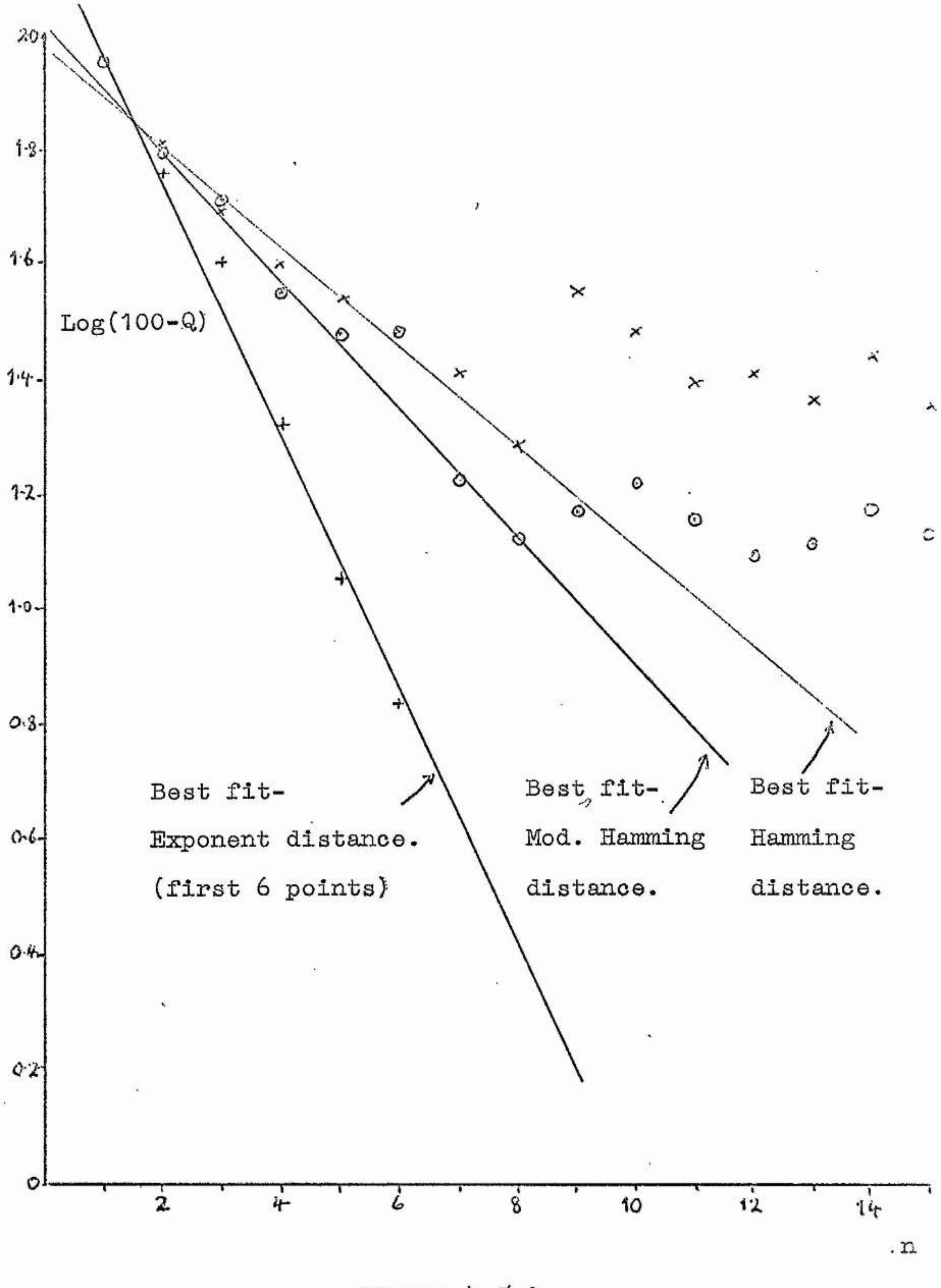


Figure 4.5.1

Graph of Log(100 - Q) against n to show

Cumulative Information Variation with number of moments used.

Calculation for first 8 moments.

	m	σ_m	1-a	C	σ_C	2-m**
Euclidean	-.0864	.0040	.180	1.979	.020	2.086
Hamming	-.0869	.0047	.181	1.980	.024	2.087
Mod. Euclidean	-.1237	.0097	.248	2.005	.049	2.124
Mod. Hamming	-.1119	.0085	.227	2.033	.043	2.112
Exponent***	-.228	.0109	.408	2.21	.055	2.23

TABLE 5

Least Mean Squares* Fit to Moments Data.

Calculation for 15 moments.

	m	σ_m	1-a	C	σ_C	2-m**
Euclidean	-.0264	.0073	.059	1.753	.066	2.026
Hamming	-.0328	.0067	.073	1.772	.061	2.033
Mod. Euclidean	-.0701	.0086	.149	1.826	.078	2.070
Mod. Hamming	-.0571	.0078	.123	1.803	.071	2.057
Exponent***	-.294	.0077	.492	2.387	.070	2.294

TABLE 6

Least Mean Squares* Fit to Moments Data.

* see Appendix 10

** see equation 4.5.4

*** Exponent figures refer to 6 and 7 moments, as opposed to 8 and 15, as these were the maximum efficiencies which were still relevant to this calculation.

C and m are the intercept and gradient respectively of the theoretically straight line graph $\text{Log}(100 - Q) = mx + C$, where x is the number of moments used. Theoretically $C \sim 2.00$ and $m = \text{Log}(a)$.

To find the theoretical value of C, we inspect equation 4.4.10.

$$\frac{x_{n+1}}{N} = 1 - (1 - m/N)^n$$

By letting $1 - m/N = a$ and $100.\frac{x_{n+1}}{N} = Q$, we obtain

$$Q_{n+1} = 100(1 - a^n) \qquad 4.5.2.$$

$$\text{Log}(100 - Q_n) = \text{Log } a^{n-1} + \text{Log } 100$$

$$= 2 + (n - 1)\text{Log } a$$

or $$\text{Log}(100 - Q_n) = (2 - \text{Log } a) + n\text{Log } a \qquad 4.5.3.$$

Hence the intercept $C = 2 - \text{Log } a$

and $m = \text{Log } a$ 4.5.4.

Theoretically, then $C = 2 - m$.

The graphs drawn show that the assumption of equal information content per moment was not very valid. The best straight line came from the exponent mechanism as expected, and the gradient displays the most information extracted per moment, and the intercept agrees closely with that predicted. From the standard deviations of C and m, it can be stated that the Hamming and Euclidean methods seem about equal in performance, whereas the modified methods are themselves equivalent though reliably better than the unmodified forms.

The form of the graphs suggests that at least for small numbers of moments, the approximation assumed holds good but breaks down the more moments used. Some reasons why this breakdown might occur are suggested in the next Section.

4.6 The Resolving Power of Moments.

Throughout the work on Character Recognition covered in Chapter 3, no attempt was made to use any special set of moments, even though Franz Alt in his research[30] selected particular moments for a decision tree. Since it was not clear that the mechanisms that were being used here demanded special moments, no such ordering was investigated. Indeed the problem of ordering is by no means simple. There is a criterion attributed to R.A.Fisher that defines the resolution R_k between two normal distributions $n(\mu_1, \sigma_1^2)$ and $n(\mu_2, \sigma_2^2)$, of the kth. element of the pattern vector so that

$$R_k^2 = (\mu_{1k} - \mu_{2k})^2 \ (1/\sigma_{1k}^2 + 1/\sigma_{2k}^2) \qquad 4.6.1.$$

This is easily extended to n distributions

$$R_k^2 = \sum_{i,j=1}^{n} \frac{(\mu_{ik} - \mu_{jk})^2}{\sigma_{ik}^2} \cdot \frac{1}{n(n-1)} \qquad 4.6.2.$$

The calculation of the R_k is shown below (table 7) for the learning set of the tape data taken over 23 alphabetic character sets.

The larger the value of R_k the better the moment is at resolving the 23 letter sets. However the emphasis is on the letters that the moment resolves well, since equation 4.6.1 has a large value for two well-separated sets. In order to emphasise those letter set pairs which do not resolve well, a root mean square reciprocal can be implemented

$$R_{k}'^{\,2} = 1 \Bigg/ \sum_{j,i=1}^{n} \frac{n(n-1)\sigma_{ik}^{2}}{(\mu_{ik}-\mu_{jk})^{2}} \qquad 4.6.3.$$

Moment	M_{30}	M_{21}	M_{40}	M_{04}	M_{31}
R_k	6.51	8.27	3.21	4.12	5.77
Moment	M_{13}	M_{12}	M_{03}	M_{50}	M_{14}
R_k	6.30	6.86	7.91	7.29	6.83
Moment	M_{22}	M_{05}	M_{41}	M_{32}	M_{23}
R_k	4.09	8.35	10.58	5.22	7.01

TABLE 7

Table 8 gives the values for R_k'.

Moment	M_{30}	M_{21}	M_{40}	M_{04}	M_{31}
R_k'	4.08	8.46	8.98	4.71	4.05
Moment	M_{13}	M_{12}	M_{03}	M_{50}	M_{14}
R_k'	3.62	3.59	5.16	5.27	3.56
Moment	M_{22}	M_{05}	M_{41}	M_{32}	M_{23}
R_k'	14.6	7.32	2.02	15.6	12.4

TABLE 8

The smaller the value of R_k', the better the moment is at resolving the 23 letter sets. An ordered comparison of R_k, R_k' and R_k(Alt) displays the best moments (see table 9).

R_k(Alt)	M_{30}	M_{21}	M_{40}	M_{04}	M_{31}	M_{13}	M_{12}	M_{03}	M_{50}	M_{14}					
R_k	M_{41}	M_{05}	M_{21}	M_{03}	M_{50}	M_{23}	M_{12}	M_{14}	M_{30}	M_{13}	M_{31}	M_{32}	M_{04}	M_{22}	M_{40}
R_k'	M_{41}	M_{14}	M_{12}	M_{13}	M_{31}	M_{30}	M_{04}	M_{03}	M_{50}	M_{05}	M_{21}	M_{40}	M_{23}	M_{22}	M_{32}
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

TABLE 9

Considering the best ten in each set, only M_{23} and M_{21} occur in the list of R_k and not in that of R_k', and only M_{04} and M_{31} occur in the list of R_k' and not in that of R_k. So the two ways of deciding which moments resolve best favourably agree. Also most of the moments in the first ten in each set are contained in the first ten of R_k(Alt).

These methods are only two of many possible ways of assessing the resolving power of different moments. As such, they represent only an introductory comparison. However, without more statistical evidence, it is probably not worthwhile going any deeper into the problems of resolution.

CHAPTER 5

THE IMPROVEMENT OF RECOGNITION EFFICIENCY AND THE MINIMIZATION OF LOSS.

The basic theory of decision analysis and loss functions has been outlined in Chapters 1 and 2, and a method for Character Recognition is suggested in Chapter 3. Some thought will now be devoted towards combining the two to observe what effect, if any, the introduction of loss functions has on the recognition efficiency.

5.1 Recognition with Minimum Expected Loss.

It should be made clear that the maximum recognition efficiency of a one-step categorizer is not improved by the introduction of loss functions as defined herein. Neither is the addition of a priori probabilities to the decision mechanism of any use when the characters presented to the categorizer themselves do not obey such probabilities. If however the sample characters are in fact taken from a standard English text, for example, then the presence of a priori probabilities may be expected to improve the recognition efficiency. But the use of loss functions is a totally different affair. Their introduction only becomes valid when one character or decision is more important than another in terms of information content or value or whichever utility the operator cares to adopt.

One elementary example of minimum expected loss is the case of optimizing the rate of transmission of information through a channel. A lot of work has been done in the field because of its industrial applications, but it suffices to mention here that the information loss due to the misrecognition of the jth. letter as the ith. letter is

$$I = - \ln \omega_j \qquad 5.1.1.$$

where ω_j is the a priori probability of the jth. letter occurring. Then, in the author's opinion, the information loss may be adopted as the loss function $L(i|j)$, and the calculation of the minimum expected loss becomes equivalent to that of the maximum information transmission rate.

The improvement on efficiency by using not just the a priori probabilities but also digraph and even trigraph probabilities should be quite considerable, especially nowadays when large storage memory banks have become available with very fast access times. This reduces substantially the cost of maintaining large matrices in core storage. A.Wood Edwards and Robert L. Chambers[39] pioneered this work, although they did not have a first class recognition system on which to apply their ideas. Indeed recognition efficiencies may be improved by a number of totally independent methods, and the use of digraphs in a noisy or otherwise suboptimal categorizer may be regarded as probably one of the best, given the required volume of computer memory.

5.2 The Forms of Error in a System.

In the type of recognition apparatus used by the author, it was a fairly simple procedure to analyse the errors produced, to determine exactly how a particular error occurred, and what could be done to redesign the categorizer to prevent that error from recurring. This analysis was carried out on the exponent method, as this gave a clearer picture of the underlying faults of the recognizer. Tables 10 and 11 show the results.

Act	Rec	D_2^*	D_1^*	** d(1)	*** $n<2\sigma$	comments	** d(Act)
Y	E	1.38	2.30	4.39	0		
I	L	1.95	1.95	0.47	>5	similar	0.65
N	B	1.35	1.35	0.52	>5	similar	0.61
W	E	1.75	>2.32	2.19	0		
R	M	5.52	5.52	0.20	4	similar	0.52
K	E	1.40	1.40	2.97	0		
Y	T	1.16	2.16	4.48	0		
V	S	1.18	>1.62	2.80	0		
W	O	1.54	>2.22	2.09	0		
Y	E	1.13	1.13	2.87	0		
I	L	4.69	4.69	0.25	>5	similar	0.54
W	O	1.19	>1.61	2.19	0		
V	S	1.06	>1.62	2.79	0		
Y	E	1.79	2.10	3.10	0		
V	S	1.22	>1.65	2.76	0		

TABLE 10

Explanatory notes to Table 10:

Act - Actual Letter

Rec - Recognized Letter

* D_2 is defined as $d(2)/d(1)$; and D_1 as $d(\text{actual})/d(1)$

** d is the exponent distance between the sample character and the mean of a given letter set. Since these were ordered, d(1) is the smallest exponent distance and applies to the chosen letter set. d(actual) is the exponent distance to the actual correct distribution mean. Units are standard deviations

$$d = 1/n^{2}{}'(\underline{X} - \underline{\mu})V^{-1}(\underline{X} - \underline{\mu})$$

n = number of dimensions in pattern vector

*** $n<2\sigma$ signifies the number of letter sets with exponent distances less than 2σ.

Out of 500 characters on the IBM tape, there were 436 alphabetic characters of which 15 were recognized incorrectly by this method. Only 4 of these would have been misrecognized if a 2σ threshold had been imposed, whereas only 2 correctly recognized letters would have been rejected. Hence with a 2σ rejection threshold, the recognition efficiency would rise from 96.56 to 99.08% with a 2.98% rejection rate. Thus the introduction of a rejection threshold, if not expensive to implement, produces a considerable increase in efficiency in this method. Furthermore, from the table above, it is apparent that such a rejection threshold is not critical, and the graph 5.2.1 shows how slowly the efficiency changes as the threshold varies.

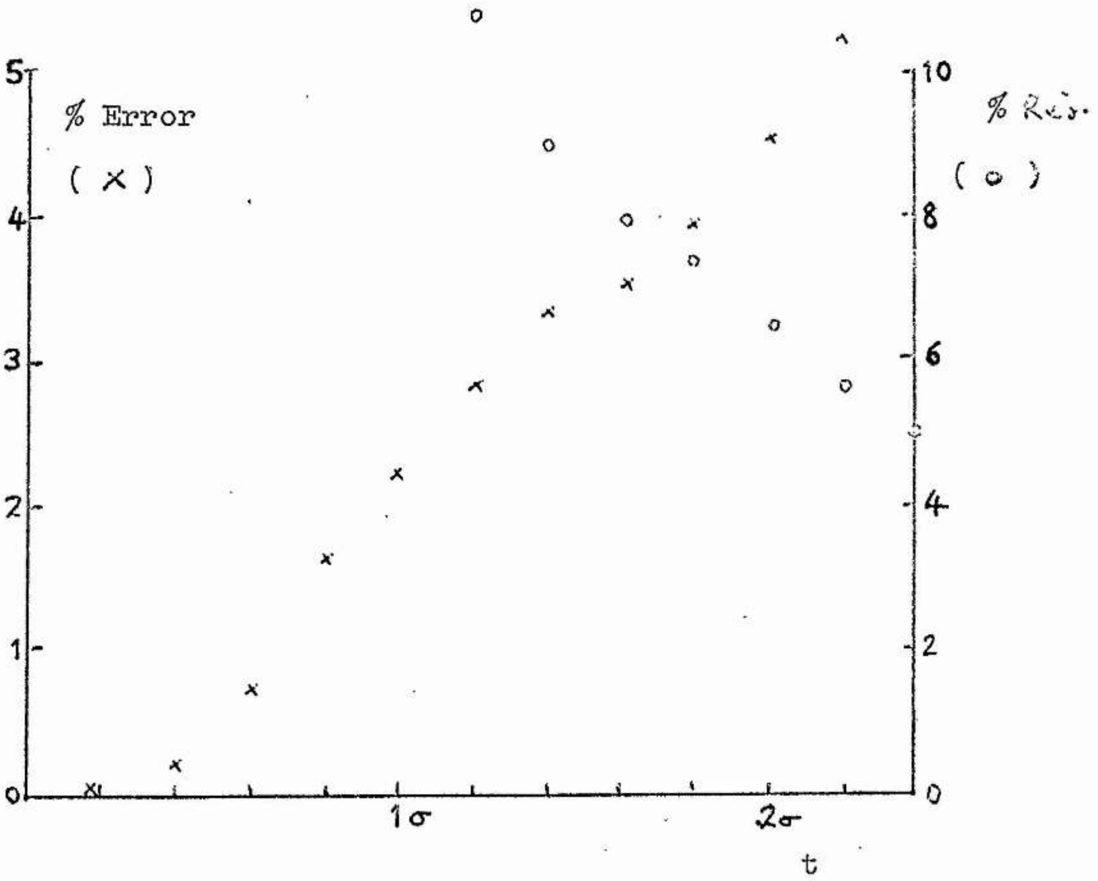


Figure 5.2.1

Variation of Error and Reject Rate with t.

```
  X                 X
  X                XX
  X                XX
                   XX
                   XX
  X                XX
 XX               XXX
 XX               XXX
 XX               XXX
XXX               XXX
XXX               XXX
XXX               XX
XXX               XX
XXX              XXX
XXX              XXX
XX               XXX
XX               XXX
XX               XXX
```

Figure 5.2.2

Comparison of Tape Set I and L.

The characters incorrectly recognized which are more than 2σ from the nearest mean are Y,K,V and W, the covariant matrices and means of which have been calculated on only about 4 samples, so that this reason alone may be sufficient to explain the frequency with which they have been misrecognized.

The characters incorrectly recognized which are not more than 2σ from the nearest mean are more interesting. The two I's which were recognized as L's can be dismissed (see figure 5.2.2), leaving the N - B and R - M as the only two real errors in the whole of the learning set. As these two letters display in no obvious way whatsoever any real difference from any of the other N's or R's that occur in the message, they must be regarded as computing errors or due to the presence of excess noise on the tape (which was otherwise singularly free from noise), and probably do not reflect upon the decision mechanism.

Total	M-N*	M-E	L-I	F-I	I-L	O-E	C-E	Tape errors
162	13	16	13	8	7	11	14	17

* M-N signifies M recognized as N

TABLE 11

Breakdown of main errors in 4000 character set with a 2σ rejection threshold.

A number of errors occurring in both learning and test sets can be attributed to capitals, which should be rejected in most cases. Also the check against which the computer-recognized letters were identified, that was a built-in part of the tape data, was not always correct; this is the reason behind the 17 tape errors. The lack of better defined distributions for letters like M,C and F are emphasised in Table 11. Two confusion matrices (figures 5.2.3 and 5.2.4) have been included to present the error display for the whole test set. The first of these was calculated without any threshold whereas the second contained a recognition threshold of 2σ.

By far the most useful possibility that can solve these assorted problems is to introduce a two-stage decision process. The first stage of this can be a simple modified Hamming distance type recognizer with a built-in rejector of either (i) letters more than a given distance from any letter mean, or (ii) letters in which the choice between the nearest and next nearest mean is insufficiently distinct. The second stage can then be the more complex though far more accurate exponent mechanism. More will be said about the virtues of a two-stage method in the following section.

It is clear that the errors are not randomly distributed throughout the test set, but heavily concentrated about particular letter pairs. Such pairs, once identified, can be weighted within the strucure of the loss function to correct those particular faults that occur due to the nature of the categorizer. Discrimination between similar letters, such as I - L, is less amenable to this treatment, but probably still of some value.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	R	S	T	U	V	W	Y
A	260	1	1		4				1						1								
B		27			21				5														
C			138		7				1														
D				138	2														1				
E					488									2	3								
F					1	56			2										6				
G					1		72																
H					2			156															
I	7								276														
J										2													
K		1			12						1												
L	1								13			137			2			2	7				
M		2			24								27	26									
N		1												256									
O	2				11										260				1				
P																96							
R													1				201		6				
S									1									227					
T																			319				
U					2										1					96			
V	1				1													30			1		
W					7										27							1	
Y					38																		8

Figure 5.2.3

Confusion Matrix for Test Set Errors

with no Rejection Threshold.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	R	S	T	U	V	W	Y
A	260														1								
B		27												1									
C			138		7																		
D				138															1				
E					488										3								
F						56													8				
G					1		72																
H								156															
I	5								276														
J										2													
K											1												
L	1								13			137			1			1	5				
M		1			15								27	10									
N		1												256									
O	2				11										260				1				
P																96							
R													1				201		6				
S									1									227					
T																			319				
U					1															96			
V																					1		
W					1																	1	
Y																							8

Figure 5.2.4

Confusion Matrix for Test Set Errors

with 2σ Rejection Threshold.

5.3 The Relation between Execution Time and Efficiency.

Whereas no absolute measure of execution time can be used to determine the cost of a particular technique in decision making, the difference between various mechanisms can be restricted in computer simulation to little more than the altering of a line or two in the entire program. In this way all the mechanisms which have been applied by the author can be compared, not only in recognition efficiency but also in execution time.

This time, for the exponent mechanism, was proportional to the square of the number of moments used, as expected. This makes it considerably slower to use although its greatly increased efficiency counteracts this drawback to a certain extent. From the graphs of efficiency against moment number, it can be seen that the maximum efficiency is reached when around eight moments are being used and that further moments alter the recognition efficiency little.

One possible two-step procedure is to reject characters further than 2σ from a mean as measured on the Hamming distance and recalculate these distances according to the exponent mechanism. Another more acceptable two-step procedure is to order, in increasing distance, the letter set means using the Hamming distance as before; then to select the three or four with the smallest distances and calculate the exponent distances to those means. This reduces the number of letter sets over which the exponent measure has to be applied.

It is interesting to plot a graph of execution time against percentage efficiency with the number of moments used as the variable parameter and to superimpose lines of the family (see figure 5.3.1)

$$(100 - E)t = \text{constant } (C) \qquad 5.3.1.$$

Naturally this family is chosen rather arbitrarily, being after all only the first term in the general expansion

$$(100 - E) = \sum_{i=1}^{\infty} a_i t^{-i} \qquad 5.3.2.$$

However it does weight the error rate equally with the execution time. From the graph it is possible to decide on the decision mechanism one wishes to use by implementing a given error/execution time measure (C in equation 5.3.1). The choice of C is, of course, related to the time equivalent to a given error rate, and hence to an estimate of how much extra time one incorrect letter will introduce. This is something only the designer can decide.

5.4 Forms of Loss Matrix.

It was mentioned in Section 5.1 that the operator could define a utility function by which certain characters could be assigned relatively more importance than average. It is suggested that, for a machine categorizer which rejects alphabetic characters to be subsequently deciphered by a human being, it is possible to weight the decision mechanism in such a way as to optimise this two-stage system. For instance, certain letters are geometrically similar, for example R and K. If the loss involved in a misrecognition of a character as another geometrically similar one is increased, then this is equivalent to reducing the rejection threshold <u>in this particular</u>

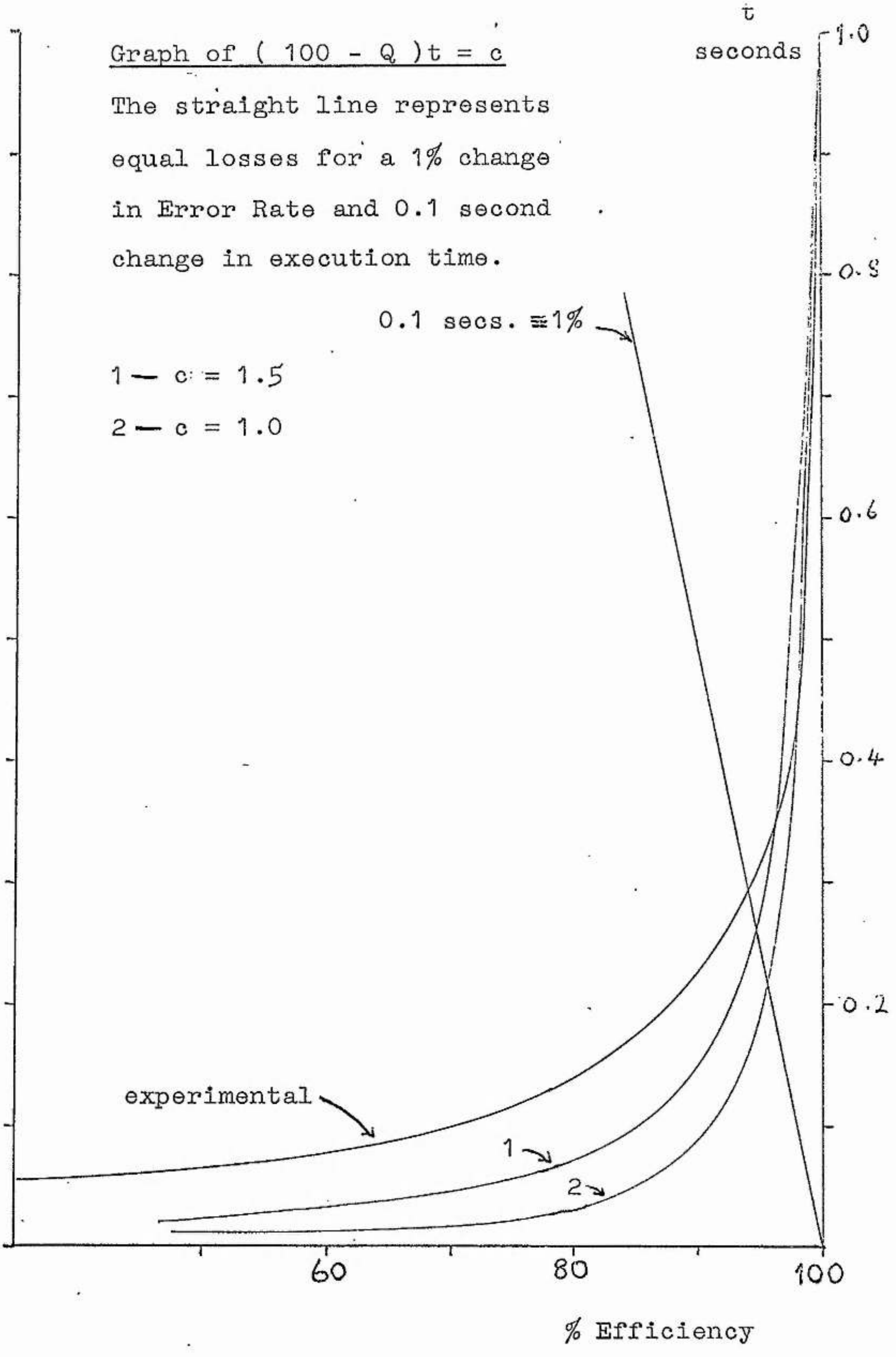


Figure 5.3.1

Variation of Execution Time with

Recognition Efficiency.

discrimination, and so the probability of rejection is increased. This is simple theoretically, but implementation is more complex. A quantitative measure of similarity (either geometrically or whatever the particular method of discrimination finds similar) needs to be introduced. The Mahalanobis distance[40] is a likely choice for this, although others may be suitable[41, 42]. Also if letters are misrecognized as being rare letters, S as Z for example, then these will be relatively more obvious to the checker, who relies more upon the context and meaning of entire words rather than upon individual characters as does the machine.

Combining these two suggestions for optimisation with the information loss suggested in Section 5.1, a loss matrix L may be formed which is related to the features mentioned above by the equation

$$L = IMS \qquad 5.4.1.$$

where (i) I is a loss matrix based on the higher costing of rarer letters because of information loss; (ii) M is a loss matrix based on the lower costing of the rarer misrecognition letters, i.e. those that the categorizer chooses which are incorrect, due to the fact that they are more obvious to a checker or reader; (iii) S is a loss matrix based on the higher costing of similar letter pairs so that the probability of rejection of this choice is increased.

This latter part is symmetric, whereas the first two are asymmetric matrices. It can be suggested, then, to break up the loss matrix into a product of three parts (i) an information loss; (ii) a misrecognition loss and (iii) a similarity loss. The first part may be omitted if the type of problem does not warrant its inclusion, for example non-textual material.

CONCLUSION.

An introduction to elementary Decision Theory has been given to explain the concept and definition of loss functions and to relate them to Bayes Theory and Optimal Boundaries. This also served to collate data and papers on the properties of losses. A few simple examples followed to show that in some cases loss functions can become important in decision analysis. An attempt was then made to compare the minimum expected loss with the loss using a suboptimal decision boundary. In the simple case chosen, the expected losses differed by only 0.5%, indicating favourably the implementation of the Euclidean distance mechanism as an alternative to the optimal method.

The work on Geometric Moments has covered recognition of hand-drawn capitals with error rates varying from 88% for very poor, elementary methods to <1% for very good methods. Also an IBM message tape was employed for character recognition using the first 500 characters as a learning set, and the whole 4000 characters as a test set. The results from these confirmed the difference in quality between the different mechanisms. These can now be ordered in increasing resolution:

(i) The Hamming and Euclidean distances (equivalent within the margin of error)

(ii) The Modified Hamming and Euclidean distances

(equivalent within the margin of error)

(iii) The exponent distance.

The modification involved has the effect of normalizing different moments, with the result that they become equally weighted in the distance measure. Error rates on the tape varied between 9.5% and 3.6% for methods (ii) and (iii) mentioned above. Various minor improvements could have been made on these figures, in particular to have thrown out all capital letters in the learning and test sets and to have increased the size of test set so that the more uncommon letters would have had better defined distributions.

The study of the moments has led to the development of the two-stage categorizer: the first stage utilizing a fast discriminator and the second stage utilizing a slower but more efficient mechanism operating either on characters not rejected by the first or on characters that the first was unable to discriminate. The type-letter set was used extensively in testing these various mechanisms due to the difficulty in obtaining a large test set for the hand-drawn characters. The tape produced a number of problems related to extracting information from a message as opposed to a selected list of alphabetic characters. However the problems were solved from a practical viewpoint and a fairly efficient discriminator was found. If anything, the tape data set was too elementary a set for the recognition system used because no real normalization in size was necessary for it. Removal of this facet of the recognizer would have led to a better definition of I's and L's. However, the general nature of the program would have been reduced and comparisons with the hand-drawn letter set less valid.

Various offshoots of this work have been investigated including a comparative study of the calculation times of the discriminants, and the effect of a rejection threshold on the efficiency of the mechanisms. This was inspected in detail, demonstrating the theoretically predicted relations between the error rate, the rejection rate and the rejection threshold. The relations appear to hold within the bounds set by the approximations. Also the discriminatory ability of the various moments has been measured and by this means the efficiency of recognition can be increased by using only the more discriminating moments.

An interesting theory has been developed to show that, given certain assumptions, the efficiency of any discriminator could be related to the number of dimensions of the pattern vector. It was shown that these assumptions are justified when the dimensionality is not too great, and reasons for the ultimate breakdown of the theory have been suggested.

The relation between loss functions and rejection threshold has been investigated in some detail. The author feels that thresholds are easier to utilize than loss functions, although conceptually farther from the truth. Some mention has been made of complete loss functions, their applications and their relation to information loss through a noiseless channel.

Summing up, the author believes that

(i) A useful decision mechanism can be forged out of

Geometric Moments and a two-stage categorizer.

(ii) The exponent mechanism, although of a quadratic nature, is sufficiently good to be considered as a worthwhile categorizer.

(iii) That loss functions should be calculated and applied in specific cases of closed discriminant systems i.e. those in which the operator or checker is included. Only after such loss functions have been considered can it be stated for certain whether or not they should be involved in the categorization process.

(iv) If this is to be done, it should be first asked if it is not easier to use a trainable classifier which is taught to minimize the losses. This bypasses an understanding of the basic theory involved but may well be economically preferable. The limits of the cost optimization process might have to be extended to the actual task of deciding whether or not to use losses at all.

APPENDIX 1.

Power and Size in the Neyman-Pearson Theory.

In this theory, experimentation is carried out in a single stage by observing the values of the first N chance variables of the sequence $\{X_i\}$. Suppose H is an hypothesis about the $\{X_i\}$ to be tested. Then the set of all sample points $x = (x_1, x_2, \ldots x_N)$ for which H is rejected is called the Critical Region. Let the hypothesis H under test be the hypothesis that $F \in \omega$ where F is the distribution of sample points and ω is the set of distributions over which H is true. Then the concept of the Power of the Critical Region is defined as the probability that H will be rejected when some F, not an element of ω, is true. The Power is thus a function of F defined for all F not in ω. Also the Size of the Critical Region is defined as the probability that H will be rejected when some F is true that is an element of ω. Thus the Size is a function of F defined for all F in ω. These concepts relate to general Decision Theory since the choice of the Critical Region is equivalent to the choice of a decision function and the notions of Size and Power are special cases of the notion of Risk.

In fact let W(F,d) be defined as follows

$$W(F,d_1) = 0 \qquad \text{when } F \in \omega$$

and $$W(F,d_1) = 1 \qquad \text{when } F \notin \omega$$

$$W(F,d_2) = 1 \qquad \text{when } F \in \omega$$

and $$W(F,d_2) = 0 \qquad \text{when } F \notin \omega$$

where d_1 and d_2 are the two possible terminal decisions.

Thus W(F,d) is a 'simple' Weight Function. We can disregard the cost of experimentation here if we restrict the choice of the experimenter to decision functions for which the expected cost of experimentation is the same constant amount. Then the simple risk corresponding to the above simple weight function is equal to the size of the critical region when $F \in \omega$ and to (1 - Power) when $F \notin \omega$.

APPENDIX 2.

A priori Probabilities and Information Content of the Alphabetic Characters.

	ω	$-\ln\omega$	$-\omega\ln\omega$	ω_{tape}
A	.0788	2.54	.200	.074
B	.0156	4.26	.066	.016
C	.0268	3.62	.097	.043
D	.0389	3.25	.126	.040
E	.1268	2.06	.261	.136
F	.0256	3.67	.094	.018
G	.0187	3.98	.074	.020
H	.0573	2.86	.164	.045
I	.0707	2.65	.187	.081
J	.0010	6.91	.007	.002
K	.0060	5.12	.031	.004
L	.0394	3.23	.127	.045
M	.0244	3.71	.090	.023
N	.0706	2.65	.187	.071
O	.0776	2.56	.199	.076
P	.0186	3.98	.074	.026
Q	.0009	7.01	.006	.002
R	.0594	2.82	.167	.059
S	.0631	2.76	.174	.063

	ω	$-\ln\omega$	$-\omega\ln\omega$	ω_{tape}
T	.0978	2.32	.227	.088
U	.0280	3.58	.100	.027
V	.0102	4.59	.047	.010
W	.0214	3.84	.082	.011
X	.0016	6.44	.010	.005
Y	.0202	3.90	.079	.013
Z	.0006	7.42	.004	.0005
Total			2.887	

ω - the a priori probability of a letter occurring in an English text (Dewey Classification of English Language.)

$-\ln\omega$ - the relative information content of each letter

ω_{tape} - the probability of a given letter occurring on the IBM tape message.

Largest ratio of information contents = $\ln \dfrac{\omega_E}{\omega_Z} = 5.36$

APPENDIX 3.

A Normality Theorem.

Theorem: The family of curves $(\underline{X} - \underline{\mu})' V^{-1}(\underline{X} - \underline{\mu}) = C$ has the property that a line normal to the surface of one of the family is normal to the whole of that familiy.

Proof: The family is of the form $f(\underline{r}) = C$

Then a line normal to the surface $f(\underline{r}) = C_0$ is

$$(\underline{r} - \underline{r}_0) \times \underline{\nabla} f = \underline{0} \qquad 1$$

where $\underline{r}_0$ lies on $f(\underline{r}) = C_0$

Similarly a line normal to $f(\underline{r}) = C_1$ is

$$(\underline{r} - \underline{r}_1) \times \underline{\nabla} f = \underline{0} \qquad 2$$

But if we define $\underline{r}_1$ to lie on line 1 then

$$(\underline{r}_1 - \underline{r}_0) \times \underline{\nabla} f = \underline{0}$$

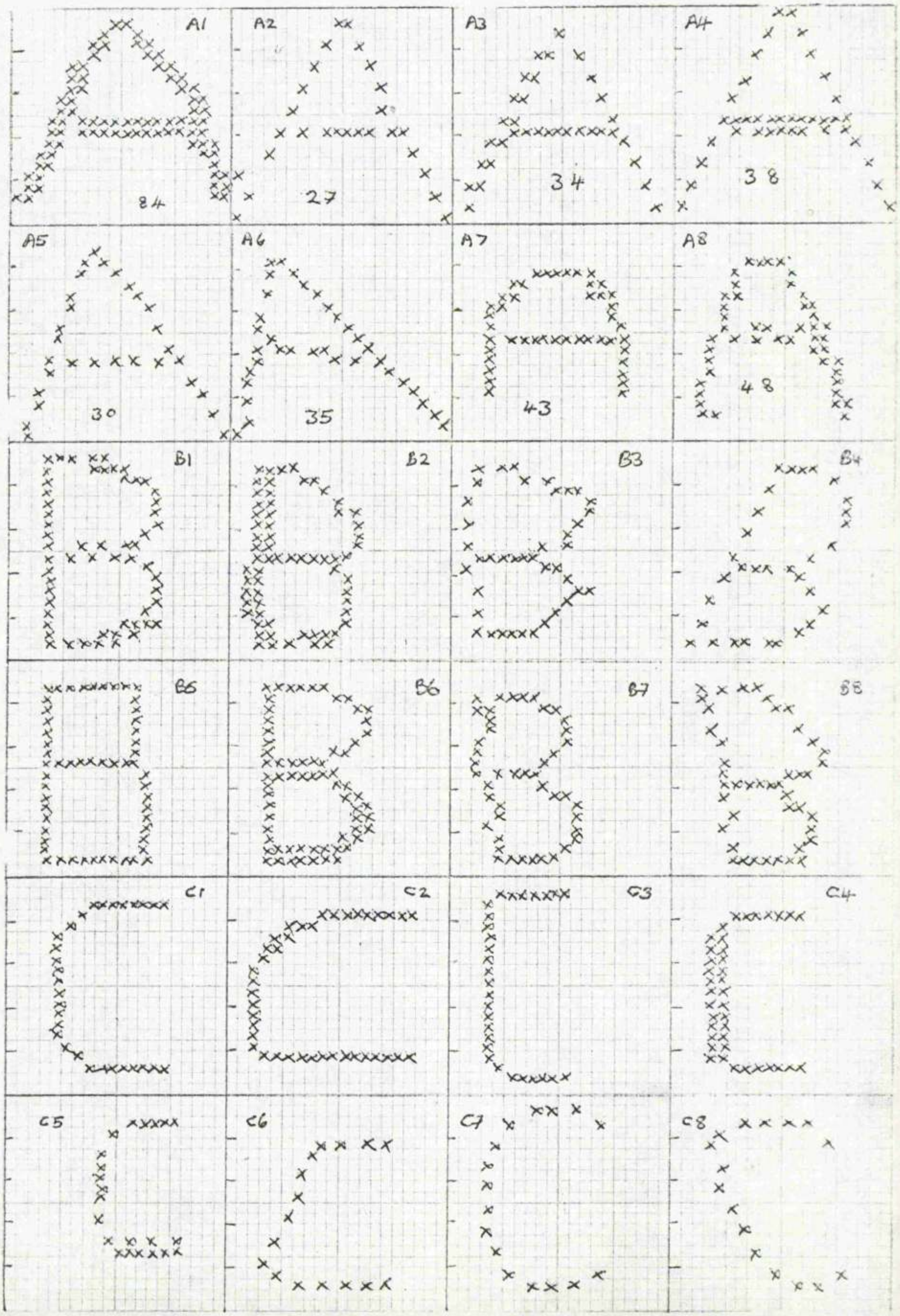
$\underline{r}_0$ lies on line 2

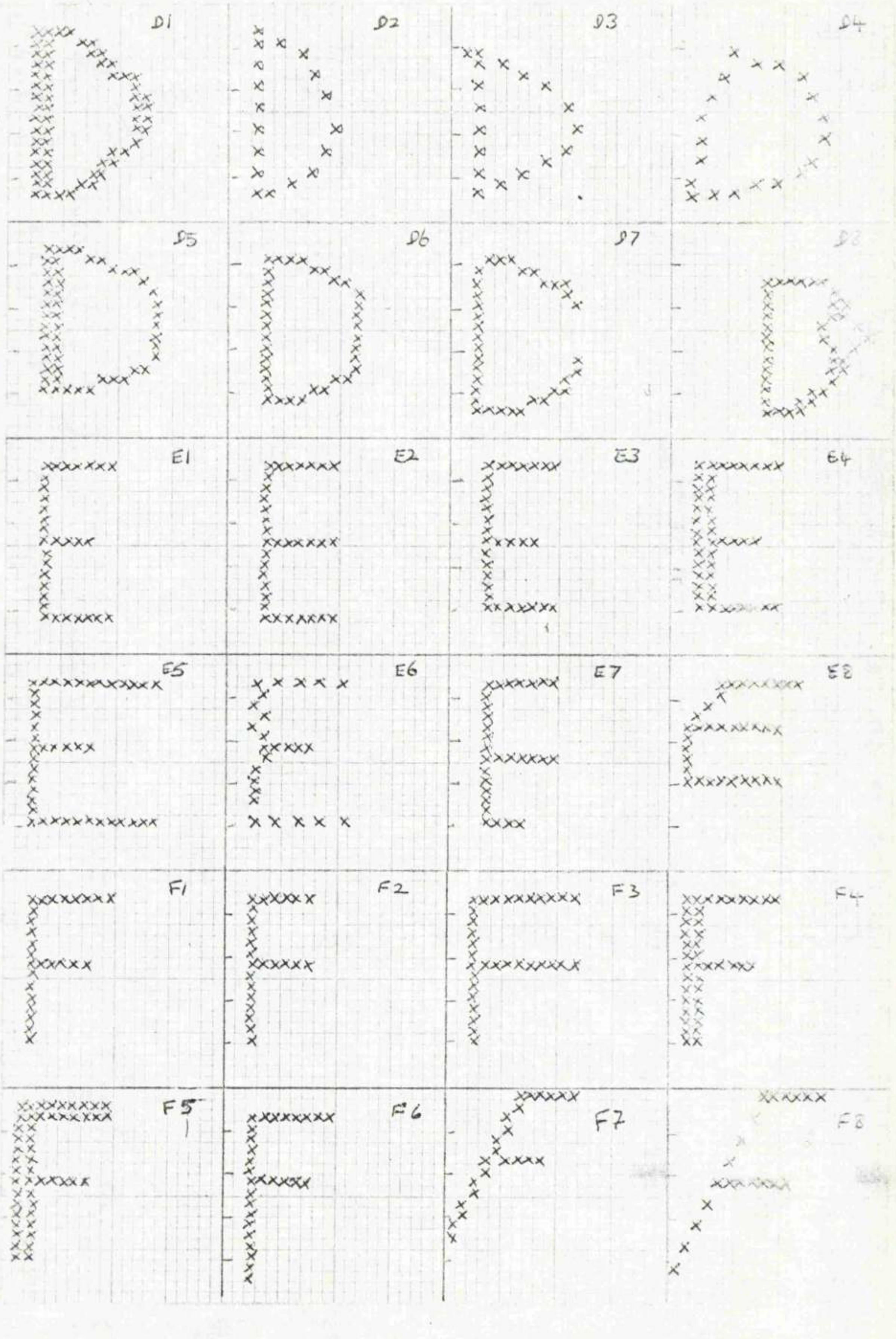
Hence $\underline{r}_1$ and $\underline{r}_0$ both lie on lines 1 and 2.

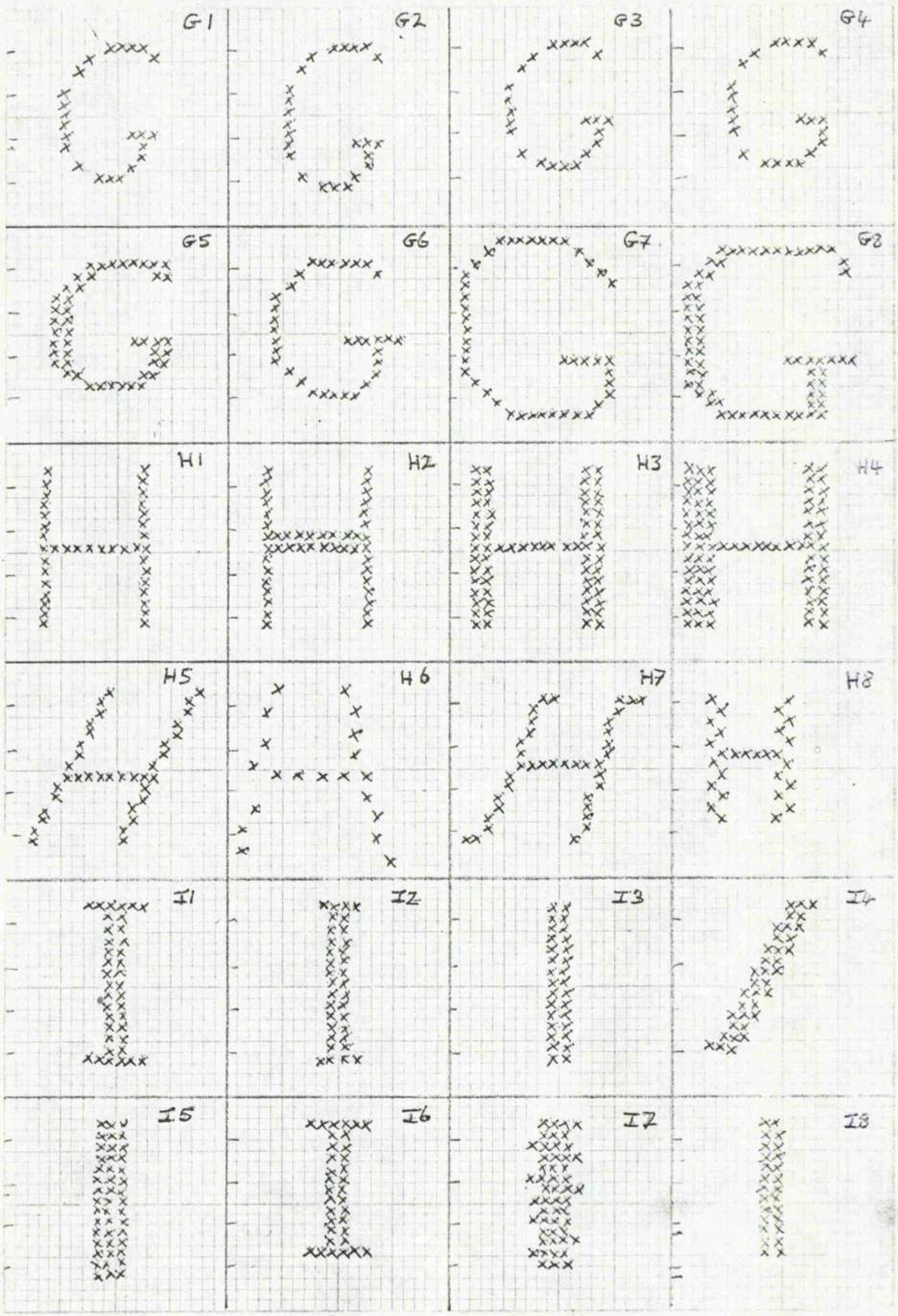
Equation 1 and equation 2 represent the same line, which is normal to each of the two surfaces.

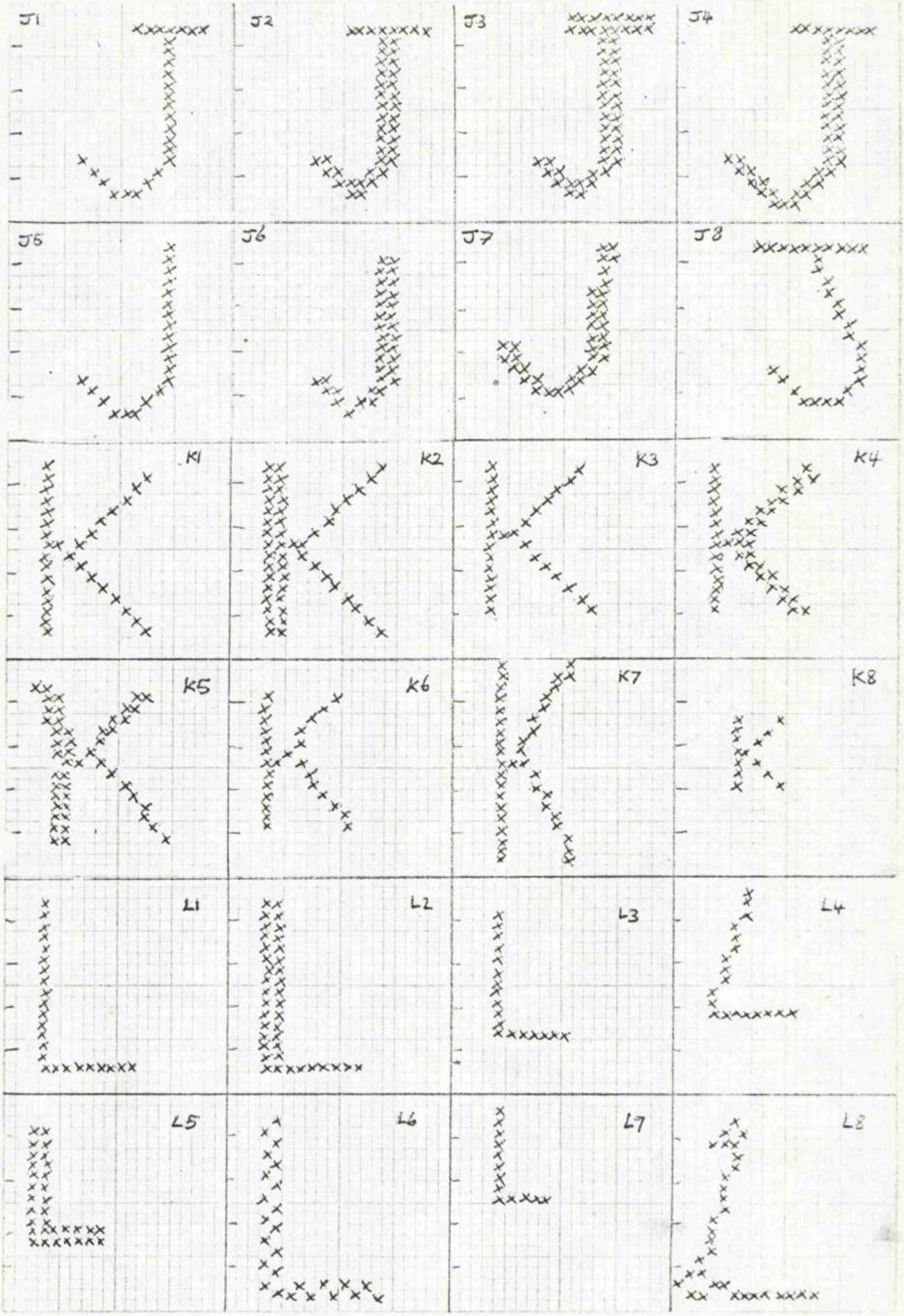
APPENDIX 4.

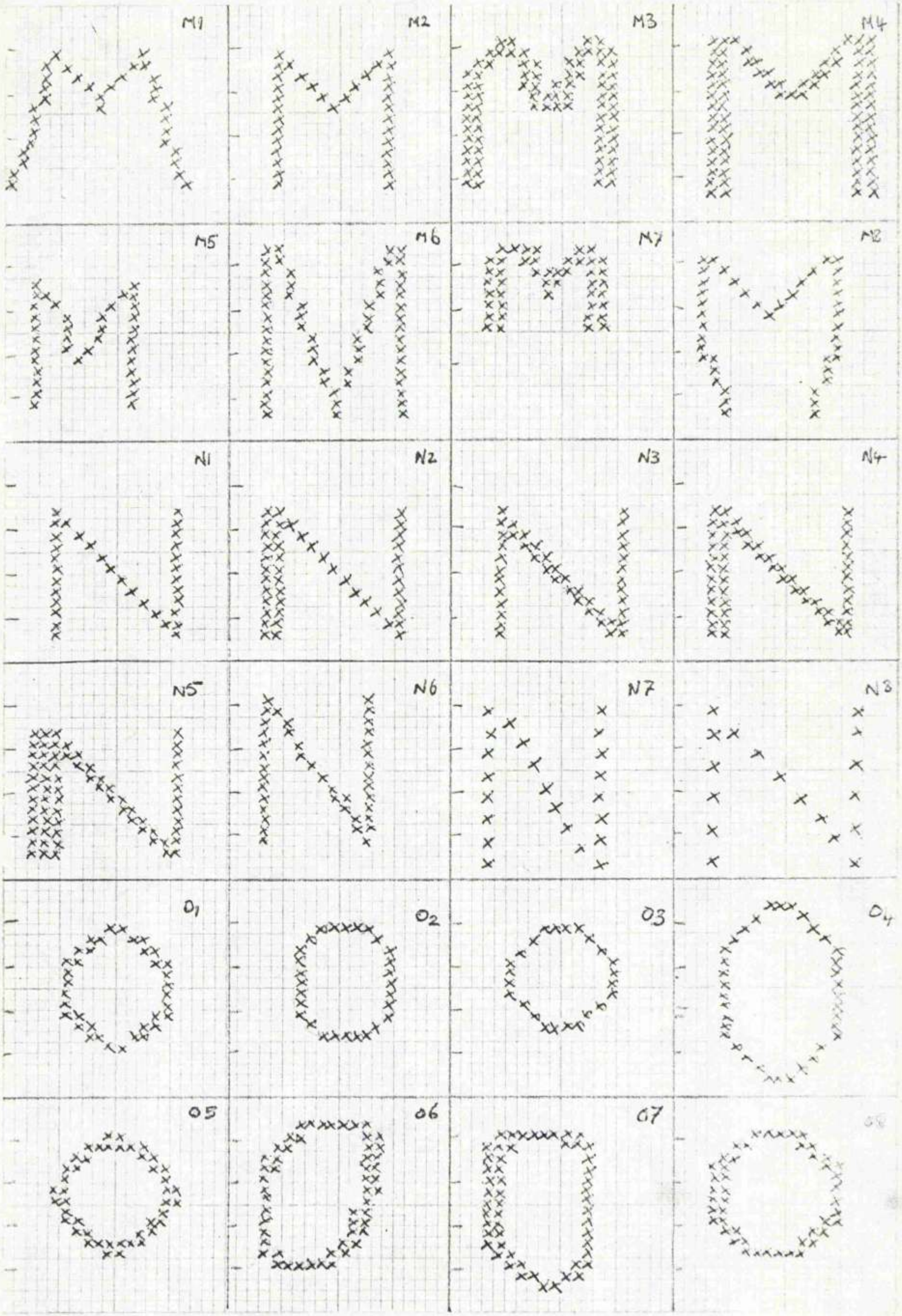
Hand-Drawn Letter Set.

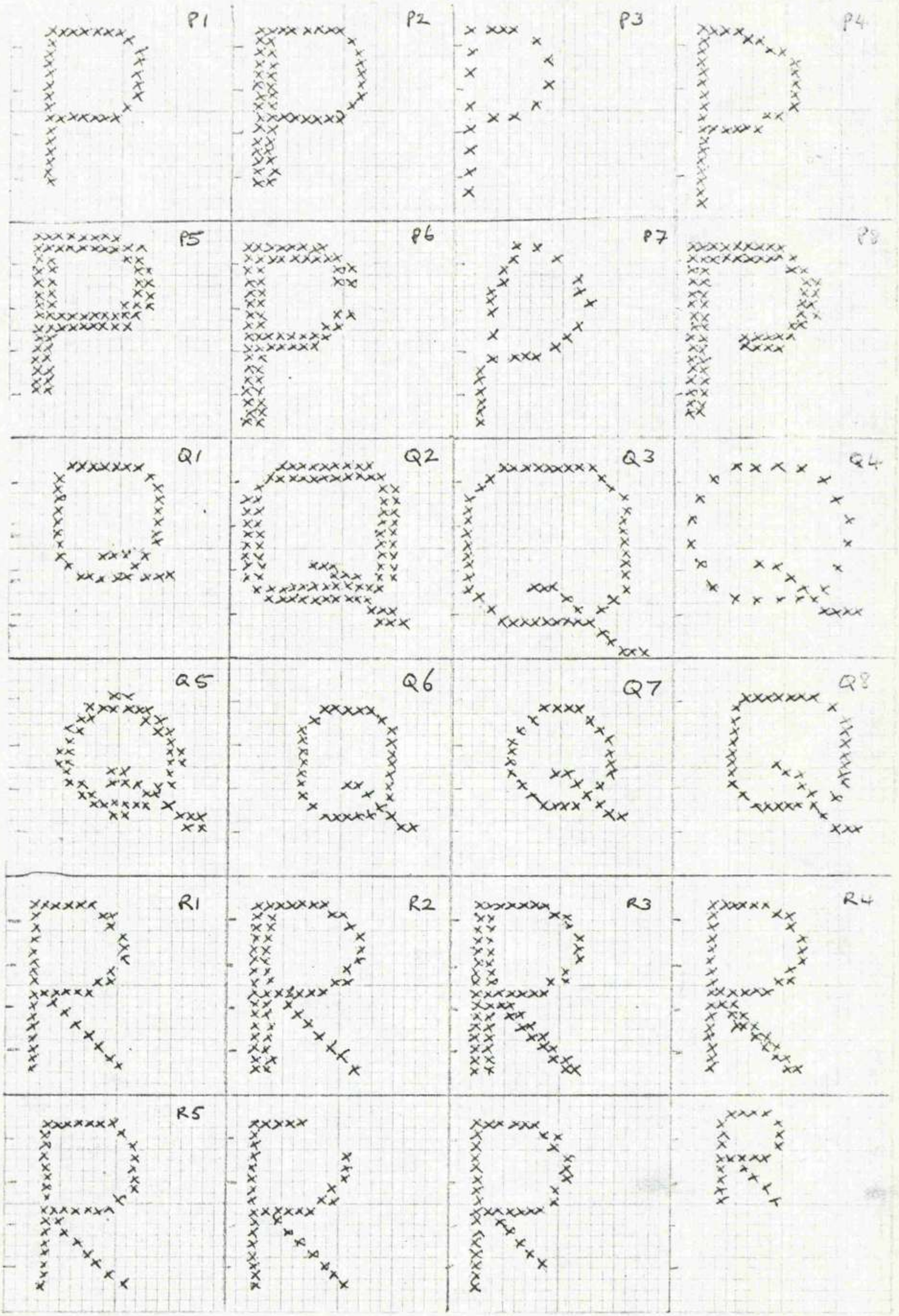


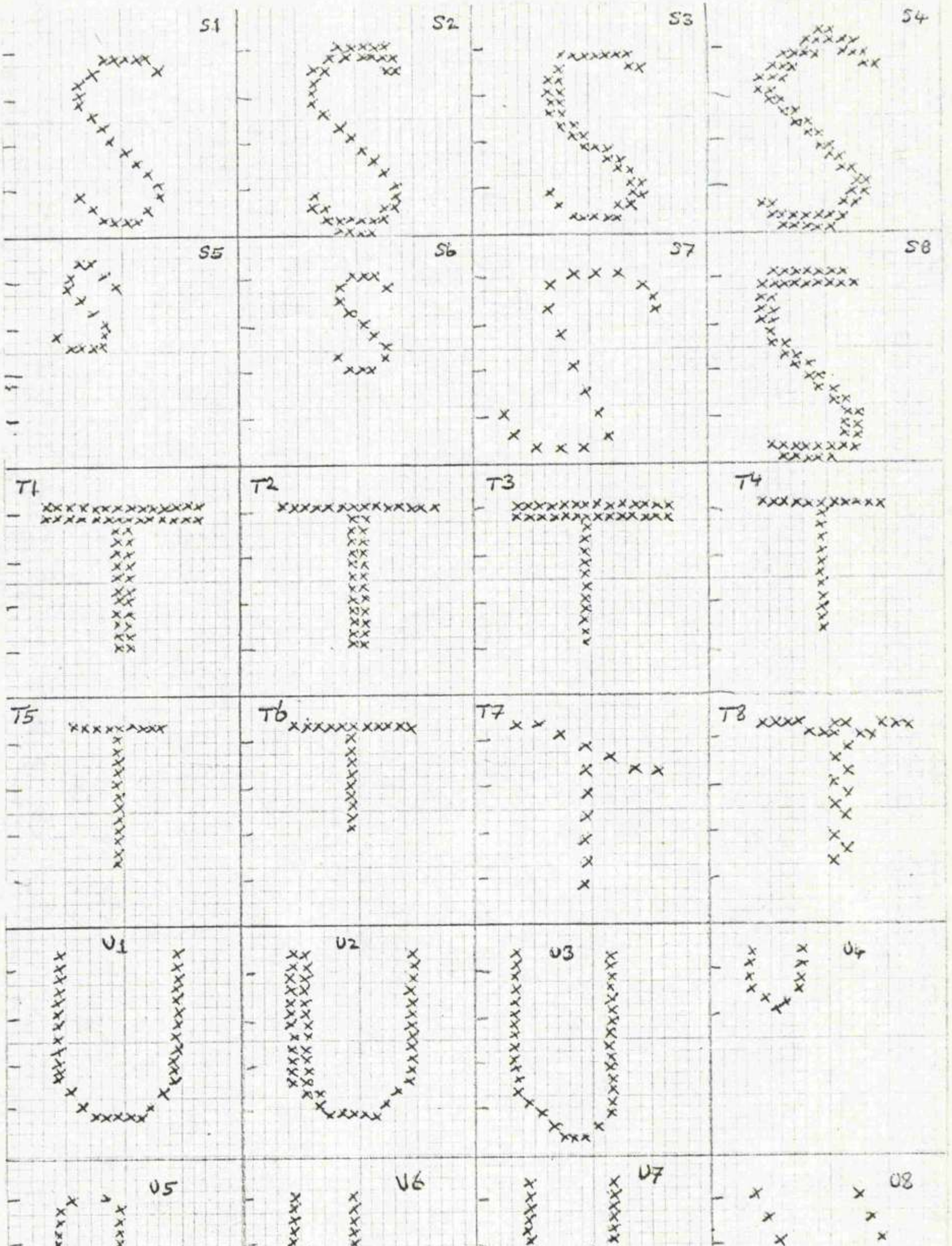

D1
D2
D3
D4
D5
D6
D7
D8
E1
E2
E3
E4
E5
E6
E7
E8
F1
F2
F3
F4
F5
F6
F7
F8

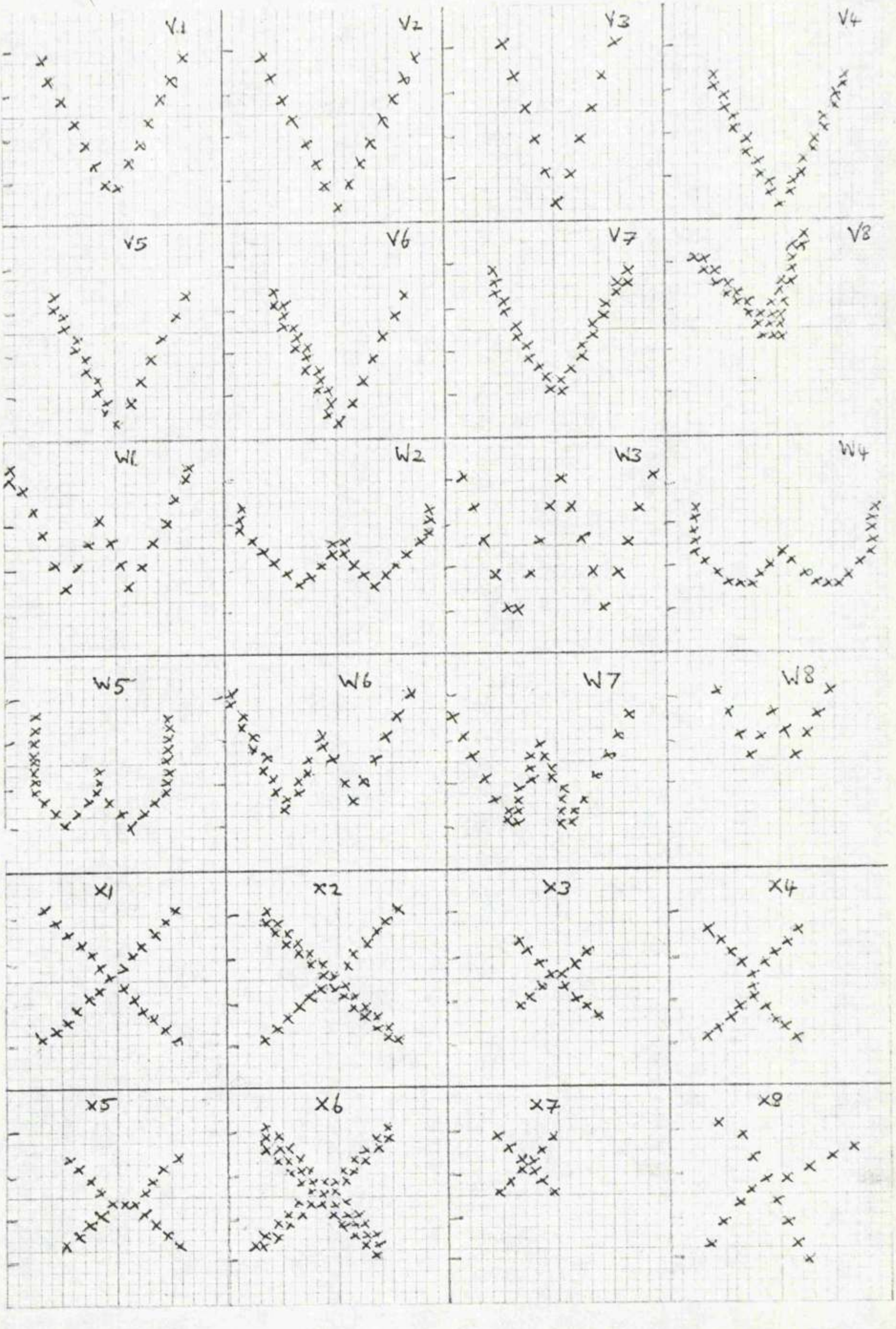

G1
G2
G3
G4
G5
G6
G7
G8
H1
H2
H3
H4
H5
H6
H7
H8
I1
I2
I3
I4
I5
I6
I7
I8

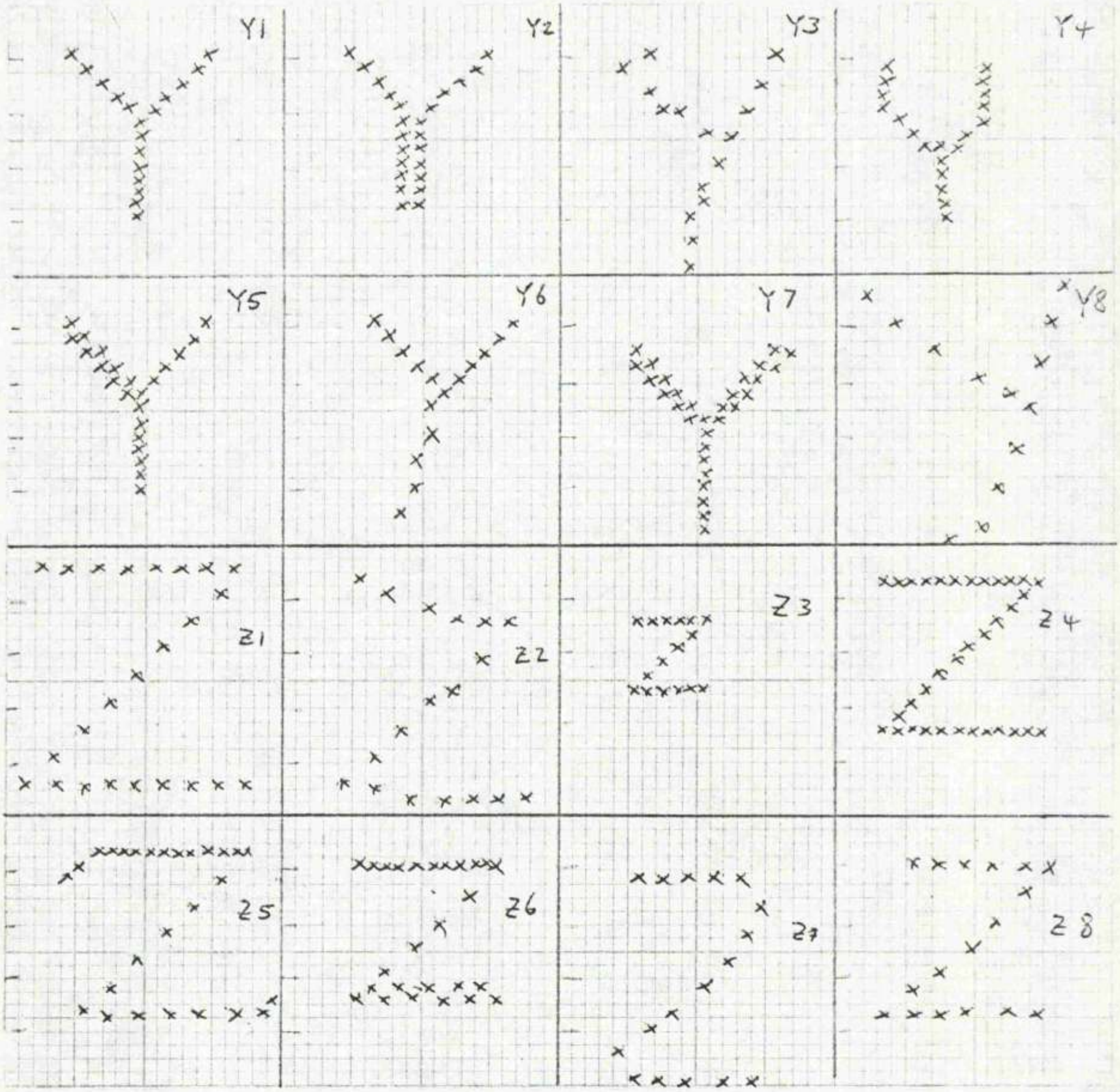

J1
J2
J3
J4
J5
J6
J7
J8
K1
K2
K3
K4
K5
K6
K7
K8
L1
L2
L3
L4
L5
L6
L7
L8


M1
M2
M3
M4
M5
M6
M7
M8
N1
N2
N3
N4
N5
N6
N7
N8
O1
O2
O3
O4
O5
O6
O7
O8


P1
P2
P3
P4
P5
P6
P7
P8
Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
R1
R2
R3
R4
R5


S1
S2
S3
S4
S5
S6
S7
S8
T1
T2
T3
T4
T5
T6
T7
T8
U1
U2
U3
U4
U5
U6
U7
U8


V1
V2
V3
V4
V5
V6
V7
V8
W1
W2
W3
W4
W5
W6
W7
W8
X1
X2
X3
X4
X5
X6
X7
X8


Y1
Y2
Y3
Y4
Y5
Y6
Y7
Y8
Z1
Z2
Z3
Z4
Z5
Z6
Z7
Z8

APPENDIX 5.

Computer Tape Message.

```
 CONDENSED COMPUTER ENCYCLOPEDIA, CO., INC., 330 W
. 42 ST., NEW YORK, N.Y. 10036, 1969; 618 PAGES, I
LLUS., £14.50. THE AVOWED PURPOSE OF TH#IS ENCYCLO
PEDIA IS ''TO DEFINE COMPUTER TERMS CLEARLY AND ME
ANINGFULLY FOR THE NONSPECIALIST.'' IT IS INTENDED
 TO BRIDGE THE GAP BETWEEN CONCISE DIC- TIONARY DE
FINITIONS AND COMPUTER MANUALS AND TEXTS, AND DOES
 A GOOD JOB OF MEETING ITS OBJECTIVES. THERE ARE A
BOUT 1000 TERMS DE- FINED OR CROSS-REFERENCED IN T
HE 574 PAGES IN THE BODY OF THIS BOOKA THIS IS AN
AVERAGE OF ABOUT ONE-HALF PAGE PER TERM, ALTHOUGH
THE LONGER POR- TIONS ARE SEVERAL PAGES IN LEN#H.
COMPUTER LANGUAGES (ABOUT 20 ARE INCLUDED) GENERAL
LY REQUIRE THREE TO FOUR PAGES SINCE A WORKED-OUT
EX- AMPLE IS GIVEN FOR EACH. THE LONGEST NONLANGUA
GE ITEM IS ''EDP CENTER'', FOR WHICH SIX PAGES ARE
 DEVOTED TO A MANAGER'S OVERVIEW OF THIS TOPIC. TH
IS ENCYCLOPEDIA IS UP TO DATE, PLACES THE EMPHASIS
 ON THE MORE IM- PO#ANT ASPECTS OF MODERN COMPUTER
 TECHNOLOGY, AND PRESENTS A FRESH (AND SOMETIMES R
EFRESHING) APP#ACH TO DESCRIBING COMPUTER TERMS AN
D CON- CEPTS. ONLY SOMEONE WHO HAS SUF- FERED A MA
JOR CARD JAM MORE THAN ONCE COULD HAVE WRITTEN THE
 ''JAM'' ITEM WITH SUCH FEELING. A COMMEND- ABLE E
FO# HAS BEEN MADE TO EXPLAIN ''COMPUTERESE'' IN LA
Y TERMS. #O EX- AMPLES ARE THE CITING OF ROBE#,S R
ULES OF ORDER TO EXPLAIN PA# OF THE CONCEPT OF REC
URSI#ON, AND THE DESCRIP- TION OF THE CDC 6600 MUL
TIPROCESSOR AS A ''V-10 DATA-PROCESSING ENGINE.''
THE EMPHASIS IS ON SO#WARE, AL- THOUGH SOMEWHAT MO
RE THAN 100 ESSENTIAL GENERAL COMPUTER HARDWARE TE
RMS ARE INCLUDED. IN ADDITION, THE HARDWARE AND OT
HER AS#CTS OF OCR (OPTICAL CHARACTER RECOGNITION)
ARE THOROUGHLY COVERED, BUT LITTLE ELSE IS CONSIDE
RED IN THE SOURCE DATA AUTOMA- TION AREA. THE BOOK
 ASLO INCLUDES A BIBLI#- RAPHY AND AN INDEX. THE B
IBLIOG#PHY IS LIMITED TO MATERIAL READILY A##- PUT
ERS. THE INDEX (PRECEDED BY THE FO#RAN PROGRAM USE
D TO PRODUCE IT) IS INTENDED FOR USE IN LOCATING T
ERMS NOT READILY FOUND IN THE ALPHABETIC PO#ION OF
 THE BOOK. HOWEVER, ONLY ABOUT ONE THIRD OF THE IN
DEX ITEMS SERVE THIS PURPOSE#HE OTHERS ARE A REPEA
T OF THE ALPHABETIC ARE NOTED. THESE INCLUDE THE I
NADEQUACY OF SOME OF THE INDEX CROSS-REFERENCING,
SUCH AS ''CNE'S COMPLEMENT'' (WHERE ''RADIX-MINUS-
ONE COMPLEMENT'' IS NOT REFERENCED) ; THE TCO BRIEF
 DISCUSSION OF ''EMULATOR'' (BETTER DISCUSSED UNDE
R ''COS'' BUT NOT REFERENCED TO IT) ; AND SMALL ITE
MS SUCH AS GROSCH'S NAME BEING MISSPELLED OR WHY O
NLY BABBA#, HOLLERITH, LEIBNITZ, VON NEUMAN#N, #SC
AL, TURING, AND WEINER RATE BIOGRAPHIES. THIS ENCY
CLOPEDIA IS HIGHLY RECOM- MENDED AS A #LUABLE REFE
RENCE FCR EVERYCNE IN THE COMPUTER FIELD. DESPITE
THE INTENT OF THE AUTHOR, IT WILL PROBABLY BE CF M
ORE USE TO THE PROFESSIONAL THAN# TO THE NONSPECIA
LIST. SOMEONE WHOSE ONLY BACKGROUND IN COMPUTERS I
S A TH#E-DAY EXECUTIVE COURSE WILL NOT BE ABLE TO
TAKE FULL ADVANTAGE OF THE INFORMATION IN THIS B#K
. IT IS AN EXCELLENT CONTRIBUTION TO THE LITERATUR
E OF THE FIELD. EVEN THOSE WHO HAVE WORKED WITH CC
M- PUTERS #R SOME TIME WILL FIND IT A GOOD REFEREN
CE FORTERMS AND COMCEPTS THAT #HER SOURCES CONSIDE
R EITHER TOO SCANTILY OR TOO THOROUGHLY. BETH##, M
D. INTE###CIRCUITS,A B ASICA COURSE FOR ENGIN## AN
D #CHNICIANS, R. G. HIE- BERD-MCG#W-HILL BOOK #.,
INC., 330 W. 42 ST., NEW #RK, N.Y. 10036, 1969; 18
3 PAGES, ILLUS., £9.95. THIS BOOK, A S#UEL TO THE
AUTHOR'S SOLID STATE EL#T#NICS, IS DI#CTED TO THE
NONELECTRONICS ENGINEER AND THE TECH- NICAL IN#ITU
TE STUDENT. THE APPROACH IS #SI#LLY QUALITATIVE, E
SSENTIAL#LY A SU#EY OF A#ILABLE INTEGRATED CIRCUIT
 T#HNOLO#. ALTHOUGH THIS REVIEWER IS PERPLEXED (BE
LIEVING THAT A SUBSTANTIALLY MORE QUANTITATIVE APP
ROACH IS REQUIRED) , THE BOOK WAS FOUND TO READ SMC
OTHLY AND TELL ITS STORY IN A READILY COMPRE- HENS
IBLE FASHION. WITH ITS MERE 172 PAGES, IT CCULD BE
 READ AS AN EVEN- ING'S #VIEW. THIS REVIEWER FOUND
 IC APPLICATIONS IN THE LAST CHA#ERS THAT WERE REV
EALINGLY DESCRI#D IN AN IN- TUITIVELYAP#ALING MANN
ER. THE EARLY CHAPTERS INDICATE THE IM- PACT OF IN
TEGRATED CIRCUITS ON T#AY'S ELECTRONICS AND REVIEW
 THE SOLID-STATE TECHNOLOGY. DIGITAL AND LINEAR CI
#UITS AND THEIR APPLICATIONS ARE REVIEWED IN THE N
EXT SEVERAL CHA#ERS. APPLICA- TIONS ##ING F#M VCLT
AGE STABILIZERS TO TELEVISION STAGES ARE INCLUDED.
 THE INTENDED PURPOSE # THE TEXT IS TO GIVE THE NC
NELECTRCNICS EX##- THE MECHANICAL AND# SYSTEMS ENG
INEER AS #LL AS THE TECHNICAL INSTITUTE G#DUATE#HE
 CA#BILITY OF UNDER- STANDING THE IC LANGUA# AND T
ECH- NOLOGY. I AM SURE THAT IT WILL SUCCEED IN THI
S O#E#I#. WHAT THE ##K WILL NOT DO IS GIVE ANY DES
IGN EXPE#ISE, IN SPITE OF ITS CLEAR DESCRIPTION CF
 THE CAPABILITIES OF ''STANDARD CATALOG IN- TEG#TE
D CI#UITS.'' RALPH W. #N#UM, JR. BELL T#PHCNE LABC
#TORIES HOLMDEL, N.J. THE ##L ENC#LO##IA # FILM. A
ND TELEVISION #CHNIQUES, R
```

APPENDIX 6.

Comparison of Tape Letters I and L.

```
  X                  X
  X                 XX
  X                 XX
                    XX
                    XX
  X                 XX
 XX                XXX
 XX                XXX
 XX                XXX
XXX                XXX
XXX                XXX
XXX                XX
XXX                XX
XXX               XXX
XXX               XXX
XX                XXX
XX                XXX
XX                XXX
```

APPENDIX 7.

Computer Programs used in Character Recognition.

```
      DIMENSION X(15),ICOUNT(50),XMU(50,15),SIG(50,15)
      LOGICAL*1 L(2),M(2),LETT(50)/50*' '/,L2(2),M2(2)
      EQUIVALENCE (L(1),LL),(L2(1),LN)
      DO 5 I=1,50
      ICOUNT(I)=0
      DO 5 J=1,15
      XMU(I,J)=0.
    5 SIG(I,J)=0.
      N=1
      ITOT=1
      READ (5,10) NTOT
   10 FORMAT (I5)
      DO 75 N=1,NTOT
      READ (3,15) L(2),X
   15 FORMAT (9X,A1,7E10.3/8E10.3)
      DO 20 I=1,ITOT
      L2(2)=LETT(I)
      IF (LL.EQ.LN) GO TO 30
   20 CONTINUE
      IF (N.EQ.1) GO TO 50
   25 ITOT=ITOT+1
      LETT(ITOT)=L(2)
      I=I+1
   30 DO 40 J=1,15
      XMU(I,J)=XMU(I,J)+X(J)
   40 SIG(I,J)=SIG(I,J)+X(J)*X(J)
      ICOUNT(I)=ICOUNT(I)+1
      GO TO 60
   50 ITOT=0
      I=0
      GO TO 25
   60 IF (ITOT.EQ.50) WRITE (6,70)
   70 FORMAT (' ITOT IS TOO LARGE')
   75 CONTINUE
      DO 80 I=1,ITOT
      DO 80 J=1,15
      XMU(I,J)=XMU(I,J)/ICOUNT(I)
   80 SIG(I,J)=SQRT(SIG(I,J)/ICOUNT(I)-XMU(I,J)*XMU(I,J))
      PUNCH 90,((XMU(I,J),J=1,15),I=1,ITOT),((SIG(I,J),J=1,15),I=1,ITCT)
   90 FORMAT (8E10.3/7E10.3)
      WRITE (6,100) (LETT(I),I=1,ITOT)
      WRITE (6,101) (ICOUNT(I),I=1,ITOT)
      WRITE (6,102) ITOT
  100 FORMAT (10A10)
  101 FORMAT (10I10)
  102 FORMAT (' ITCT IS ',I10)
      STOP
      END
```

```
C     CALCULATION OF CHARACTER SET MOMENTS 2500*15 AND KEYWORDS
      DIMENSION LETTER(32,67),X(15)
      LOGICAL*1 LOG(292)
      DO 20 I=1,4000
      II=(I/100)*100-I
      IF (II.EQ.0) WRITE (6,25) I
   25 FORMAT (I5)
      READ (4,10) LOG
   10 FORMAT (146A1,146A1)
      CALL CHANGE(LOG,LETTER)
      CALL MOMENT(LETTER,67,32,X)
      WRITE (3,15) LOG(9),X
   15 FORMAT (9X,A1,7E10.3/8E10.3)
   20 CONTINUE
      STOP
      END
      SUBROUTINE MOMENT(LET,M,N,X)
      DIMENSION LET(N,M),X(15)
      A=0.
      B=0.
      C=0.
      D=0.
      E=0.
      DO 25 I=1,M
      DO 25 J=1,N
      IF (LET(J,I).EQ.0) GO TO 25
      A=A+1.
      B=B+I
      C=C+J
      D=D+I*I
      E=E+J*J
   25 CONTINUE
      IF (A.LE.0.001) RETURN
      XBAR=B/A
      YBAR=C/A
      SIGMAX=1./SQRT(D/A-XBAR*XBAR)
      SIGMAY=1./SQRT(E/A-YBAR*YBAR)
      DO 35 L=1,15
   35 X(L)=0.
      DO 40 I=1,M
      DO 40 J=1,N
      IF (LET(J,I).EQ.0) GO TO 40
      XX=(I-XBAR)*SIGMAX
      YY=(J-YBAR)*SIGMAY
      XSQ=XX*XX
      XCUBE=XSQ*XX
      XFOUR=XCUBE*XX
      XFIVE=XFOUR*XX
      YSQ=YY*YY
      YCUBE=YSQ*YY
      YFOUR=YCUBE*YY
      YFIVE=YFOUR*YY
      X(1)=X(1)+XCUBE
      X(2)=X(2)+XSQ*YY
      X(3)=X(3)+XFOUR
      X(4)=X(4)+YFOUR
      X(5)=X(5)+XCUBE*YY
      X(6)=X(6)+YCUBE*XX
      X(7)=X(7)+YSQ*XX
      X(8)=X(8)+YCUBE
      X(9)=X(9)+XFIVE
      X(10)=X(10)+YFOUR*XX
      X(11)=X(11)+XSQ*YSQ
      X(12)=X(12)+YFIVE
      X(13)=X(13)+XFOUR*YY
      X(14)=X(14)+XCUBE*YSQ
      X(15)=X(15)+XSQ*YCUBE
   40 CONTINUE
      RETURN
      END
      REAL MU
      DIMENSION XSUM(23),XMU(23,8),SIG(23,64),X(15),EFF(12),REJR(12)
      DIMENSION REJT(12),ERR(12),MU(8),SI(64),NUM(41)
      LOGICAL*1 LETR(23),LETER(23),L1(4)/4*' '/,L2(4)/4*' '/,L3(4)/4*' '
     1/,LET
      EQUIVALENCE (L1(4),LET),(L2(4),LETR(1)),(L1(1),P),(L2(1),Q),(L3(1)
     1,R)
      READ (5,5) (LETER(I),I=1,23)
    5 FORMAT (80A1)
      READ (5,6) NUM
    6 FORMAT (20(I2,2X))
      DO 14 I=1,41
      READ (4,10) MU
   10 FORMAT (8E10.3)
```

```
      J=NUM(I)
      IF (NUM(I).LT.24) GO TO 12
      GO TO 14
   12 DO 13 K=1,8
   13 XMU(J,K)=MU(K)
   14 CONTINUE
      DO 114 I=1,41
      READ (4,110) SI
  110 FORMAT (8E10.3)
      J=NUM(I)
      IF (NUM(I).LT.24) GO TO 112
      GO TO 114
  112 DO 113 K=1,64
  113 SIG(J,K)=SI(K)
  114 CONTINUE
      DO 212 I=1,12
      REJT(I)=0.
      ERR(I)=0.
  212 EFF(I)=0.
      ITOT=0
      ICOUNT=0
      NN=8
      RN=0.04*NN*NN
      DO 102 N=1,4000
      READ (3,15) LET,(X(J),J=1,15)
   15 FORMAT (9X,A1,7E10.3/8E10.3)
      DO 16 IL=1,23
      L3(4)=LETER(IL)
      IF (P.EQ.R) GO TO 17
   16 CONTINUE
      GO TO 102
   17 ITOT=ITOT+1
      DO 20 I=1,23
      LETR(I)=LETER(I)
   20 XSUM(I)=0.
      DO 26 I=1,23
      DO 25 J=1,8
      DO 25 JJ=1,8
      JJJ=8*(J-1)+JJ
   25 XSUM(I)=XSUM(I)+(X(J)-XMU(I,J))*(X(JJ)-XMU(I,JJ))*SIG(I,JJJ)
   26 XSUM(I)=ABS(XSUM(I))
      CALL ORDER(XSUM,LETR,23)
      IF (P.NE.Q) GO TO 45
      ICOUNT=ICOUNT+1
      DO 40 IR=1,12
      IL=13-IR
      REJ=RN*IL*IL
      IF (XSUM(1).GT.REJ) GC TO 100
      EFF(IL)=EFF(IL)+1
   40 CONTINUE
      GO TO 100
   45 WRITE (6,46)
   46 FORMAT (10X,'ERROR')
      DO 50 IR=1,12
      IL=13-IR
      REJ=RN*IL*IL
      IF (XSUM(1).GT.REJ) GO TO 100
   50 ERR(IL)=ERR(IL)+1
  100 CONTINUE
      WRITE (6,101) LET,(LETR(I),I=1,5),(XSUM(I),I=1,5),DIFF
  101 FORMAT (1X,6A4,5E10.3,F8.3)
  102 CONTINUE
      DO 105 IR=1,12
      REJT(IR)=(ITOT-ERR(IR)-EFF(IR))*100./ITOT
      ERR(IR)=ERR(IR)*100./ITOT
      EFF(IR)=EFF(IR)*100./ITOT
  105 REJR(IR)=0.2*IR
      EFFMAX=ICOUNT*100./ITOT
      WRITE (6,35) REJR,EFF,EFFMAX,ERR,REJT
   35 FORMAT (7X,'REJECTION RATE'/7X,12(F3.1,5X)/1X,'EFF',1X,12(F5.1,3X)
     1/7X,'MAX EFF=',F7.1/1X,'ERR',1X,12(F5.1,3X)/1X,'REJ',1X,12(F5.1,3X
     1))
      STOP
      END
      SUBROUTINE ORDER(A,B,LONG)
C     ASCENDING ORDER SORTER FOR VECTOR A, WITH LABEL B
      DIMENSION A(LONG)
      LOGICAL*1 B(LONG),BB,TEST
      II=LONG-1
    5 TEST=.TRUE.
      DO 15 I=1,II
      IF (A(I).LE.A(I+1)) GO TO 10
      TEST=.FALSE.
      AA=A(I)
      A(I)=A(I+1)
      A(I+1)=AA
      BB=B(I)
      B(I)=B(I+1)
      B(I+1)=BB
   10 CONTINUE
   15 CONTINUE
      IF (.NOT. TEST) GO TO 5
      RETURN
      END
```

```
*        SUBROUTINE CHANGE
*        ********** ******
*
CHANGE   CSECT
WORD1    EQU   2
WORD2    EQU   3
START    EQU   4
LIMIT    EQU   5
COLBASE  EQU   6
ROW      EQU   7
COUNT    EQU   8
POSITION EQU   9
RECORD   EQU   10
         USING CHANGE,15
         STM   14,12,12(13)
*
*        FETCH PARAMETERS
*
         L     RECORD,0(1)   INPUT
         L     START,4(1)    OUTPUT
*
*        FETCH 11TH HALFWORD INTO LIMIT
*
         LH    LIMIT,20(RECORD)
         C     LIMIT,=F'134' IS IT LEGAL
         BNH   INIT          YES
         LA    LIMIT,134     NO,PUT IN MAX POSS LEGAL VALUE ANYWAY
*
*        INITIALIZE
*
INIT     LR    COLBASE,START COLUMN BASE REGISTER
         XR    ROW,ROW       FIRST ROW
         XR    COUNT,COUNT   NUMBER OF HALFWORDS EXPANDED.
         LA    POSITION,24   DISPLACEMENT FOR 13TH HALFWORD.
*
*        EXPAND NEXT WORD.
*
NEXTWORD CR    COUNT,LIMIT   HAVE WE DONE ENOUGH YET?
         BNL   ZERO
         L     WORD1,0(RECORD,POSITION)
         XR    WORD2,WORD2   FOR SHIFTING LOW 16 BITS INTO.
         SRDL  WORD1,16
         OR    WORD1,WORD2   REASSEMBLE WORD IN CORRECT ORDER.
*
```

```
*        BREAK BITS OFF.
*
NEXTBIT  C      ROW,=F'124'      END OF CURRENT COLUMN?
         BH     INC              YES.
         SRDL   WORD1,1          GET A BIT
         SRL    WORD2,31         MAKE IT A 1 OR A 0
         ST     WORD2,0(COLBASE,ROW)      STORE IT.
         LA     ROW,4(ROW)       NEXT ROW DOWN.
         B      NEXTBIT
*
*        INCREMENT THE VARIOUS COUNTERS
*
INC      LA     COLBASE,128(COLBASE)      NEW COLUMN
         XR     ROW,ROW          FIRST ROW.
         LA     COUNT,2(COUNT)            2 MORE HALFWORDS DONE
         LA     POSITION,4(POSITION)      NEXT PAIR ON INPUT RECORD.
         B      NEXTWORD         AND CARRY ON.
*
*        SET REMAINDER OF OUTPUT AREA TO ZERO
*
ZERO     SLL    COUNT,6          NUMBER OF BYTES GENERATED,
         XR     WORD1,WORD1      A ZERO TO STORE.
         L      LIMIT,=F'8572'            ADDR OF LAST WORD IN OUTPUT ARE
STORE    CR     COUNT,LIMIT      IS ALL THE REST ZERO?
         BH     EXIT             YES.
         ST     WORD1,0(START,COUNT)      STORE ANOTHER ONE.
         LA     COUNT,4(COUNT)            POINT TO NEXT WORD.
         B      STORE
*
*        RETURN TO MAIN PROGRAMME.
*
EXIT     LM     14,12,12(13)
         BR     14
         END
```

APPENDIX 8.

Calculation of the Expected Overlap of $(A_n \cap g_{n+1})$.

Probability of overlap of 1 = P_1

Probability of overlap of 2 = P_2

Probability of overlap of 3 = P_3

$$\text{Expected volume of overlap} = \sum_{r=0}^{m} r . P_r$$

The probability of overlap of r bits is (the number of ways (m - r) bits can be taken from (N - y)) . (the number of ways r bits can be taken from A_n)/(the total number of ways m bits can be taken from N)

$$= \frac{{}^{N-y}C_{m-r} \; . \; {}^{y}C_r}{{}^{N}C_m}$$

Hence the expected volume of the overlap is

$$\frac{\sum_{r=0}^{m} r \; . \; {}^{N-y}C_{m-r} \; . \; {}^{y}C_r}{{}^{N}C_m}$$

APPENDIX 9.

Proof of $\sum\limits_{r=0}^{m} {}^{y}C_r \, . \, {}^{N-y}C_{m-r} = {}^{N}C_m$

By the binomial expansion

$$(a + b)^y (a + b)^{N-y} = \sum_{r=0}^{y} \binom{y}{r} a^r b^{y-r} \sum_{s=0}^{N-y} \binom{N-y}{s} a^s b^{N-y-s}$$

$$= \sum_{r,s} \binom{y}{r} \binom{N-y}{s} a^{r+s} \, b^{N-(r+s)}$$

Putting m = r + s and summing over r

$$(a + b)^N = \sum_{m=0}^{N} \binom{N}{m} a^m \, b^{N-m}$$

$$\therefore \binom{N}{m} = \sum_{r=0}^{m} \binom{y}{r} \binom{N-y}{m-r}$$

APPENDIX 10.

Best straight line fit of a set of points $\{x_i, y_i\}$ to $y = mx + c$ is given by

$$c = \left\{\sum x_i^2 - \sum y_i - \sum x_i \sum x_i y_i\right\} / \left\{n \sum x_i^2 - (\sum x_i)^2\right.$$

and $m = \left\{n \sum x_i y_i - \sum x_i \sum y_i\right\} / \left\{n \sum x_i^2 - (\sum x_i)^2\right.$

$$\sigma_m = \sigma_y \sqrt{n/(n \sum x_i^2 - (\sum x_i)^2)} \quad \text{and } \sigma_y = \sqrt{\sum (\delta y_i)^2/(n - 2)}$$

where $\delta y_i = y_i - (mx_i + b)$; σ_y is the standard deviation of the points from the expected y values; and σ_m is the standard deviation of m about the mean.

REFERENCES.

1. Neyman, J., 'Sur la verification des hypotheses statistiques composees', Bull. Soc. Math. France, 63,(1935).

2. Neyman, J. and Pearson, E.S., 'Contributions to the theory of testing statistical hypotheses', Stat. Res. Mem. Pts. 1 and 2 (1936).

3. Wald, A., Sequential Analysis, John Wiley, New York,(1947).

4. Wald, A., Statistical Decision Functions, John Wiley, New York, (1950).

5. Anderson, T.W., 'Classification by Multivariate Analysis', Psychometrika, 16, (1951), 31-50.

6. Luce, R.D. and Raiffa, H., Games and Decisions, John Wiley, New York, (1958).

7. De Groot, M.H., Optimal Statistical Decisions, Mc.Graw-Hill, New York, (1970).

8. Raiffa, H. and Schlaifer R., Applied Statistical Decision Theory, Harvard University Press, (1961).

9. Highleyman, W.H., 'Linear Decision Functions with Application to Pattern Recognition, Summary of Doctoral Dissertation', Optical Character Recognition, ed. Fischer et al, (1962).

10. Highleyman, W.H., 'Linear Decision Functions with Application to Pattern Recognition'., Doctoral Dissertation, University Microfilms, (1962).

11. Sage, A.P. and Melsa, J.L., System Identification, Mathematics in Science and Engineering, Vol. 80, Academic Press, New York, (1971).

12. Hays, W.L. and Winkler, R.L., Statistics : Probability, inference and decision, Vols. 1 and 2, International Series in Decision Processes, Holt, Rinehart and Winston, New York, (1970).

13. Fukunaga, K., Introduction to Statistical Pattern Recognition, Academic Press, New York, (1972).

14. Chow, C.K., 'On Optimum Recognition Error and Reject Trade-off', IEEE Transactions on Information Theory, IT-16, (Jan. 1970), 41-46.

15. Highleyman, W.H., 'A note on Optimum Pattern Recognition Systems', IRE Transactions on Electronic Computers, EC-10, (June 1961), 287-288.

16. Mendel, J.M., and Fu, K.S., Adaptive, Learning and Pattern Recognition Systems : Theory and Applications, Academic Press, New York, (1970).

17. Johnson, N.L. and Smith, H., editors, New Developments in Survey Sampling, John Wiley, New York, (1969).

18. Fukunaga, K., and Krile, T.F., 'Calculations of Bayes Recognition Error for two multivariate Gaussian distributions', IEEE Transactions on Computers, C-18, (1969), 220-229.

19. Ihm, P., 'Numerical Evaluation of certain Multivariate Normal Integrals', Sankhya, 21, (1959), 363-366.

20. John, S., 'On some classification statistics', Sankhya, 22, (1960), 309-316.

21. Bomba, J.S., 'Alphanumeric Character Recognition using Local Operations', Proc. Eastern Joint Computer Conference, (1959), 218.

22. Grimsdale, R.L. et al., 'A System for the Automatic Recognition of Patterns', Proc. IEE, 106 Pt. B, No. 26, (1959), 210.

23. Coombs, A.W.M., 'The Special Case of Postcode reading for the automatic sorting of machine printed mail', IOP conference on Machine Perception of Pattern and Pictures (1972), 62.

24. Ullman , J.R., 'Experiments with the n-tuple method of Pattern Recognition', IEEE Transactions on Computers, C-18 No. 12, (Dec. 1969), 1135-1137.

25. Rosenfeld, A., Picture Processing by Computer, Academic Press, New York, (1969).

26. Leifer, I., Rogers, G.L., and Stephens, N.W.F., 'Incoherent Fourier Transformations : a new approach to Character Recognition', Optica Acta, 16, No. 5 (1969), 535-553.

27. Jones, C.M. et al., 'Fourier Transform Methods for Pattern Recognition', Research Note AFCRL-62-512, Project 4691, (July 1962).

28. Hu, M.K., 'Pattern Recognition by Moment Invariants', (Letter), Proc. IRE, 49, (Sept. 1961), 1428.

29. Hu, M.K., 'Visual Pattern Recognition by Moment Invariants', IRE Transactions on Information Theory, IT-8, (Feb.1962), 179-187.

30. Alt, F.L., 'Digital Pattern Recognition by Moments', Optical Character Recognition, editors Fischler and Pollock (1962).

31. Taylor, W.K., 'Pattern Recognition by Means of Automatic Analogue Apparatus', Proc. IEE, 106 Pt. B, No. 26, 198.

32. Wada, H., 'An electronic reading machine', Proc. UNESCO conference on Information Processing, Olderburg Verlag, Munich, Butterworth, London.

33. Hosking, K.W., 'A Contour Method for the Recognition of Hand-Printed Characters', Marconi Review, 34, No. 181, (1971), 92-112.

34. Minsky, M.L., 'Self-Educating Pattern Recognition Schemes', Proc. IRE, 50 Pt.2, (1962), 1707.

35. Hu, M.K. et al., Theory of Adaptive Mechanisms, Syracuse University AD-429935, (Dec. 1963).

36. Chow, C.K., 'An Optimum Character Recognition System Using Decision Functions', IRE Transactions on Information Theory, EC-6, (Dec. 1957), 247-254.

37. Chow, C.K., 'Statistical Independence and Threshold Functions', IEEE Transactions on Electronic Computers, EC-14, (Feb. 1965), 65-68.

38. Bowman, R.M. and McVey, E.S., 'A Method for the optimal design of a class of pattern recognition systems', Pattern Recognition, Vol. 2 No. 3, (Sept. 1970), 181-198.

39. Edwards, A. Wood and Chambers, Robert L., 'Can A Priori Probabilities Help in Character Recognition?', JACM, 11 No. 4, (Oct. 1964), 465-470.

40. Das Gupta, S., Optimum Classification Rules for Classification into two multivariate normal populations', Ann. Math. Stat., 36, (1965), 1174-1184.

41. Chitti Babu, C., 'On the Probability of Error and the Expected Bhattacharyya Distance in Multiclass Pattern Recognition', Proc. IEEE (USA), 60, (Nov. 1972), 1451-1452.

42. Chitti Babu, C., 'On the Relationship Between the Equivocation and other Probabilistic Distance Measures Used for Feature Selection', Proc. IEEE (USA), 60, (Sept. 1972), 1098-1099.

43. Helstrom, C.W., Statistical Theory of Signal Detection, Pergamon Press, London, (1968).

44. Ullmann, J.R., Pattern Recognition Techniques, Butterworths, London, (1973).

45. Chhikara, R.S. and Odell, P.C., 'Discriminant analysis using certain normed exponential densities with emphasis on remote sensing application', Pattern Recognition, Vol. 5 No. 30, (Sept. 1973), 259-272.