# MACHINE RECOGNITION OF THAI TEXT

M. Gray. B.Sc. (Hons.)

Department of Computing Science,
University of Adelaide.

February, 1970.

# TABLE OF CONTENTS

Chapter Page

# SUMMARY

A character recognition system may be divided into four systems; namely, an input device, a character separator and preprocessor, a feature extractor and a categorizer. Each of these subsystems is investigated in the development of a recognition system for Thai text.

As a starting point for the investigation, a description of Thai printing and of its large and unusual character set is given.

The input device to the CDC6400 computer is a modified piece of office equipment, a Gestetener ES390 scanning machine. These modifications and the method of input from the scanner to the computer are described with a detailed consideration of timing requirements.

The isolation of each character from the scanned image of a page is complex in the case of Thai printing. A detailed description of a method for isolation is given, together with an investigation into the implications of the angle of tilt on the page in the scanner on this proposed method. Preprocessing of the binary image of each character after isolation, necessary to remove

random noise, is considered and character defects that may affect the result of the feature extraction subsystem are described. Some account is given of other work on pre-processing and of the adopted method.

It is found convenient to transform the binary image of a scanned character to a point in p-dimensional space for identification. A method for computing such points, called 'feature vectors' is outlined with emphasis being placed on the selection of those feature elements most suitable for use in the recognition system. Having selected a subset of the feature elements, a principal component analysis is used to reduce the dimension of the feature vectors. Experimental results illustrating the effectiveness of the selection technique are presented.

Many characters of the Thai alphabet are very similar in shape, but one out of each group of similar characters tends to occur more frequently than the others. This property is exploited in the categorizer by using a statistical model, optimal in the Bayes sense, which is briefly outlined. This model requires knowledge of the probability density functions for each category (or character), which are not easily estimated.

A method for approximating these functions is described and results presented.

Finally the effect of introducing rejection as a possible decision in the categorizer is investigated and experimental results given using both simulated and scanner data.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any University. To the best of my knowledge and belief, it contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

## Acknowledgements

I wish to thank the following persons with whom I have discussed either the content or form of this thesis: Professor J.A. Ovenstone, Dr. I.N. Capon, Messr. J.N. Weadon and my co-students.

# Chapter 1. Introduction

## 1.1 Background

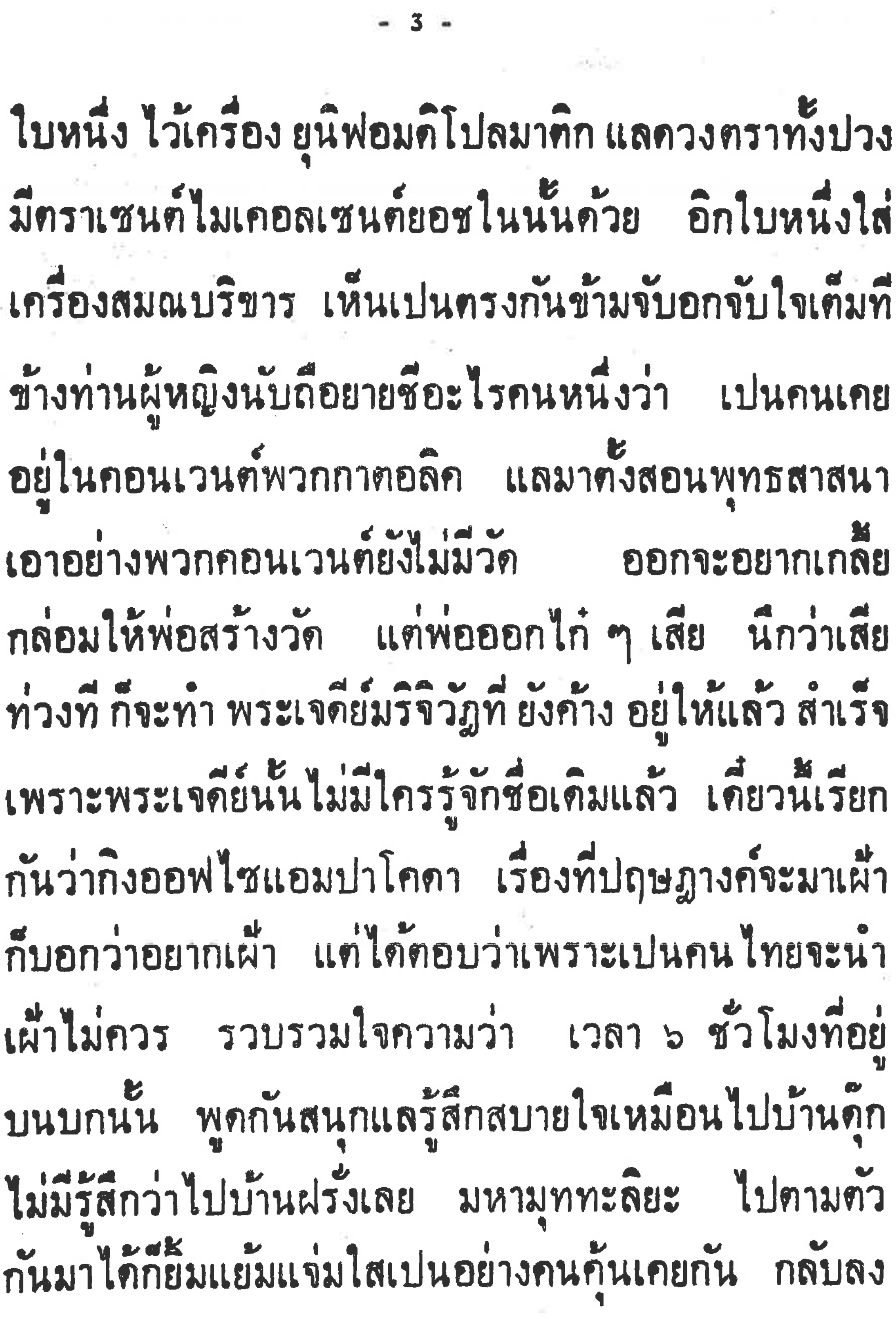
In recent years there has been an upsurge in the interest in the automatic translation of languages. For example, extensive work has been carried out on both Russian-English and Chinese-English translation and to a lesser extent on French-English and German-English translation [1,2]. In keeping with this current interest, Strehlow and Perry [3] have made a study of the automatic translation of Thai text into English, since little formal analysis has been made of this language and a knowledge of the structure and grammar of the Thai language is of some importance to Australia generally.

The latter authors wished to acquire data efficiently and economically in a form suitable for computer use as is the case for the automatic translation of any language. The first proposal they proposed was to purchase a card or teletype punch enabling a suitably trained operator to transcribe Thai into cards or teletype for subsequent analysis. However, the expense of buying or leasing such equipment was prohibitive and, in addition, an operator would have to be trained to interpret and punch the large and unusual character set. This method of data input would also be slow and there is no assurance that an operator even after training would be able to transcribe the text accurately. To enable the effort on the automatic translation of Thai text to continue, it became necessary to seek some other method of acquiring the data for computer use with the minimum of human intervention and interpretation during transcription and at the same time minimising costs. Thus an automated system for encoding Thai text was proposed and it is with this problem that the present study is concerned.

The need for machine recognition of Thai text for automatic translation was not the only reason for this study. The large and unusual character set of the Thai language which is intermediate and complexity between English and Chinese, makes this an extremely interesting character recognition problem. An example of Thai text is presented in Figure 1.1.1.

## 1.2 Thai Text

As outlined by Allison [4], the Thai language is printed as a sequence of characters left to right along a line with the lines running down the page as in English. There is no upper or lower case in Thai, each character always being the same size irrespective of its use. Normally



- 3 -

ใบหนึ่ง ไว้เครื่อง ยุนิฟอมดิโปลมาติก แลดวงตราทั้งปวง
มีตราเซนต์ไมเคอลเซนต์ยอชในนั้นด้วย อิกใบหนึ่งใส่
เครื่องสมณบริขาร เห็นเปนตรงกันข้ามจับอกจับใจเต็มที
ข้างท่านผู้หญิงนับถือยายชีอะไรคนหนึ่งว่า เปนคนเคย
อยู่ในคอนเวนต์พวกกาตอลิค แลมาตั้งสอนพุทธสาสนา
เอาอย่างพวกคอนเวนต์ยังไม่มีวัด ออกจะอยากเกลี้ย
กล่อมให้พ่อสร้างวัด แต่พ่อออกไก๋ ๆ เสีย นึกว่าเสีย
ท่วงที ก็จะทำ พระเจดีย์มริจิวัฏที่ ยังค้าง อยู่ให้แล้ว สำเร็จ
เพราะพระเจดีย์นั้นไม่มีใครรู้จักชื่อเดิมแล้ว เดี๋ยวนี้เรียก
กันว่ากิงออฟไซแอมปาโคดา เรื่องที่ปฤษฎางค์จะมาเฝ้า
ก็บอกว่าอยากเฝ้า แต่ได้ตอบว่าเพราะเปนคนไทยจะนำ
เฝ้าไม่ควร รวบรวมใจความว่า เวลา ๖ ชั่วโมงที่อยู่
บนบกนั้น พูดกันสนุกแลรู้สึกสบายใจเหมือนไปบ้านดุ๊ก
ไม่มีรู้สึกว่าไปบ้านฝรั่งเลย มหามุททะลิยะ ไปตามตัว
กันมาได้ก็ยิ้มแย้มแจ่มใสเปนอย่างคนคุ้นเคยกัน กลับลง

Figure 1.1.1. Thai Text.

Thai words do not have spaces between them, the usual spaces only being between groups of words. Apart from punctuation marks, at each character position there is always a symbol on the line, but in addition there may be detached symbols above and/or below the line. It is therefore convenient to consider a character position as being divided into three areas which will be called upper, middle and lower areas. The suffices, U, M and L respectively, will be used to distinguish symbols in these areas. Certain upper symbols may be placed above but between two middle symbols and the term "dual character position" is used to describe this case. (See Figure 1.2.1.).

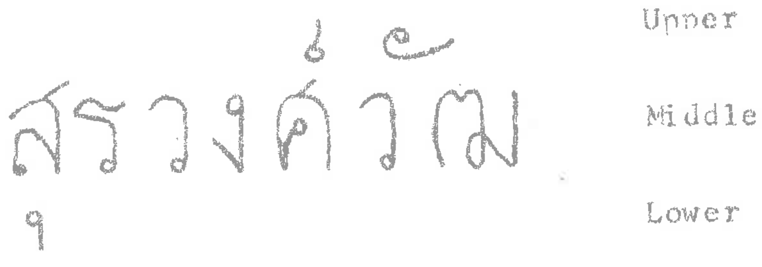


Figure 1.2.1. A sample from Thai text showing the three areas and the "dual character".

The distinct symbols can readily be divided into classes which consist of consonants (C), vowels (V), numerals (N), combined vowel-tone symbols to be called towels (W) and special symbols (S). There are 45 consonants in the Thai alphabet which may only appear in the middle area [5]. However, only slightly more than half of them are used extensively (some are even obsolete [4]). The vowels may be subdivided into those which occur in the lower area ($V_L$) of which there are two, those which occur in the middle area ($V_M$) of which there are seven, and five appearing in the upper area ($V_U$) plus one which occurs in the upper area between two middle area symbols ($V_S$) - the dual character case. There are ten numerals (N) which may occur only in the middle area. There are six tones which always appear in the upper area. Two of these have limited combinations with vowels. Towels occur only in the upper area and consist of combinations of the five members of $V_U$ with five of the tones, or combinations of the "dual character" with four of the tones, denoted by $W_V$ and $W_S$ respectively. Note that each tone is printed directly above the vowel with which it is associated. Finally, there are seven special symbols, six appearing in the middle area and one in the upper area. They can be summarized in the following way;

| | | |
|---|---|---|
| Middle Area ($S_M$): | ๆ | word repetition |
| | ฯ | abbreviation sign |
| | . | full stop |
| | ( | opening parenthesis |
| | ) | closing parenthesis |
| | - | horizontal dash |
| Upper Area ($S_U$): | " " | quotation marks |

Figure 1.2.2. illustrates all the individual shapes of the Thai character set. Table 1.2.1. is a summary of the vertical combinations that are possible in the Thai language.

When considering the problem of machine recognition of Thai text, one immediately considers the possibility of regarding each vertical combination as one character, but from Table 1.2.1. it is apparent that the number of possible vertical combinations makes this a formidable proposition, if not impossible. It is necessary to consider characters falling into the three natural areas, U, M, and L separately, but even then it is still necessary for the recognition of some combinations of symbols, in particular the towels.

Figure 1.2.2. The Thai Alphabet

(a) The Middle Area Characters. Nos. 1-43 are consonants (C), 44-50 vowels ($V_M$), 51 and 52 special symbols ($S_N$), 53-62 numerals (N), 63-66 special symbols ($S_M$). The two marked * are no longer in use.

(a) Middle Area

| | | | | | | |
|---|---|---|---|---|---|---|
| ก | ข | ค | จ | ฆ | ง | ฉ |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| ช | ซ | ฎ | ฏ | ฌ | ญ | ฐ |
| 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| ฑ | ฒ | ณ | ด | ต | ถ | ฤ |
| 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| ท | ธ | น | บ | ป | ผ | ฝ |
| 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| พ | ฟ | ภ | ม | ย | ร | ล |
| 29 | 30 | 31 | 32 | 33 | 34 | 35 |

Figure 1.2.2. (a) Continued.

(a) Middle Area (cont)

| | | | | | | |
|---|---|---|---|---|---|---|
| ว | ศ | ษ | ส | ห | ฬ | อ |
| 36 | 37 | 38 | 39 | 40 | 41 | 42 |
| ฮ | ฃ* | ฅ* | ะ | า | เ | แ |
| 43 | | | 44 | 45 | 46 | 47 |
| โ | ใ | ไ | ๆ | ฯ | | |
| 48 | 49 | 50 | 51 | 52 | | |
| ๑ | ๒ | ๓ | ๔ | ๕ | ๖ | ๗ |
| 53 | 54 | 55 | 56 | 57 | 58 | 59 |
| ๘ | ๙ | ๐ | | | | |
| 60 | 61 | 62 | | | | |

Figure 1.2.2. (b) The Upper Area Characters.
Nos. 1 - 6 are vowels, (Vu)
7 - 12 are tones (T) and
13 is a special symbol (Su)

Figure 1.2.2. (c) Lower Area Vowels $(V_L)$

(b) Upper Area

Vowels (Vu)

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| ิ | ี | ึ | ื | ํ | ั |
| 1 | 2 | 3 | 4 | 5 | 6 |

Tones (T)

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| ่ | ้ | ๊ | ๋ | ็ | ์ |
| 7 | 8 | 9 | 10 | 11 | 12 |

(c) Lower Area Vowels ($V_L$)

|  |  |
|---|---|
| ุ | ู |
| 1 | 2 |

Table 1.2.1. A Summary of the number of possible vertical combinations in the Thai language.

| | Single character positions | | | | | | | | | | Dual Character Positions | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Upper | | | T | $V_U$ | T<br>$V_U$ | | T | | | $S_U$ | $V_S$ | T<br>$V_S$ |
| Middle | N | C | C | C | C | C | C | $V_M$ | $S_M$ | | C C | C C |
| Lower | | | | | | $V_L$ | $V_L$ | | | | | |
| No. of Vertical Combinations | 10 | 45 | 44<br>x6 | 44<br>x5 | 44<br>x25 | 44<br>x2 | 44x<br>2x5 | 7 | 5 | 1 | 44x<br>44x1 | 44x<br>44x4 |

The symbols falling into these natural groups are summarized in Table 1.2.2.

Table 1.2.2. The number of symbols occuring in the three natural areas.

| Area | Type | Number | Total |
|---|---|---|---|
| Middle | N | 10 | |
| | C | 45 | |
| | $V_M$ | 7 | |
| | $S_M$ | 6 | 68 |
| Upper | $W_U\begin{pmatrix}T\\V_U\end{pmatrix}$ | 25 | |
| | $W_S\begin{pmatrix}T\\V_S\end{pmatrix}$ | 4 | |
| | $S_U$ | 1 | |
| | $V_U$ | 5 | |
| | $V_S$ | 1 | |
| | T | 6 | 42 |
| Lower | $V_L$ | 2 | 2 |

It becomes clear therefore, that there is a total of 112 symbols to be recognised, which can be divided into three distinct groups consisting of 68, 42 and 2 members. It was mentioned earlier that some consonants are rarely used and some are even obsolete. Indeed, it appears that there are 2 consonants which are not used in modern Thai writing [4]. Thus only 66 characters need be considered in the middle area group and the total number of symbols to be recognised is reduced to 110.

Before an automatic system of encoding the data was proposed, a considerable amount of Thai text from various sources was coded manually, punched into cards, and stored on magnetic tape. In excess of 10,000 characters were available from which, unfortunately, numerals and special symbols were excluded. Using this data a frequency count of character occurrence was made, the results of which are presented in Table 1.2.3. in the form of probability for both overall and within area occurrence. The overall probability for the upper area only takes into account the individual character occurrence, for example, a towel is counted as one vowel and one tone. However, for the within area probability a towel is considered as

Table 1.2.3. The Probabilities of character occurrence in Thai text.

| Char. No. | Overall Prob. (X10$^{+2}$) | Within Area Prob. (X10$^{+2}$) | Char No. | Overall Prob. (X10$^{+2}$) | Within Area Prob. (X10$^{+2}$) |
|---|---|---|---|---|---|
| | | Middle Area | | | |
| 1 | 3.83 | 5.16 | 28 | 0.04 | 0.05 |
| 2 | 1.09 | 1.47 | 29 | 1.17 | 1.57 |
| 3 | 2.38 | 3.21 | 30 | 0.14 | 0.18 |
| 4 | 1.88 | 2.53 | 31 | 0.23 | 0.31 |
| 5 | 0.01 | 0.02 | 32 | 3.60 | 4.85 |
| 6 | 3.26 | 4.39 | 33 | 2.49 | 3.36 |
| 7 | 0.09 | 0.12 | 34 | 5.39 | 7.26 |
| 8 | 0.89 | 1.20 | 35 | 2.15 | 2.89 |
| 9 | 0.31 | 0.42 | 36 | 2.45 | 3.30 |
| 10 | 0.00 | 0.00 | 37 | 0.52 | 0.70 |
| 11 | 0.02 | 0.03 | 38 | 0.19 | 0.26 |
| 12 | 0.05 | 0.07 | 39 | 1.88 | 2.53 |
| 13 | 0.45 | 0.67 | 40 | 1.94 | 2.61 |
| 14 | 0.10 | 0.14 | 41 | 0.02 | 0.03 |
| 15 | 0 03 | 0.04 | 42 | 3.97 | 5.35 |
| 16 | 0.05 | 0.06 | 43 | 0.01 | 0.01 |
| 17 | 0.80 | 1.08 | 44 | 0.00 | 0.00 |
| 18 | 1.82 | 2.45 | 45 | 0.00 | 0.00 |
| 19 | 1.75 | 2.36 | 46 | 2.31 | 3.12 |
| 20 | 0.47 | 0.63 | 47 | 6.44 | 8.67 |
| 21 | 0.04 | 0.06 | 48 | 3.09 | 4.16 |
| 22 | 2.05 | 2.76 | 49 | 1.11 | 1.50 |
| 23 | 0.23 | 0.31 | 50 | 0.41 | 0.58 |
| 24 | 5.51 | 7.42 | 51 | 0.81 | 1.09 |
| 25 | 2.12 | 2.85 | 52 | 2.05 | 2.76 |
| 26 | 1.72 | 2.32 | 53 | 0.10 | 0.14 |
| 27 | 0.68 | 0.92 | 54 | 0.04 | 0.06 |
| | | Lower Area | | | |
| 1 | 0.96 | 38.40 | 2 | 1.54 | 61.60 |

Table 1.2.3. (cont).

| Char.No. | Overall Prob. $(X10^{+2})$ | Within Area Prob $(X10^{+2})$ | Char.No | Overall Prob. $(X10^{+2})$ | Within Area Prob. $(X10^{+2})$ |
|---|---|---|---|---|---|
| | Upper Area (Vowels & Tones) | | | | |
| 1 | 3.56 | 11.47 | 7 | 3.79 | 15.24 |
| 2 | 1.64 | 12.10 | 8 | 3.57 | 15.10 |
| 3 | 2.71 | 3.01 | 9 | 0.01 | 0.01 |
| 4 | 0.50 | 1.12 | 10 | 0.06 | 0.01 |
| 5 | 0.89 | 3.99 | 11 | 0.63 | 5.45 |
| 6 | 0.61 | 12.59 | 12 | 0.73 | 5.80 |
| | Upper Area (Towels) | | | | |
| | Vowel Tone Combination | | | Vowel Tone Combination | |
| 13 | 1 - 7 | 0.28 | 28 | 4 - 7 | 1.61 |
| 14 | 1 - 8 | 0.28 | 29 | 4 - 8 | 0.42 |
| 15 | 1 - 9 | 0.00 | 30 | 4 - 9 | 0.00 |
| 16 | 1 - 10 | 0.00 | 31 | 4 - 10 | 0.00 |
| 17 | 1 - 12 | 0.14 | 32 | 4 - 12 | 0.00 |
| 18 | 2 - 7 | 3.43 | 33 | 5 - 7 | 0.07 |
| 19 | 2 - 8 | 2.66 | 34 | 5 - 8 | 0.07 |
| 20 | 2 - 9 | 0.00 | 35 | 5 - 9 | 0.00 |
| 21 | 2 - 10 | 0.00 | 36 | 5 - 10 | 0.00 |
| 22 | 2 - 12 | 0.00 | 37 | 5 - 10 | 0.00 |
| 23 | 3 - 7 | 1.82 | 38 | 6 - 7 | 0.56 |
| 24 | 3 - 8 | 0.14 | 39 | 6 - 8 | 2.66 |
| 25 | 3 - 9 | 0.00 | 40 | 6 - 9 | 0.00 |
| 26 | 3 - 10 | 0.00 | 41 | 6 - 10 | 0.00 |
| 27 | 3 - 12 | 0.00 | | | |

one symbol, hence the reason for apparently no relation between the two probabilities for each upper area character. Note that character 10 of the middle area group did not occur in the sample as was also the case with many towels. All that may be concluded is that these are very rare symbols and special account must be made of these in the recognition system. The attention of the reader is drawn to the numbering system for the characters presented in Figure 1.2.2. and Table 1.2.3. which will be used in what follows.

Special mention should be made of the characters in the upper area. Character 6 is the "dual character" which may be used by itself or with any of the four tones 7 - 10 inclusive. All of the other upper vowels may occur alone or with any one of the tones excluding 11.

Included in the upper vowels is the symbol "o", which in fact is not a true vowel [5]. It occurs with the middle area character "า" and together they form a vowel. This middle area character may also appear alone in which case it is a consonant, but when it appears with "o" which is always in the upper area of the <u>preceding</u> character, the combination is a vowel. This vowel will be referred to as the "special upper vowel". Note that the preceding character in addition to "o" may still have a tone in the upper area. For purposes of machine recognition, this special upper vowel is considered as two distinct symbols and consequently the probability of occurrence of symbol 45 (middle area only) takes into account the occurrence of 45 and the special upper vowel in Table 1.2.3.

## 1.3. The Recognition System

It is convenient to consider a Character Recognition System as a number of subsystems (see Figure 1.3.1.). For the proposed system it is necessary to examine each of these subsystems in turn, bearing in mind that the overall objective of this study is to develop a system for the automatic reading of Thai text which operates cheaply.

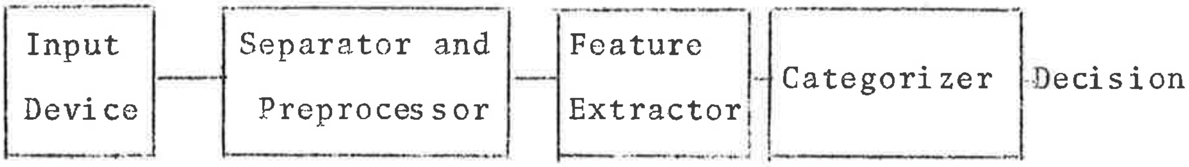


Figure 1.3.1. The subsystems of a Character Recognition System.

The Input Device or Scanner, is a standard office Gestetener ES390 copying machine, modified to produce a digitized version of the page being scanned. These modifications, the input from the scanner to the computer and the subsequent timing considerations are presented in Chapter 2.

The isolation of each character for its individual recognition from the scanned image of a page is complex in the case of Thai printing. It is with this aspect of the recognition system with which the Separator and Preprocessor are concerned. A possible method for isolation is described in Chapters 3 and 4 together with the preprocessing needed to remove random noise from the scanned image of each character to improve the recognition rate. It was found convenient to simulate the output of the Separator and Preprocessor subsystem to obtain experimental results in the development of the recognition system, a description of which is also given in Chapter 4.

The purpose of the Feature Extractor is to receive the preprocessed patterns and to derive a number of feature elements or measurements, thus effectively reducing the dimension of the data. If there are 'n' such measures then a pattern can be represented by a point in n-dimensional space and, hopefully, points belonging to the same category will cluster near one another. In this study, normalized bivariate central moments are used to form the elements of the feature vector for each pattern and a method is presented for selecting those most beneficial (although not necessarily optimal) to the recognition system. Results are presented for experiements with simulated data in Chapter 6.

Some groups of characters of the Thai alphabet are very similar in shape. However, the frequency count (see section 1.2) has shown that for these groups one character is far more likely to occur than the other; for example, character 11 has a probability of 0.02 of occurring while character 10 was not found in the sample taken. It was decided to exploit this property in the Categorizer, and thus a statistical decision theory model optimal in the Bayes' sense was adopted. A review is given in Chapter 5. A problem associated with this type of Categorizer is the unknown form of the probability density functions associated with each category. A method for approximating these functions is discussed in Chapter 7 together with results, once again obtained using simulated data.

Finally, in Chapter 8, results using both simulated data and the scanner are presented, together with a discussion of the difficulties encountered when applying the scanner to the recognition of this kind of text.

# Chapter 2. The Page Scanner

In this chapter a page-scanning device which provides suitable input to the character recognition system is described. The equipment used is essentially the same as that described by Jessup and Wallace [6] with one modification; eight grey levels are available instead of two. Basically the scanner consists of a Gestetener ES390, a piece of equipment that can easily be modified to provide a slow but cheap page scanner suitable as an input device for any computer, in particular the CDC6400 machine. The ES390 has a high resolution and is very effective for experimental work in character recognition.

## 2.1. General Description

The page of text to be scanned must be on a loose sheet of paper which is placed around a rotating drum and fixed in position by a flexible clear plastic cover. When operating, the drum (which is about 5 inches in diameter) rotates at approximately 240 r.p.m. while a scanning carriage, mounted on a lead screw, moves slowly past the drum parallel to the axis of rotation. The page is scanned in a spiral path but the pitch of the screw is small and adjacent scans may be considered to be both vertical and parallel with an error of less than

0.03 percent.

A basic requirement of the scanner is that a digitized version of a page must be independent of any variation in the drum rotation speed, and thus it was necessary to attach a mechanical clocking device to the scanner. A clear plastic disc with approximately 1600 fine markings around the edge was fixed to one end of the rotating drum. A focussed light beam, passing through a portion of the edge of the disc, is directed at two photoelectric cells mounted at such an angle that each marking on the edge causes two distinct pulses. Each pulse causes the intensity reading of the scanning carriage to be converted to a discrete grey level in the range 0 to 7. Special markings on the disc cause about 20 percent to be effectively blanked out and so no grey level readings are made. These markings are such that this blanked portion coincides with the clip on the rotating drum passing the scanning carriage. A photograph of the modified Gestetener is presented in Figure 2.1.1.

## 2.2. Input to the CDC6400 computer and System Operation

Input to the computer is via an interface unit built for the "Computer Assisted Instruction" project [7], the scanner being connected to one of the 64 local controllers (see Figure 2.2.1.) The

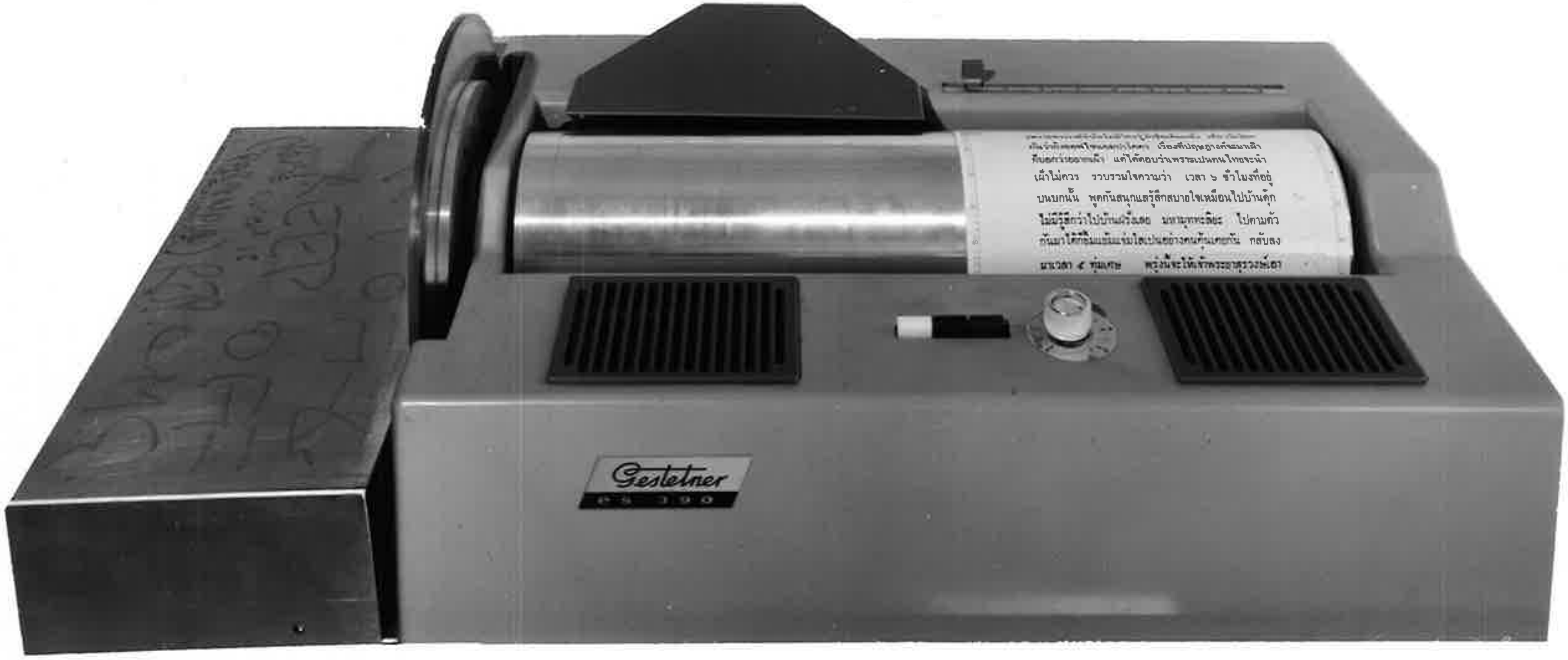


Figure 2.1.1. The Page Scanner.

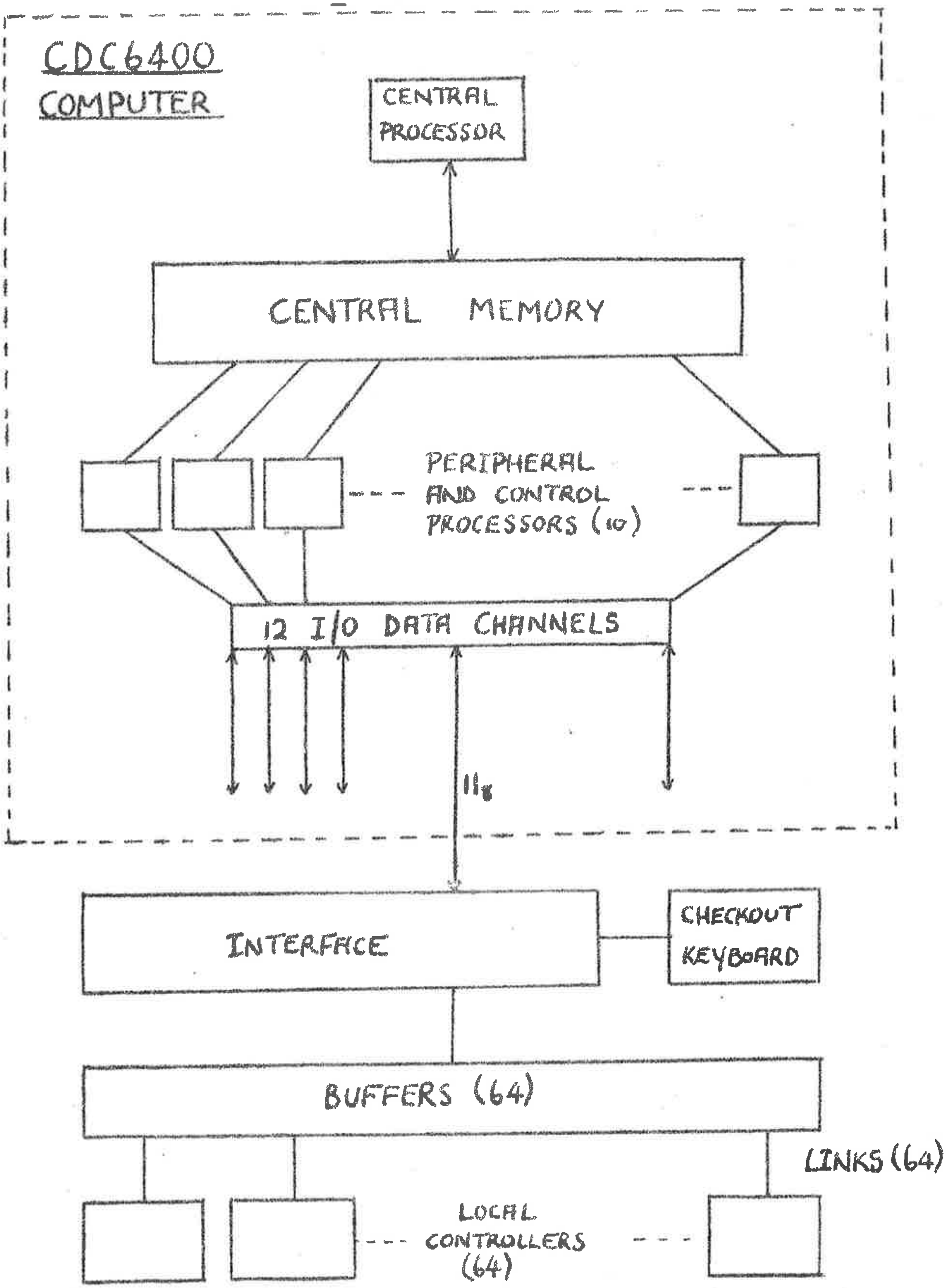

CDC6400
COMPUTER
CENTRAL PROCESSOR
CENTRAL MEMORY
PERIPHERAL AND CONTROL PROCESSORS (10)
12 I/O DATA CHANNELS
$11_8$
INTERFACE
CHECKOUT KEYBOARD
BUFFERS (64)
LINKS (64)
LOCAL CONTROLLERS (64)

Figure 2.2.1.

input is performed under the control of a driving program residing in one of the 10 available peripheral processors. A small buffer area of 512 (60 bit) words is required in the computer central memory (65K) and a second peripheral processor is used to write the contents of this buffer area onto magnetic tape when requested.

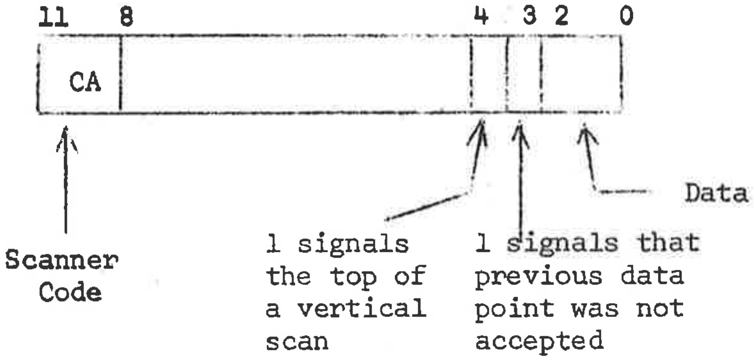

11
8
4
3
2
0
CA
Data
Scanner Code
1 signals the top of a vertical scan
1 signals that previous data point was not accepted

Figure 2.2.2. A Data Word

When operating, data for one vertical scan is stored in a peripheral processor as it becomes available, each data point being stored in the bottom 3 bits of one 12 bit word, with status information stored in the other 9 bits (see Figure 2.2.2). This rather extravagant use of storage is a direct result of the CDC6400 system. Firstly, the 12 bit data word is basic to the design of the computer which transfers words of this size through its data channels. Secondly, there is insufficient time to process each data point (check status information and take the appropriate action) and pack the data into the buffer area. It may be verified from the specifications given in the previous section that a data point becomes available at a rate of approximately 1 every 83 μ seconds (assuming no variation in drum speed.) To interrogate and receive a reply from the scanner takes 53 μ seconds, thus leaving a maximum of 30 μ seconds to process each data point. However, for safety, this must be reduced to allow for variation in the drum speed or distance between markings on the disc. The time required to simply store each data word in the buffer area is 11 μ seconds and thus it becomes apparent that there is insufficient time for any processing as a point becomes available.

When the clip is detected the contents of the peripheral processor buffer area are transferred to the central memory buffer area and the second peripheral processor is signalled that the contents of the buffer area are to be written on magnetic tape. The cycle is repeated when the first data point of the next scan is detected. Finally, when the scanner switches off, or is switched off, an "end-of-file" mark is written on the magnetic tape and the process is complete, the time taken for a complete scan being approximately 6 minutes.

One may ask, why not process the data words for 1 vertical scan when the clip is detected or before the data is written on magnetic tape? In the first case, when the time taken to transfer the data to central memory is taken into account, about 19 μ seconds are available to process each word; exactly the same situation as discussed above arises, that is, insufficient time. Secondly, the magnetic tape driver, with the large buffer area required for storing the information for 1 scan, only just fits into a peripheral processor, with only 26 locations not being used. Far more storage would be needed to enable the instructions required for processing to be added to the program.

Thus the information for a complete scan must be processed at a later point in time.

The scanner may be used during normal computer operation, but, while it is in use, one data channel and a magnetic tape unit become unavailable to other users, in addition to the two peripheral processors and buffer area. This is not a great demand on the system and as the central processor is not required during the scanning of a page the relative slowness of the ES390 is not costly.

## 2.3. Discussion

It was mentioned earlier that the intensity level picked up by the scanning carriage is converted to one of eight discrete grey levels, 0 corresponding to white and 7 to black for each pulse. Experiment showed that these levels were not reliable and, in particular, when scanning a uniform black line, unexpected variation in the grey level occurred, although white background was always recorded as 0, with little or no variation. Consequently the usefulness of this attribute of the scanner was limited at the time and prompted the following action. It has been found possible to construct a frequency diagram (Figure 2.3.1.) of the number of readings against each grey level for a complete scan and, using this to establish a reliable threshold which enables a decision to be made for each point of the scan whether it is black or white. The number of readings with grey levels 0 and 1 was usually large, which was to be espected since black printing on a white background was being used. At other levels the number was considerably lower with a slight increase for grey levels of 6 and 7. By setting the threshold above the level which is obviously white,(that is above 1), good results were obtained. Isolated areas of noise are of course introduced by choosing such a low threshold but these are removed in the preprocessing stage of the system (see Chapter 4).

For **the purpose of character recognition it is** natural to expect a point to be black or white and so this problem is not as serious as was first anticipated. However, it did prevent possible use of grey levels in determining the reliability of a reading which could be incorporated in the recognition process. This technique requires the processing of the magnetic tape on completion of a scan, but, as processing must be carried out to check status information, this was not a real problem, since the threshold can be found at the same time, although it was nearly always found to be 1.

One important feature of the scanner is its high resolution. Approximately 1800 scan lines of 2500 points each are available, giving a resolution in both the horizontal and vertical direction of approximately 0.005 inches, thus making it ideal for experimental work. A "small" middle area character generates about 32 bits vertically and 24 bits horizontally at this resolution.

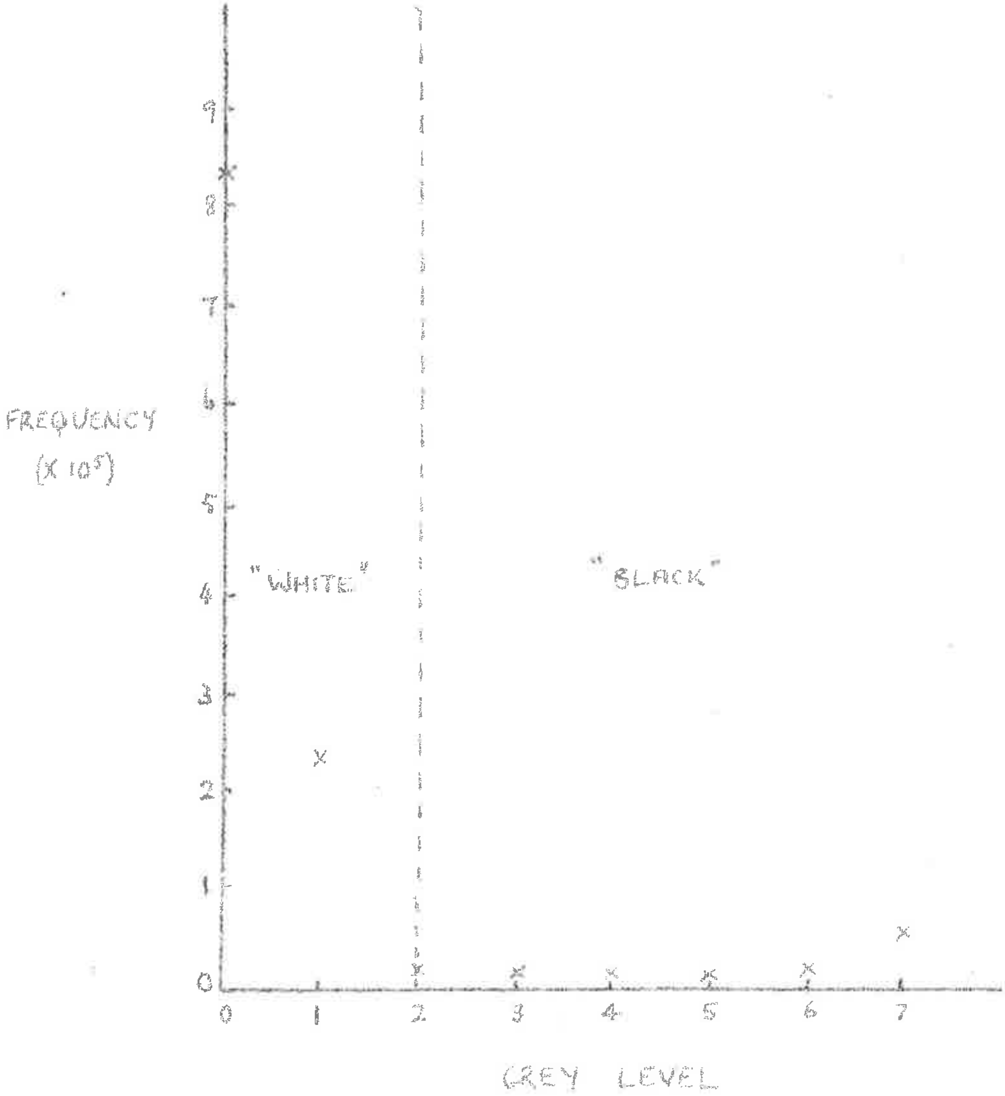


Figure 3.3.1. Frequency Diagram showing the Threshold Value

To process the magnetic tape takes about 12 minutes of central processor time using a simple FORTRAN program which occupies the minimum amount of central memory required by the computer system for a FORTRAN program. This program is used to check the status information associated with each data point, to take the appropriate action if error conditions occur, and write the processed scan information on another magnetic tape.

For each data point only three checks are needed, namely,

(1) the scanner code,

(2) the acceptance of the previous data point, and

(3) the first data point for each vertical scan.

If a data point does not have the scanner code, then it is assumed that the data originated from another piece of equipment connected to the same local controller, and the point is neglected.
In the event of a data point being missed, a local average of the intensity is computed and a dummy point with this intensity is inserted. If the first data point for a vertical scan is missed, an error message is printed indicating the scan number. The action taken in this case is to insert a series of zero intensity data points at the beginning of

the scan line to increase the number of points in the vertical scan to the same as the others. However, this condition rarely occurred. The input stage of the recognition system is presented schematically in Figure 2.3.2.

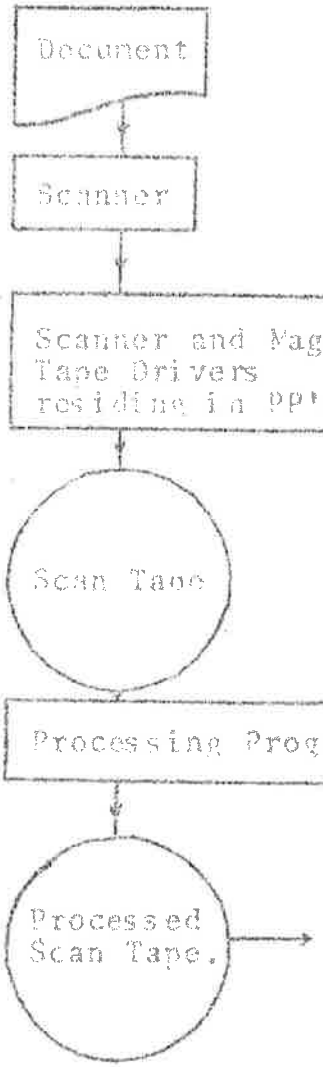


Figure 2.3.2. Flow diagram of programs in the Input Stage of the System.

# Chapter 3. Line Position

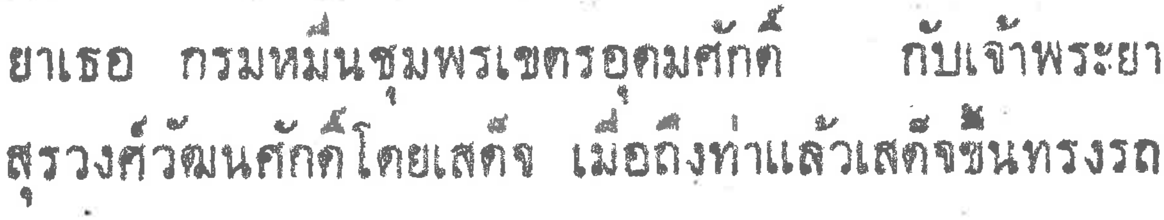
Object isolation is all too often ignored in laboratory studies, yet this is a major problem in automatic reading of machine printed text. In section 3.1 the isolation problem is presented, pointing out the difficulties in applying usual techniques to Thai printing, and thus establishing the need to find a point within the left hand bounds of the middle area of each printed line. A method for finding such a point with a study of the effect of page misregistration (or tilt) on it, is described in the next section.

Throughout this study it is assumed that each page of Thai text is scanned from top to bottom and from left to right across the page. Thus any horizontal coordinate on the page is specified by the number of data points from the beginning of a scan line, and a horizontal line across the page is defined by data points, with the same horizontal coordinate from successive (vertical) scan lines.

## 3.1. The Isolation Problem with Thai Printing.

A usual method adopted for the isolation of characters for their individual recognition is by the so called segmentation process [8]. Briefly, the approximate position of each line of print is found and a window of fixed dimension is stepped across the page, one scan at a time, searching for vertical white lines separating the individual characters. To enable this process to be effective a certain condition must hold:- namely, the continuous presence across the width of the page of gaps between successive lines of print. These gaps must be specified so that there are at least two bits which are blank across the width of the page. It would be expected that this process could be extended to Thai printing, to separate "character blocks", with a secondary horizontal segmentation being necessary to separate the different area characters from each resulting block. A minor technical difficulty would, of course, arise with the dual character, but this might not be too great a problem.

However, these are two reasons why this method does not suffice for Thai text. Firstly, on close examination of Thai printing, it is found that in some cases there is little or no separation between the lower area of one line and the upper area of the next. For example, if the sample of printing of Figure 3.1.1. was placed in the scanner exactly horizontally, it is unlikely that there would be a clear white gap of at least two data points between the two lines of print. This is due to the towel (upper vowel-tone combination) of the lower line almost extending into the lower area of the line above.



ยาเธอ กรมหมื่นชุมพรเขตรอุดมศักดิ์ กับเจ้าพระยา
สุรวงศ์วัฒนศักดิ์โดยเสด็จ เมื่อถึงท่าแล้วเสด็จขึ้นทรงรถ

Figure 3.1.1. Two lines of print not separable by a clear white line.

Even if these lines were separated sufficiently, then only a very small "angle of tilt," say θ, on the page in the scanner could be tolerated. This "angle of tilt" is the angle of the page in the scanner to the horizontal position. It should always be expected that the page will be placed in the scanner at some angle and so the method is not practical. Secondly, even if there is sufficient separation between the lines, then since the line width, 'h', is large compared with the inter-character distance, 'd', an impractical restriction on 'θ' is once again necessary, if character block separation is to be possible. The parameters are illustrated in Figure 3.1.2. In fact, using results from the scanner, 'h' and 'd'

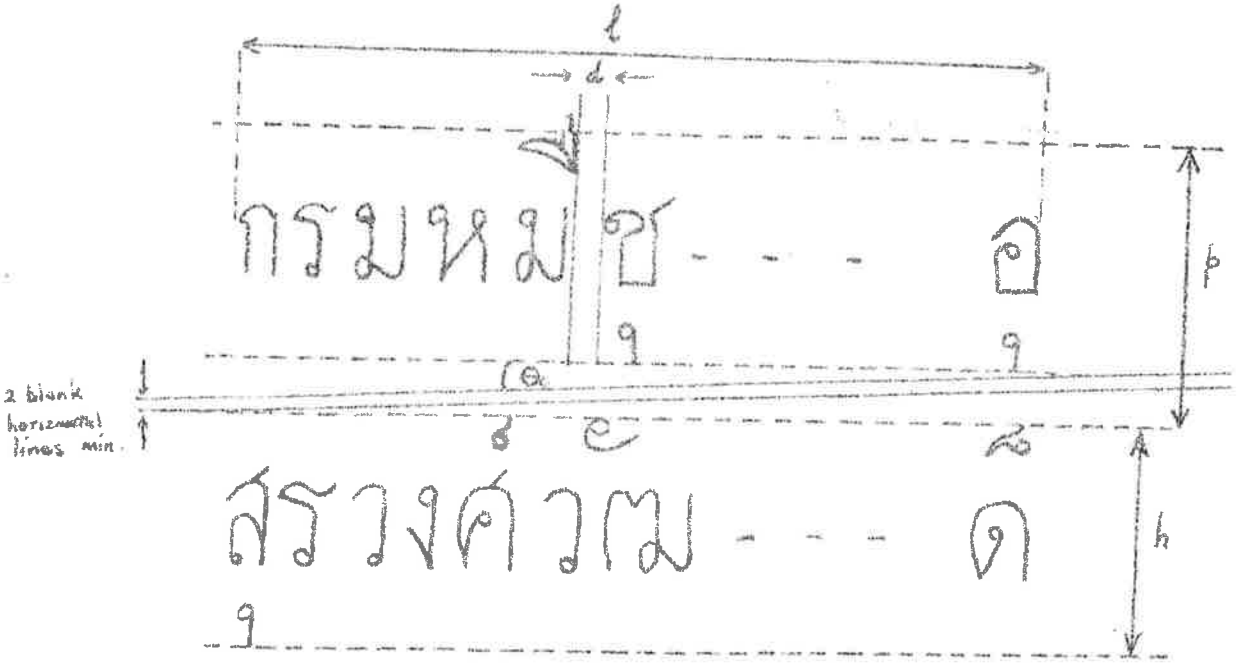


Figure 3.1.2. Tilt due to misplacement of the page in the scanner.

d = distance between adjacent characters

p = line pitch

h = line width

l = line length

$\theta_c$ = critical angle of tilt

$\tan \theta_c = \frac{p-h}{l}$

were found to be 0.5 inches and 0.04 inches, respectively. Bearing in mind that each character block must be separated by a vertical white line of fixed width, say three scans, 'd' is effectively reduced to approximately 0.01 inches thus making the critical angle of tile $\theta_c$ less than 1 degree. (The relationship between h, $\theta$ and d is derived in Chapter 4.) To add to the problem the upper area characters are not always printed exactly above the middle area character with which they are associated, thus reducing d even more.

It becomes apparent therefore, that a direct application of the normal separation process would not suffice for Thai printing. This led to considering the separation of the characters of each area separately (so reducing h and increasing the allowable tilt) which is a complex problem as some of the middle area characters extend into the upper area and others into the lower area. However, if the characters of the middle area for a line of print are isolated, then information obtained in this process can be used to separate the upper and lower area characters for the same line of print, since for Thai printing there is a fixed distance between the middle, and upper and lower areas. The author has observed that this distance is approximately equal to half the height of the middle area, the height of each of the upper and lower area characters also being equal to this height, a towel being larger. (see Figure 3.1.2.) Although a severe restriction on $\theta$ is still necessary, it is at least possible to ensure that it takes a practical value. The new procedure is not simple but it is considered necessary because of the complexity of Thai printing.

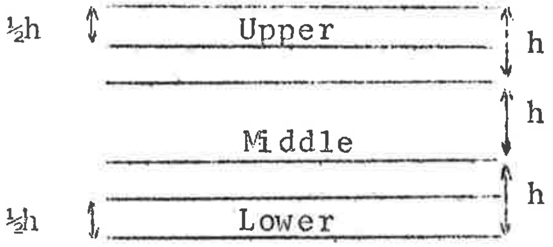


Figure 3.1.2. The relationship of inter-area distance h ≑ 0.167 inches.

## 3.2. Middle Area Position

Before any attempt can be made to isolate the character patterns it is necessary to establish the approximate horizontal coordinates of each line of print. Once a reliable threshold value has been found to distinguish between black and white (see Section 2.3), it is necessary to make another pass over the scan tape to convert each data point to 0 or 1 representing white or black, respectively. During this conversion a count is made of the black points across each horizontal line of the scan. By plotting a graph of count (which ideally would be intensity) against horizontal coordinate, the approximate position of each line of print may be found by locating the regions of maximum 'intensity'. However, because of the need to know to which area each isolated character belongs, it is necessary for the line finding process to yield a single point at the left of each line within the bounds of the middle area, which immediately implies a restriction on 'θ' for the page in the scanner.

### 3.2.1. The Model for Thai Printing

Intuitively, a study of the line finding technique and the implication of the above condition on the angle of tilt can be made by considering a model of a line of Thai printing. The model consists of three distinct horizontal blocks in which all data points are set to 1, that is, it is assumed that no middle area characters extend into either the upper or lower area, and both upper and lower characters occur with each middle area character. Both of these assumptions are not strictly correct. Firstly, some middle area characters do extend into the other areas, but the probability of a large middle area character occurring is much less than an ordinary sized character (see Table 1.2.3.) and so this assumption is made. Secondly,a count of frequency of occurrence of middle, upper (towels counting as one symbol) and lower area characters over ten pages of text showed that, if the probability of occurrence of a middle area character is assumed to be 1, the probability of an upper area character occurring in any position is 0.184 and that of the lower area 0.019. Since the intensity along horizontal lines is of interest, any contribution of the upper or lower area to the intensity is weighted according to the above probabilities to make the model more general. The model used in this investigation is illustrated in Figure 3.2.1. for various values of θ. The critical angle of tilt, $\theta_c$, is defined as arctan (h/1), where h is the perpendicular height of the middle area and 1 the length of the model line.

The effect of θ on the graph of intensity against horizontal coordinate for the model line may be investigated by plotting the graph for various values of θ (Figure 3.2.2.). It is found for θ less than $\theta_c$, the maximum intensity is not unique, that is, the same maximum intensity occurs for several horizontal coordinates. However, these maxima are all within the horizontal bounds of the left end of the middle area. For θ equal to $\theta_c$ a maximum intensity results at the horizontal coordinate corresponding to one of the horizontal bounds of the middle area. Finally, for $\theta$ greater than $\theta_c$, the maximum intensity is not well defined, but in this case all the corresponding horizontal coordinates lie outside of the bounds for the left of the middle area. To summarize, a point can be found within the left horizontal bounds of the middle area for $\theta$ less than or equal to $\theta_c$ and for $\theta$ greater than $\theta_c$ it is impossible to find such a point.

Figure 3.2.2. The Model for Thai Printing.

(a) $\theta = 0$

(b) $\theta = 1/2\ \theta_c$

(c) $\theta = \theta_c$

(d) $\theta > \theta_c$

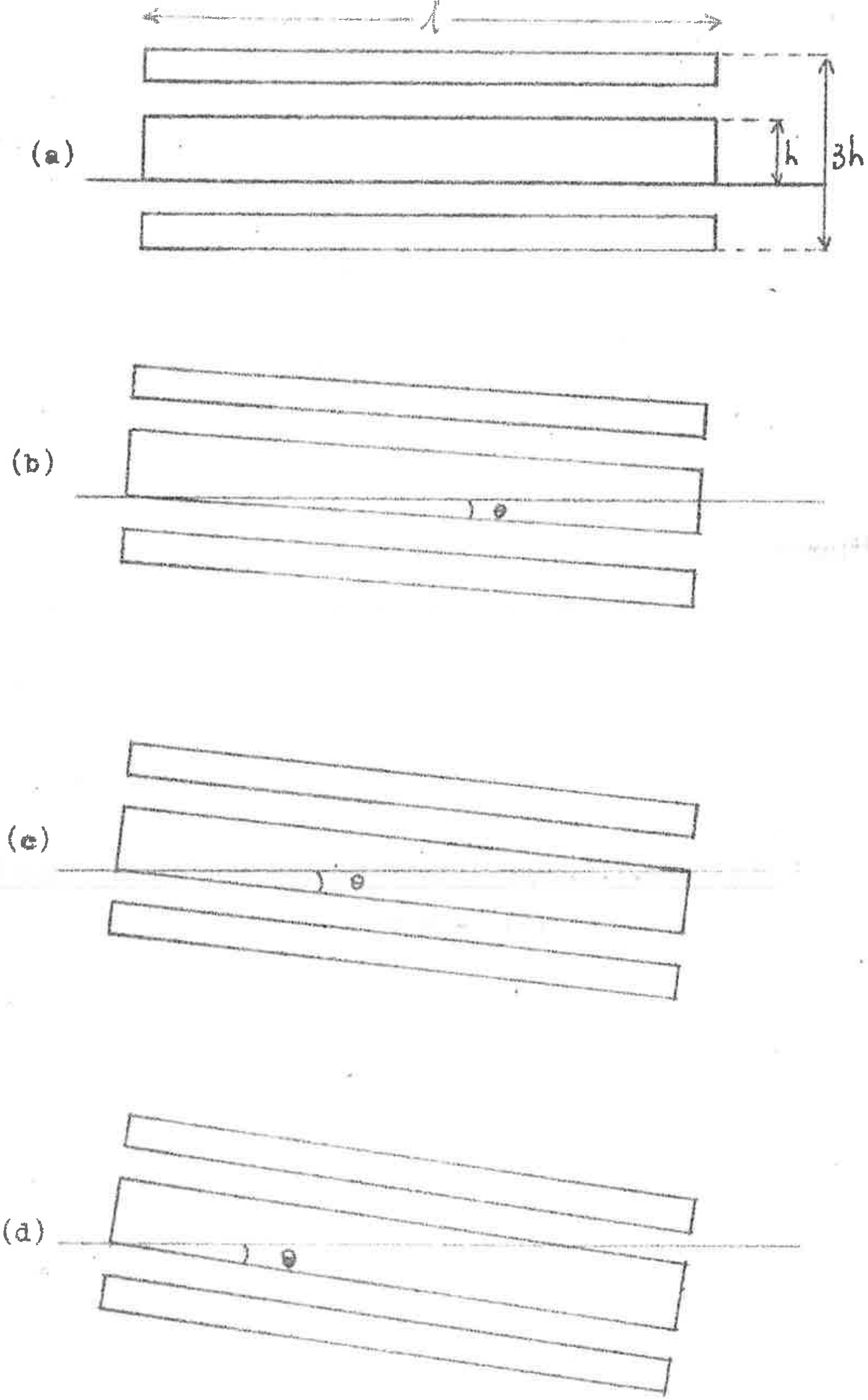

λ
(a)
h
3h
(b)
θ
(c)
θ
(d)
θ

Figure 3.2.2. The Intensity Graphs for the Model Line.

(a) $\theta = 0$

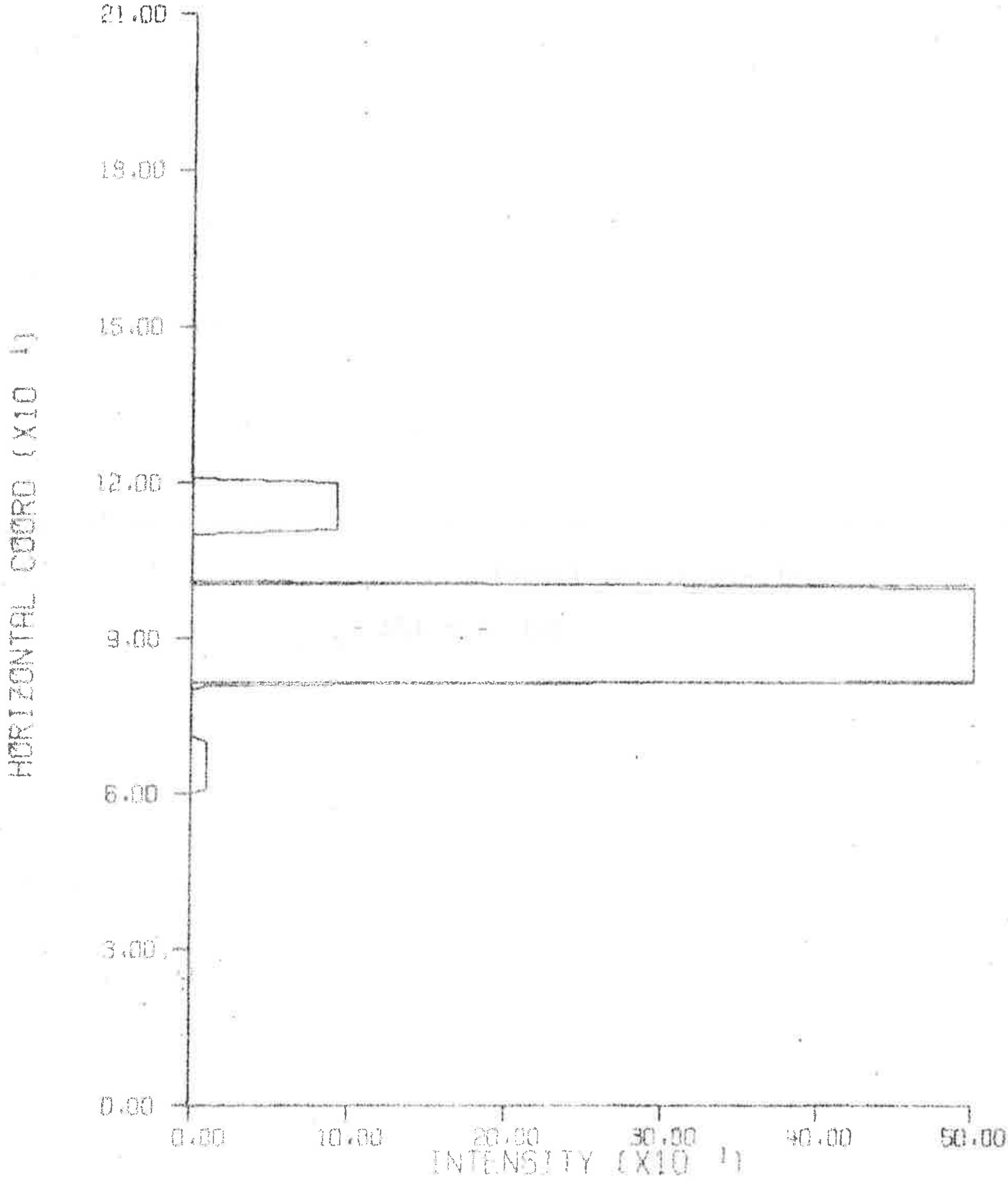

21.00
18.00
15.00
12.00
9.00
6.00
3.00
0.00
HORIZONTAL COORD (X10 1)
0.00
10.00
20.00
30.00
40.00
50.00
INTENSITY (X10 1)

Figure 3.2.2. (cont)

(b) $\theta = 1/2\ \theta_c$

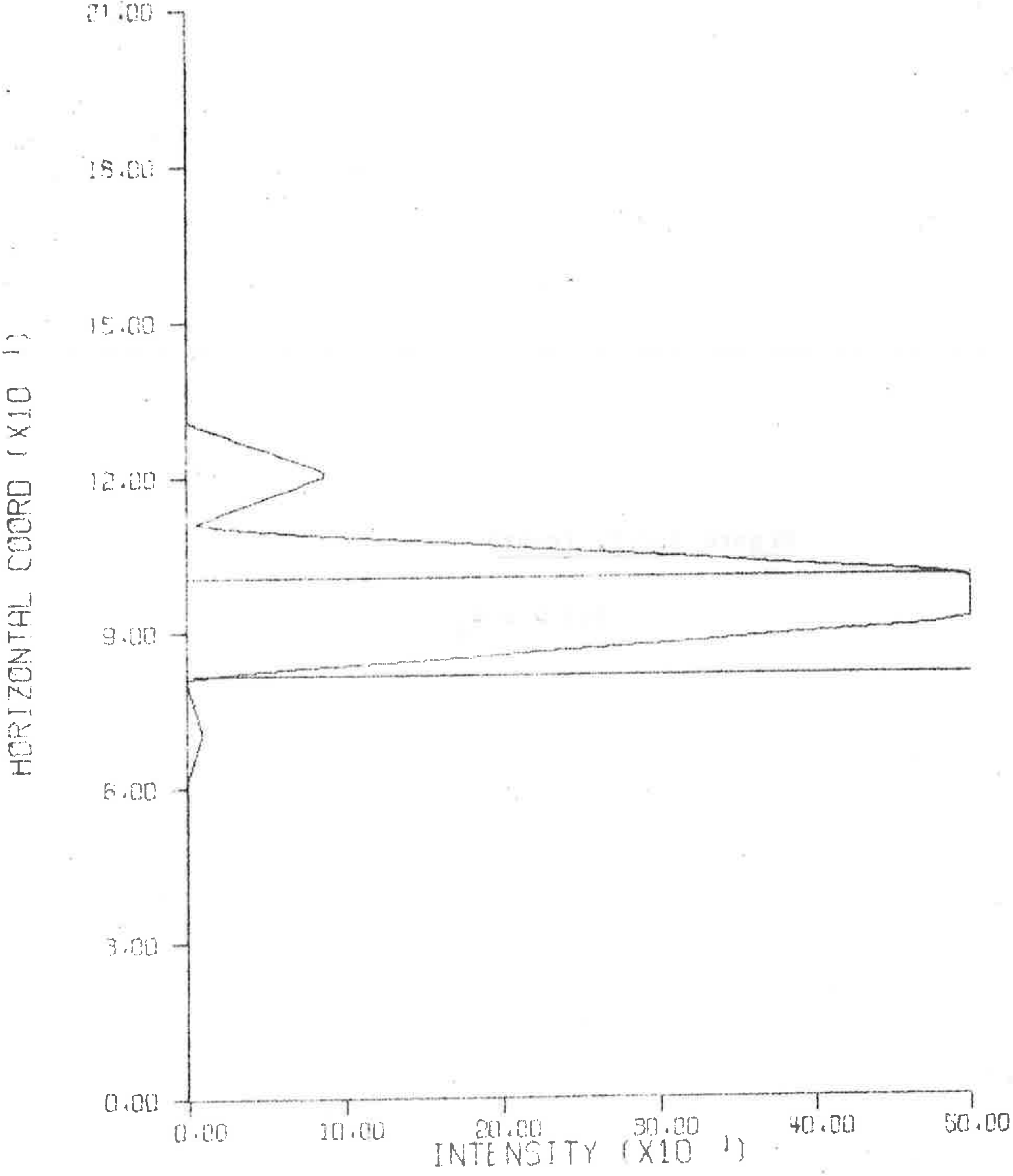

21.00
18.00
15.00
12.00
9.00
6.00
3.00
0.00
HORIZONTAL COORD (X10 1)
0.00
10.00
20.00
30.00
40.00
50.00
INTENSITY (X10 1)

Figure 3.2.2. (cont)

(c) $\theta = \theta_c$

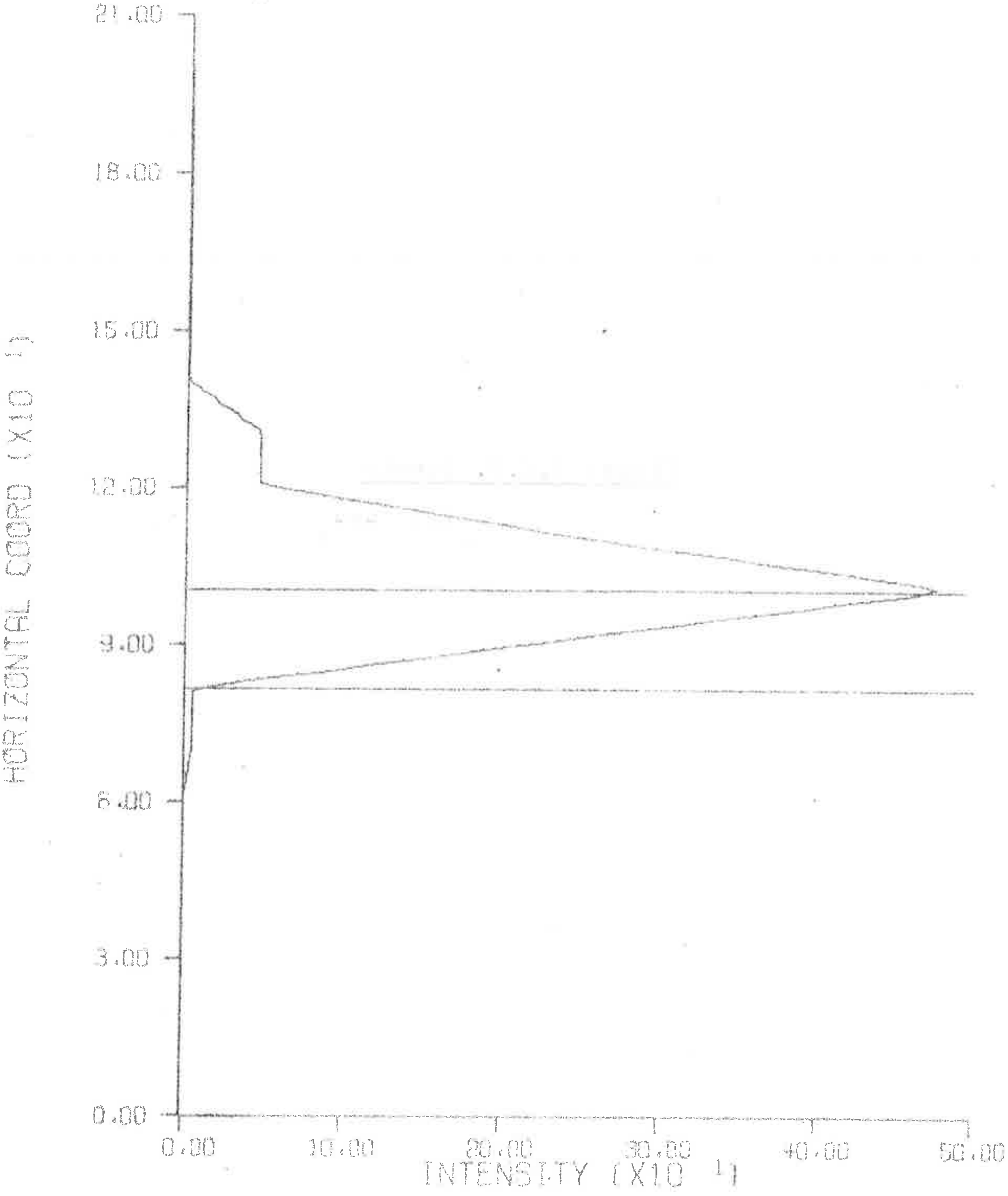
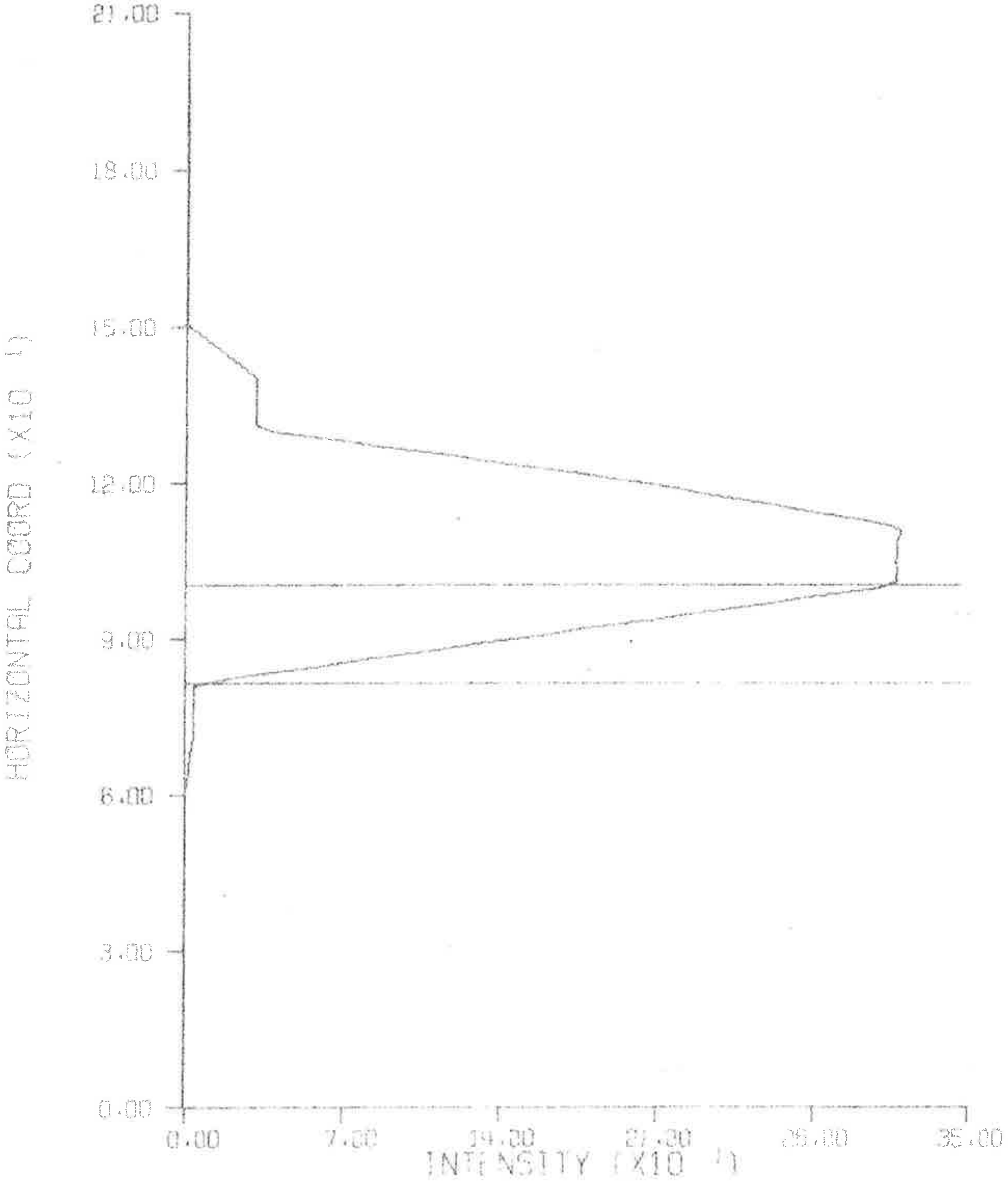

21.00
18.00
15.00
12.00
9.00
6.00
3.00
0.00
HORIZONTAL COORD (X10 1)
0.00
10.00
20.00
30.00
40.00
50.00
INTENSITY (X10 1)

Figure 3.2.2. (cont)

(d) $\theta > \theta_c$


21.00
18.00
15.00
12.00
9.00
6.00
3.00
0.00
HORIZONTAL COORD (X10$^{1}$)
0.00
7.00
14.00
21.00
28.00
35.00
INTENSITY (X10$^{1}$)

### 3.2.2. Line Position for Thai Printing

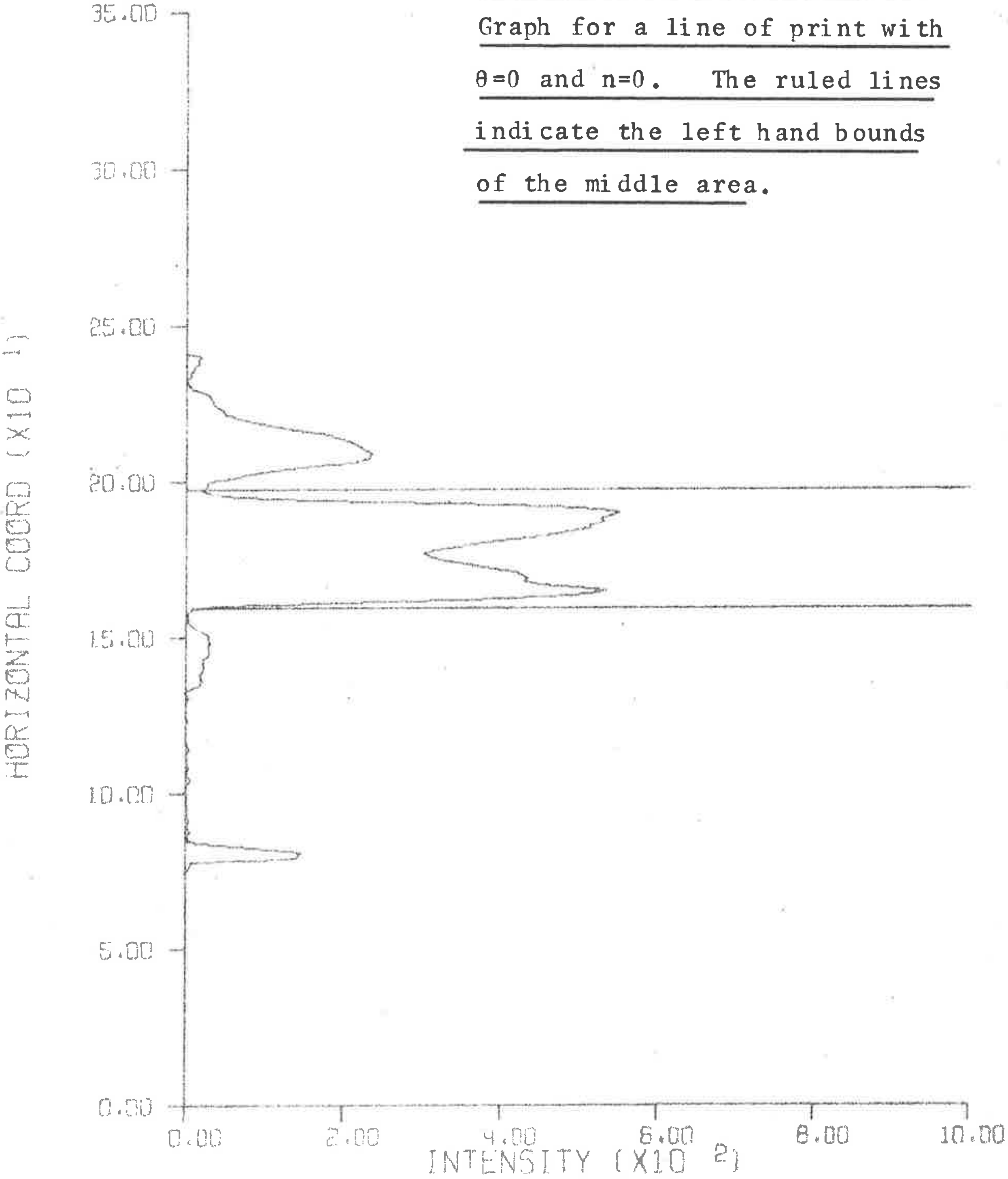
It would be expected, and it is indeed the case, that the preceding results can be extended to Thai printing,but, due to the assumptions made in constructing the model line, minor modifications are necessary.

The intensity plot for each line of Thai print does not form a smooth curve since the areas are not uniformly black as in the model. Many local maxima occur, but in the region of each line of print the intensities are relatively high. From the computational view point, it is desirable that a single maximum intensity value indicates a point within the bounds of the left end of each middle area, and thus there is a need for a smoothing process to eliminate the local maxima. The following "moving average" process is well known: for each intensity count, the adjacent 'n' original intensities both above and below, and the one under consideration, are averaged with the average replacing the original intensity level. If this procedure is repeated for the intensity levels for all horizontal coordinates, the intensity graph is smoothed with the amount of smoothing depending on the chosen value of 'n'. For a scanner with resolution 0.005 inches and the text used, a sufficiently large value of 'n' was found to be 12. However, for this value the maxima indicating the upper (and to a lesser extent the lower area) still exist. It is found by increasing the value of 'n' still further, these maxima dissappear, and then the only maxima remaining indicate the position of the middle areas which is a desirable feature. In general all the local maxima are eliminated for 'n' equal to 16, and the single maxima appear within the bounds of the middle area for each line of print. There is little advantage to be gained by increasing the value of 'n' beyond 16, the only effect being to spread the intensity curve for each line with a decrease in the maximum intensity values. In Appendix A the effect of increasing the value of 'n' on an intensity graph for a single line of print with θ equal to zero is illustrated.

In the model it is assumed that the three areas of Thai printing are distinct, but this is not strictly true because of the existence of large middle area characters which extend into either the upper or lower regions. However, since a normal-size, middle-area, character is more likely to occur than a large one, the only difference in the intensity graph for a real line of print and the model is a slight variation in the regions on either side of the peaks corresponding to the middle area. For small θ, distinct maximum intensities appear in both the upper and lower regions before smoothing, provided characters from these groups occur in the line of print under consideration. (It is possible a lower area character will not.) However, these maxima are smaller in magnitude than those for the middle area and there is no chance of confusing them. In any case they are eliminated with smoothing. These maxima are shown quite clearly in Figure 3.2.3., in which the separation points between the areas are clear. It should be noted that at the separation point between the upper and middle areas, the intensity is not zero because some middle-area characters extend into the upper area in the line of print for which the graph is constructed.

Previously it was stated that the value of $\theta_c$ for the model line was arctan (h/1). To discover whether or not this result can be extended to printing, lines of Thai text taken individually, were rotated for various values of θ by computer program and the intensity graphs plotted for each (see Appendix A.) It was found that the maximum intensity for the lines always occur between the left

Figure 3.2.3. The intensity Graph for a line of print with θ=0 and n=0. The ruled lines indicate the left hand bounds of the middle area.



horizontal bounds of the middle area for θ approximately in the range ±2 degrees without smoothing. Results from the scanner show that the height of the middle area, h, and the length of the line, l, to be approximately 0.2 inches and 5.5 inches, respectively, thus giving a theoretical value of θ for the model of

$$\theta_c = \arctan (0.2/5.5) \doteq 2^o,$$

which is in agreement with the experimental result. However, with smoothing (n=16), it is found that the range of θ is reduced with the new maximum for each line occurring within the required bounds for θ in the range of -2 degrees and +1.5 degrees (see Appendix A.) Thus to guarantee the maximum indicating the required point, 'θ' should be restricted to these limits.

This small value for θ could have been improved by using less of each line of print to establish the horizontal point in the middle area. That is, l may be reduced but care must be taken to ensure that it is long enough to provide sufficient information for each line of print. To show how $\theta_c$ may be increased, consider the model of a page of text presented in Figure 3.2.4. Suppose all information up to the vertical scan AB may be used in the line finding process and $l_1$ and

$l_2$ are the lengths of the top and bottom lines on the page that are to be used to "find" the line. If $l_1$ is such that enough information is available to establish the position of the top line, $l_2$ is assumed sufficient for the bottom line. But the critical angle for the top line exceeds that of the bottom line. Consequently $\theta_c$ for the page is determined by the length of the bottom line portion $l_2$. Now $l_2$ is less than the length of the line, $l$, and thus $\theta_c$ which is now equal to arctan $(h/l_2)$ is increased. For $\theta$ opposite in sign the critical angle is determined by $l_1$. It can however, be dangerous to adopt this technique, since the length of line necessary to give sufficient information will vary because of the gaps which may occur in a line of print. As a consequence, throughout this study care was taken when placing a page in the scanner to ensure that it was as horizontal as possible, and the full length of the lines used to determine the required horizontal point.

An alternative process to finding the approximate position of each line of print is to find the position of the first and second lines accurately by the method described, and then make use of the variable dimension window (see Chapter 4.) This gives an accurate estimate of the interline distance and the

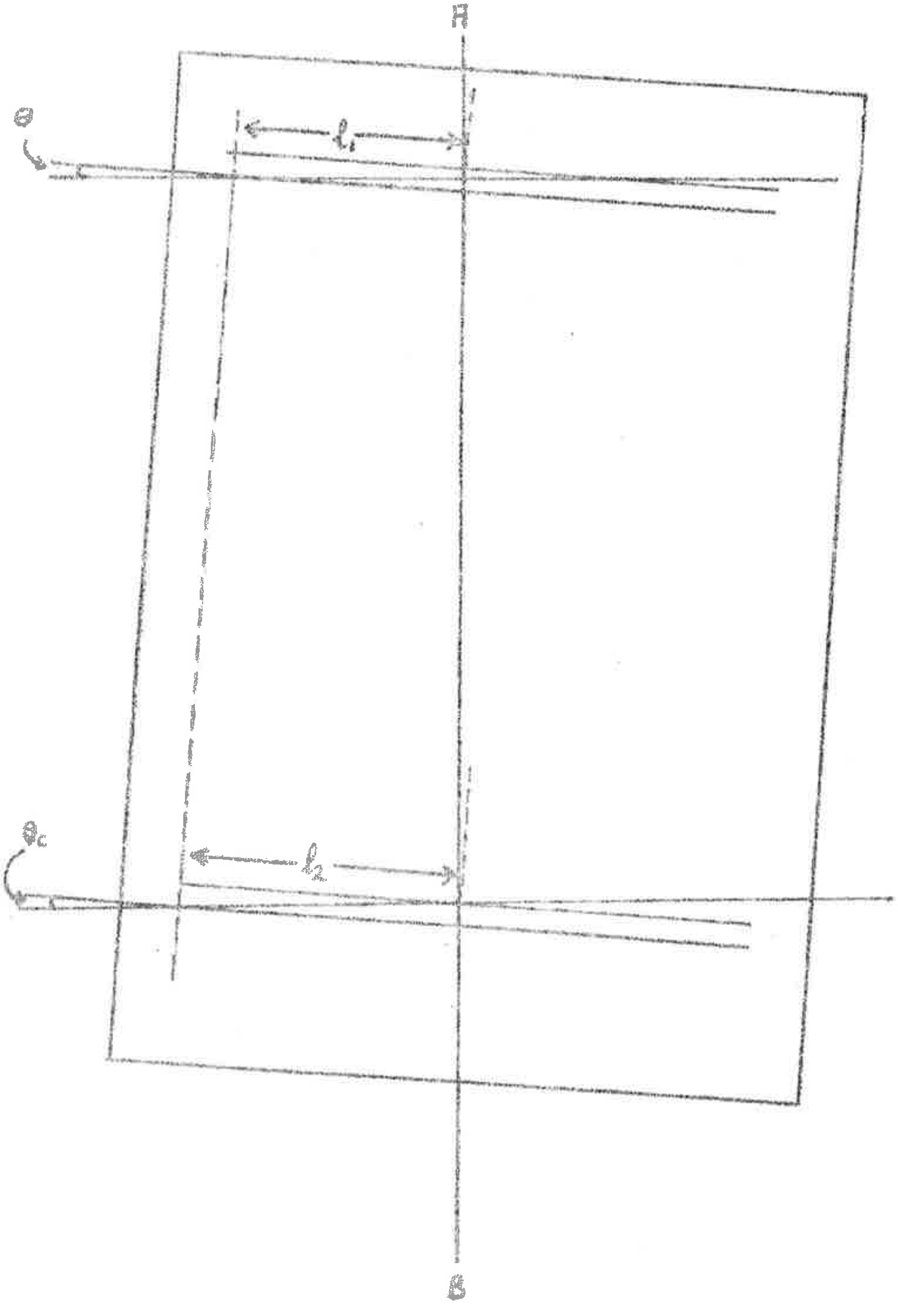


Figure 3.2.4. $\theta_c$ = arctan ($h/l_2$)

angle of tilt from which the position of successive lines of print can be calculated. However, since the interline gaps may vary where paragraphing occurs, this method is not satisfactory generally.

### 3.3. The Conversion and Line Position Program

As mentioned earlier, horizontal intensity counts for a complete scan, (1800 vertical scans), can be made at the same time as the conversion of each data point to zero or one. This depends on the threshold value found by the first of the scan processing programs (see Chapter 2). The storage requirements of this program are not great. Two arrays of 2,500 locations are needed to store the progressive and the subsequent smoothed intensity counts, one of 125 and another of 42 locations to store the processed and the converted information of one vertical scan, respectively. The vertical scans are converted one at a time, with the resulting data points written on a second magnetic tape, having been packed 60 points to a computer word.

At the completion of the conversion, the intensity counts are smoothed and the local maxima indicating line position are found and printed for use in the character isolation program (see Chapter 4.) The total computing time for this processing for a complete scan is about 120 seconds. A block diagram of this stage of the system is presented in Figure 3.2.5. which continues on from page 31.

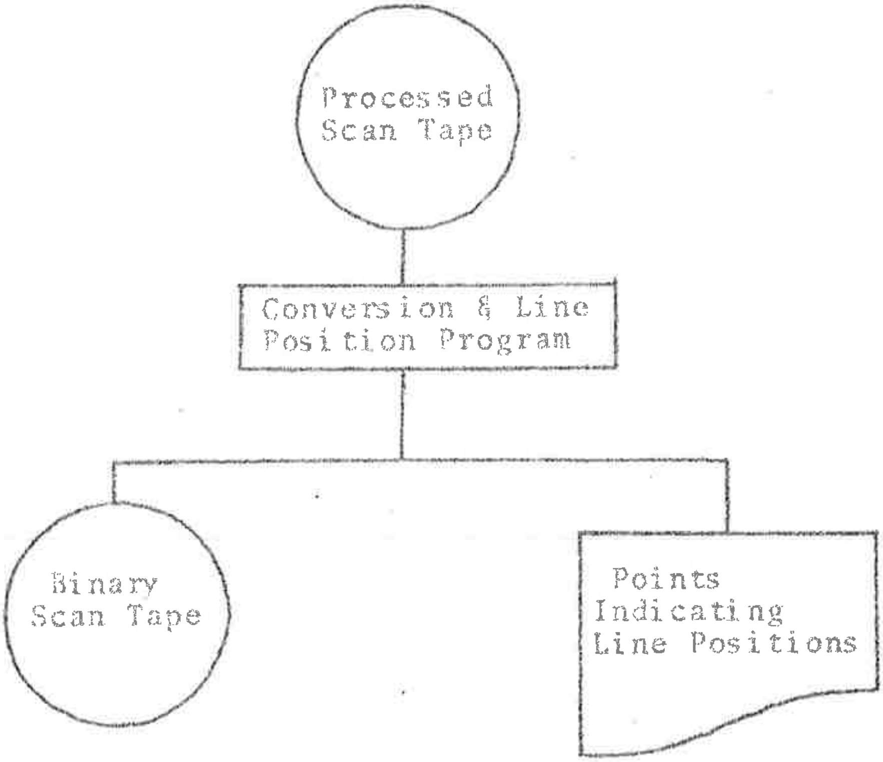

Processed Scan Tape
Conversion & Line Position Program
Binary Scan Tape
Points Indicating Line Positions

Figure 3.2.5. Block diagram of Conversion and Line Position Programs.

## Chapter 4. Character Isolation and Preprocessing

In this Chapter there are two problem areas that are considered, namely:

(a) the isolation of characters, and

(b) the removal of isolated areas of noise from each resulting character pattern.

Because of the nature of the output from the page scanner, these two areas were investigated by simulating a possible reading machine for Thai printing by computer program. This investigation gave an indication of the complexity of an automatic reading machine that would be needed, if such a machine is to be constructed.

In section 4.1 a method for isolating each character from the scan image of a page is described. Since it is not expected that any pageof print will be placed in the scanner exactly horizontally, the effect of the tilt angle of the page on the proposed technique must be considered.

Once the character patterns are isolated, it is necessary to remove noise to increase the chance of correctly identifying each one. A review of previous work in this field, the method used, and results are presented in section 4.2.

Before the scanner became available it was necessary to simulate the output of the preprocessing stage of the character recognition system. This simulated data proved to be extremely useful in obtaining preliminary results, which are incorporated in the final system. A description of the preparation of this data is given in section 4.3.

## 4.1. Isolation of Characters

Once a point is found within the horizontal bounds of the middle area of each line of print, the procedure to separate the characters for individual recognition may begin. The most difficult problem to overcome is the feature that some of the middle area characters extend into either the upper or lower area. Because of this, it is necessary to implement a "window" (or a photodiode array in the case of a reading machine). A "window" may be considered as an aperature through which a portion of the scan image of the page may be viewed. This window is fixed at 3 scans in width and has a variable height (called its dimension) which can be adjusted to enable it to extend over each character pattern in turn. The procedure is probably best described by simply stating that, as it moves right across the character, this window of fixed width is expanded vertically and manoeuvered until it just exceeds the vertical bounds of the character pattern by two clear horizontal lines both above and below. The right hand limit of each character is assumed when the window reaches a "white" area. An "area" is considered "white" when the bit count within the window is less than a fixed threshold, which is chosen according to the noise level of the scan image.

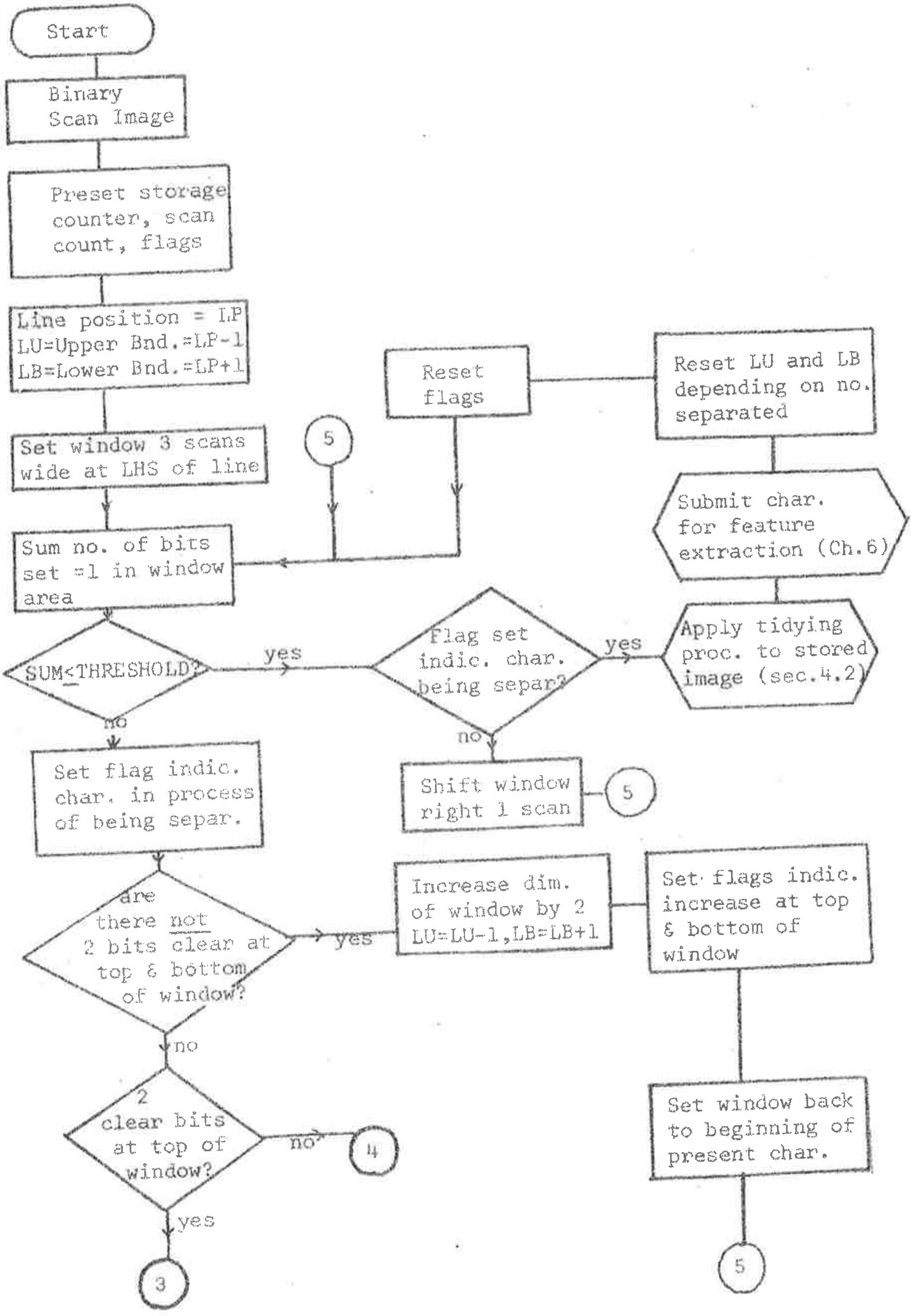
This isolation technique is implemented by computer program. The portion of the scan image of the page encompassing the line of print, (indicated by the line position coordinate), is stored in central memory, having been read from the binary scan tape. As each character is being revealed by the moving window on each step to the right, the contents of the left-most column of the window are stored in a column of a 2-dimensional array, the current column being indicated by a counter. If it is found that the window is not sufficiently expanded to extend beyond the bounds of the character, it is set back to its "most recent" starting point, and the indicating counter for the 2-dimensional array restored to its initial value (unity). The "most recent" starting point for the window is determined by the end of the previous character or, if it is the first character of a line, by the first vertical scan line. Many variations of this situation of insufficient window expansion arise. The appropriate action for each situation is best explained by the flow diagram of Figure 4.1.1.

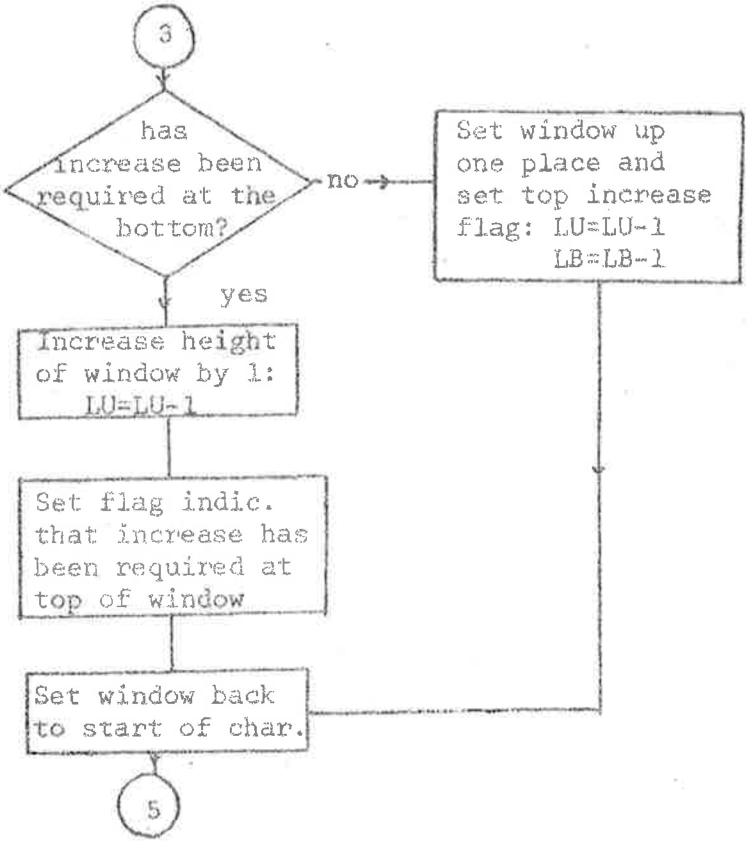
### 4.1.1. Middle Area Separation

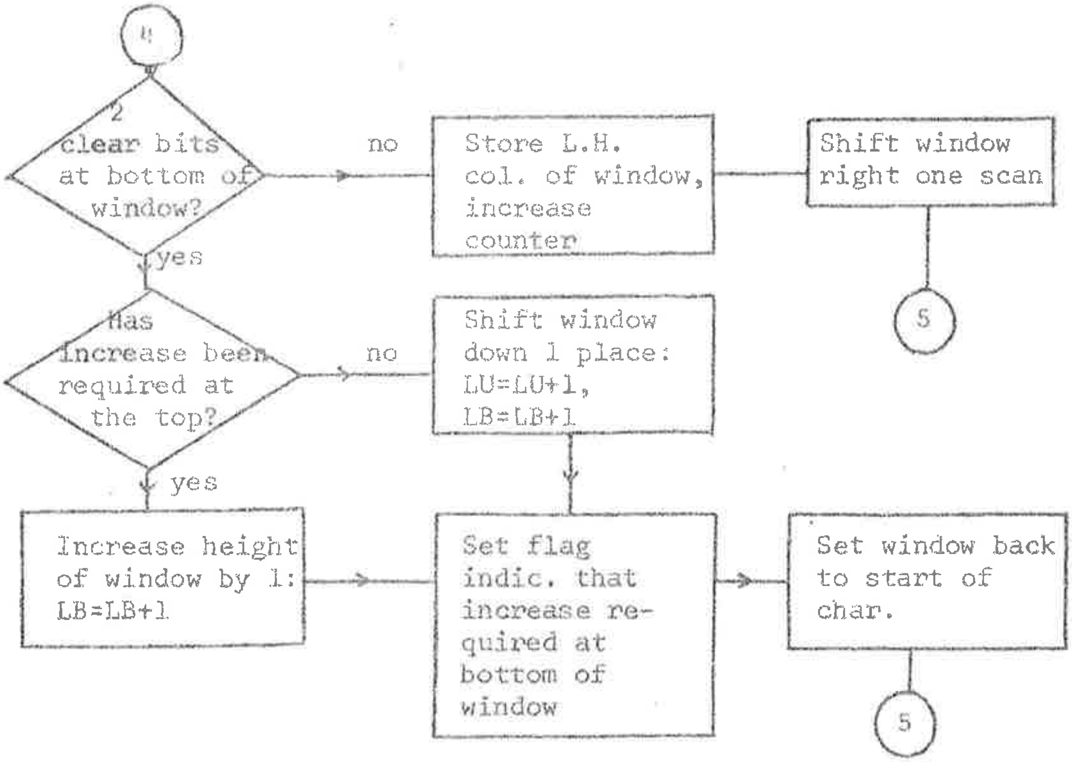
To isolate the characters of the middle area, the window initially with a dimension of 3, is set at the left end of a line of print about the point found within the bounds of the middle area, (see Chapter 3.) It is then manoeuvered, its dimension necessarily being increased, to isolate the first character. After isolating the first character and assuming that no information is available about the area height, the extreme upper and lower horizontal coordinates of the window obtained for this first character, are stored. The dimension and horizontal position of the window are then restored to their initial values, the window moved right, and the next character separated in a like manner. This process is repeated for say, the first five characters of the line.

A line of print always begins with a word and reference to a Thai dictionary [2] reveals it is thus assured that at least one ordinary size character will occur in the first five of a line. Consequently, after five characters have been isolated, the stored upper and lower horizontal coordinates are examined

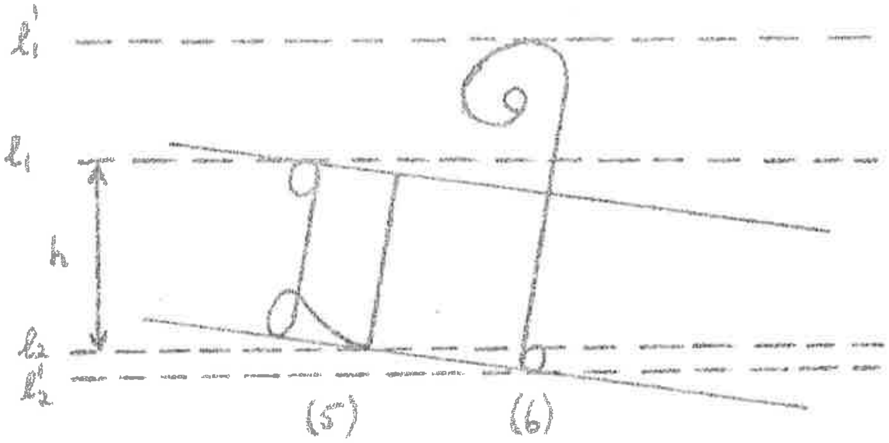
Figure 4.1.1. The isolation technique employing a variable dimension window.


Start
Binary Scan Image
Preset storage counter, scan count, flags
Line position = LP
LU=Upper Bnd.=LP-1
LB=Lower Bnd.=LP+1
Set window 3 scans wide at LHS of line
Sum no. of bits set =1 in window area
SUM≤THRESHOLD?
yes
no
Flag set indic. char. being separ?
yes
no
Apply tidying proc. to stored image (sec.4.2)
Submit char. for feature extraction (Ch.6)
Reset LU and LB depending on no. separated
Reset flags
5
Shift window right 1 scan
5
Set flag indic. char. in process of being separ.
are there not 2 bits clear at top & bottom of window?
yes
no
Increase dim. of window by 2
LU=LU-1,LB=LB+1
Set flags indic. increase at top & bottom of window
Set window back to beginning of present char.
5
2 clear bits at top of window?
no
4
yes
3


3
has increase been required at the bottom?
no
Set window up one place and set top increase flag: LU=LU-1
LB=LB-1
yes
Increase height of window by 1:
LU=LU-1
Set flag indic. that increase has been required at top of window
Set window back to start of char.
5


4
2 clear bits at bottom of window?
no
Store L.H. col. of window, increase counter
Shift window right one scan
5
yes
Has increase been required at the top?
no
Shift window down 1 place:
LU=LU+1,
LB=LB+1
yes
Increase height of window by 1:
LB=LB+1
Set flag indic. that increase required at bottom of window
Set window back to start of char.
5

to establish the approximate upper and lower bounds for the next character of the line. It is possible to find these bounds since only a small angle of tilt on the page in the scanner is permissible to satisfy the requirements of section 3.2. As a result, there is no chance of confusion between the bounds of a large and ordinary character. For example, suppose that it is established that the fifth character of the line is of normal size with height, h, and upper and lower bounds, $l_1$ and $l_2$, respectively, as illustrated in Figure 4.1.2. Then the approximate horizontal bounds for the next character are assumed to be $l_1$ and $l_2$ and the window is shifted right with dimension h(equal to $l_2$-$l_1$), and horizontal bounds $l_1$ and $l_2$, to isolate the next character. If the next character is large, then the window must be expanded and shifted to fit over the character. Suppose that the upper and lower limits are found to be $l_1'$ and $l_2'$ respectively, then $l_2$ and $l_2'$ will be approximately equal and the difference in $l_1$ and $l_1'$ relatively large. It is therefore established that the character is large and extends into the upper area. As a result, to separate the next character, the window is set in the limites, ($l_2'$ -h), and $l_2'$, with a dimension of 'h'. Naturally, if the sixth character is not large, then $l_1$ and $l_1'$ would



$h = l_2 - l_1$

Figure 4.1.2. Finding approximate horizontal bounds

be approximately equal and the window would be stepped right with limits $l_1'$ and $l_2'$. If the fifth character happens to be large, then the approximate bounds are found by using 'h', which is known from at least one ordinary size character, plus the appropriate bound for the fifth character. The bound chosen depends on whether it extends into the upper or lower area.

The only difficulty encountered in this procedure arises when the character, & , occurs in the first five characters. Only half of this character is generally isolated by the window, and this fact is revealed by its horizontal bounds when they are compared to the bounds of the other four characters. A way to overcome this problem is to step the window back in order to isolate the complete character after the approximate bounds for the middle area have been computed.

If either the value of h is known beforehand, or, after the characters of one line have been isolated, it is possible to use the first two and not the first five characters to evaluate the approximate horizontal bounds for the next line. It is necessary to use two characters as it is possible, (but rare) , that the first character of a line may be large and extending into the lower area, in which case the second character will be of normal dimension, or large, extending into the upper area [2]. At least one of the first two characters will have a lower bound which is approximately equal to the common lower bounds for line. Using this fact and 'h', the approximate horizontal bounds for the next character can be evaluated. In this study the lower horizontal bound for the first character was used in conjunction with 'h' to establish the approximate bounds of the second character to reduce computing time. This can be done because a middle area character rarely extends into the lower area, particularly at the beginning of a line [2]. To ensure that this assumption did not introduce error, a brief inspection was required of the scanned page and a flag in the isolating program to signal whether or not a character extends into the lower area in the first position.

An alternative method for window-shift may be used only after the characters of one line have been separated. This entails taking into consideration the angle of tilt, θ, and requires that the window be shifted up or down one position at fixed intervals across the page. The direction and interval of shift depends on θ. This tilt angle, θ, may be evaluated from the limiting horizontal bounds of the line already isolated.

### 4.1.2. The Upper and Lower Areas

From the results found for the middle area, a reasonably accurate estimate of θ can be made by using the horizontal bounds at the beginning and end of a line. The variable dimension window is set at the left end of the area concerned with its dimension and horizontal coordinates fixed according to the area height and the inter-area distance, respectively. These values are fixed (see Section 3.2). The window is then stepped across the page taking θ into account while progressing, as outlined at the end of the previous section.

When using the variable dimension window for the upper and lower areas there are two minor modifications that are necessary to the procedure used for the middle area. Firstly, if the window has to be extended into the middle area to fit over a character, it is assumed that a large middle area character has been encountered. Secondly, if the window is extended significantly below the lower bound for the lower area it is assumed that a towel has been encountered from the line below. If either of these conditions occur, the window is simply moved on :- that is, these characters are ignored.

To isolate each character takes about 0.6 seconds computing time. This time is excessive, but this is a direct consequence of the complexity of Thai printing and the versatility needed in the process to isolate the characters. Note that the time required for preprocessing the scan tape is ignored in this estimate.

It is found convenient to combine the preprocessing and feature extraction programs required for the recognition system, with the isolation program. Thus the programming consideration for the isolation process are given at the end of Chapter 6.

### 4.1.3. Restrictions on Tilt Angle.

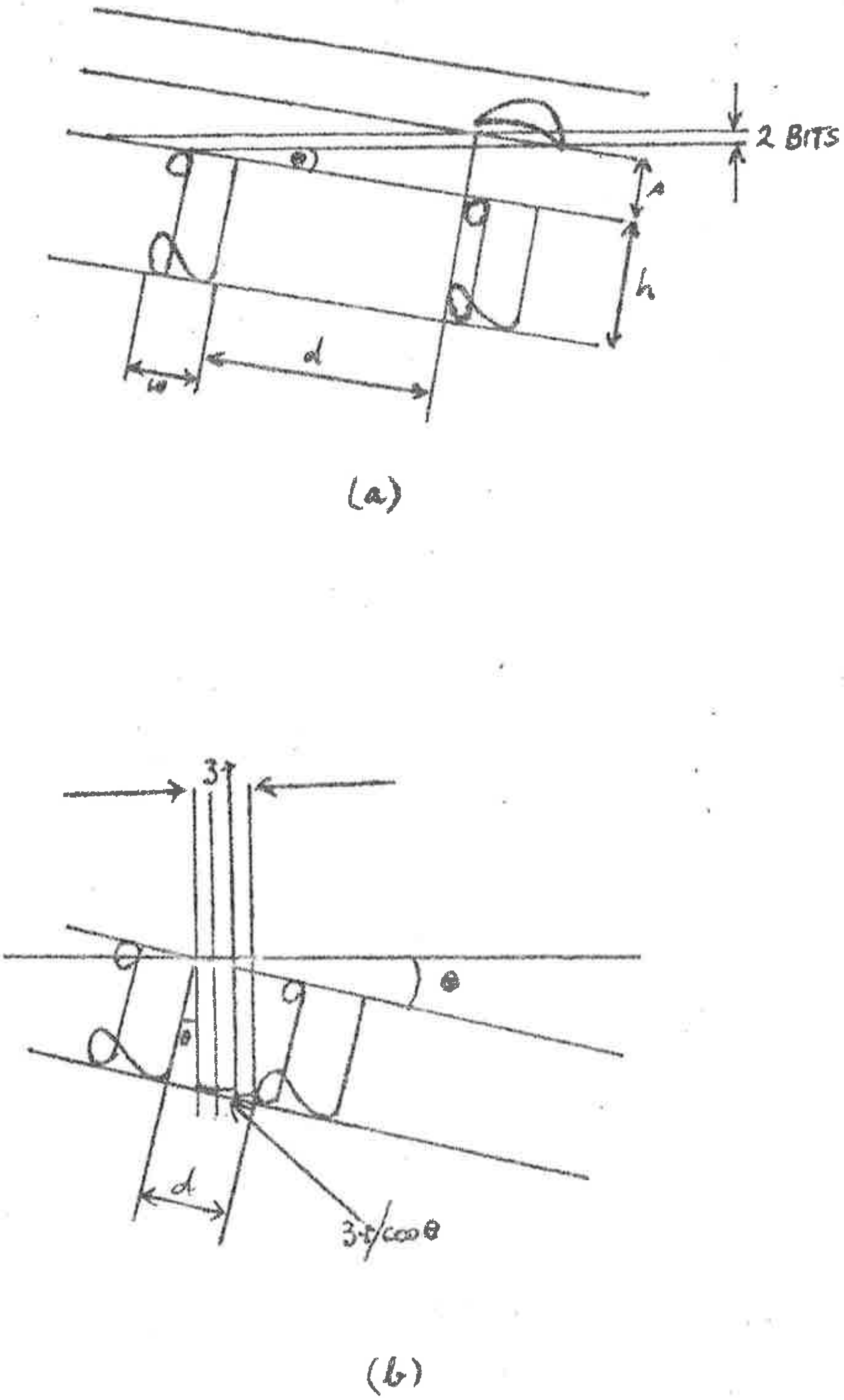
For the character isolation procedure to function without error it is necessary that:-

(a) in the middle area the window encounters each successive character of a line before it moves into either the upper or lower area depending on the sign of '$\theta$', and

(b) at least 3 vertical scans separate the characters of all areas.

The implication of these two requiremenets on $\theta$ is considered below.

Case 1. If the width of the last isolated character is w, d is the intercharacter distance, h the height of the middle area, and s the distance between the areas (see Figure 4.1.3 (a)), then the



(a)

(b)

Figure 4.1.3. The restrictions on the angle of tilt necessary for the isolation process to function without error (a) Case 1 (b) Case 2

maximum allowable angle of tilt is given by

$$\tan \theta_{max} = \frac{s}{d + w}$$

Typically s = h/2, where h is the perpendicular height of the middle area, w = 3h/4, and d = h/4, giving a maximum angle of arctan (0.5), which is approximately 25 degrees. The dependence of $\theta_{max}$ on the intercharacter distance, d, should be noted. In particular, if d is increased, then $\theta_{max}$ is decreased, and if there is a distance of reasonable magnitude in a line of print then the restriction on θ is quite severe. For example, if d is some 2 inches (which is quite possible for an interphrase gap) then d is approximately equal to 5 h and $\theta_{max}$ is restricted to approximately 3 degrees. However, this is still greater than the critical angle for the line finding process and the isolation method does not introduce a further restriction on the maximum allowable angle of tilt.

<u>Case 2</u>. If d and h are the same as above and r is the resolution in the horizontal direction (the distance between adjacent scans) then, from Figure 4.1.3. (b):-

$$\tan \theta_{max} = \frac{d-3\ r/\cos\ \theta_{max},}{h}$$

whence $h\sqrt{1-\cos^2\theta_{max}}-d\ \cos\theta_{max}-3r$, and

$(h^2+d^2)\cos^2\theta_{max}-6rd\ \cos\ \theta_{max}+9r^2-h^2=0$, and

$$\cos \theta_{max} = \frac{6rd \pm \sqrt{(6rd)^2 - 4(h^2+d^2)\ (9r^2-h^2)}}{2(h^2+d^2)}$$

On an average, d was found to be approximately equal to 5r, and for a large middle area character, h equal to 40 r, giving a maximum allowable tilt angle of approximately 4 degrees. Once again this is greater than the restriction imposed by the line finding technique and so, in theory, it is possible to separate all characters in the proposed method without introducing a further restriction on the angle of tilt. The restrictions imposed by the isolation technique on the angle of tilt do however, become significant if part of each line of print is used in the line finding process instead of the full line.

### 4.1.4. Touching Characters

Ideally a white line one bit wide is sufficient in both the horizontal and vertical direction to separate each character, but, in order that a false character end or edge is not generated by a broken stroke for example, the minimum gap has been set at 3 bits for vertical and 2 bits for horizontal separation. If a space meeting these requirements is not found between adjacent characters (that is if they are touching one another or the angle of tilt is excessive) there does not appear to be an infallible method for separating them.

To a human reader each set of two characters of Figure 4.1.4 should be separated along the dotted line, but the isolation process would regard each combination as one shape. There seems to be two possible solutions to this problem. One is to use a trial and error technique to divide the shape into portions at various points until the two shapes become recognisable. In other words, until the position of the dotted line is found [9]. This increases the complexity of the system considerably and in view of the few times in which this problem arose , the inclusion of such a process was considered too costly. The second, and adopted solution, is to require the recognition method to indicate that the shape is unrecognisable.

## 4.2. Preprocessing

The purpose of the preprocessor is to reduce the intraclass variety amongst the patterns presented in such a way that the probability of correct recognition is increased. Mason and McFall [10] have formulated four ways in which the preprocessor can act to achieve this:-

(a) Countering the effects of noise by filling in gaps and removing isolated areas of noise.

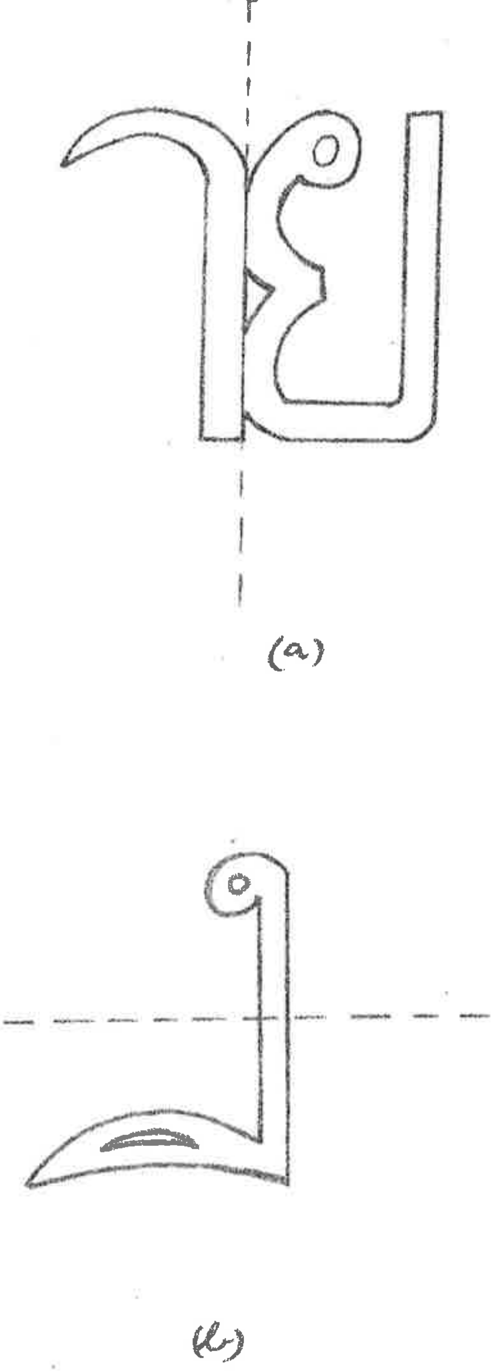


Figure 4.1.4. Touching Characters

(a) Middle Area (b) Upper & Lower Areas.

(b) Removing differences which occur between separate examples of the same character, (e.g. rotation and height variation)

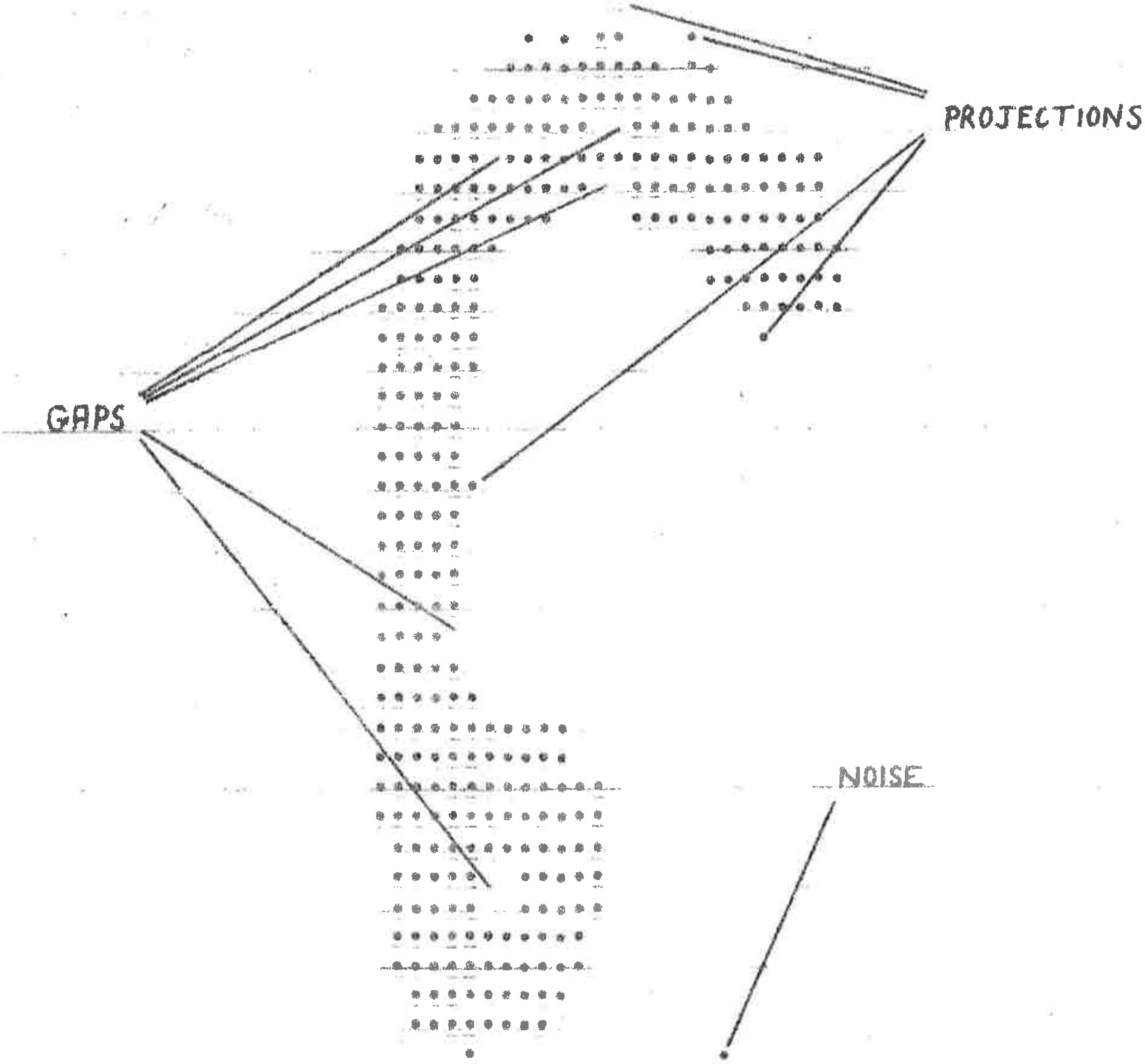
(c) Removing redundant information. (This includes resolution reduction by local averaging or line thinning).

(d) Removing information irrelevant to the classification criterion.

The most important of these four requirements is the first. Requirement (b) is adequately dealt with in the evaluation of suitable characteristic vectors to describe the character, and because of the choice of characteristics, (c) and (d) are not a problem and need not be considered here (see Chapter 6.) Thus it is only necessary for the preprocessor to reduce the effects of noise, but ensure that no distortions are introduced which could make the character difficult to recognise.

The possible defects of a character from the scanner are shown in Figure 4.2.1. and are enlarged upon here. In general the gaps must be detected and filled in,but ideally, they should be detected and then a decision made on their validity. In some Thai printing the numerous loops which occur are sometimes solid and other times show a distinct white spot in the centre due to variation in print

Figure 4.2.1. Character Defects.



quality. For the sake of consistency these areas are required to be always filled in. Other defects are in the form of small projections in the horizontal and vertical directions, and isolated noise spots both of which should be removed.

### 4.2.1. Some Previous Work on Preprocessing.

One method is to consider each point of the binary input matrix in turn and reach a decision whether it should be "one" or "zero" by considering the neighbouring points. Unger [11] reaches a decision by considering the neighbouring points in a 3 x 3 window centred on the target point. However, his insert and delete algorithms only deal with small defects affecting one or two matrix points on the edges of a character stroke. Sherman [12] and Deutsch [13] both describe techniques designed for both smoothing and thinning of character strokes, a feature which is not required.

Perhaps a more important technique is that described by Dineen [14]. With a character matrix of dimension 90 x 90 he uses a 5 x 5 local area. The operation is performed by observing the contents of the 5 x 5 window centred on each element in turn. A count of the number of one's in the window is compared with some threshold, T. If the count is greater than or equal to T, then the corresponding element in the new image is one, otherwise it is zero. Use of a low threshold eliminates scattered "ones" and fills in holes, but for high threshold corner and junction points are isolated. In general for low thresholds, the character is thickened and as the threshold is increased it is thinned. Dineen suggests the use of a larger window, say 15 x 15 around the smaller window to determine the threshold for the smaller window. For a dense window, a high threshold would be used, causing thinning, and for a sparse window a low threshold causing smoothing.

Alcorn and Hoggar [15] suggest a method for implementing Dineen's method. Using a 24 x 24 character matrix the thresholds for the local 3 x 3 window are determined by a 7 x 3 larger window. The possible bit counts of the 7 x 3 window are divided into three regions (by gates) and a pair of insert and delete thresholds associated with each region. When a new matrix point is chosen, the large window count is first evaluated and then the appropriate thresholds selected for use in the local area decision.

### 4.2.2. Experimental Results

Binary character matrixes with the approximate dimensions 40 x 40, were isolated from a scan image of a page and recorded on magnetic tape. This data was then used to test some of the above mentioned techniques.

The Dineen method was the first to be tested. Using a 5 x 5 window and trying various thresholds, (not set by a larger window), good results were obtained. However, there was difficulty in deciding what value to allot to the threshold. The processing time required for this method was about 1.3 seconds per character.

The method proposed by Alcorn and Hoggar was then applied to the same data. The large window was chosen to be of dimension 7 x 7 with a smaller 3 x 3 window for the local area. Once again good results were obtained, but some difficulty was found in choosing the values for the gate and threshold values. In addition the processing time was increased by about 50 percent over the time required by the Dineen method.

It was finally decided to use the method suggested by Bomba [16] in which a 3 x 3 local window is used with an insert threshold of 5 and a delete threshold of 4. The processing time was reduced to approximately 0.2 seconds per character, the process removing isolated "ones" and inserting "ones" into gaps. If insufficient smoothing is obtained in one application then the process may be repeated.

Figure 4.2.2. illustrates the effectiveness of this smoothing process on the character of Figure 4.2.1. applied once and then twice.

## 4.3. Simulated Data

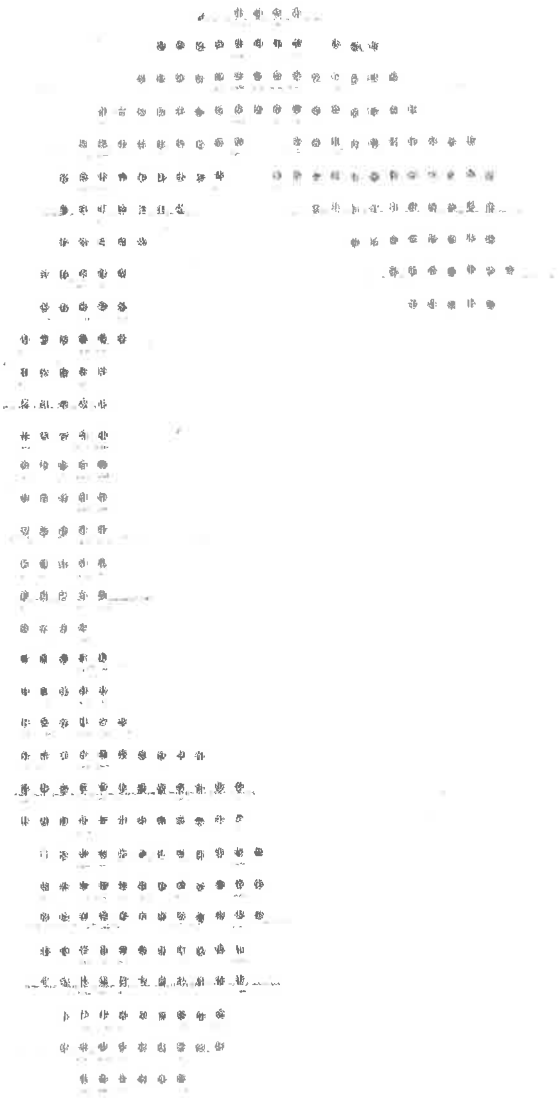
In order to test programs and experiment before the scanner was built, test characters were made up and punched into cards. Later this simulated output of the preprocessing section of the Character Recognition System, proved to be extremely useful in finding preliminary results which were incorporated in the final system.

The binary matrices simulating the output of the preprocessor were constructed by taking large characters from a Thai Primer [17], laying graph paper over each character in turn, and marking the black squares. The resulting binary matrices were punched into cards. To fully simulate the output from the scanner, noise was introduced into the binary matrices by using a pseudo-random number generator. The noise was introduced by stepping through the binary matrices element by element, a random number generated after each step. If the random number exceeded a present level, the element was changed to white if it was black, and if it was white and in the vicinity of a black point, it was changed to black. Otherwise it remained unaltered.

Figure 4.2.2. The result of smoothing.

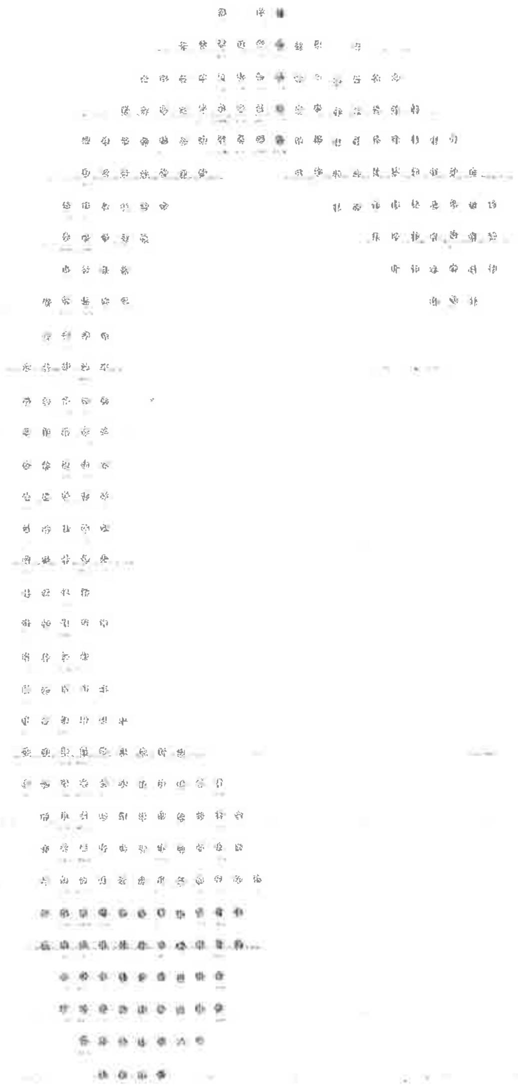
(a) After applying the technique once

(b) After applying the technique twice.



(a)

Figure 4.2.2. (cont)



(b)

This fully simulated the output of the preprocessor in which isolated black points were eliminated.

A measure of the noise level was available from the threshold value. The random number generated numbers in the range of 0 - 1 and thus the threshold must be in this range, say 0.9 for example, giving a noise level defined at 10 percent. Figure 4.3.1. gives an example of a maually constructed character and the effect of noise on the same character for various noise levels.

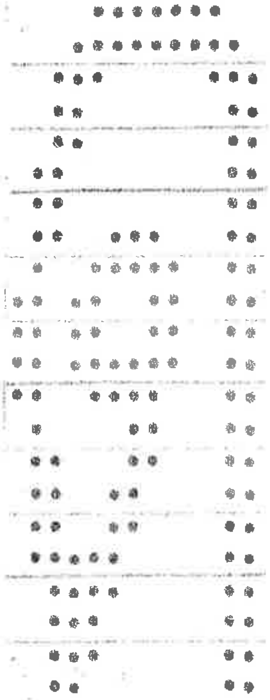
Figure 4.3.1. Manually Constructed Characters

(a) No noise

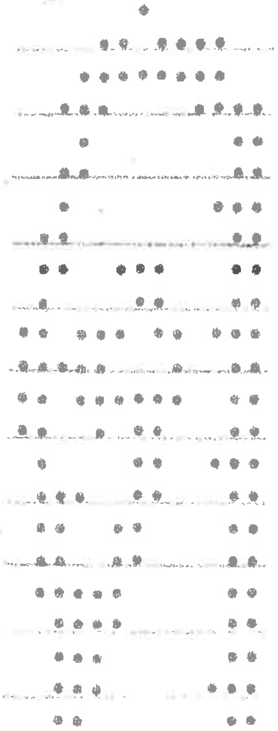
(b) Noise Level = 10%

(c) Noise Level = 20%

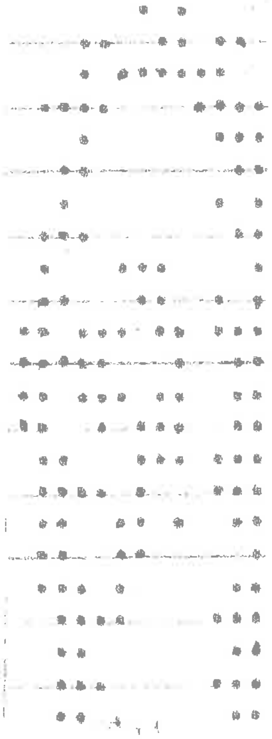
(d) Noise Level = 30%



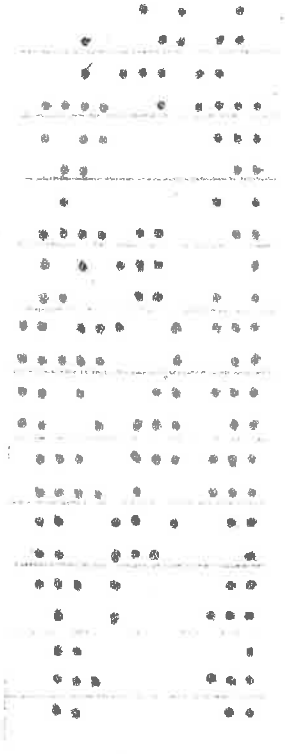
(a)



(b)



(c)



(d)

## Chapter 5. The Categorizer

For each binary pattern presented to the feature extractor, an n-dimensional vector, y, was computed which was used as the input to the categorizer. The latter was a device which applied some decision procedure to assign each vector to one of a finite number of categories, say r. It was required that the output of the categorizer be an integer i(=1,2....r) under the convention that an output j was to mean that the system had assigned an input to the j th category of the r possible ones. In effect the categorizer could be considered as a procedure to compute i=R(y), where R(y) is the recognition function. It was assumed that each sample presented to the categorizer for recognition belonged to one of the r categories but if rejection is recommended, that is, the system cannot make a decision as to which category an input belongs with certainty, the system would not be considered in error. However, if the system assigned an input to category j (when in fact it belonged to some other category), then it was considered to have made an error.

The use of statistical decision functions is one of the many possible decision procedures used in the implementation of a categorizer. They were chosen for this study because they provided an opportunity to make use of the a priori probabilities that are available (section 1.2) for character occurrence. Decision functions minimizing the "average risk" and the probability of error for a given rejection rate were first considered by Chow [18]. Many other workers have done work in this area in more recent times. (See [19] to [23]). The theory of Bayes' decision functions are reviewed here.

## 5.1. Bayes' Decision Functions

Suppose that $S = \{s_1, s_2, \dots s_r\}$ is the set of r categories and each input $y = (y_1, y_2, \dots, y_n)$ belongs to one and only one category $s_i$. Assume that the a priori probability of an input vector belonging to category $s_i$ is $p_i$ for $i=1,2,\dots,r$ and

$$\sum_{i=1}^{r} p_i = 1. \qquad (1)$$

Because of noise in the original binary pattern of an unknown character, y is subject to random variation and thus may be considered as a random variable. Let us assume that the probability distribution of y is determined by the categories and is given in the form of a conditional probability density function on the measurement space Y'. Thus if the category is $s_i$, then the probability of an input y is $F(y|s_i)$.

Given an input vector, y, the problem is to decide to which category it belongs. The system may make any one of (r + 1) decisions $d_o$, $d_1$,... ,dr, for any input y, where $d_o$ is the decision to reject y as being unrecognisable, and $d_i$, i=1,2,... ,r, is the decision that y belongs to category $s_i$. A decision function or decision rule $\emptyset(d_i|y)$ is defined for every y in Y' and decision $d_i$, i=0,1,... ,r, such that

$$\sum_{j=0}^{r} \emptyset(d_i|y) = 1, \quad \text{and}$$

$$\emptyset(d_i|y) \geq 0, \quad j = 0,1,....r. \qquad (2)$$

Note that $\emptyset(d_i|y)$ is the probability the categorizer will make the decision $d_i$ given a random vector y in Y'. The problem is to find a decision function which is 'optimal' in some sense.

Define a loss (or cost) function, $W(s_i,d_j)$, written $w_{ij} \leq 0$, such that $w_{ij}$ is the loss incurred by making the decision $d_j$ when the input y belongs to category $s_i$. Since it is required that the decision be $d_i$ if y belongs to $s_i$, we have

$$w_{ij} \leq w_{i0} \leq w_{ii} \quad , \quad i \neq j; \quad i,j=1,2,....r, \qquad (3)$$

where $w_{i0}$ is the loss incurred on the system by rejection if y belongs to $s_i$, and $w_{ii}$ is the loss associated with correct recognition.

The probability of making a decision $d_j$, when the input is from category $s_i$ is

$$P[d_j|s_i]=\int_{Y'} F(y|s_i)\emptyset(d_j|y)dy \tag{4}$$

where the integration is over the whole of the measurement space.

Now the loss or risk incurred when the random vector belongs to category $s_i$ and the decision rule $\emptyset$ is used is

$$R(s_i,\emptyset)=\sum_{j=0}^{r} w_{ij}\int_{Y'} F(y|s_i)\emptyset(d_j|y)dy. \tag{5}$$

Taking into account the a priori probability of occurrence $p=(p_1,p_2,\ldots.p_r)$, the 'average risk' for the whole system is $R(p,\emptyset)$, where

$$R(p,\emptyset)=\sum_{i=1}^{r} p_i R(s_i,\emptyset),$$

$$=\sum_{i=1}^{r}\sum_{j=0}^{r} w_{ij}p_i\int_{y1} F(y|s_i)\emptyset(d_j|y)dy \tag{6}$$

The optimum categorizer is defined to be the implementation of the decision rule which minimizes the average risk $R(p,\emptyset)$ given by equation (6).

The optimum categorizer can be found without difficulty [18] for equation (6) and can be rewritten as

$$R(p,\emptyset)= R_0+R_1(p,\emptyset), \tag{7}$$

where

$$R_0(p)=\sum_{i=1}^{r} p_i w_{i0},$$

$$R_1(p,\emptyset=\int_{Y'} \sum_{j=0}^{r} \emptyset(d_j|y)Y_j(y)dy,$$

$$Y_j(y)=\left[\begin{array}{ll}\sum_{i=1}^{r}(w_{ij}-w_{i0})p_i F(y|s_i) & ;j=1,\ldots,r, \\ & \\ 0 & ;j=0\end{array}\right.$$

Note that $R_0(p)$ is the average risk when rejection is made for all inputs and $R_1(p,\emptyset)$ can obviously be adjusted through $\emptyset$. Now

$$R_1(p,\emptyset)\geq\int_{Y'}\min_j\ [Y_j(y)]dy, \qquad (8)$$

and equality holds if, and only if, the decision rule $\emptyset$ is chosen as

$$\left.\begin{array}{lll}\emptyset(d_k|y)=\delta(d_k|y)=1 & \text{for} & j=k, \\ \emptyset(d_j|y)=\delta(d_j|y)=0 & \text{all} & j\neq k,\end{array}\right] \qquad (9)$$

whenever

$$\min_j\ [Y_j(y)]=Y_k(y)$$

Equations (9) yield the optimum categorizer when the criterion of minimum risk is adopted. The expected loss is given by

$$R(p,\delta)=\ \sum_{i=1}^{r} p_i w_{i0}+\int_{Y'}\min_j[Y_j(y)]dy. \qquad (10)$$

To assign an unknown point, $y$, to one of the categories, or reject it, as being unrecognisable it is necessary to find the value of $j$ for which $Y_j(y)$ has its minimum value, $j=0,1,\dots,r$, assuming that $w_{i0}\geq 0, i=1,2,\dots,r$. Then the aim is to find the value of $j$ for which

$$\sum_{i=1}^{r} w_{ij}p_i F(y|s_i), \quad j=0,1,\dots,r, \tag{11}$$

takes its minimum value. Having found such a $j$(say $k$) $y$ is assigned to $s_k$. If $j$ is found to be zero then $y$ is rejected as being unrecognisable.

## 5.2. Probabilities of Error and Rejection for Equal Losses

Let $w_{ij}=w$, $w_{i0}=w_0$ and $w_{ii}=0$, the costs of misrecognition, rejection, and correct recognition, respectively. These are taken to be independant of the pattern. Then equation (6) may be rewritten as

$$\begin{aligned} R(p,\emptyset) &= \sum_{i=1}^{r} \int_{Y'} w_{i0}\emptyset(d_0|y)F(y|s_i)p_i\,dy \\ &\quad + \sum_{i=1}^{r} \sum_{j=1}^{r} \int_{Y'} w_{ij}\emptyset(d_j|y)p_iF(y|s_i)\,dy \\ &= w_0 \sum_{i=1}^{r} \int_{Y'} \emptyset(d_o|y)p_iF(y|s_i)\,dy \\ &\quad + w \sum_{i=1}^{r} \int_{Y'} \sum_{\substack{j=1 \\ j\neq i}}^{r} \emptyset(d_j|y)p_iF(y|s_i)\,dy, \end{aligned} \tag{12}$$

but

$$\sum_{\substack{j=1\\ j\neq i}}^{r} \phi(d_j|y)=1-\phi(d_0|y)-\phi(d_j|y), \qquad (13)$$

and thus

$$R(p,\phi)=wP_e(\phi)+w_0P_r(\phi),$$

where

$$\left.\begin{aligned} P_r(\phi)&=\sum_{i=1}^{r}\int_{Y'}\phi(d_0|y)p_iF(y|s_i)\,dy=\text{probability or rejection,}\\ P_c(\phi)&=\sum_{i=1}^{r}\int_{Y'}\phi(d_i|y)p_iF(y|s_i)\,dy=\text{probability of correct recognition}\\ P_e(\phi)&=1-P_r(\phi)-P_c(\phi)=\text{probability of error}\end{aligned}\right]\ 15$$

## 5.3. Minimum Error Rate

Chow [18] has shown that, for a given rejection rate, the error rate in a recognition system is minimized if the following decision criterion is used. Choose category, k, for the input, y, if

$$p_kF(y|s_k)\geq p_jF(y|s_j)\ \text{ for all } j\neq k,$$

$$p_kF(y|s_k)\geq \beta\sum_{i=1}^{r}p_iF(y|s_i),\quad 0\leq\beta\leq 1,$$

and reject y if

$$p_jF(y|s_j)<\beta\sum_{i=1}^{r}p_iF(y|s_i)\ \text{for all } 1\leq j\leq r. \qquad (16)$$

The constant $\beta$ is chosen to force the system to meet a given rejection rate determined by the condition $P_r(\emptyset) = \alpha$. It has been shown by Highleyman [19], that the optimal system minimizing the average risk and errors for a given rejection rate are equivalent (in the case of constant costs) if

$$\beta = 1 - \frac{w_o}{w} = 1 - c. \tag{17}$$

The decision rule given by equations (16) is the one used for all experimental work described below.

## Chapter 6. Selecting Features for the Feature Vectors

Printed characters would appear, at first sight, to be of fixed size and shape and to lend themselves to a simple "Template Matching" recognition technique at least on prima facie examination. A machine should therefore look for a mask among its set of masks which is "something like" the character being examined. "Something like" is, however, far too vague a condition to be implemented directly; it is usually human assessment which is by no means understood that performs this process. How do we measure the degree of "something likeness"? The fixed size and shape of printed characters are more apparent than real; they vary in thickness, angle of presentation, spacing, height and style, to which can be added different fonts, ability and cryptographic capability.

A more reasonable approach is to break the patterns down into sub-patterns of "features", which are less variable within categories, and to re-describe the patterns in terms of these features. One can consider the input patterns in matrix form, mapped into a vector feature space of, say, n-dimensions. An ideal choice of features should provide feature vectors which:

(a) form closely pack clusters in feature space for any one category with clusters representing different categories widely separated,

(b) are independent of the position of the character on the page,

(c) are independent of the size and orientation of the character,

(d) are independent of the boldness of the print, and

(e) vary only slightly in the presence of a reasonable amount of "noise" allowing for other minor differences such as small changes of proportions, slight bending of lines and so on.

Many methods have been used to extract features to form feature vectors, some of which are mentioned below. Bomba [16] described a cross-correlation technique to extract features from a character. The method may be visualised as the comparison of the 'two-dimensional' input pattern with a set of standard patterns, or templates, representing the features to be detected. For optimal performance, the templates should be matched with the unknown pattern in the correct orientation and position, and the templates should be the same size and shape. This technique has been extended to auto-correlation for feature extraction (see [24] - [26] ). The basic principle underlying feature detection by auto-correlation is the use of features actually present in the input pattern as their own templates, making it unnecessary to position and orientate the pattern. Another well-known technique is to use geometrical properties of the input pattern for recognition purposes (see [27] - [29] ). In this case, recognition is made by the explicit use of the relative positions of continuous line segments of specific shape and orientation of each character.

The main difficulty with these methods of feature extraction is the need for extensive pre-processing of the input patterns. In all of the above cases it is necessary to remove redundant information by the thinning of lines and, in addition, to cross-correlate rotational and height variations. However, characters may be recognised by using features far more abstract than those outlined above and which may be selected so as to eliminate the need for line thinning and other . normalisation processes in the preprocessing stage of the recognition system. Such a method has been described by Giuliano, Jones, Kimball, Meyer and Stein [30] and Alt [31]. These authors use higher bivariate moments of patterns (blackness being analogous to mass) to derive features with the properties (a) - (c). For this study this method of feature extraction was chosen because very little preprocessing is required on the input binary patterns and each moment is easily evaluated. The remainder of this chapter is devoted to the derivation of the bivariate moment invariants and to a method of choosing those suitable for use in the feature vector to give the best discrimination between classes.

It should be noted it was stated in (c) that the feature vector for each binary pattern should be independent of orientation of the unknown character. Normalisation with respect to rotation has been omitted but "slanting" such as that which occurs in italic type is included in its place. This decision was motivated by the fact that only small angles of tilt of the scanned page are admissable so that separation of characters for their individual recognition is possible. (Chapter 3). Consequently any rotational variation in characters of the same category would have little effect on the moments for each character. However, the derivation of moment invariants taking rotation into account may be found in a paper by Hu [32].

## 6.1. Derivation of Bivariate Moment Invariants

The binary matrix of an input character can be represented by a real, two-dimensional density function, $\rho(x,y)$, $0\leq\rho(x,y)\leq 1$, where $\rho(x,y)=1$ for completely black points, and $\rho(x,y)=0$ for completely white points. It was mentioned in Chapter 2 that the scanner should be capable of assigning one of eight intensity readings to each data point depending on the grey level. This feature could be conveniently used here, with the density function $\rho(x,y)$ taking one of eight values in the range 0 - 7. However, since these intensity readings were unreliable at the time and it was necessary to convert a data point reading to either 0 or 1, then $\rho(x,y)$ is either 0 or 1.

In general the transformation of the input pattern into feature space can be done by computing the correlation between the unknown pattern $\rho(x,y)$ and a set of filtering (or discriminant functions), $\rho_i(x,y)$, to yield a set of n measurements $\{a_i, i=1,2,\ldots,n\}$. The correlation integral is given by

$$a_i = \iint \rho_i(x,y)\rho(x,y)\,dx\,dy, \qquad i=1,2,\ldots,n. \quad (1)$$

The set of n measurements thus defined constitute the output of the feature extractor. The functions $\rho_i(x,y)$ should be chosen on the basis that the measurements

(1) satisfy condition (a), and

(2) may be transformed so that the feature vector satisfies conditions (b) to (e).

Firstly, good discrimination can be obtained by selecting the filtering functions $\rho_i(x,y)$ so that the measurements, $a_i$, represent the leading coefficients in a series expension approximating the density function, $\rho(x,y)$, of the unknown character. If the filtering functions are set as

$$\left.\begin{aligned} \rho_1(x,y) &= 1, \\ \rho_2(x,y) &= x, \\ \rho_3(x,y) &= y, \\ \rho_4(x,y) &= x^2, \\ \rho_5(x,y) &= xy, \\ \rho_6(x,y) &= y^2, \\ \text{etc}; & \end{aligned}\right] \qquad (2)$$

then the set of measurements $\{ a_i , i = 1,2,\ldots,n \}$ represents the moments of the pattern, a black point being considered as unit mass. The set of discriminant functions, $\rho_i(x,y)$, $i=1,2,\ldots,n$ are the Taylor's series coefficients of the Fourier transform of the density function of the pattern. Other series expansions of the density function, $\rho(x,y)$, or of some transform of it, are possible (e.g. Fourier coefficients). The measurements

$a_i$, $i=1,2,\ldots,n$, must be characteristic of the function $\rho(x,y)$, since $\rho(x,y)$ can be approximated as accurately as is wished by taking sufficient terms of the approximating series expansion. However, to discriminate between characters of different classes it is not necessary to have an accurate expansion but merely a few low order terms.

Secondly, the feature measurements must be independent of "mass" (or boldness of print), size, position on the page, and slant of the character under consideration. Moments can be conveniently normalised with respect to all of these variables.

If the filtering functions $\rho_i(x,y)$, $i=1,\ldots,n$, are the ascending powers in x and y, then the measurements given by equation (1) represent the moments of the pattern, and may be rewritten as

$$M_{ij} = \iint x^i y^j \rho(x,y)\,dxdy, i,j=0,1,2\ldots, \qquad (3)$$

where the order of each moment is defined as $(i+j)$. It is assumed that $\rho(x,y)$ is constant over small finite areas (equal to 0 or 1). Thus (3) may be rewritten as

$$M_{ij} = \sum_{\rho(x,y)\neq 0} cx^i y^j, \qquad (4)$$

where c is the "mass" of each cell.

The zero order moment is equal to the total "mass" of the pattern

$$M_{00} = c\ M, \qquad (5)$$

where M is the number of black cells. Normalisation with respect to mass is achieved by dividing all quantities by $M_{00}$. The centre of gravity of the pattern then has the coordinates.

$$\bar{x}=c\sum_{\rho\neq 0} x/M_{00}=M_{10}/M_{00}, \qquad (6)$$

$$\bar{y}=c\sum_{\rho\neq 0} y/M_{00}=M_{01}/M_{00},$$

and the central moments (i.e. moments independent of position on the page) are given by

$$\overline{M}_{ij}=c\sum_{\rho\neq 0}(x-\bar{x})^{i}(y-\bar{y})^{j}. \qquad (7)$$

The variances

$$\sigma_{x}=\sqrt{M_{20}/M_{00}},\quad \sigma_{y}=\sqrt{M_{02}/M_{oo}}, \qquad (8)$$

can be used to normalise the coordinates (thus making each coordinate independent of line width) by setting

$$x^{*}=(x-\bar{x})/\sigma_{x}, \qquad y^{*}=(y-\bar{y})/\sigma_{y}. \qquad (9)$$

The moments must be normalised with respect to "mass", and the new moments may then be written as

$$m^{*}_{ij}=c\sum_{\rho\neq 0}x^{*i}y^{*j}/M_{00} \qquad (10)$$

We finally set

$$\lambda=\sum_{\rho\neq 0}x^{*}y^{*}/\sum_{\rho\neq 0}y^{*2},$$

$$b=\frac{x^{*}-\lambda y^{*}}{\sqrt{1-\lambda^{2}}},\quad d=y^{*}, \qquad (11)$$

and refer the moments to these coordinates. Then

$$m_{ij}=c\sum_{\rho\neq 0} b^{i}d^{j}/M_{00}, \qquad (12)$$

are invariant under conditions (b), (c) and (d).

Because of the transformations involved in the derivation of equation (12),

$$m_{00}=m_{20}=m_{02}=1, \text{ and}$$

$$m_{10}=m_{01}=m_{11}=0,$$

which are of no use as feature vector elements since they provide no information about the input patterns. If the moments are arranged in descending i values to form the feature vector:-

$$X=\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} = \begin{bmatrix} m_{30} \\ m_{21} \\ m_{12} \\ \cdot \\ \cdot \\ m_{ij} \end{bmatrix},$$

We can refer to this as the natural order of the features in later discussion.

If an input pattern is affected by noise, or there is some other minor difference (e.g. the slight bending of a line) then a few of the cells which should be black are white and vice versa. For our purposes, any difference between an input pattern and the perfect sample of the same character may be regarded as due to noise of some kind. If the noise level is small then only a few terms in the summation of equation (12) will be affected and this will make little difference to the numerical result. However, the effect of noise on the individual moments for a given pattern is not readily solved. Algebraically the task is formidable and about the only two conclusions that can be made are that, firstly, the effect of noise on individual moments depends on its distribution and, secondly, the high-order moments are affected more by noise than the low-order moments.

## 6.2. Selection of the Moments for the Feature Vectors

Knowing that noise will be present in the patterns from the scanner, even after preprocessing, and that an infinite number of moments could be computed, for each pattern, one is faced with the problem of deciding which moments should be used in the feature vector to give the "best" results. Using statistical decision functions in the Bayes sense, the "best" set of feature measurements that can be used in a recognition system are those which minimize the 'average risk' or, in the case of equal costs of misrecognition and rejection, those which minimize the error rate for a given rejection rate. In practice, however, this procedure is difficult to apply for the following reasons:

(a) The underlying probability distributions associated with the pattern classes may not be known, or cannot be easily calculated to allow an analytic calculation of the 'average risk' or rejection rate, and

(b) The analytic expression for the 'average' risk or rejection rate may not be suitable for the minimization required in a specific decision structure.

It is therefore necessary to find a method for selecting the moments without computing the rejection rate, or having a complete knowledge of the probability structure, for the input patterns under consideration. An attempt can be made to select a subset of contributing moments by excluding those which are 'detrimental' to the recognition system. Here 'detrimental' is defined in the sense that noise causes variation in the numerical value of a moment, and the inclusion of such a moment introduces more error than information into the system. The 'best set' of moments found by this screening procedure is not necessarily a unique set, but a systematic procedure is proposed for selecting a "reliable" set of moments.

A "reliable" moment is defined as one for which there is little variation in its numerical value.

From a computational point of view it is not feasible to deal with all the 'reliable' moments. Thus a selection of those containing significant information is required. In this study the selection is determined by performing a *principal component* analysis.

The results described in the following text were obtained using the simulated input characters (see section 4.3.), for which moments of orders 3 to 7 inclusive were calculated for the "perfect" images and for 20 images of each character for each noise level. (Moments for the perfect images can be found in Appendix B). All recognition experiments were carried out using the special case of the Bayes' "minimum risk" system with equal misrecognition costs and equal *a priori* probabilities of occurrence of each category. (see Chapter 5.) The conditional probability density functions, $F(X|s_i), i=1,2,\ldots,r$, were approximated by

$$F(X|s_i)=\frac{1}{(2\pi)^{n/2}h^n}\exp\left[-\frac{(X-X_i)'(X-X_i)}{2h^2}\right], \quad (13)$$

where

$n$ = the dimension of the feature vector,

$X_i$ = the training vector for the category $s_i$ and in this case is the feature vector for the 'perfect' simulated pattern, and

h = a "smoothing parameter".

It is well known that moments are not independent and that the variances for different moments do differ, this is not implied by the form of equation (13), the density function for a multivariate normal distribution of n independent variates with equal variances. It was decided to use an approximating function for the probability density functions (see Chapter 7) given by

$$F(X|s_i) = \frac{1}{(2\pi)^{n/2}h^n}\,\frac{1}{m}\sum_{k=1}^{m} \exp\left[-\frac{(X-X_{ik})'(X-X_{ik})}{2h^2}\right] \quad (14)$$

where

$X_{ik}$=k$^{th}$ training vector for category $s_i$,

and the other parameters are the same as for equation (13). Equation (13) is the first approximation, and suffices for the following discussion.

### 6.2.1. Selection of Feature Elements

It would be expected that the "best" set of moments or feature vector elements would be those with minimum intraclass and maximum inter-class scatter or spread. A method for choosing elements according to this criterion was programmed and tested on a computer, but it was found that this method did not give the best results. However, it is profitable to consider this method to enable comparison with the adopted method.

Suppose, as before, there is a set of r categories denoted by $\{s_1, s_2,\ldots,s_r\}$, and there are $n_1, n_2,\ldots n_r$ simple vectors of n dimensions from each category respectively. Thus all the sample vectors can be represented by

$$\{x^j_{1i_j}, x^j_{2i_j},\ldots,x^j_{ni_j};\ i_j=1,\ldots,n_j;\quad j=1,\ldots,r\}.$$

Let $\{\underline{u}^j=(u^j_1,\ldots,u^j_n),\ \ j=1,\ldots,r\}$, be the set of vector means for each of the r categories, and $\underline{u} = (u_1,\ldots,u_n)$ be the vector of all sample means.

Within each of the r categories compute the sum of the square deviations about the group mean for each element of the feature vector using

$$C^j_k = \sum_{l_j=1}^{n_j} (x^j_{kl_j} - u^j_k)^2, \quad k=1,\ldots,n. \qquad (15)$$

Thus for a particular element of the feature vector, the pooled within category sum of squared deviations about the group mean, $A_k$, can be determined by

$$A_k = \sum_{j=1}^{r} C_k^j, \; k=1,\dots,n. \qquad (16)$$

The total sum of the squared deviations about the grand mean may be found from

$$S_k = \sum_{j=1}^{r} \sum_{1_j=1}^{n_j} (x_{k1_j} - u_k)^2, k=1,\dots,n. \qquad (17)$$

The sum of the squared deviations between group means and the grand mean may be obtained by subtracting the pooled within group sum of squares, $A_k$, from the total sum of squares, $S_k$ or

$$B_k = S_k - A_k, \quad k=1,\dots,n. \qquad (18)$$

The criterion for selecting the first element, say $x_t$, is

$$\frac{B_t}{A_t} \geq \frac{B_k}{A_k}, k=1,\dots,n; \qquad (19)$$

and the second, say $x_v$, is

$$\frac{B_v}{A_v} \geq \frac{B_k}{A_k}, \; \begin{array}{l} k=1,\dots,n, \\ k \neq t \end{array} \qquad (20)$$

and so on. In the rare event that equality occurs, the procedure arbitarily selects the first feature in the order 1,...,n.

An alternative method, and one yielding better results than the one described above, is to simply compute the variance within each category for each of the n feature vector elements, find the maximum variance over all categories for each element and select that one with the smallest maximum variance, the second with the second smallest maximum, and so on. That is, the first element is chosen such that it has little intra-class variation, the second a little more, and so on, disregarding the inter-category variation.

The variances for each element within each category are computed by using the usual formula for variance, namely,

$$V_k^j = \frac{1}{(n_j-1)} \sum_{l_j=1}^{n_j} (x_{kl_j}^j - u_k^j)^2 \quad , \quad \begin{matrix} k=1,\ldots,n, \\ j=1,\ldots,r, \end{matrix} \qquad (21)$$

Since there exists a finite number of categories, r, for a fixed value of k there will be a j, say $a_k$, such that $V_k^{a_k}$ is a minimum. The criterion for selecting the first variable, say $x_t$, is to choose that value of k(=t) such that $V_t^{a_t}$ is a minimum, i.e. choose t such that

$$V_t^{a_t} = \min_{\text{all } k} \{ \max_{\text{all } j} (V_k^j) \} \qquad (22)$$

The second variable, say $x_v$, is chosen according to the rule

$$V_v^{a_v} = \min_{\text{all } k \neq t} \{ \max_{\text{all } j} (V_k^j) \} , \qquad (23)$$

and so on. In the case of equal maximum variances, once again the procedure arbitary selects the first feature in the order 1,...,n.

To test these two methods of feature element ordering the test data for each area was generated with a noise level of 10 percent and the feature elements recorded according to the criterion firstly, defined by the inequalities (19) and (20) and secondly, by equations (22) and (23). The new orders found by these two methods is presented in Table 6.2.1. for all three areas together with the ratios $B_k/A_k$ and $V_k^{a_k}$ for k = 1,...,n.

Table 6.2.1. The recorded feature vectors using the inequalities (19) and (20) and equations (22) and (23) for the Middle, Upper and Lower Areas using test data with a 10 percent noise level.

(a) The Middle Area

| Reordered Element Number | Original Feature Element (k) | Ratio $B_k/A_k$ | Original Feature Element | Max$^m$ Variance |
|---|---|---|---|---|
| 1 | 2 | 61.02 | 2 | 0.00 |
| 2 | 3 | 57.83 | 3 | 0.01 |
| 3 | 11 | 45.71 | 7 | 0.01 |
| 4 | 14 | 38.27 | 4 | 0.02 |
| 5 | 12 | 36.14 | 8 | 0.02 |
| 6 | 27 | 32.70 | 6 | 0.03 |
| 7 | 5 | 28.31 | 1 | 0.03 |
| 8 | 7 | 24.71 | 12 | 0.04 |
| 9 | 1 | 21.65 | 13 | 0.04 |
| 10 | 4 | 21.13 | 9 | 0.06 |
| 11 | 29 | 20.79 | 20 | 0.06 |
| 12 | 18 | 20.01 | 19 | 0.14 |
| 13 | 20 | 19.01 | 14 | 0.14 |
| 14 | 10 | 18.86 | 27 | 0.22 |
| 15 | 15 | 17.95 | 5 | 0.27 |
| 16 | 26 | 16.33 | 11 | 0.30 |
| 17 | 13 | 16.28 | 18 | 0.35 |
| 18 | 25 | 16.22 | 28 | 0.47 |
| 19 | 24 | 15.45 | 26 | 0.55 |
| 20 | 6 | 15.13 | 21 | 0.60 |
| 21 | 8 | 14.41 | 15 | 0.66 |
| 22 | 30 | 13.42 | 22 | 2.92 |
| 23 | 9 | 12.99 | 29 | 3.34 |
| 24 | 16 | 12.81 | 25 | 3.95 |
| 25 | 28 | 12.50 | 17 | 4.21 |
| 26 | 17 | 12.12 | 10 | 5.03 |
| 27 | 21 | 11.47 | 30 | 18.32 |
| 28 | 22 | 9.44 | 24 | 50.62 |
| 29 | 23 | 8.95 | 16 | 63.05 |
| 30 | 19 | 7.73 | 23 | 864.11 |

Table 6.2.1. (Cont.)

(b) The Upper Area.

| Reordered Element Number | Original Feature Element (k) | Ratio $B_k/A_k$ | Original Feature Element | $\text{Max}^m$ Variance |
|---|---|---|---|---|
| 1 | 3 | 54.21 | 3 | 0.00 |
| 2 | 2 | 53.44 | 2 | 0.00 |
| 3 | 7 | 51.32 | 7 | 0.00 |
| 4 | 11 | 45.40 | 8 | 0.01 |
| 5 | 1 | 41.70 | 12 | 0.01 |
| 6 | 14 | 34.71 | 6 | 0.01 |
| 7 | 10 | 33.27 | 1 | 0.01 |
| 8 | 4 | 32.52 | 4 | 0.02 |
| 9 | 15 | 29.51 | 13 | 0.02 |
| 10 | 9 | 29.49 | 11 | 0.04 |
| 11 | 12 | 29.09 | 19 | 0.04 |
| 12 | 18 | 26.27 | 18 | 0.05 |
| 13 | 20 | 25.70 | 14 | 0.05 |
| 14 | 27 | 25.39 | 5 | 0.05 |
| 15 | 6 | 24.63 | 9 | 0.06 |
| 16 | 17 | 23.48 | 20 | 0.07 |
| 17 | 24 | 22.76 | 27 | 0.15 |
| 18 | 5 | 21.09 | 26 | 0.16 |
| 19 | 25 | 20.18 | 25 | 0.33 |
| 20 | 30 | 20.01 | 21 | 0.33 |
| 21 | 22 | 19.87 | 17 | 0.40 |
| 22 | 23 | 19.43 | 28 | 0.44 |
| 23 | 29 | 19.29 | 15 | 0.66 |
| 24 | 26 | 18.66 | 10 | 0.80 |
| 25 | 13 | 14.59 | 29 | 1.97 |
| 26 | 8 | 14.38 | 24 | 3.42 |
| 27 | 19 | 13.49 | 22 | 4.92 |
| 28 | 16 | 12.88 | 16 | 7.01 |
| 29 | 21 | 11.25 | 30 | 29.62 |
| 30 | 28 | 8.76 | 23 | 72.69 |

Table 6.2.1. (Cont.)

(c) The Lower Area

| Reordered Element Number | Original Feature Element(k) | Ratio $B_k/A_k$ | Original Feature Element | Max$^m$ Variance |
|---|---|---|---|---|
| 1 | 5 | 67.24 | 3 | 0.00 |
| 2 | 16 | 41.88 | 2 | 0.00 |
| 3 | 11 | 24.89 | 7 | 0.00 |
| 4 | 24 | 24.72 | 8 | 0.00 |
| 5 | 4 | 17.99 | 6 | 0.01 |
| 6 | 2 | 16.76 | 5 | 0.01 |
| 7 | 15 | 15.89 | 12 | 0.01 |
| 8 | 3 | 11.58 | 4 | 0.01 |
| 9 | 30 | 11.43 | 13 | 0.01 |
| 10 | 14 | 8.49 | 1 | 0.01 |
| 11 | 7 | 7.70 | 20 | 0.01 |
| 12 | 8 | 7.62 | 19 | 0.01 |
| 13 | 9 | 7.42 | 14 | 0.01 |
| 14 | 22 | 6.45 | 9 | 0.01 |
| 15 | 21 | 6.07 | 18 | 0.02 |
| 16 | 28 | 5.75 | 11 | 0.02 |
| 17 | 29 | 5.07 | 27 | 0.02 |
| 18 | 19 | 4.63 | 26 | 0.06 |
| 19 | 20 | 3.11 | 21 | 0.06 |
| 20 | 13 | 1.56 | 28 | 0.08 |
| 21 | 18 | 1.32 | 25 | 0.09 |
| 22 | 23 | 1.06 | 17 | 0.14 |
| 23 | 27 | 0.87 | 15 | 0.16 |
| 24 | 12 | 0.80 | 10 | 0.24 |
| 25 | 10 | 0.40 | 16 | 0.24 |
| 26 | 26 | 0.21 | 29 | 0.26 |
| 27 | 6 | 0.11 | 24 | 0.26 |
| 28 | 25 | 0.10 | 22 | 0.53 |
| 29 | 17 | 0.07 | 30 | 3.45 |
| 30 | 1 | 0.01 | 23 | 4.96 |

Note that the scatter ratio method selects several high order moments on which noise has an adverse affect, before lower order moments. However, using variance ordering, the general trend is for the higher order moments to be selected last.

To test the effectiveness of these ordering procedures, the reordered features were submitted for recognition according to the criterion described in the previous section. There is no real reason for choosing the elements in the natural order. To compare the results with a standard, the feature elements were rearranged in a random order by making use of a random number generator and were also submitted for recognition. Thus recognition of the test data was attempted using natural, random, scatter ratio and variance ordering, by firstly attempting recognition based on the first two feature elements, secondly on the first four, and so on until all 30 feature elements were used. The percentage of correct recognition was computed after each experiment. A simplified flow diagram showing the computer simulated recognition system using all four inputs is presented in Figure 6.2.1. In each case there were 15 recognition experiments, the results of which are given in Table C.1. of Appendix C and presented graphically in

Figure 6.2.1. The Recognition Experiments Simplified Flow Diagram

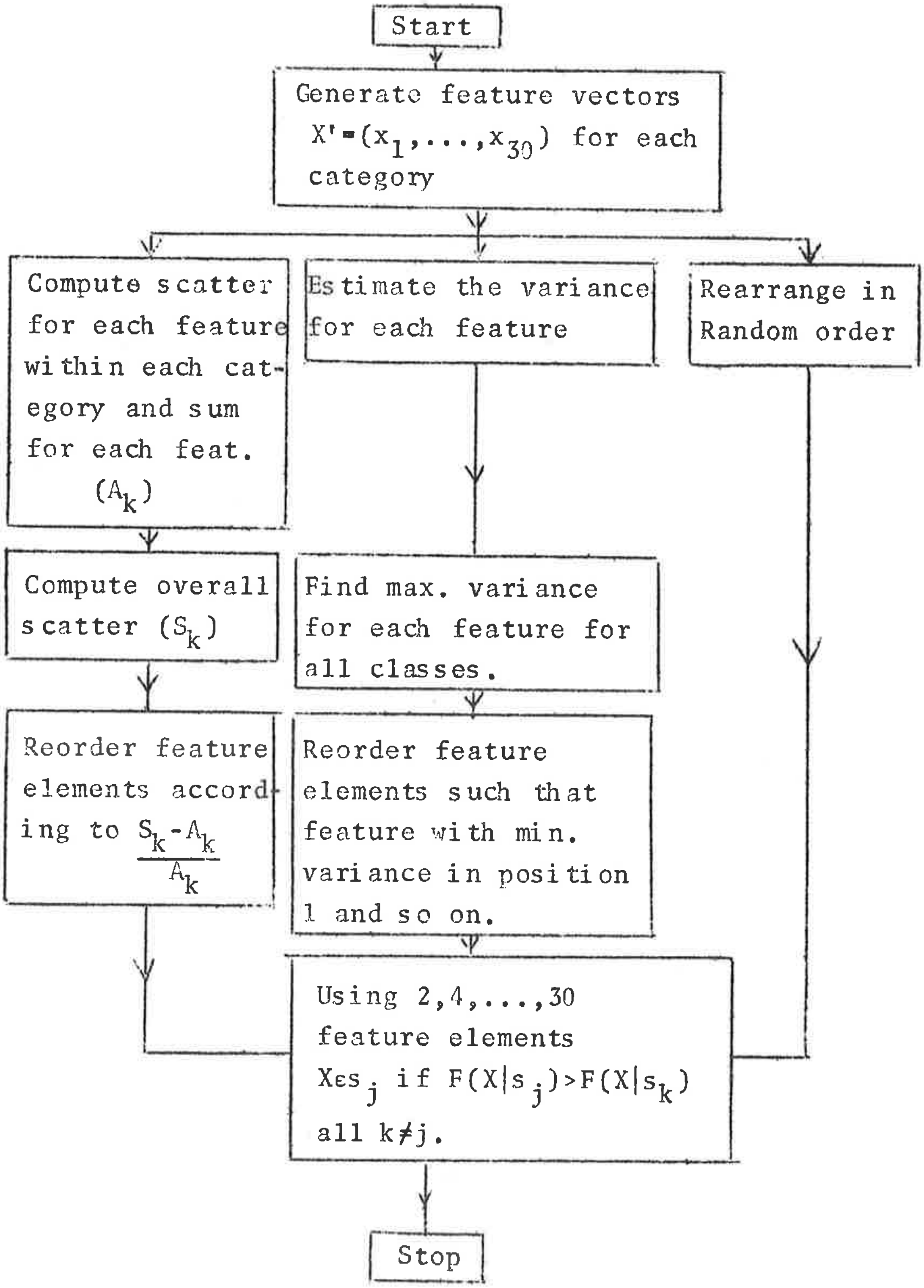

Start
Generate feature vectors $X'=(x_1,\ldots,x_{30})$ for each category
Compute scatter for each feature within each category and sum for each feat. $(A_k)$
Estimate the variance for each feature
Rearrange in Random order
Compute overall scatter $(S_k)$
Find max. variance for each feature for all classes.
Reorder feature elements according to $\frac{S_k-A_k}{A_k}$
Reorder feature elements such that feature with min. variance in position 1 and so on.
Using 2,4,...,30 feature elements $X\epsilon s_j$ if $F(X|s_j)>F(X|s_k)$ all $k\neq j$.
Stop

Figure 6.2.2. It was found for the Lower Area that 100 percent correct recognition was achieved irrespective of the order of the feature elements for this noise level, although it is suspected that if elements 23 and 30 were the first two elements used for recognition, then this would not be the case. However, for the present discussion, the Lower Area was not the most important.

From Figure 6.2.2. it is apparent that variance ordering of feature elements is more advantageous than the other two ordering methods, all three being better than random ordering. For both the Upper and Middle Areas the variance ordered elements yield maximum recognition rates using the first 16. However, as the number of feature elements is increased beyond 16 in the recognition system the recognition rates decrease steadily, showing that the inclusion of features with large variation is detrimental to the system.

Special note should be made of the natural order recognition rates. In this case it is expected, and it is indeed the case, that the recognition curve would be close to the corresponding curve for the variance ordered elements since the low order moments exhibit less variation than those of higher order in the presence of noise. It appears that

Figure 6.2.2. Recognition Curves using 2,4,...,30 feature elements comparing variance, scatter ratio, natural and random ordering.

(a) The Middle Area.

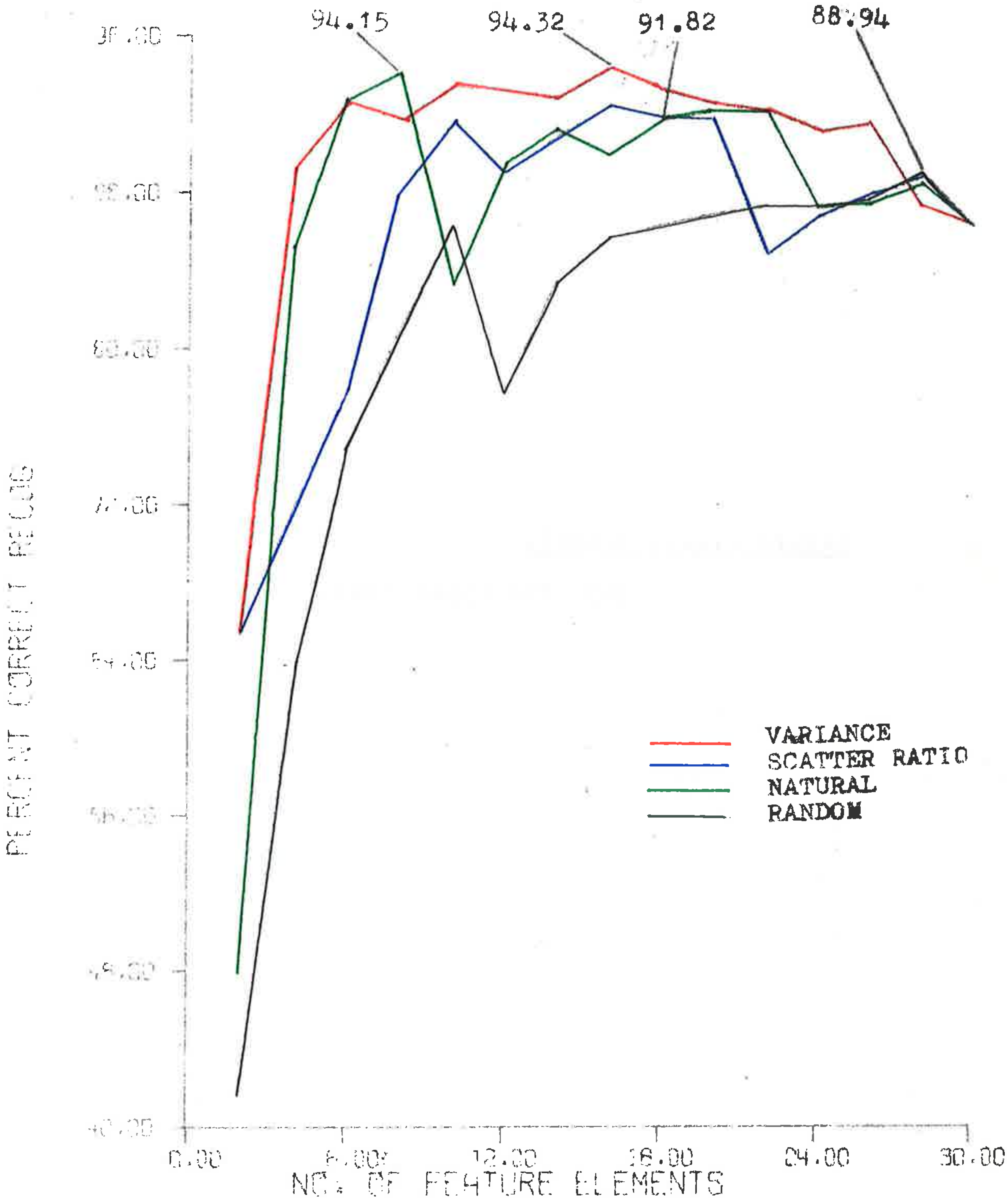
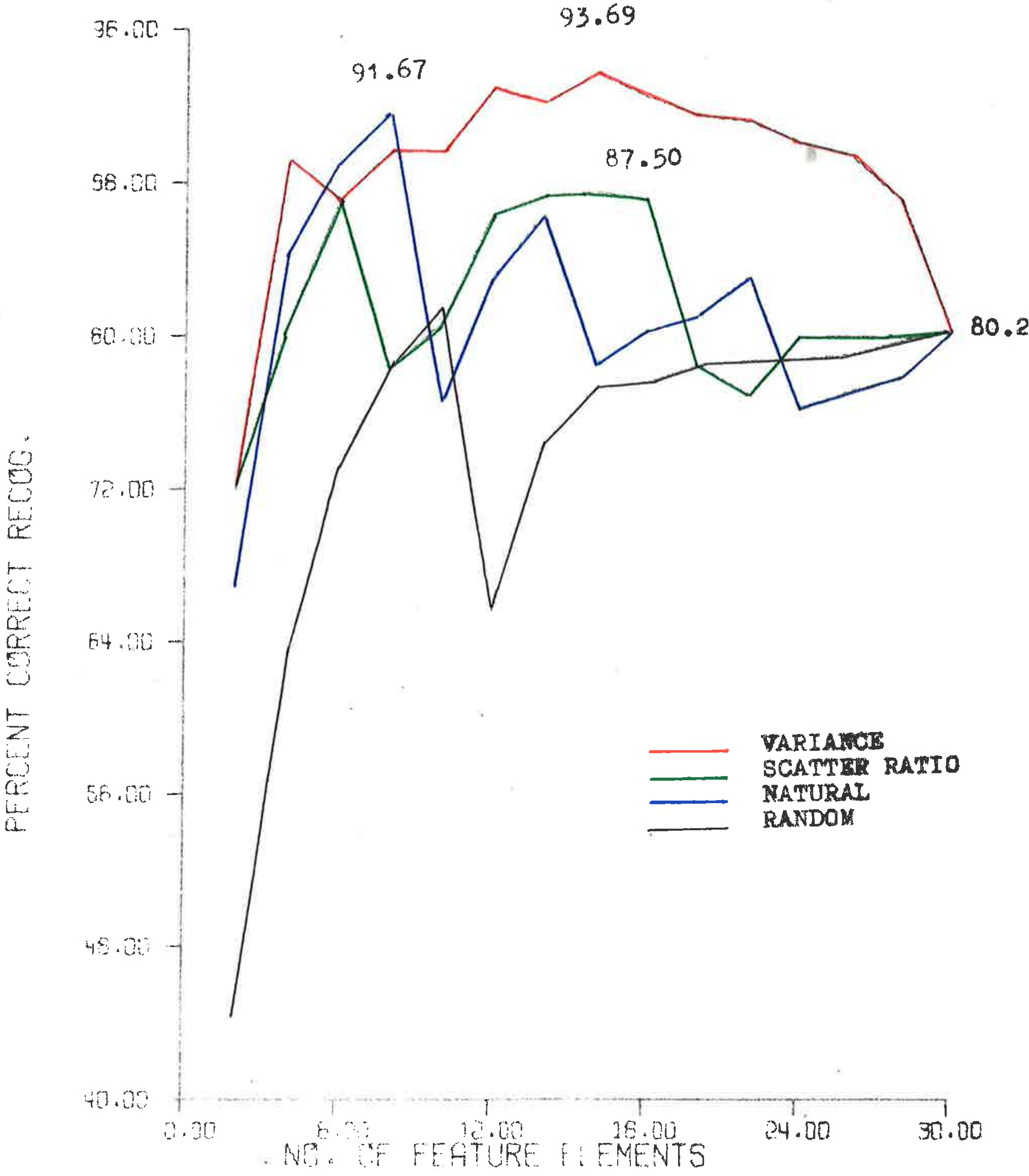

94.15
94.32
91.82
88.94
96.00
88.00
80.00
72.00
64.00
56.00
48.00
40.00
PERCENT CORRECT RECOG
0.00
6.00
12.00
18.00
24.00
30.00
NO. OF FEATURE ELEMENTS
VARIANCE
SCATTER RATIO
NATURAL
RANDOM

Figure 6.2.2. (cont).

(b) The Upper Area.


93.69
91.67
87.50
80.2
96.00
88.00
80.00
72.00
64.00
56.00
48.00
40.00
PERCENT CORRECT RECOG.
0.00
6.00
12.00
18.00
24.00
30.00
NO. OF FEATURE ELEMENTS
VARIANCE
SCATTER RATIO
NATURAL
RANDOM

the recognition rates using the first 6 or 8 natural order elements yield better results than the first 6 of 8 ordered by the other methods. However, as the number of feature elements is increased a higher recognition rate is obtained using variance ordering.

The recognition curve using variance-ordered feature elements is reasonably smooth but there are small fluctuations in the gradient, the inclusion of more elements sometimes being detrimental and at others improving the system. This phenomenon can be explained by observing that some feature elements may have a small intra-category and inter-category spread, which do not provide much information for the recognition system. As a result the small variation which occurs in the element's value is sufficient to cause a slight decrease in the recognition rate instead of it remaining constant as would perhaps be expected. On the other hand, the recognition rate may be increased slightly with the inclusion of elements with a large intra-category variance, when the inter-category spread is sufficiently large. For example, in the variance-ordered case when the number of elements used in the recognition system for the Middle Area is increased from 6 to 8 thus including elements 5 and 7, there is a decrease in the recognition rate of 0.83 percent and when the number of elements is increased from 24 to 26 (including elements 17 and 10) there is an increase of 0.38 percent in the recognition rate.

So far the experimental work has been restricted to moments of orders 3 to 7 inclusive and a question that one might ask is: "Can the performance of the recognition system be improved by taking higher order moments ?" It was shown earlier that variance-ordered moments yield maximum recognition rates using the first 16 feature elements for both the Upper and Middle Areas. For the Upper Area in this 16, there are no moments of order 7 and, since noise has more effect on higher order moments than low order, it is apparent that the inclusion of moments of order higher than 6 is of no value in the recognition system. However, for the Middle Area ordered element 14 is a moment of order 7, and thus there is a possibility that the performance of the system can be improved by including moments of order greater than 7. Consequently a recognition experiment was carried out for the Middle Area including moments or order 8. Using variance ordering it was found that the same maximum recognition rate was obtained using the first 16 feature elements and they were the same elements as in the previous case. Hence the inclusion of higher order moments did not improve the performance of the system.

To this stage, characters with a 10 percent noise level have only been considered - but what of other noise levels? To investigate this question, the procedure for variance ordering was repeated for characters of all areas with noise levels and 20 and 30 percent and the recognition experiment repeated using the first 2 feature elements, the first 4 and so on, and the percent correctly recognised calculated for each. The results are presented in Table 6.2.2.

Table 6.2.2. Reordered feature elements, the maximum variances and percent correct recognition using 2,4,...,30 elements for recognition with noise levels 20 and 30 percent

(a) The Middle Area

| Noise Level 20% | | | Noise Level 30% | | |
|---|---|---|---|---|---|
| Feat. El. | Max$^{m}$ Var. | % correct recog. | Feat. El. | Max$^{m}$ Var. | % correct recog. |
| 2 | 0.005 | | 2 | 0.006 | |
| 3 | 0.008 | 53.11 | 7 | 0.008 | 35.38 |
| 7 | 0.008 | | 3 | 0.009 | |
| 4 | 0.015 | 77.73 | 8 | 0.024 | 70.83 |
| 8 | 0.022 | | 4 | 0.024 | |
| 9 | 0.035 | 85.23 | 9 | 0.028 | 77.88 |
| 6 | 0.039 | | 6 | 0.044 | |
| 1 | 0.042 | 87.65 | 1 | 0.053 | 78.79 |
| 12 | 0.053 | | 12 | 0.055 | |
| 13 | 0.062 | 86.52 | 13 | 0.083 | 78.25 |
| 20 | 0.126 | | 20 | 0.110 | |
| 14 | 0.164 | 88.11 | 14 | 0.123 | 79.62 |
| 19 | 0.182 | | 19 | 0.259 | |
| 18 | 0.327 | 89.24 | 18 | 0.332 | 79.92 |
| 11 | 0.331 | | 11 | 0.367 | |
| 27 | 0.331 | 87.80 | 21 | 0.407 | 82.50 |
| 5 | 0.335 | | 27 | 0.429 | |
| 15 | 0.492 | 85.68 | 5 | 0.441 | 81.06 |
| 21 | 0.661 | | 15 | 0.544 | |
| 26 | 0.667 | 86.06 | 26 | 0.980 | 76.44 |
| 28 | 0.836 | | 28 | 1.090 | |
| 22 | 1.760 | 84.02 | 22 | 1.474 | 74.92 |
| 29 | 3.399 | | 29 | 1.859 | |
| 25 | 4.930 | 83.11 | 25 | 4.363 | 74.55 |
| 17 | 5.174 | | 17 | 6.228 | |
| 10 | 6.887 | 84.47 | 10 | 9.486 | 75.38 |
| 30 | 12.112 | | 30 | 10.782 | |
| 24 | 63.936 | 76.06 | 24 | 84.342 | 65.15 |
| 16 | 92.358 | | 16 | 127.479 | |
| 23 | 1371.547 | 75.00 | 23 | 1902.151 | 62.42 |

Table 6.2.2. (Cont).

(b) The Upper Area

| Noise Level 20% | | | Noise Level 30% | | |
|---|---|---|---|---|---|
| Feat. El. | $\text{Max}^{m}$ Var. | % correct recog. | Feat. El. | $\text{Max}^{m}$ Var. | % correct recog. |
| 2 | 0.005 | | 2 | 0.010 | |
| 3 | 0.008 | 59.05 | 3 | 0.010 | 46.90 |
| 7 | 0.008 | | 7 | 0.011 | |
| 1 | 0.016 | 83.21 | 8 | 0.015 | 73.81 |
| 4 | 0.018 | | 4 | 0.024 | |
| 8 | 0.019 | 85.83 | 1 | 0.026 | 78.45 |
| 6 | 0.020 | | 6 | 0.037 | |
| 12 | 0.024 | 84.40 | 12 | 0.037 | 77.14 |
| 13 | 0.051 | | 13 | 0.079 | |
| 11 | 0.071 | 83.33 | 11 | 0.079 | 77.14 |
| 14 | 0.095 | | 18 | 0.119 | |
| 19 | 0.098 | 81.79 | 5 | 0.132 | 79.17 |
| 18 | 0.104 | | 14 | 0.132 | |
| 5 | 0.116 | 85.48 | 9 | 0.147 | 80.60 |
| 9 | 0.176 | | 19 | 0.200 | |
| 20 | 0.185 | 87.62 | 20 | 0.313 | 81.31 |
| 26 | 0.332 | | 25 | 0.562 | |
| 27 | 0.412 | 85.48 | 26 | 0.696 | 75.12 |
| 25 | 0.622 | | 17 | 0.722 | |
| 17 | 0.642 | 80.12 | 27 | 0.854 | 72.86 |
| 21 | 0.726 | | 15 | 0.951 | |
| 15 | 0.944 | 82.02 | 21 | 1.082 | 74.17 |
| 28 | 1.086 | | 28 | 1.727 | |
| 10 | 1.599 | 81.07 | 10 | 2.024 | 73.45 |
| 29 | 3.953 | | 24 | 5.422 | |
| 24 | 5.995 | 79.52 | 29 | 8.098 | 71.90 |
| 22 | 14.666 | | 22 | 10.041 | |
| 16 | 15.683 | 76.19 | 16 | 10.067 | 69.40 |
| 30 | 54.841 | | 30 | 62.002 | |
| 23 | 155.856 | 66.19 | 23 | 200.893 | 55.95 |

Table 6.2.2. (Cont).

(c) The Lower Area

| Noise Level 20% | | | Noise Level 30% | | |
|---|---|---|---|---|---|
| Feat. El. | Max$^{m}$ Var. | % correct recog. | Feat. El. | Max$^{m}$ Var. | % correct recog. |
| 8 | 0.001 | | 8 | 0.003 | |
| 3 | 0.002 | 100.00 | 3 | 0.003 | 100.00 |
| 2 | 0.003 | | 2 | 0.005 | |
| 7 | 0.004 | 100.00 | 7 | 0.007 | 100.00 |
| 4 | 0.006 | | 4 | 0.010 | |
| 13 | 0.010 | 100.00 | 5 | 0.018 | 100.00 |
| 5 | 0.011 | | 6 | 0.018 | |
| 6 | 0.011 | 100.00 | 13 | 0.020 | 100.00 |
| 20 | 0.014 | | 20 | 0.021 | |
| 12 | 0.015 | 100.00 | 9 | 0.022 | 100.00 |
| 14 | 0.015 | | 14 | 0.023 | |
| 1 | 0.015 | 100.00 | 1 | 0.024 | 100.00 |
| 7 | 0.017 | | 12 | 0.027 | |
| 19 | 0.021 | 100.00 | 19 | 0.031 | 100.00 |
| 11 | 0.034 | | 21 | 0.047 | |
| 27 | 0.034 | 100.00 | 11 | 0.048 | 100.00 |
| 18 | 0.042 | | 27 | 0.064 | |
| 21 | 0.043 | 100.00 | 18 | 0.082 | 100.00 |
| 26 | 0.068 | | 28 | 0.095 | |
| 28 | 0.087 | 100.00 | 26 | 0.148 | 100.00 |
| 29 | 0.186 | | 29 | 0.196 | |
| 17 | 0.188 | 100.00 | 15 | 0.244 | 97.50 |
| 15 | 0.192 | | 17 | 0.447 | |
| 25 | 0.215 | 100.00 | 25 | 0.486 | 97.50 |
| 10 | 0.337 | | 10 | 0.610 | |
| 16 | 0.412 | 100.00 | 16 | 0.738 | 100.00 |
| 24 | 0.426 | | 24 | 0.801 | |
| 22 | 0.732 | 100.00 | 22 | 0.851 | 100.00 |
| 30 | 4.627 | | 30 | 5.027 | |
| 23 | 6.348 | 100.00 | 23 | 11.321 | 100.00 |

The maximum recognition rates for the Middle Area are 94.32, 89.24 and 82.50 percent for noise levels 10, 20 and 30 percent, using the first 16, 14 and 16 feature elements, respectively. In all cases the first 13 elements are found to be the same with slight variation in the order. There is not a fixed set of elements giving a maximum rate, but elements 5, 11, 18 and 27 are all included in the next 5 reliable measurements. On inspection of the results, however, it is found that whenever element 27 is added to the recognition system, (irrespective of the element it is paired with), the recognition rate is decreased. This element was therefore excluded and the other three added to the first 13. In addition it is found that element 21 is associated with an increase in the recognition rate for two of the three noise levels and so this element was added to the other 16 to make a total of 17 to be submitted for a principal component analysis (see next section).

For the Upper Area maximum recognition rates of 93.69, 87.62 and 81.31 percent occurred using the same first 16 feature elements for all three noise levels. With the inclusion of elements 15 and 21 for noise levels 20 and 30 percent, there is a sharp increase in the recognition rate. Thus a recognition experiment using 18 feature elements (including 15 and 21) was carried out for the samples with a noise level of 10 percent. The expected increase in the recognition rate was not evident, in fact the recognition rate was 90.02 percent.- a decrease of 3.67 percent. Thus only the first 16 were submitted for a principal component analysis.

Although the selection of feature elements for the two areas was carried out independently, the first 16 selected for each area were the same, namely, $\{x_1-x_9, x_{11}-x_{14}, x_{18}-x_{20}\}$ and in addition $x_{21}$ was selected for the middle area.

There is little point in submitting ordered variables for the Lower Area for a principal component analysis since 100.00 percent correct recognition is obtained for all noise levels using the first two elements, which however, are not the same in all cases. Feature elements 8 and 3 are found to be the most reliable for the higher noise levels and so all recognition of lower area characters is based on these two elements.

### 6.2.2. Dimension Reduction

It was pointed out in the previous section that some feature elements provide more information than others for the recognition system. For example, the first 8 elements of the "natural" order provide a better recognition result than the first 8 of the variance ordered feature vector for the middle area characters using a noise level of 10 percent. Obviously it would be an advantage if the elements of the feature vector selected by variance ordering could be reordered or weighted according to their relative importance in describing the patterns.

Some work has been done on this type of selecting and ordering of feature elements. But most of the methods proposed require a knowledge or make an assumption about the underlying probability distribution associated with each category. For example, Kullback [33] proposed a method for selection based on the principal of divergence which was later investigated in detail by Marill and Green [34]. For this method, if the conditional probability density functions of the categories are Gaussian with equal covariance matrices the divergence is uniquely related to the recognition error of the Bayes' decision theory classifier. However, when the covariance matrices are different, the divergence is neither uniquely nor monotonically related to the recognition error. Thus, Marill and Green provided the upper and lower bounds on the recognition error as a function of the divergence by using a Monte-Carlo type simulation. The feature elements are selected to maximize the divergence between a pair of categories and at the same time minimizing the recognition error rate. Other methods vary from the simple approach of weighting the feature elements according to the "goodness" of the measurement [35] to the more complex method of using an information theory measure to eliminate the less useful elements [36] - [37].

In this work, a method for selecting and ordering the feature elements which does not require a knowledge of the probability structure of the categories under consideration was used. Essentially the procedure is that of pre-weighting the feature elements according to their relative importance in describing the pattern, regardless of the decision structure of the recognition system by the method of principal components [38].

Suppose that the elements chosen by variance ordering for the recognition system, form the general m-dimension vectors, $Z_i=(z_{1i},x_{2i},\ldots,z_{mi})$, and the overall mean vector is $V=(v_1,v_2,\ldots,v_m)$ (no mention is made of which category each $Z_i$ originated). Then the procedure is to find a set of principal components, $Y_j$, in which all the variation in the system is summarized in fewer variables. The $j^{th}$ principal component of a sample of m-variate observations is defined to be the linear compound

$$Y_j = a_{j1} + a_{j2}z_2 + \ldots + a_{jm}z_m \,, \qquad (24)$$

whose coefficients are the elements of the characteristic vector of the sample covariance matrix, S, corresponding to the $j^{th}$ largest characteristic root, $\lambda_j$, where S is defined in the normal way, e.g.

$$S_{ij} = \frac{1}{(N-1)} \sum_{k=1}^{N} (z_{ik} - v_i)(z_{jk} - v_j). \qquad (25)$$

If $\lambda_i \neq \lambda_j$ then the coefficients of the $i^{th}$ and $j^{th}$ components are necessarily orthogonal. If $\lambda_i = \lambda_j$ , the elements can be chosen to be orthogonal although the infinite number of such orthogonal vectors exist. The sample variance of the $j^{th}$ component is $\lambda_j$, and as a consequence of the orthogonality conditions for the coefficient vectors, the total variance of the feature elements is

$$\lambda_1 + \lambda_2 + \ldots + \lambda_m = \text{trace } S.$$

The importance of the $j^{th}$ component is measured by

$$\frac{\lambda_j}{\text{trace } S}$$

The algebraic sign and magnitude of $a_{ji}$ indicates the direction and importance of the contribution of the i$^{th}$ response to the j$^{th}$ component.

The importance and usefulness of each component is measured by the proportion of the total variance attributed to it. For example, if 90 percent of the variation in a system of 10 feature elements could be accounted for by a simple weighted average of the element values, it would appear that almost all the variation could be expressed along a single continuum rather than a 10 dimensional space. This would be most useful in itself, but, in addition, the coefficients of the 10 feature elements would indicate the relative importance of each in the new derived moment. This may be summarized by stating that the importance of the principal component technique is that of summarizing most of the variation in the system in fewer variables.

If the dimension of the original feature vectors is m, and less than m principal components are taken, then some variance will always be unexplained. How then should one decide how many components provide an adequate description of the system of feature elements? To gain further insight into this question a principal component analysis was performed using the selected feature elements of the sample data used in the previous section (10 percent noise level,) for the Middle and Upper Areas. The selected elements of the training vectors used in the recognition experiments of the previous section were transformed using the calculated coefficient vectors and used to identify the principal components of the test data, recognition being based on the first 2,4,...,16 components. These recognition results are tabulated in Table 6.2.3. together with each characteristic root and the relative importance of each component in describing the variation in the feature elements. These results are also presented graphically in Figure 6.2.3. in which the recognition curve using variance ordered features is included for comparison. The eigenvectors corresponding to each of the eigenvalues may be found in Appendix D.

The property that each successive principal component contains less information is illustrated clearly in Figure 6.2.3. as the addition of more components to the recognition system result in a progressively smaller increase in the recognition rate. Each component adds information to the system which is shown by the curve of recognition rate against the number of components being strictly increasing but, in the case of variance ordering, the addition of some features may be detrimental.

Table 6.2.3. The roots of the characteristic equation $|S-\lambda I|=0$, the amount of variance in each component and the recognition rates using 2,4,6,... principal components.

(a) Middle Area

| j | $\lambda_j$ | % of total Variance | % correct recog. |
|---|---|---|---|
| 1 | 1.230E+00 | 37.77 | |
| 2 | 6.384E-01 | 19.61 | 62.42 |
| 3 | 3.966E-01 | 12.18 | |
| 4 | 3.092E-01 | 9.50 | 89.92 |
| 5 | 2.711E-01 | 8.33 | |
| 6 | 1.442E-01 | 4.43 | 94.24 |
| 7 | 1.002E-01 | 3.08 | |
| 8 | 7.625E-02 | 2.34 | 94.62 |
| 9 | 4.268E-02 | 1.31 | |
| 10 | 2.068E-02 | 0.64 | 94.70 |
| 11 | 1.267E-02 | 0.39 | |
| 12 | 8.102E-03 | 0.25 | 94.77 |
| 13 | 3.014E-03 | 0.09 | |
| 14 | 9.496E-04 | 0.03 | 94.77 |
| 15 | 7.767E-04 | 0.02 | |
| 16 | 4.577E-04 | 0.01 | 94.77 |
| 17 | 2.990E-04 | 0.01 | 94.77 |

(B) Upper Area

| j | $\lambda_j$ | % of total Variance | % correct recog. |
|---|---|---|---|
| 1 | 1.115E+00 | 38.88 | |
| 2 | 7.408E-01 | 25.58 | 64.29 |
| 3 | 4.612E-01 | 16.08 | |
| 4 | 1.944E-01 | 6.78 | 88.33 |
| 5 | 1.354E-01 | 4.72 | |
| 6 | 7.840E-02 | 2.73 | 91.07 |
| 7 | 5.209E-02 | 1.82 | |
| 8 | 3.577E-02 | 1.25 | 91.90 |
| 9 | 2.286E-02 | 0.80 | |
| 10 | 1.486E-02 | 0.52 | 93.10 |
| 11 | 8.249E-03 | 0.29 | |
| 12 | 4.899E-03 | 0.17 | 93.69 |
| 13 | 2.864E-03 | 0.10 | |
| 14 | 7.429E-04 | 0.03 | 93.69 |
| 15 | 4.802E-04 | 0.02 | |
| 16 | 1.814E-04 | 0.01 | 93.69 |

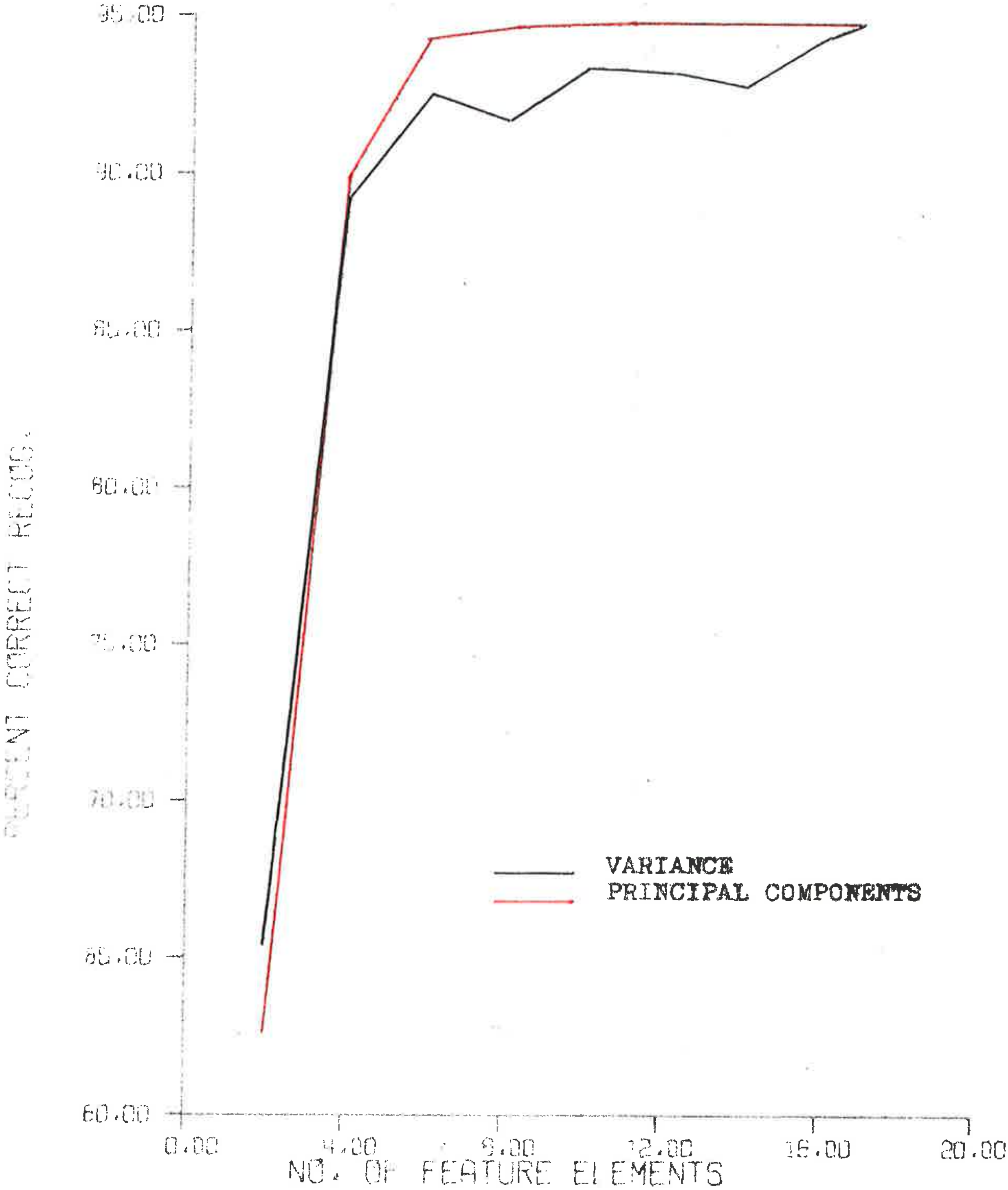
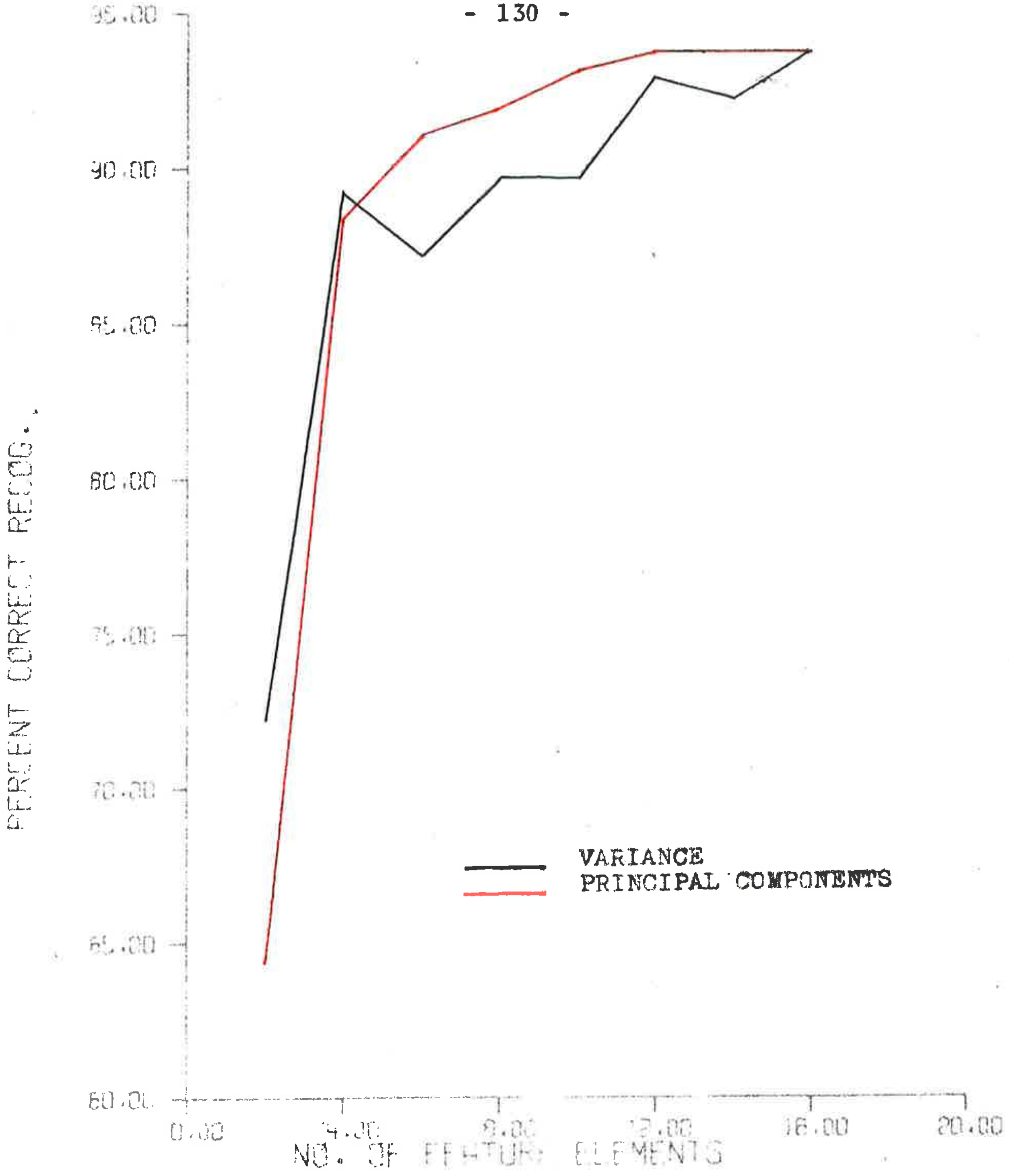


Figure 6.2.3. Comparison of Recognition Rates using Variance Ordering and Principal Components

(a) Middle Area for 2,4,...,16,17 feature elements.



<u>**Figure 6.2.3. (cont)**</u>

**(b) Upper Area for 2,4,...,16 feature elements.**

From the view point of developing an efficient recognition system therefore, almost all of the variance is contained in the first 8 principal components for both the Upper and Middle Areas since they account for 97.84 and 98.55 percent, respectively. Thus by using the first 8 components in the recognition system almost all the variance in the system is accounted for and the dimension of the feature vector reduced dramatically.

It should be noted that it is only necessary to perform one principal component analysis for each area, in the Recognition System. To be more explicit, the moments chosen by variance-ordering are generated for a set of "training data" and a principal component analysis performed on these; for each area. The resulting eigenvectors may then be used to form linear combinations of the moments generated for characters submitted for recognition at some later point in time. The "training data" in the experimental system of this chapter was the set of moments for the character images actually submitted for recognition. However, equally as good results were obtained by using the moments generated for the perfect images. These moments also form the training points for the conditional probability density functions. It was found convenient to store the eigenvectors in cards so that they could be used in later recognition experiments. The principal component program used to compute these eigenvectors simply computes the covariance matrix for the data, and the corresponding eigenvalues and eigenvectors by the standard Jacobi method. The system for preparing the eigenvectors is presented schematically in Figure 6.2.4.

Using the scanner, the same technique is applied with only minor modifications being necessary to the system design. The simulated data is replaced by a binary scan image of training characters (on magnetic tape), together with the line position points. The simulating program is replaced by a corresponding program used for scanner input. It was mentioned earlier (section 4.1.) that character isolation, preprocessing and feature generation can all be conveniently included in the one computer program. It is this program which replaces the simulating program. The program is basically an implementation of the isolation process described in section 4.1. The preprocessing technique (see section 4.2.) and the generation of the selected moments are applied as each isolated image becomes available, the moments being recorded on magnetic tape. This complete procedure requires an average time of about one second per character. For the experimental investigation it was found convenient to record the information for each of the areas separately, that is, disregarding the order of character occurrence.

The system for generating the coefficient vectors using the scanner is shown in Figure 6.2.5. It should be noted at this stage that the principal component results using the simulated data was used in the final system for reasons given in section 8.2.

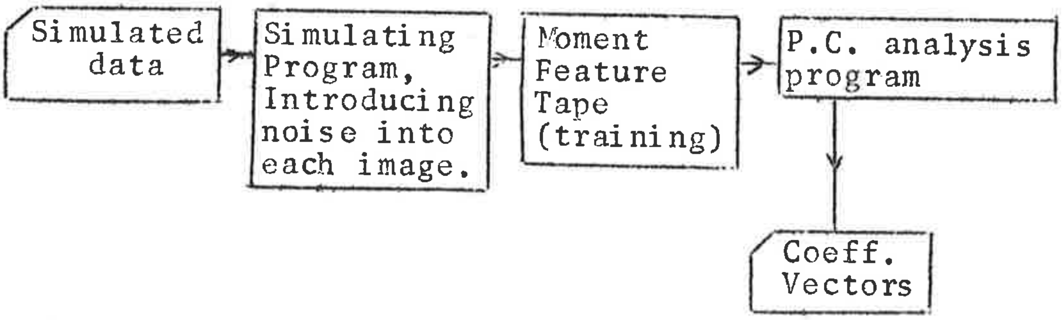

Simulated data
Simulating Program, Introducing noise into each image.
Moment Feature Tape (training)
P.C. analysis program
Coeff. Vectors

Figure 6.2.4. The simulated system of producing Coefficient Vectors

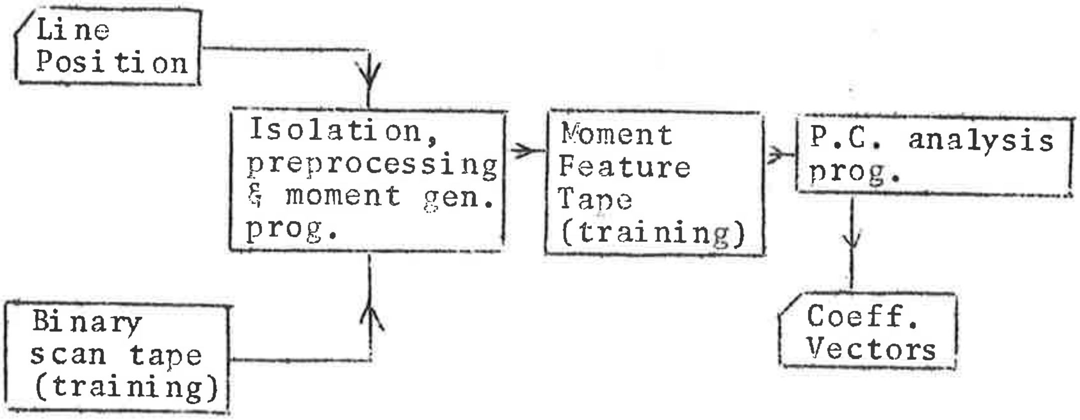

Line Position
Isolation, preprocessing & moment gen. prog.
Moment Feature Tape (training)
P.C. analysis prog.
Binary scan tape (training)
Coeff. Vectors

Figure 6.2.5. The scanner system for producing Coefficient Vectors.

# Chapter 7. Estimates of Conditional Probability Densities.

In order to use the Bayes' decision rules for character recognition it is necessary to evaluate the *a priori* probabilities $p_i$, the losses (or costs) $w_{ij}$, and the conditional probability densities $F(y|s_i)$; $i,j=1,2,\ldots,r$. The $p_i$ can be estimated by making a frequency count of characters from text and the $w_{ij}$ evaluated by trial. However, the conditional probability densities are usually unknown, and thus a categorizer based upon the optimum decision function is not practically realisable. There are at least three possible ways to overcome this problem.

1) Assume a certain form for the conditional density functions. It is common to assume normal distributions and independence for each category [39,40].

2) Make no assumptions about the conditional densities involved but rather make certain restrictions on the structure of the categorizer [40].

3) Approximate the conditional probability densities by using an interpolation function.

It is the third of these alternatives that is used and the method of approximation is discussed in this chapter.

## 7.1. Background

Parzen [41] was one of the first to derive the asymptotic properties of a class of estimates, $f_n(x)$, of a univariate probability density function $f(x)$ based on the random sample $x_1, x_2, \ldots, x_n$ from $f(x)$. Later Murthy [42] and Cacoullos [43] extended this theory to the multivariate case which is of direct importance here.

The estimate, $f_n(X)$, of the true probability density $f(X)$ is of the form

$$f_n(X) = \int \frac{1}{h^p(n)} K\left(\frac{x-y}{h(n)}\right) dF_n(y),$$

$$= \frac{1}{nh^p(n)} \sum_{k=1}^{n} K\left(\frac{x-X_k}{h(n)}\right), \qquad (1)$$

where $F_n(X)$ denotes the empirical distribution function based on the sample of n independent observations $X_1, X_2, \ldots, X_n$ of the random p-dimensional vector X with density $f(X)$; $K(y)$ is a kernel which is chosen to satisfy suitable conditions and $\{h(n)\}$ is a sequence of positive constants satisfying

$$\lim_{n\to\infty} h(n)=0. \qquad (2)$$

The integration is over the entire range of the integral variable. Note that the contribution of one pattern to the overall estimate is not dependent on the other points in the training set.

If the kernel $K(y)$ is a real valued function in p-dimensions satisfying the conditions

$$\left.\begin{array}{ll}
\text{(a)} & K(y) \geq 0 \ , \\
\text{(b)} & \sup\limits_{\text{all } y} K(y) \geq \infty \ , \\
\text{(c)} & \lim\limits_{|y|\to\infty} |y|^{p} K(y) = 0 \ , \\
\multicolumn{2}{l}{\text{where } |y| \text{ denotes the length of vector } y, \text{ and}} \\
\text{(d)} & \int K(y)\,dy = 1 \ ,
\end{array}\right] \qquad (3)$$

and $h(n)$ satisfies (2), then the following asymptotic properties can be found for the estimate $f_n(X)$ [43].

1) The estimator (1) is consistent (asymptotically approaches the true density function $f(X)$)at all points $X$ at which the true density function is continuous, providing

$$\lim_{n\to\infty} nh^{p} = \infty \qquad (4)$$

2) If $f(X)$ is uniformly continuous then the estimate is uniformly consistent (approaches $f(X)$ everywhere) if

$$\lim_{n\to\infty} nh^{2p} = \infty \qquad (5)$$

3) The theoretical solution for the value of $h$ which minimizes $E[f_n(X)-f(X)]^2$ for particular values of $X$ and $n$ is found to be

$$h = [p(nI^2)^{-1} f(X) \int K^2(y)\,dy]^{1/(p+4)}, \qquad (6)$$

where $I = \int \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^2 f(X)}{\partial x_i \partial x_j} y_i y_j K(y)\,dy$.

Note that the value of h cannot be evaluated and thus all that can be concluded is that the optimum choise of h is $0(n^{1/(p+4)}$. However, a sequence satisfying (2), (4) and (5) with $h \sim 0(n^{-(1/p+4)})$ for a fixed value of X is

$$h(n)=c\ n^{-\alpha} \qquad \begin{array}{l} \alpha = (p+4)^{-1} \text{ and} \\ c = \text{constant} > o \end{array} \qquad (7)$$

It is desirable that the overall estimated density function should be smooth and continuous over the domain of X, and that the kernel approaches zero as the Euclidean distance from the training samples tends to infinity. Parzen [41] presents a table of kernels which satisfy all the properties of (3) of which

$$K(y) = \frac{1}{(2\pi)^{p/2}h^p} \exp\left[-\frac{y'y}{2}\right] \qquad (8)$$

is the one with the most desirable characteristics. Thus if n random p-dimensional vectors are available from f(X), namely $X_1, X_2, \ldots, X_n$, then the estimator may be written as

$$f_n(X) = \frac{1}{(2\pi)^{p/2}h^p} \frac{1}{n} \sum_{i=1}^{n} \exp\left[-\frac{(X-X_i)'(X-X_i)}{2h^2}\right] \qquad (9)$$

Using this estimator or approximating function, the conditional probability density functions for the r categories in the p-dimensional principal component space, may be written as

$$F_{n_i}(y|s_i) = \frac{1}{(2\pi)^{p/2} h^p(n_i)} \; \frac{1}{n_i} \sum_{j=1}^{n_i} \exp\left[- \frac{(y-y_{ij})'(y-y_{ij})}{2h^2(n_i)}\right] \tag{10}$$

$$i = 1,2,\ldots,r,$$

where $n_i$ = the number of training points for category $s_i$ and, $y_{ij}$ = $j^{th}$ training point for category $s_i$

$$= \begin{bmatrix} y_{ij_1} \\ \cdot \\ \cdot \\ y_{ij_p} \end{bmatrix}$$

## 7.2. Experiments

Specht [44, 45] has made use of such approximating functions as given in (10) in a recognition system, but in doing so has restricted himself to two categories (r=2), and chosen h independent of the number of training points for each category. In the present study r is increased to a maximum of 66 and h is chosen according to (7). Note that for a fixed value of n, h may be varied by changing c and thus as c tends to zero the decision rule becomes "nearest neighbour" and as it tends to infinity the rule is "minimum distance" [44].

### 7.2.1. Data

To test the effectiveness of the approximations for the conditional probability density functions, it was found convenient to "construct" a piece of Thai "text" by making use of simulated characters (see section 4.3.). The text was constructed by simply generating a noisy image of each character from a section of text. Because of relative sizes, the images from the scanner of characters of the upper and lower areas are in general noisier than those of the middle area. Thus the generated images were given noise levels of 10, 15 and 15 percent for the middle, upper and lower areas, respectively. The selected moments were calculated for each image and linear combinations of these taken according to the principal component results described in Chapter 6. Each feature vector was written on magnetic tape. In the sample of text used, there were 2,402, 573 and 34 characters from the middle, upper and lower areas, respectively.

Twenty training feature vectors were generated for all categories and also written on magnetic tape. The noise levels were the same as for the "text" and the same principal component results were used.

### 7.2.2. Results

Recognition of the "text" was performed by using the Bayes' decision rule with forced decision, equal misrecognition costs, and equal and *a priori* probabilities for each character. Note that for characters which did not occur in the sample of text from which the *a priori* probabilities were evaluated (see section 1.2.) which,however,are still used in the Thai language, a small positive probability was assigned to each.

Using 5 training points for each category, and choosing h according to (7), the text was submitted for recognition with a fixed value of c, making use of the approximating functions (10). A flow diagram of the experiments is presented in Figure 7.2.1. The recognition experiment was repeated several times, varying c (and thus h) on each occasion, and the correct recognition rates computed for equal and *a priori* probabilities. These results are presented in Table 7.2.1. and Figure 7.2.2. for the three areas. Generally it is found when using *a priori* probabilities, as the value of c is increased the correct recognition rates drop dramatically, but for small c the results are quite good. On the other hand when assuming equal probabilities the recognition rates are good for a large range of c, with only a small decrease in the rates as c is increased.

*A priori* probabilities are indeed an advantage to the recognition system for the upper area. Using both equal and *a priori* probabilities a

Figure 7.2.1. Recognition experiment to test conditional probability density function approximations.

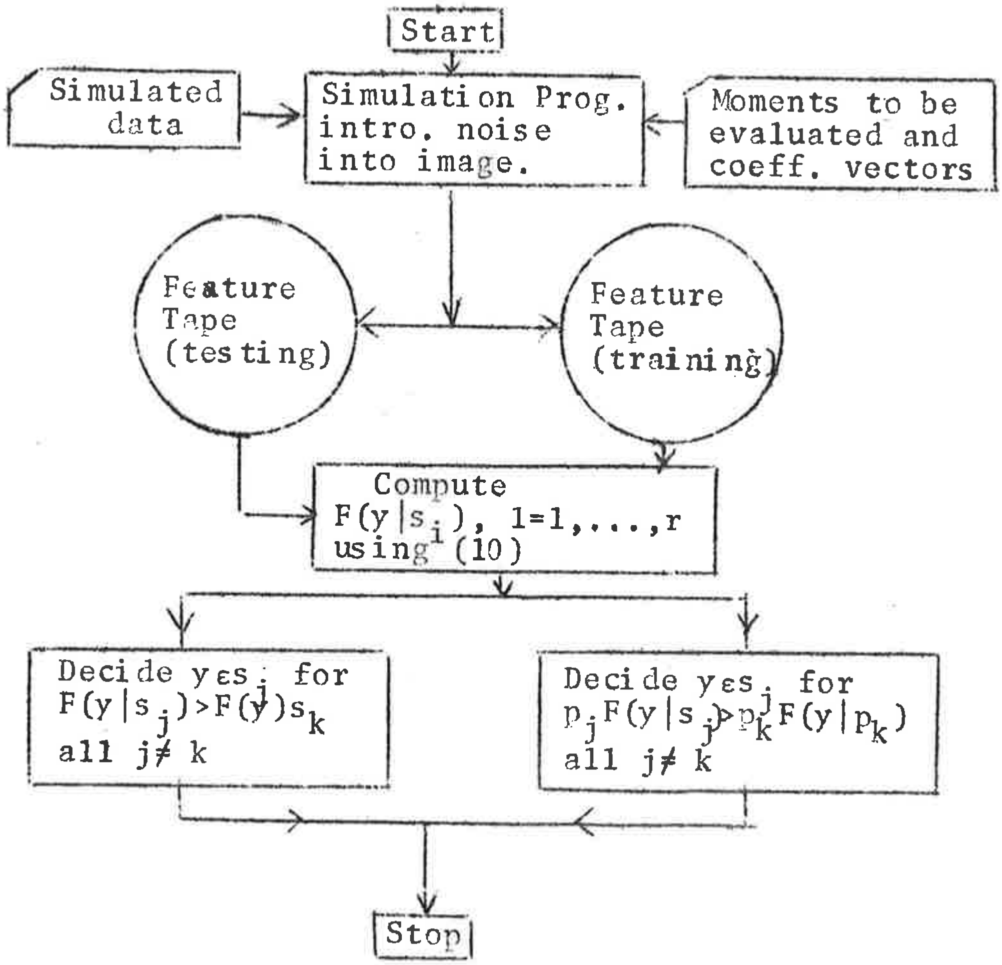

Start
Simulated data
Simulation Prog. intro. noise into image.
Moments to be evaluated and coeff. vectors
Feature Tape (testing)
Feature Tape (training)
Compute $F(y|s_i)$, 1=1,...,r using (10)
Decide $y\varepsilon s_j$ for $F(y|s_j)>F(y)s_k$ all $j\neq k$
Decide $y\varepsilon s_j$ for $p_jF(y|s_j)>p_kF(y|p_k)$ all $j\neq k$
Stop

Table 7.3.1. Recognition results for simulated text using 5 training samples for each category and varying h(n) according to $h(n)=cn^{-\alpha}, \alpha=(p+4)^{-1}$, where n = the number of training points and p the dimension of the feature vectors.

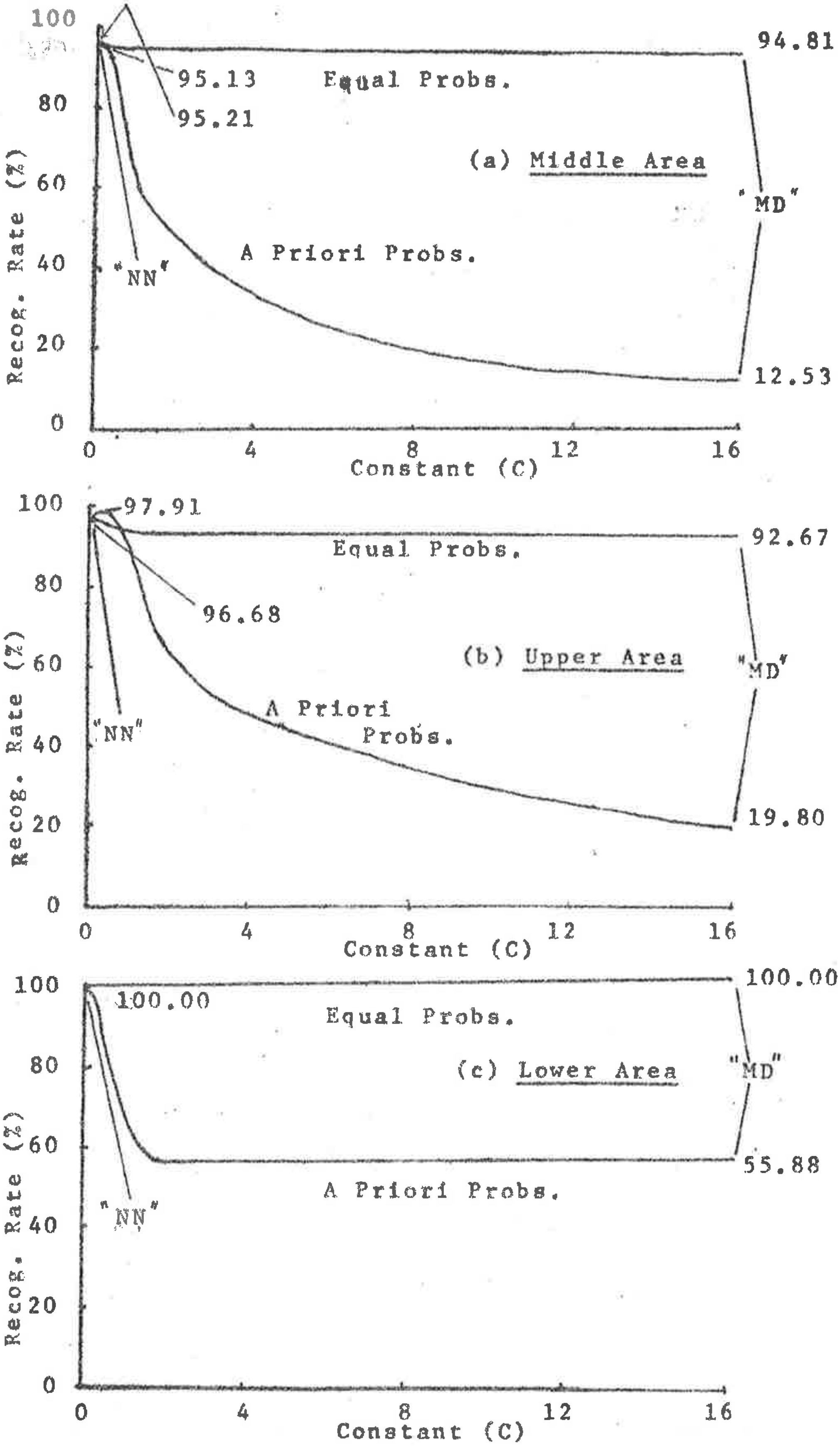
Percent Correctly Recognised

| c | Equal Probabilities | A Priori Probabilities |
|---|---|---|
| | Middle Area | |
| 0.063 | 95.13 | 95.21 |
| 0.125 | 95.37 | 95.21 |
| 0.250 | 95.21 | 93.14 |
| 0.500 | 94.97 | 87.71 |
| 1.000 | 94.81 | 63.21 |
| 2.000 | 94.81 | 46.77 |
| 4.000 | 94.81 | 34.48 |
| 8.000 | 94.81 | 19.87 |
| 16.000 | 94.81 | 12.53 |
| | Upper Area | |
| 0.063 | 96.63 | 97.38 |
| 0.125 | 96.68 | 97.91 |
| 0.250 | 96.51 | 97.91 |
| 0.500 | 94.76 | 94.59 |
| 1.000 | 93.02 | 88.83 |
| 2.000 | 92.84 | 61.08 |
| 4.000 | 92.67 | 48.69 |
| 8.000 | 92.67 | 36.30 |
| 16.000 | 92.67 | 19.90 |
| | Lower Area | |
| 0.062 | 100.00 | 100.00 |
| 0.125 | 100.00 | 100.00 |
| 0.250 | 100.00 | 94.12 |
| 0.500 | 100.00 | 64.71 |
| 1.000 | 100.00 | 55.88 |
| 2.000 | 100.00 | 55.88 |
| 4.000 | 100.00 | 55.88 |
| 8.000 | 100.00 | 55.88 |
| 16.000 | 100.00 | 55.88 |

Figure 7.2.2. Recognition Curves using Equal and A Priori Probabilities for Simulated Text. The smoothing parameter h(n) $= cn^{-\alpha}, \alpha = cn^{-\alpha}(p+4)^{-1}$, for n=5 and p=8.

NN= Nearest Neighbour Classifier

MD =Minimum Distance Classifier.


100
94.81
95.13
Equal Probs.
80
95.21
(a) Middle Area
60
"MD"
Recog. Rate (%)
40
A Priori Probs.
"NN"
20
12.53
0
0
4
8
12
16
Constant (C)
100
97.91
92.67
Equal Probs.
80
96.68
(b) Upper Area
60
"MD"
"NN"
A Priori Probs.
40
Recog. Rate (%)
20
19.80
0
0
4
8
12
16
Constant (C)
100
100.00
100.00
Equal Probs.
80
(c) Lower Area
"MD"
60
55.88
A Priori Probs.
"NN"
40
Recog. Rate (%)
20
0
0
4
8
12
16
Constant (C)

maximum recognition rate is obtained for c equal to 0.125. However, there is an increase of 1.22 percent from 96.63 to 97.91 percent in the recognition rate when a priori probabilities are used.

The same cannot be said for the middle area. With c equal to 0.125 maximum recognition rates of 95.37 and 95.21 percent are obtained using equal and a priori probabilities, respectively. In addition as c is increased the recognition rate using equal probabilities is decreased by only 0.32 percent. Bearing these rates in mind, it would seem a considerable advantage to assume equal probabilities for the middle area characters, choose a large value of c thus enabling the use of the polynomial expansion of the approximating functions (10) (see Specht [44]). This polynomial expansion is far more efficient than the series form from the point of view of storage and computing considerations. However, it has been domonstrated by Edwards and Chambers [46] that as noise in the feature vector (or binary pattern) is increased, then a priori probabilities become increasingly useful in a recognition system. Thus a small value of c (0.125) and a priori probabilities are used in the system for the middle area.

Assuming equal probabilities for both characters in the lower area 100 percent correct recognition is obtained for all values of c, but using a priori probabilities 100 percent recognition results for c equal to 0.063 and 0.125 only. Thus it would seem that there is no advantage to be gained by using a priori probabilities for this area, but, it was decided to use them for the sake of consistency.

It would be expected that increasing the number of training points for each category would yield more accurate estimates of the conditional probability density functions, thus improving the recognition rate. Since there could not be any improvement for the lower area, further recognition experiments were carried out on the simulated text for the upper and middle areas only, but in this case using 10 training points for each category. For the upper area with c set to 0.125 there was an increase of 0.52 and 0.70 percent in the recognition rates using equal and a priori probabilities, respectively. With c equal to 1.0 there were corresponding increases of 1.68 and 1.52 percent for the middle area. In general as n is increased the curves of Figure 7.2.1. are shifted upward.

To obtain these increases in the recognition rate, the computing time for recognition was approximately doubled. Using 5 training points for each category, the samples are recognised at a rate of approximately 10, 9 and 20 per second for the upper, middle and lower areas, respectively.

Because of the necessity to store all training points in central memory there is a need to place a restriction on the number of training points for each category. A restriction of this nature immediately restricts the amount of computing time required to identify each sample. Consideration must also be made of the significant increase in the recognition rate for the middle area using 10 training samples. Thus it was decided to restrict the upper and lower areas to a maximum of 5 and the middle area to a maximum of 10 training points for each category. With these restrictions, if the first 8 principal components are used to form thé feature vector for the upper and middle areas, and the feature vector of the lower area is of dimension 2 then about 7,000 central memory words are required to store the training points.

### 7.2.3. Discussion

When the approximation functions for the conditional probability densities are used in a recognition system, dimension reduction is particularly important. For example, if the dimension of the feature vectors for the upper and middle areas was increased to 10, then with the number of training samples specified above, a further 1,740 central memory locations would be required by the training points.

The results presented in section 7.2.2. show that for recognition in all three areas it is an advantage to use a priori probabilities in the recognition system with c equal to 0.125. This indicates that the optimal decision surfaces between the categories are highly non-linear [44]. Thus it can be seen that it is an advantage to approximate the conditional probability densities by (10) rather than assume normal distribution for each category, say. (Normal distributions with unequal covariance matrices only yield quadratic surfaces).

Note that the range of h for which the decision rule yields optimal results is small, and thus it must be selected with care.

## Chapter 8 - Results

In this chapter the recognition results obtained using simulated text and the page scanner are discussed. Of particular interest is the introduction of rejection as a possible decision in the categorizer, and its subsequent effect on the recognition system. The difficulties encountered using the page scanner as the input device to the recognition system are also explained.

For convenience the Bayes' decision rule for equal costs of misrecognition and rejection, which was derived in Chapter 5 is restated here.

$$y \varepsilon s_k \text{ for } p_k F(y|s_k) \geq p_j F(y|s_j) \text{ for all } j \neq k$$

$$p_k F(y|s_k) \geq \beta \sum_{i=1}^{r} p_i F(y|s_i), \quad 0 \leq \beta \leq 1,$$

and reject y as being unrecognisable if

$$p_j F(y|s_j) < \beta \sum_{i=1}^{r} p_i F(y|s_i), j=1,...,r \qquad (1)$$

The constant, $\beta$, is chosen to force the system to meet a given rejection rate, and for equal costs for rejection, $w_o$, and equal costs for misrecognition, $w$ ,

$$\beta = 1 - w_o/w \ .$$

### 8.1. Simulated Text

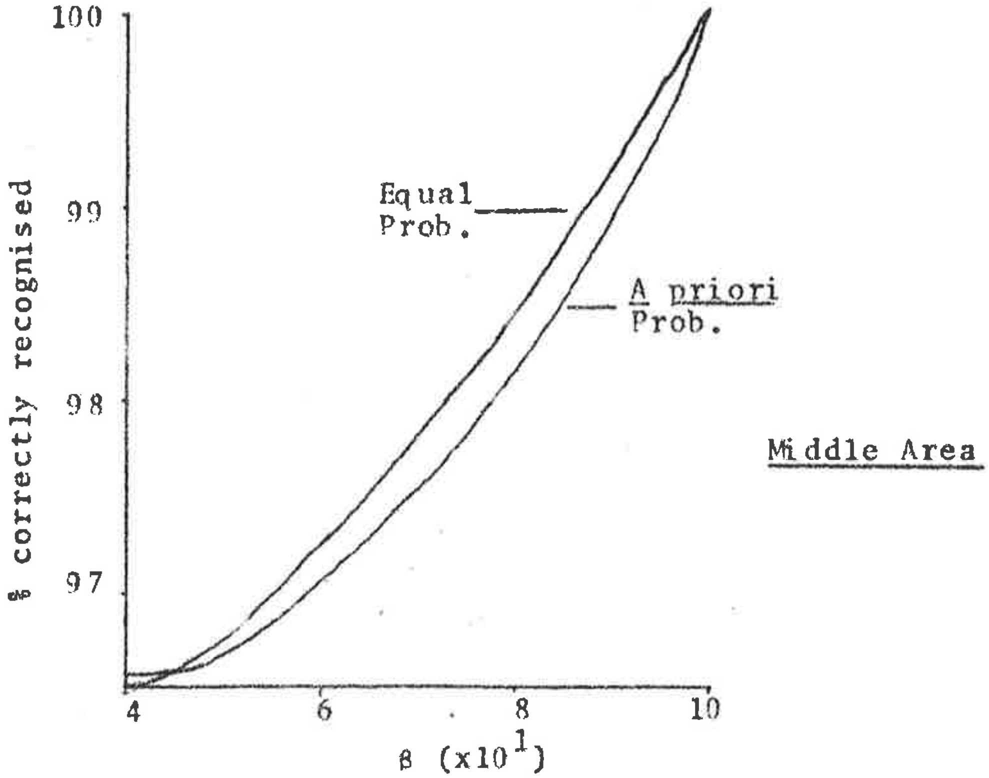
Using the simulated text, with the constant c set to 0.125 and using 10, 5 and 5 training samples for the middle, upper and lower areas respectively, the recognition experiment described in section 7.2 was repeated. However, in this case the identified category, $s_i$, $p_i F(y|s_i)$ and $\sum_{j=1}^{r} p_j F(y|s_j)$, using both *a priori* and equal $p_i$, were recorded on magnetic tape for each input, y. This tape will be referred to as the decision tape. A series of experiments were then performed on these results by implementing (1) in a computer program and using a different value of $\beta$ for each experiment. The percents of text correctly recognised and rejected were computed for each area for equal and *a priori* probability of character occurrence. These results are summarized in Figures 8.1.1. and 8.1.2, the former showing the correct recognition rate, the latter showing the percent of the samples rejected plotted against $\beta$. Table 8.1.1. shows the confusion tables for forced decision.

Note that there is no advantage to be gained by introducing rejection as a possible decision for the lower area for this data. With forced recognition 100 percent correct recognition is obtained (see Chapter 7). Thus this area is not considered here.

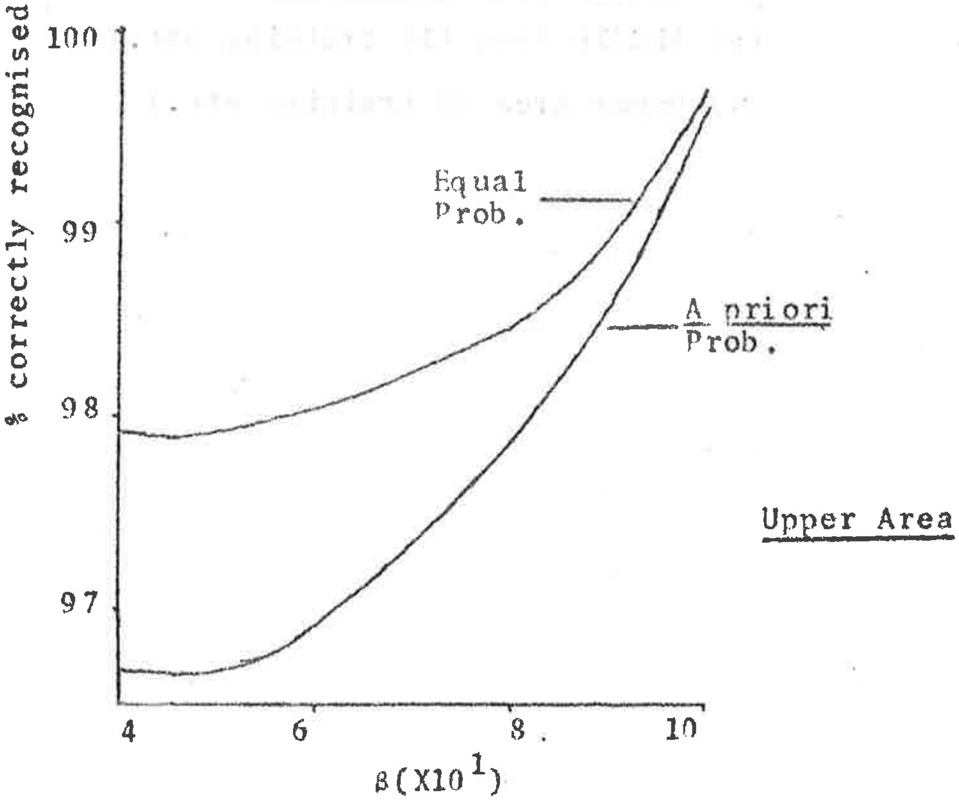
Figure 8.1.1. The Percent of the Simulated Text correctly recognised plotted against $\beta$ using *a priori* and equal probabilities $c = 0.125$.

(a) Middle Area using 10 training pts.

(b) Upper Area using 5 training pts.

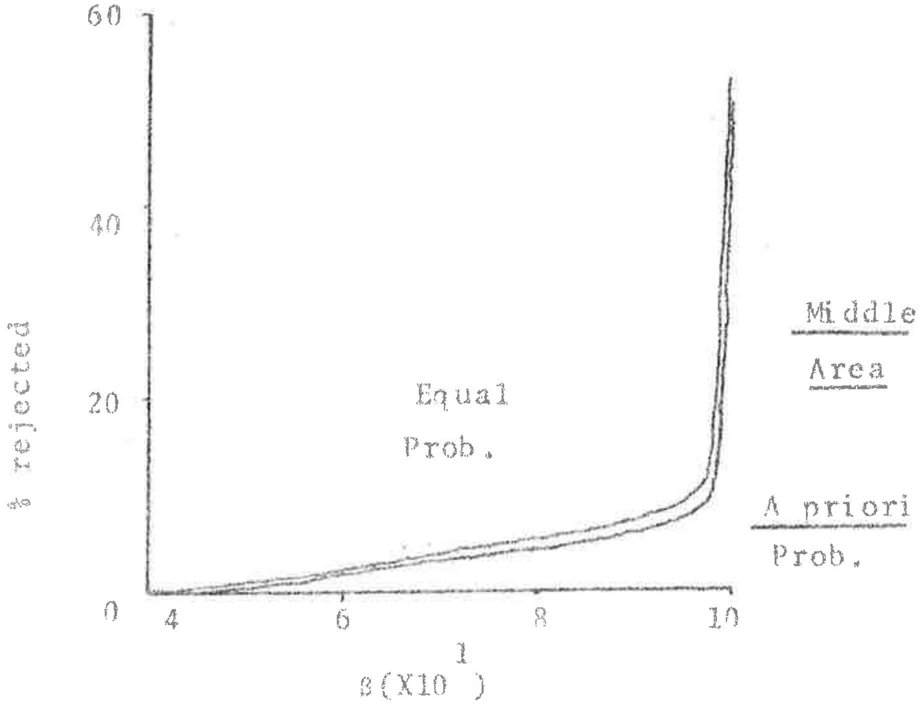

100
99
98
97
% correctly recognised
Equal Prob.
A priori Prob.
Middle Area
4
6
8
10
β (x10$^1$)

(a)

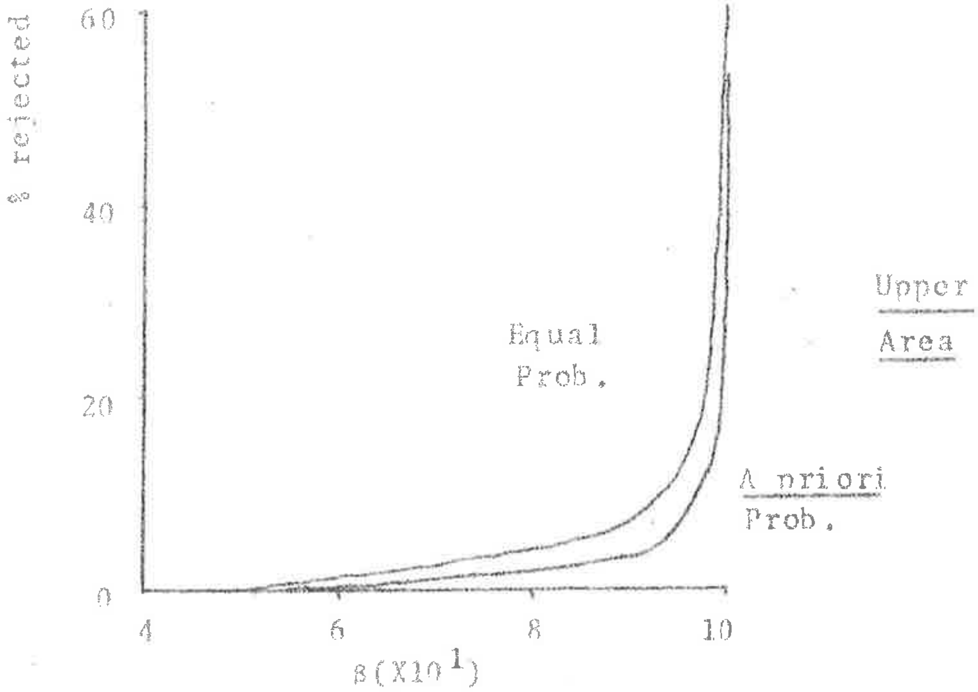

100
99
98
97
% correctly recognised
Equal Prob.
A priori Prob.
Upper Area
4
6
8
10
β(X10$^1$)

(b)

Figure 8.1.2. Percent of Simulated Text rejected,
plotted against β using a priori
and equal probs., c=0.125.
(a) Middle Area (10 training pts.)
(b) Upper Area (5 training pts.)


60
40
20
0
% rejected
4
6
8
10
β(X10$^1$)
Equal
Prob.
A priori
Prob.
Middle
Area

(a)


60
40
20
0
% rejected
4
6
8
10
β(X10$^1$)
Equal
Prob.
A priori
Prob.
Upper
Area

(b)

Table 8.1.1. The confusion table for forced decision using equal and a priori probabilities using simulated text. i(j) means that i characters were misrecognised as character j.

(a) Middle Area for Noise Level 12.5 percent, c=0.125 and using 10 training patterns for each category

| Char. | Total No. | No. of Errors |  | Distribution of Errors |  |
|---|---|---|---|---|---|
|  |  | Equal Prob. | A Priori Prob. | Equal Prob | A Priori Prob |
| 1 | 141 | 1 | 1 | 1(3) |  |
| 2 | 35 | 0 | 0 |  |  |
| 3 | 56 | 0 | 0 |  |  |
| 4 | 39 | 0 | 0 |  |  |
| 5 | 0 | 0 | 0 |  |  |
| 6 | 106 | 0 | 0 |  |  |
| 7 | 3 | 0 | 0 |  |  |
| 8 | 63 | 7 | 4 | 7(9) | 4(9) |
| 9 | 9 | 1 | 1 | 1(8) | 1(8) |
| 10 | 0 | 0 | 0 |  |  |
| 11 | 4 | 1 | 0 | 1(10) |  |
| 12 | 0 | 0 | 0 |  |  |
| 13 | 9 | 0 | 0 |  |  |
| 14 | 14 | 0 | 0 |  |  |
| 15 | 0 | 0 | 0 |  |  |
| 16 | 2 | 0 | 0 |  |  |
| 17 | 8 | 2 | 0 | 2(12) |  |
| 18 | 48 | 13 | 13 | 13(19) | 13(19) |
| 19 | 54 | 32 | 34 | 1(3), 30(18), 1(43) | 1(3), 33(18) |
| 20 | 8 | 0 | 0 |  |  |
| 21 | 2 | 0 | 0 |  |  |
| 22 | 71 | 8 | 12 | 1(29), 7(40) | 1(29), 11(40) |
| 23 | 7 | 0 | 0 |  |  |
| 24 | 182 | 7 | 5 | 1(12), 1(25), 5(32) | 5(32) |
| 25 | 52 | 0 | 0 |  |  |
| 26 | 76 | 0 | 0 |  |  |
| 27 | 19 | 0 | 0 |  |  |
| 28 | 0 | 0 | 0 |  |  |
| 29 | 31 | 0 | 0 |  |  |
| 30 | 0 | 0 | 0 |  |  |
| 31 | 18 | 0 | 0 |  |  |
| 32 | 71 | 1 | 2 | 1(24) | 2(24) |

Table 8.1.1. (cont.)

| | | | | | |
|---|---|---|---|---|---|
| 33 | 77 | 0 | 0 | | |
| 34 | 191 | 0 | 0 | | |
| 35 | 56 | 0 | 0 | | |
| 36 | 87 | 1 | 1 | 1(45) | 1(45) |
| 37 | 8 | 0 | 0 | | |
| 38 | 23 | 0 | 0 | | |
| 39 | 45 | 0 | 0 | | |
| 40 | 49 | 5 | 3 | 3(22),2(29) | 2(22,1(29) |
| 41 | 0 | 0 | 0 | | |
| 42 | 106 | 2 | 0 | 1(7),1(55) | |
| 43 | 0 | 0 | 0 | | |
| 44 | 70 | 0 | 0 | | |
| 45 | 291 | 0 | 0 | | |
| 46 | 129 | 1 | 1 | 1(56) | 1(56) |
| 47 | 54 | 0 | 0 | | |
| 48 | 3 | 0 | 0 | | |
| 49 | 35 | 0 | 0 | | |
| 50 | 44 | 0 | 0 | | |
| 51 | 4 | 2 | 2 | 2(52) | 2(52) |
| 52 | 0 | 0 | 0 | | |
| 53 | 0 | 0 | 0 | | |
| 54 | 0 | 0 | 0 | | |
| 55 | 0 | 0 | 0 | | |
| 56 | 1 | 0 | 0 | | |
| 57 | 0 | 0 | 0 | | |
| 58 | 0 | 0 | 0 | | |
| 59 | 0 | 0 | 0 | | |
| 60 | 0 | 0 | 0 | | |
| 61 | 0 | 0 | 0 | | |
| 62 | 1 | 0 | 0 | | |
| 63 | 0 | 0 | 0 | | |
| 64 | 0 | 0 | 0 | | |
| 65 | 0 | 0 | 0 | | |
| 66 | 0 | 0 | 0 | | |
| TOTAL | 2402 | 84 | 79 | | |

Table 8.1.1.(b) Upper Area for Noise Level 15 percent, c=0.125 and using 5 training patterns for each category.

| Char. | Total No. | No. of Errors | | Distribution of Errors | |
|---|---|---|---|---|---|
| | | Equal Prob. | A Priori . | Equal Prob. | A Priori Prob. |
| 1 | 49 | 0 | 0 | | |
| 2 | 71 | 0 | 0 | | |
| 3 | 14 | 4 | 2 | 4(4) | 2(4) |
| 4 | 8 | 1 | 3 | 1(3) | 3(3) |
| 5 | 17 | 0 | 0 | | |
| 6 | 65 | 0 | 0 | | |
| 7 | 94 | 8 | 2 | 1(14),1(40), 6(42) | 1(14),1(40) |
| 8 | 114 | 0 | 0 | | |
| 9 | 0 | 0 | 0 | | |
| 10 | 0 | 0 | 0 | | |
| 11 | 47 | 2 | 2 | 2(38) | 2(38) |
| 12 | 7 | 0 | 0 | | |
| 13 | 3 | 0 | 0 | | |
| 14 | 1 | 0 | 0 | | |
| 15 | 0 | 0 | 0 | | |
| 16 | 0 | 0 | 0 | | |
| 17 | 2 | 0 | 0 | | |
| 18 | 24 | 1 | 1 | 1(28) | 1(28) |
| 19 | 8 | 1 | 0 | 1(24) | |
| 20 | 0 | 0 | 0 | | |
| 21 | 0 | 0 | 0 | | |
| 22 | 0 | 0 | 0 | | |
| 23 | 13 | 0 | 0 | | |
| 24 | 1 | 0 | 0 | | |
| 25 | 0 | 0 | 0 | | |
| 26 | 0 | 0 | 0 | | |
| 27 | 0 | 0 | 0 | | |
| 28 | 11 | 1 | 1 | 1(18) | 1(18) |
| 29 | 2 | 0 | 0 | | |
| 30 | 0 | 0 | 0 | | |
| 31 | 0 | 0 | 0 | | |
| 32 | 0 | 0 | 0 | | |
| 33 | 1 | 1 | 1 | 1(12) | 1(12) |
| 34 | 1 | 0 | 0 | | |
| 35 | 0 | 0 | 0 | | |
| 36 | 0 | 0 | 0 | | |
| 37 | 0 | 0 | 0 | | |
| 38 | 7 | 0 | 0 | | |
| 39 | 13 | 0 | 0 | | |
| 40 | 0 | 0 | 0 | | |
| 41 | 0 | 0 | 0 | | |
| 42 | 0 | 0 | 0 | | |
| TOTAL | 573 | 19 | 12 | | |

## 8.2. Scanner

In order to obtain reasonable recognition results using the scanner, it was found necessary to photograph each page of text, for reasons enlarged upon below.

A primary requirement of the scanner is that the pages of text to be scanned must be on a loose sheet of paper. It was found that print of reasonable quality suitable for machine recognition, could only be found in expensive books, and the removal of individual pages for recognition was out of the question. Cheaper magazines and journals, from which pages could be easily (and cheaply) removed, contain print of poor quality, with many characters being distorted and smudged. Another common fault with magazine printing was found with large middle area characters, which have a single stroke extending into the upper area. These were quite often printed in two distinct parts, with one portion in the middle and the other in the upper area, thus appearing to be an ordinary middle area character with a tone above it.

To add to this difficulty a tone may be combined with the portion in the upper area (see Figure 8.2.1.). The identification of these characters, and those which are distorted (particularly middle area characters which are similar), by a human reader is difficult, and in many cases only context allows identification. Thus there is little point in submitting such text for machine recognition.

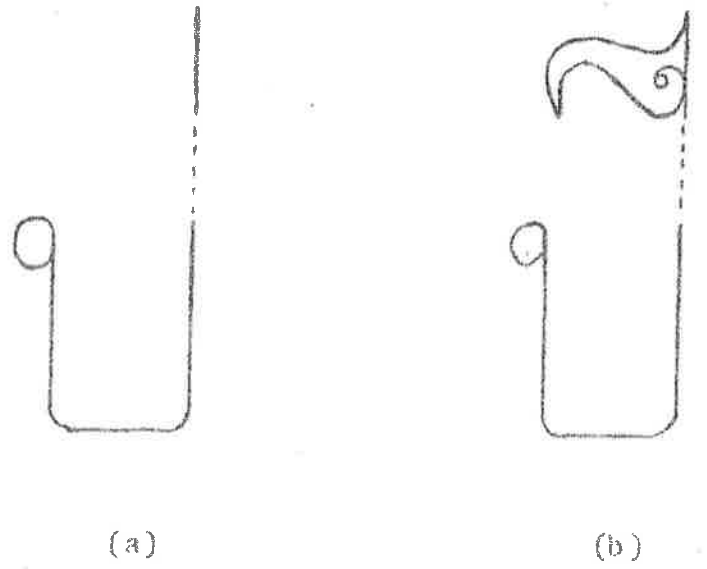


Figure 8.2.1. A broken large middle area character

(a) without a tone

(b) with a tone.

Finally, and perhaps the most significant requirement, was the need to enlarge the pages of text, (effectively increasing the resolution of the scanner), to enable a reasonable correct recognition rate to be obtained. Without enlargement, difficulty was also encountered in isolating the individual characters since many were found to be "touching" (see section 4.2). The sample of text illustrated in Figure 1.1.1. is of the size actually used to obtain the results described below.

As was to be expected, the collection of training data was a problem, particularly as the necessity for photographing restricted the number of pages that could be processed. The method adopted was to simply scan a few pages of text and gather suitable training samples from these. Since some characters are rarely used and because of the limited number of pages available, it was not possible to obtain the recommended number of training samples for all categories (see Chapter 7). In all the recognition experiments using the scanner, the categories for which no training samples were available were automatically given zero probability of occurring.

Since training samples were not available for some categories and a variable number of others, a principal component analysis of the moments for the available data would not yield realistic results. The categories with more training samples would have the most influence, which is an undesirable condition, since, ultimately members of all categories must be identified. Thus linear combinations of the generated moments for each character were taken according to the principal component analysis of Chapter 6, with the resulting feature vectors being stored on magnetic tape. These combinations are obviously not optimal, but they are better than those that could be obtained using the available data.

The system used to construct the training points in p-dimensional principal component space for the scanner input is presented in Figure 8.2.1.

The same recognition experiment as was described in section 8.1. was performed for the scanned version of the same piece of Thai text from which the simulated text was constructed. In this case, however, instead of 10, 5 and 5 training samples being available for each category of the middle, upper and lower areas respectively, a variable number were available for each. The number of training samples were restricted to maxima of 10, 5 and 5 for each category of the respective areas. This experimental system is presented diagramatically in Figure 8.2.2. with the results being given in Figures 8.2.3. and 8.2.4. Table 8.2.1. shows the confusion table for each area in the case of forced recognition together with the number of training samples that were available for each category. Note that 100 percent correct recognition was once again obtained for the lower area characters with forced recognition and so there is no point in including this result diagramatically.

## 8.3. Discussion

A possible alternative for constructing a training set, would be to cut specimens of each character from the pages of a text book, photograph and enlarge them and then scan the print. However, once again this would involve the destroying of a text book and it does not overcome the necessity for photographing.

A second alternative would be to purchase a typewriter, preferably with large letters and use this to arrange training sets. However, since the present study is not concerned with recognition of different fonts it would be necessary to transcribe Thai text onto loose sheets of paper. This would of course defeat the purpose of the project.

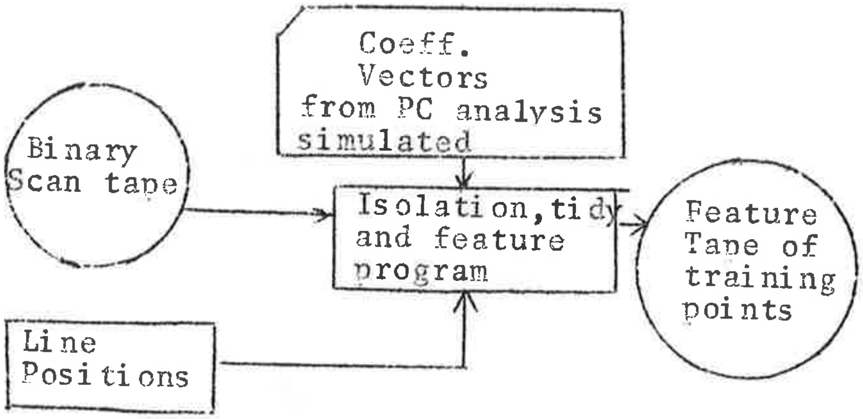


Figure 8.2.1. The experimental system for gathering training points.

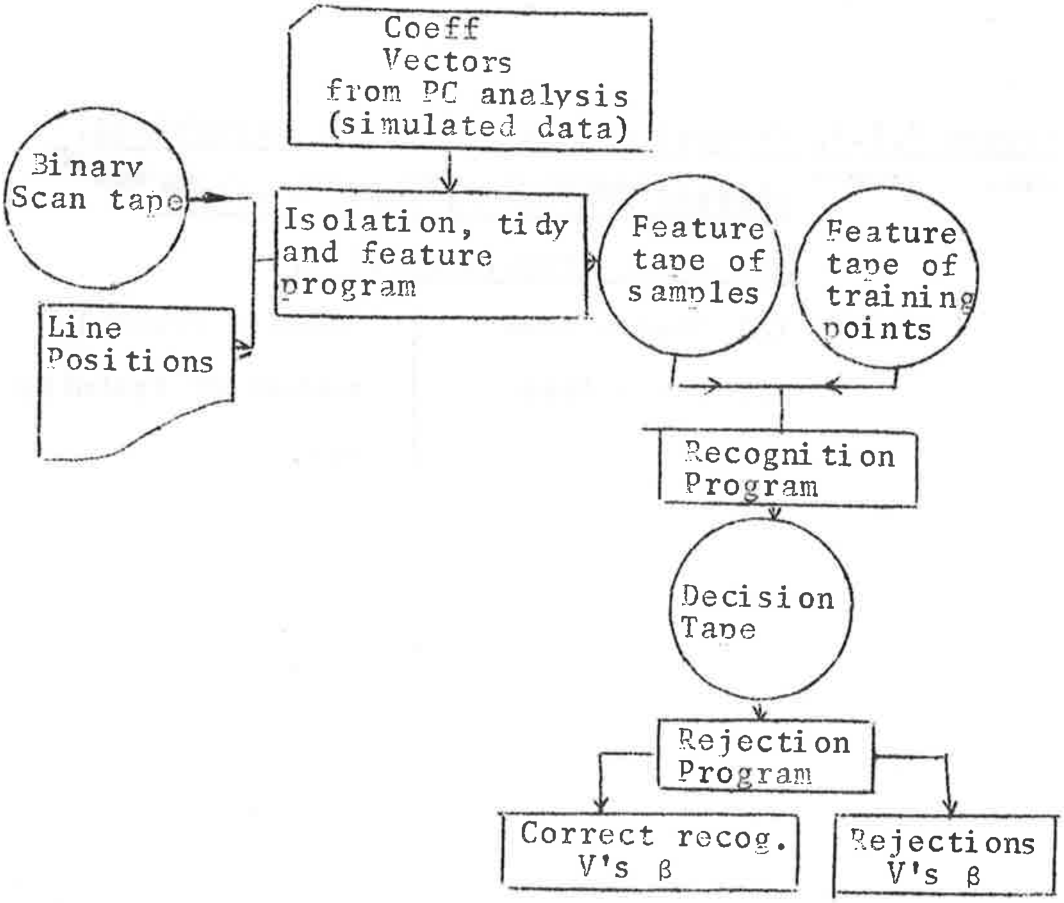


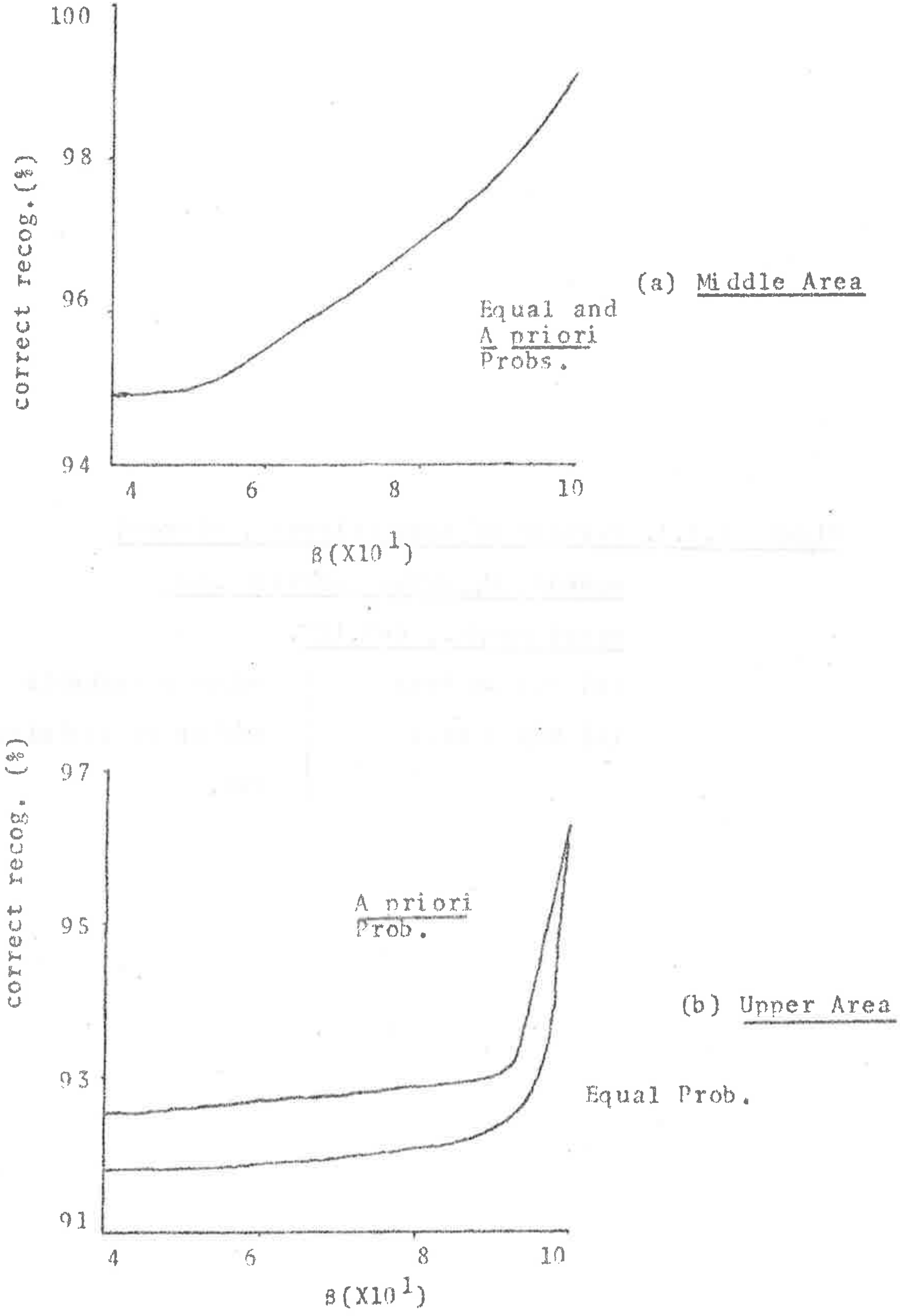
Figure 8.2.2. The completion of the recognition system, which follows from Figure 3.2.5.

Figure 8.2.3. Percent of text correctly recognised, plotted against $\beta$, using a priori and equal probs., c=0.125.

(a) Middle Area

(b) Upper Area

Using a variable number of training pts.

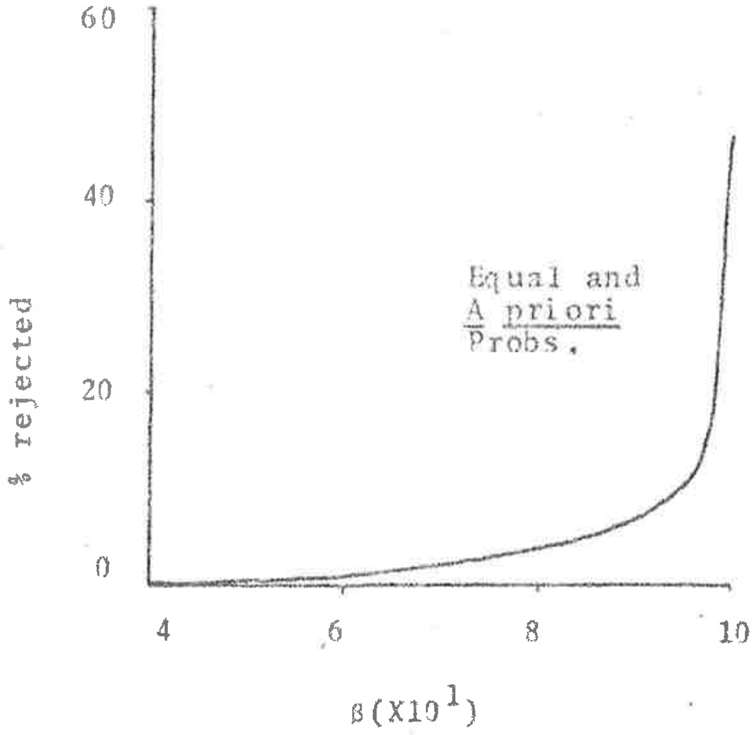


(a) Middle Area

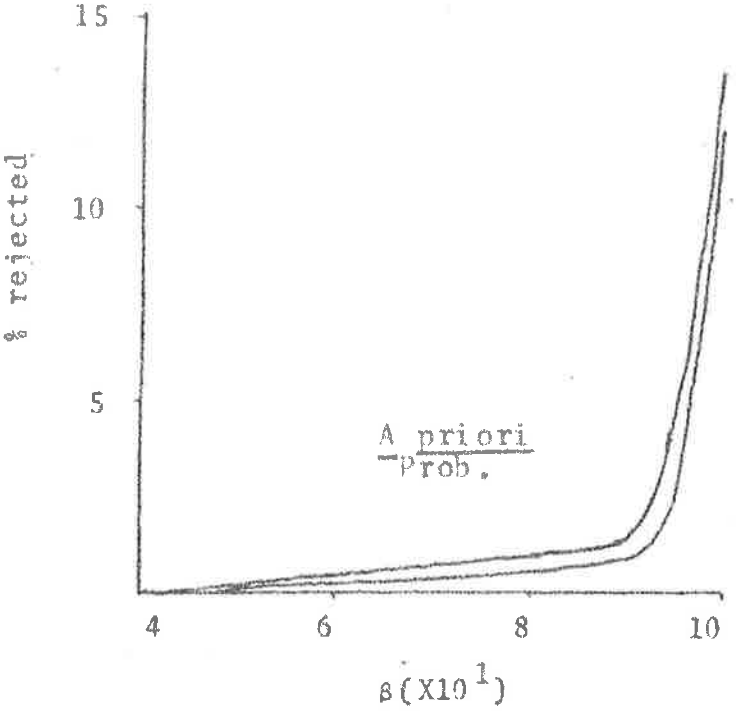
(b) Upper Area

Figure 8.2.4. Percent of text rejected, plotted against β, using a priori and equal probs., c=0.125.

| | |
|---|---|
| (a) Middle Area<br>(b) Upper Area | Using a variable number of training pts. |



(a) Middle Area



(b) Upper Area

Table 8.2.2. The confusion table for forced decision using equal and a priori probabilities using scanned text. i(j) means that i characters were misrecognised as character j.

(a) Middle Area for c=0.125 and using a maximum of 10 training samples for each category.

| Char. | No. of Training Samples | No. for Recog. | No.of Errors | | Distribution of Errors | |
|---|---|---|---|---|---|---|
| | | | Equal Prob. | A Priori Prob. | Equal Prob | A Priori Prob |
| 1 | 10 | 141 | 2 | 2 | 2(3) | 2(3) |
| 2 | 10 | 35 | 0 | 0 | | |
| 3 | 10 | 56 | 0 | 0 | | |
| 4 | 10 | 39 | 0 | 0 | | |
| 5 | 1 | 0 | 0 | 0 | | |
| 6 | 10 | 106 | 0 | 0 | | |
| 7 | 1 | 3 | 3 | 3 | 2(27),<br>1(42) | 1(27)<br>1(33)<br>1(42) |
| 8 | 10 | 63 | 0 | 0 | | |
| 9 | 3 | 9 | 9 | 9 | 9(8) | 9(8) |
| 10 | 0 | 0 | 0 | 0 | | |
| 11 | 2 | 4 | 4 | 4 | 3(27),<br>1(42) | 3(27),<br>1(42) |
| 12 | 0 | 0 | 0 | 0 | | |
| 13 | 3 | 9 | 0 | 0 | | |
| 14 | 9 | 14 | 0 | 0 | | |
| 15 | 0 | 0 | 0 | 0 | | |
| 16 | 4 | 2 | 0 | 0 | | |
| 17 | 4 | 8 | 2 | 2 | 1(27),<br>1(47) | 1(24),<br>1(47) |
| 18 | 10 | 48 | 13 | 12 | 13(19) | 12(19) |
| 19 | 10 | 54 | 32 | 32 | 32(18) | 32(18) |
| 20 | 5 | 8 | 0 | 1 | | 1(42) |
| 21 | 3 | 2 | 0 | 0 | | |
| 22 | 10 | 71 | 10 | 12 | 3(29),<br>7(40) | 3(29),<br>9(40) |
| 23 | 2 | 7 | 7 | 7 | 1(33),<br>6(42) | 1(33),<br>6(42) |

Table 8.2.2. (cont)

| | | | | | | |
|---|---|---|---|---|---|---|
| 24 | 10 | 182 | 11 | 11 | 2(25),<br>9(32) | 2(25),<br>9(32) |
| 25 | 10 | 52 | 0 | 0 | | |
| 26 | 10 | 76 | 0 | 0 | | |
| 27 | 10 | 19 | 0 | 0 | | |
| 28 | 0 | 0 | 0 | 0 | | |
| 29 | 10 | 31 | 3 | 2 | 1(22),<br>2(40) | 2(40) |
| 30 | 0 | 0 | 0 | 0 | | |
| 31 | 10 | 18 | 0 | 0 | | |
| 32 | 10 | 71 | 3 | 4 | 3(24) | 4(24) |
| 33 | 10 | 77 | 0 | 0 | | |
| 34 | 10 | 191 | 5 | 5 | 5(36) | 5(36) |
| 35 | 10 | 56 | 1 | 1 | 1(42) | 1(42) |
| 36 | 10 | 87 | 1 | 1 | 1(1) | 1(1) |
| 37 | 2 | 8 | 6 | 6 | 3(3),<br>2(18),<br>1(19) | 3(3),<br>2(18),<br>1(19) |
| 38 | 10 | 23 | 0 | 0 | | |
| 39 | 10 | 45 | 0 | 0 | | |
| 40 | 10 | 49 | 6 | 5 | 3(22),<br>3(29) | 2(22),<br>3(29) |
| 41 | 0 | 0 | 0 | 0 | | |
| 42 | 10 | 106 | 0 | 0 | | |
| 43 | 0 | 0 | 0 | 0 | | |
| 44 | 10 | 70 | 0 | 0 | | |
| 45 | 10 | 291 | 1 | 1 | 1(1) | 1(1) |
| 46 | 10 | 129 | 0 | 0 | | |
| 47 | 10 | 54 | 0 | 0 | | |
| 48 | 3 | 3 | 0 | 0 | | |
| 49 | 10 | 35 | 0 | 0 | | |
| 50 | 10 | 44 | 0 | 0 | | |
| 51 | 2 | 4 | 2 | 2 | 2(2) | 2(2) |
| 52 | 0 | 0 | 0 | 0 | | |
| 53 | 1 | 0 | 0 | 0 | | |
| 54 | 2 | 0 | 0 | 0 | | |
| 55 | 1 | 0 | 0 | 0 | | |
| 56 | 3 | 1 | 0 | 0 | | |
| 57 | 1 | 0 | 0 | 0 | | |
| 58 | 1 | 0 | 0 | 0 | | |
| 59 | 1 | 0 | 0 | 0 | | |
| 60 | 1 | 0 | 0 | 0 | | |
| 61 | 2 | 0 | 0 | 0 | | |
| 62 | 5 | 1 | 0 | 0 | | |
| 63 | 1 | 0 | 0 | 0 | | |
| 64 | 1 | 0 | 0 | 0 | | |
| 65 | 0 | 0 | 0 | 0 | | |
| 66 | 3 | 0 | 0 | 0 | | |
| TOTAL | | 2402 | 119 | 120 | | |

Table 8.2.1.(Cont)

(b) Upper Area for c=0.125 and using a maximum of 5 training samples for each category.

| Char. | No. of Training Samples | No. for Recog. | No.of Errors | | Distribution of Errors | |
|---|---|---|---|---|---|---|
| | | | Equal Prob. | A Priori Prob. | Equal Prob. | A Priori Prob. |
| 1 | 5 | 49 | | | | |
| 2 | 5 | 71 | 1 | 1 | 1(3) | 1(3) |
| 3 | 5 | 14 | 7 | 7 | 1(2),<br>6(4) | 1(2),<br>5(4) |
| 4 | 5 | 8 | 3 | 3 | 3(3) | 3(3) |
| 5 | 5 | 17 | | | | |
| 6 | 5 | 65 | | | | |
| 7 | 5 | 94 | | | | |
| 8 | 5 | 14 | | | | |
| 9 | 1 | 0 | | | | |
| 10 | 1 | 0 | | | | |
| 11 | 5 | 47 | 3 | 3 | 1(12),<br>2(38) | 1(12),<br>2(38) |
| 12 | 5 | 7 | | | | |
| 13 | 2 | 3 | | | | |
| 14 | 1 | 1 | 1 | 1 | 1(19), | 1(19), |
| 15 | 0 | 0 | | | | |
| 16 | 0 | 0 | | | | |
| 17 | 2 | 2 | | | | |
| 18 | 5 | 24 | 2 | 2 | 1(23),<br>1(28) | 1(23)<br>1(28) |
| 19 | 5 | 8 | | | | |
| 20 | 0 | 0 | | | | |
| 21 | 1 | 0 | | | | |
| 22 | 0 | 0 | | | | |
| 23 | 5 | 13 | 2 | 2 | 2(18) | 2(18) |
| 24 | 1 | 1 | 1 | 1 | 1(19) | 1(19) |
| 25 | 0 | 0 | | | | |
| 26 | 0 | 0 | | | | |
| 27 | 0 | 0 | | | | |
| 28 | 5 | 11 | | | | |
| 29 | 2 | 2 | 1 | 1 | 1(19) | 1(19) |
| 30 | 0 | 0 | | | | |
| 31 | 0 | 0 | | | | |
| 32 | 0 | 0 | | | | |
| 33 | 1 | 1 | 1 | 1 | 1(12) | 1(12) |
| 34 | 1 | 1 | 1 | 1 | 1(42) | 1(42) |
| 35 | 0 | 0 | | | | |
| 36 | 0 | 0 | | | | |
| 37 | 0 | 0 | | | | |

Table 8.2.1. (Cont)

| | | | | | | |
|---|---|---|---|---|---|---|
| 38 | 5 | 7 | | | | |
| 39 | 5 | 13 | | | | |
| 40 | 1 | 0 | | | | |
| 41 | 0 | 0 | | | | |
| 42 | 5 | 0 | | | | |
| TOTAL | | 573 | 47 | 43 | | |

Because of the availability of training data, the results obtained using the simulated text are more meaningful than those obtained for the scanner. However, the results obtained using the scanner at least show that the proposed recognition system is feasible.

From the confusion tables for both simulated and scanner data, it is observed that most of the errors are accounted for by several confusion pairs. It would seem that context may be the only way to resolve these pairs.

Generally, as the value of β is increased, the percent of samples correctly recognised increases. This is because the confusing samples are rejected as being unrecognisable by the recognition system. Notice for the middle area that the percent of characters correctly recognised using equal probabilities just exceeds the corresponding percent using *a priori* probabilities, for all values of β.

This feature is more apparent for the simulated text, the two recognition curves being almost identical for the scanner data. In the case of the upper area, a priori probabilities are more of an advantage. As the value of β is increased, however, a priori probabilities become less important with the two recognition curves approaching a common value. For β equal to 1, nearly 100 percent of the samples are correctly recognised for each area.

From the rejection curves, it is observed that for all values of β the number of samples rejected is greater when using equal rather than a priori probabilities. That is, recognition is attempted for more samples when a priori probabilities are used and in the case of the upper area a better recognition rate is obtained. There is a sharp increase in the number of samples rejected as the value of β is increased beyond 0.9. However, the increase in the number of samples rejected is not warranted by the correspondingly small increase in the recognition rates. Thus the value of β should be chosen to give the best recognition rate while keeping the number of rejections to a reasonable limit. A reasonable choice for the value of β is in the range 0.6 to 0.8.

## Chapter 9. Conclusions

A recognition system for Thai text has been developed. The limited results obtained by using the scanner show that it is possible to automatically read Thai text, however, the cost is prohibitive.

Cost of photographing each page of text to be scanned was negligible compared with the computing cost. A breakdown of the central processor computing time required for the recognition of about 800 characters in one average page of Thai text is given below:-

(a) 12 minutes for processing (Chapter 2),

(b) 2 minutes for conversion and finding each line position (Chapter 3),

(c) 8 minutes for the isolation of each character's binary image (Chapter 4),

(d) 3 minutes for preprocessing (Chapter 4),

(e) 2½minutes for feature vector generation (Chapter 6), and

(f) 2½minutes for the recognition of each feature vector (Chapter 7).

This is a total of 30 minutes and this excludes peripheral processor time.

A comparable amount of peripheral processor time is required by the system, with at least 12 minutes taken up when the scanner is operated.

The cheapest rates that are available for the CDC6400 computerare $A100 and $A20 per hour for central processor and peripheral processor times, respectively. These costs rise to $A400 and $A100, respectively, these being the rates applicable to people outside of this University. Thus the minimum cost for recognition of one average page of Thai text is about $A60. This cost together with the cost of translation makes the proposed automatic translation system a very expensive proposition. It is the author's opinion that it would perhaps be cheaper to employ a linguist as a translator.

From the breakdown of times shown above, it can be seen that 40 percent of the computing time is taken in the processing of the original scan tape. A considerable saving of cost could be made if this processing could be handled as each data point becomes available and thus the 12 minutes central processor time would be eliminated. There is no way apparent to the author by which the source of this cost can be eliminated due to the characteristics of the CDC6400 computer (see Chapter 2).

From an experimental rather than a commercial point of view, this cost of processing the scan information at the completion of a scan is partially offset by the cheapness of the scanning equipment.

It can also be seen from the times above that another 30 percent of central processor time is taken to isolate the characters from the scan' image of a page of Thai text. This considerable time is a direct consequence of the versatility required by the isolation procedure. (see Chapter 4). This versatility indicates the complexity that would be required of an automatic machine for reading Thai text. Since the cost of construction is undoubtedly increased by the versatility, the cost to build such an automatic machine would be forbidding.

To transform the binary pattern of each character to a point in n-dimensional feature space, normalised bivariate moments most advantageous to the recognition system were selected and later used, as described in Chapter 6. From the experimental results described in this Chapter, it can be concluded that the best of the methods tried was based on "variance ordering" for selecting those moments to be used in the feature vectors. This "variance ordering" method for selection is a distinct advantage for the recognition of the upper area characters. At the time of these experiments it was considered more advantageous to gain a maximum recognition rate for characters of the middle area and so the "variance order" method was adopted. However, on review it is the author's opinion that perhaps comparable results with less computing effort could be obtained by using the first 8 "natural order" moments as described in Chapter 6.

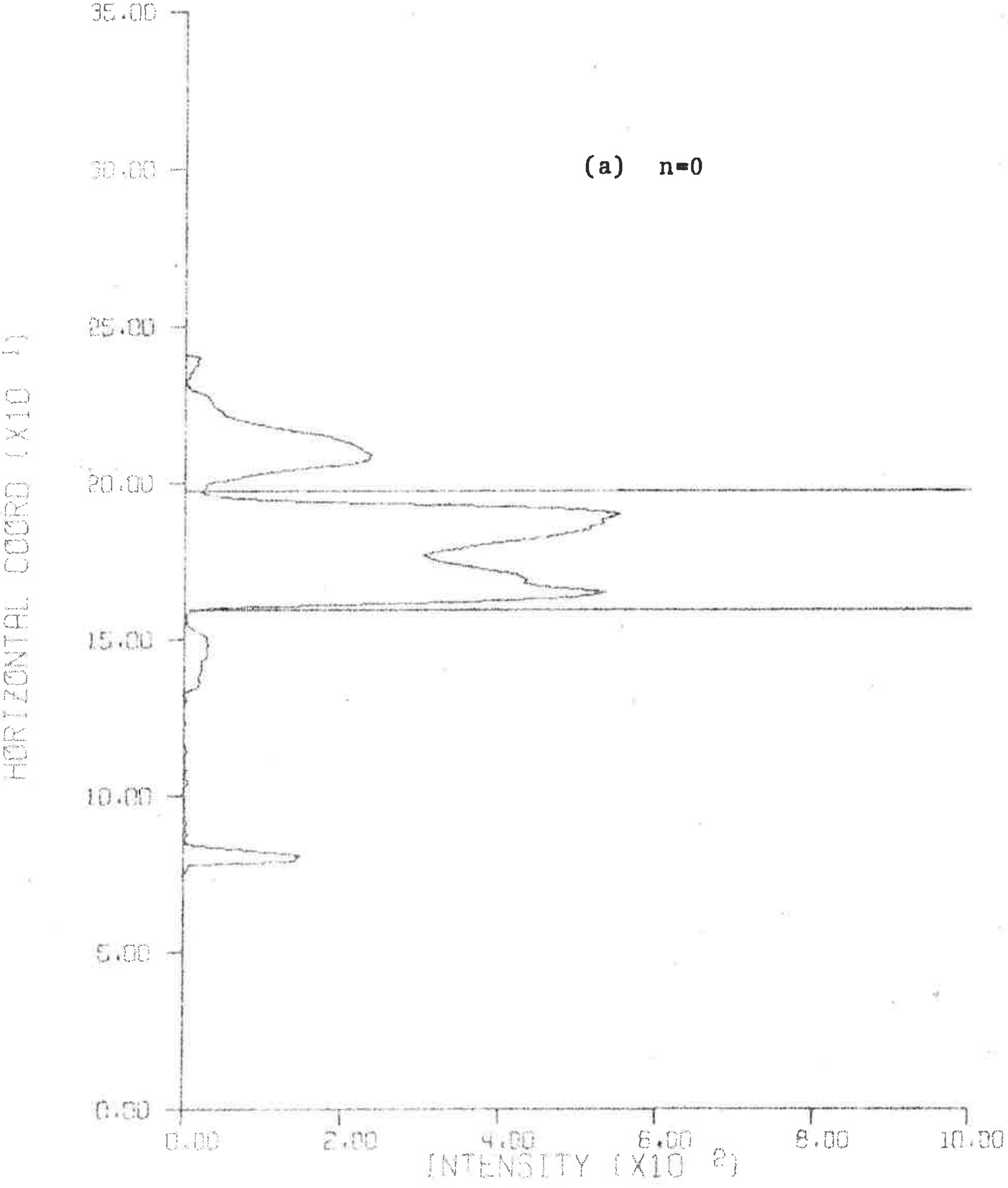
The advantage of reducing the dimension of the feature vectors is apparent in view of the approximating functions used to estimate the conditional probability densities for each category, and the subsequent need for storing all training points. To reduce the dimension principal, components were chosen because of their simplicity and ease of implementation.

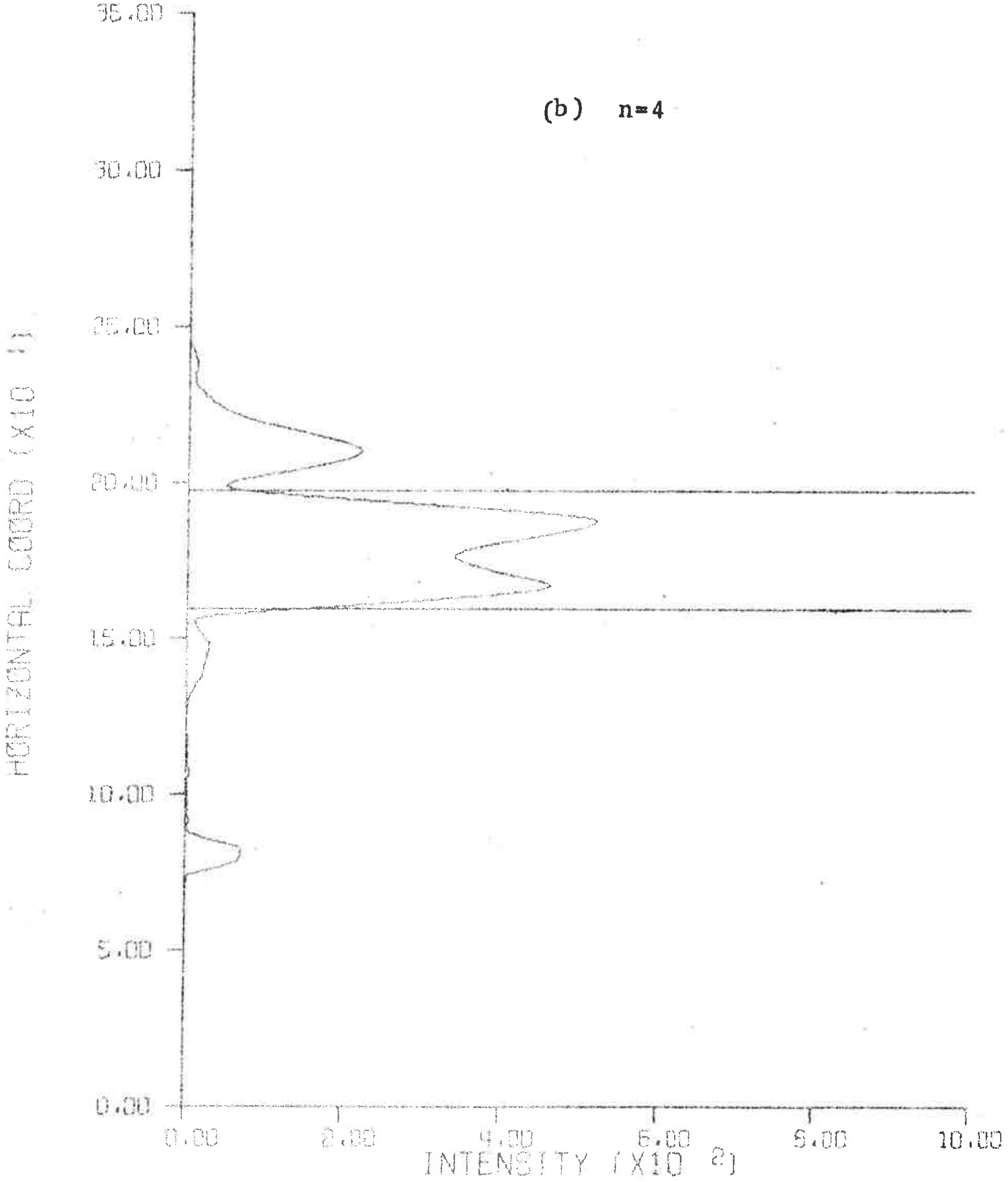
The experimental results described in Chapter 7 show that the best results are obtained for a small range of the smoothing parameter "h", which is in contrast to the results described by Specht [45]. From the results described in Chapter 8, it can be seen that the approximating functions for the conditional probability densities have been applied quite successfully to the problem of recognition of Thai characters.

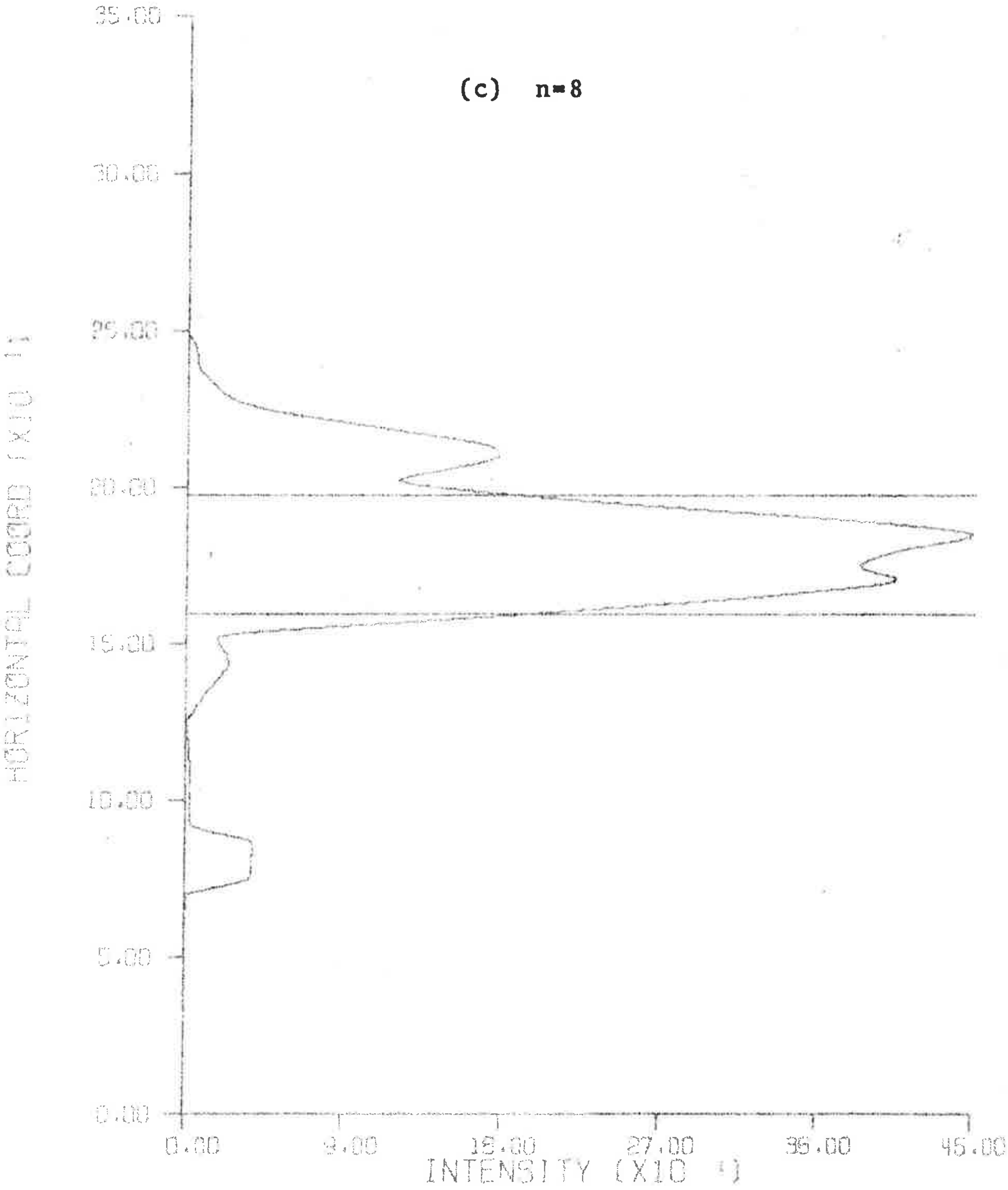
Finally, the results obtained have revealed that a priori probabilities are of little use to the recognition system, when it attempts to distinguish between some of the middle area confusion pairs of characters. These pairs account for most of the errors made by the system. In addition a human, when reading Thai text can only distinguish between these pairs by context in many cases. The author feels that future work on the recognition of Thai text, could be carried out incorporating contextual information in the decision process. That is, balance appropriately the information which is obtain from contextual considerations and the information from the measurements on the character and arrive at a decision using both.

Figure A.1. The smoothing of an intensity plot using the "moving average" technique. The ruled lines indicate the left end bounds of the Middle Area.

## APPENDIX A

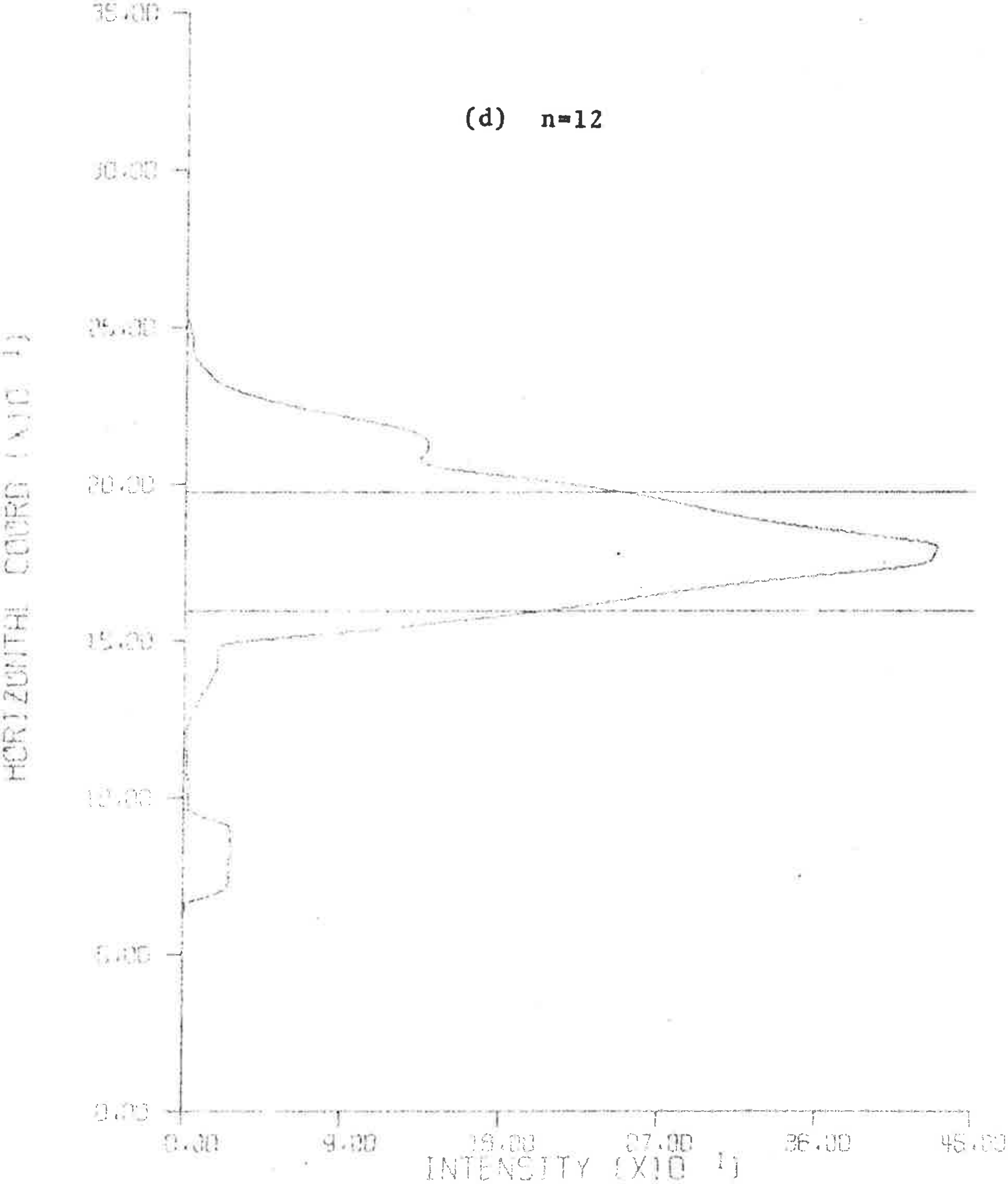

(a) n=0
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
HORIZONTAL COORD (X10 1)
0.00
2.00
4.00
6.00
8.00
10.00
INTENSITY (X10 2)


(b) n=4
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
HORIZONTAL COORD (X10 1)
0.00
2.00
4.00
6.00
8.00
10.00
INTENSITY (X10 2)



(c) n=8

35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
HORIZONTAL COORD (X10 1)
0.00
9.00
18.00
27.00
36.00
45.00
INTENSITY (X10 1)


35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
HORIZONTAL COORD (X10 1)
0.00
9.00
18.00
27.00
36.00
45.00
INTENSITY (X10 1)

(d) n=12

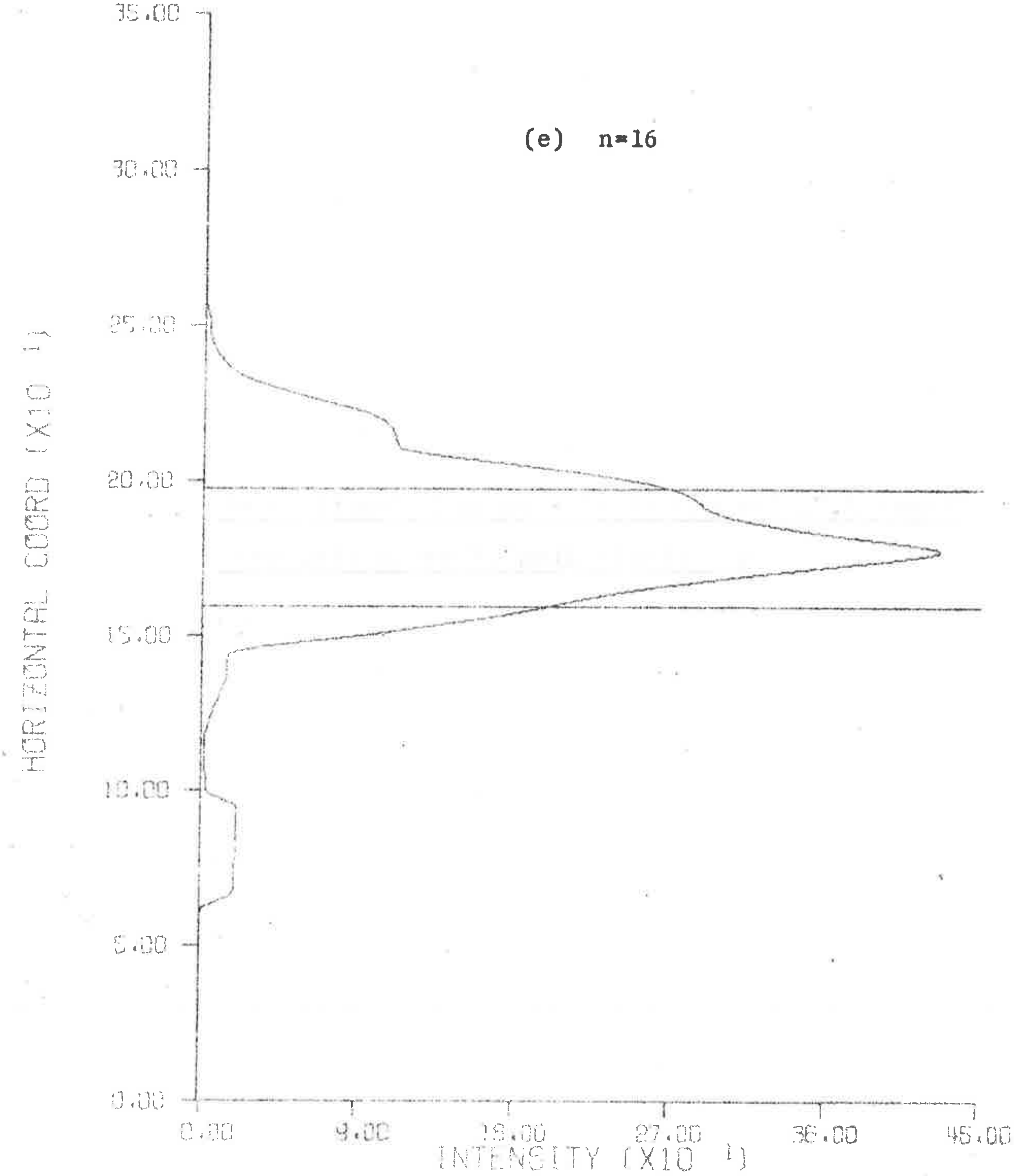
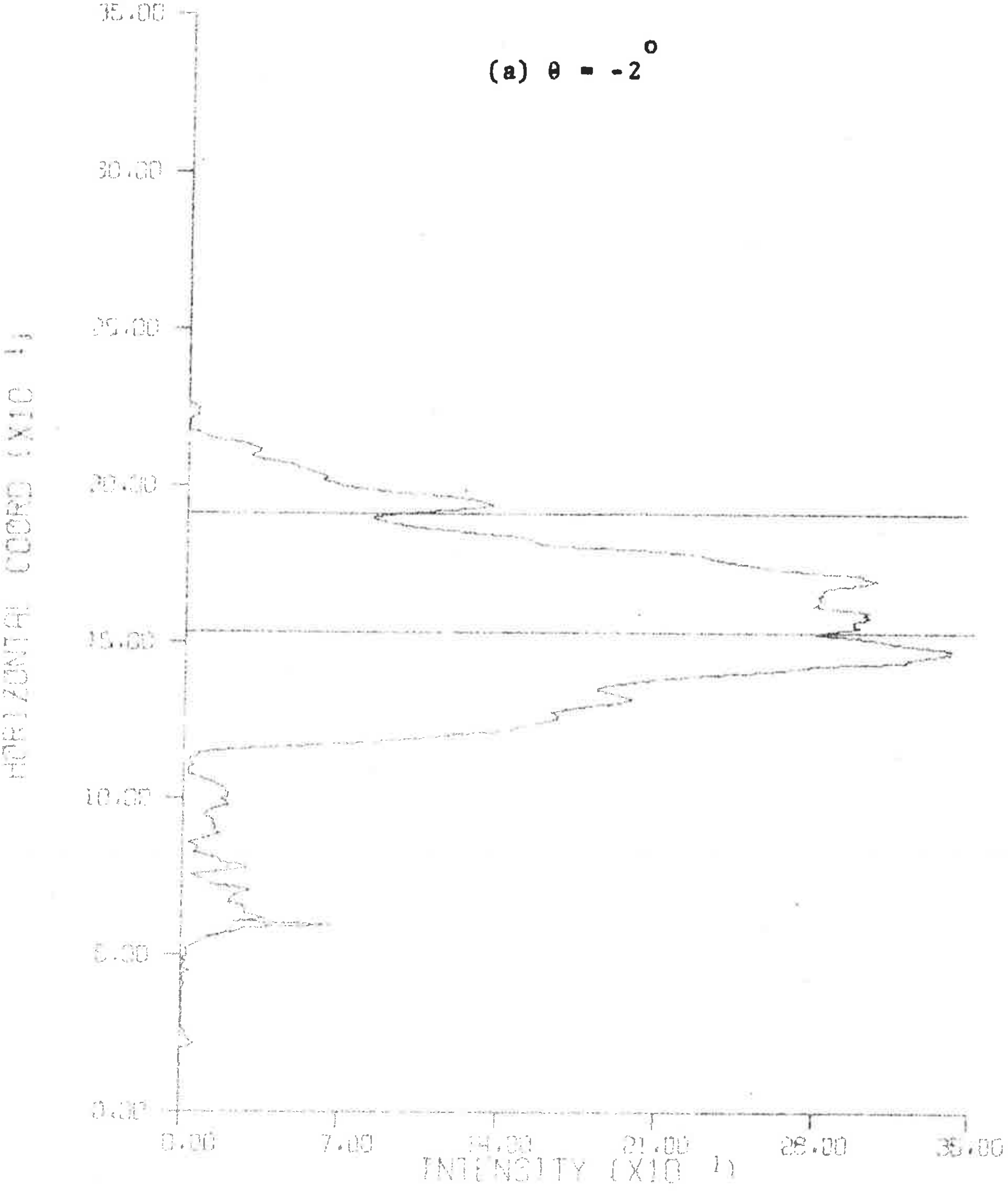
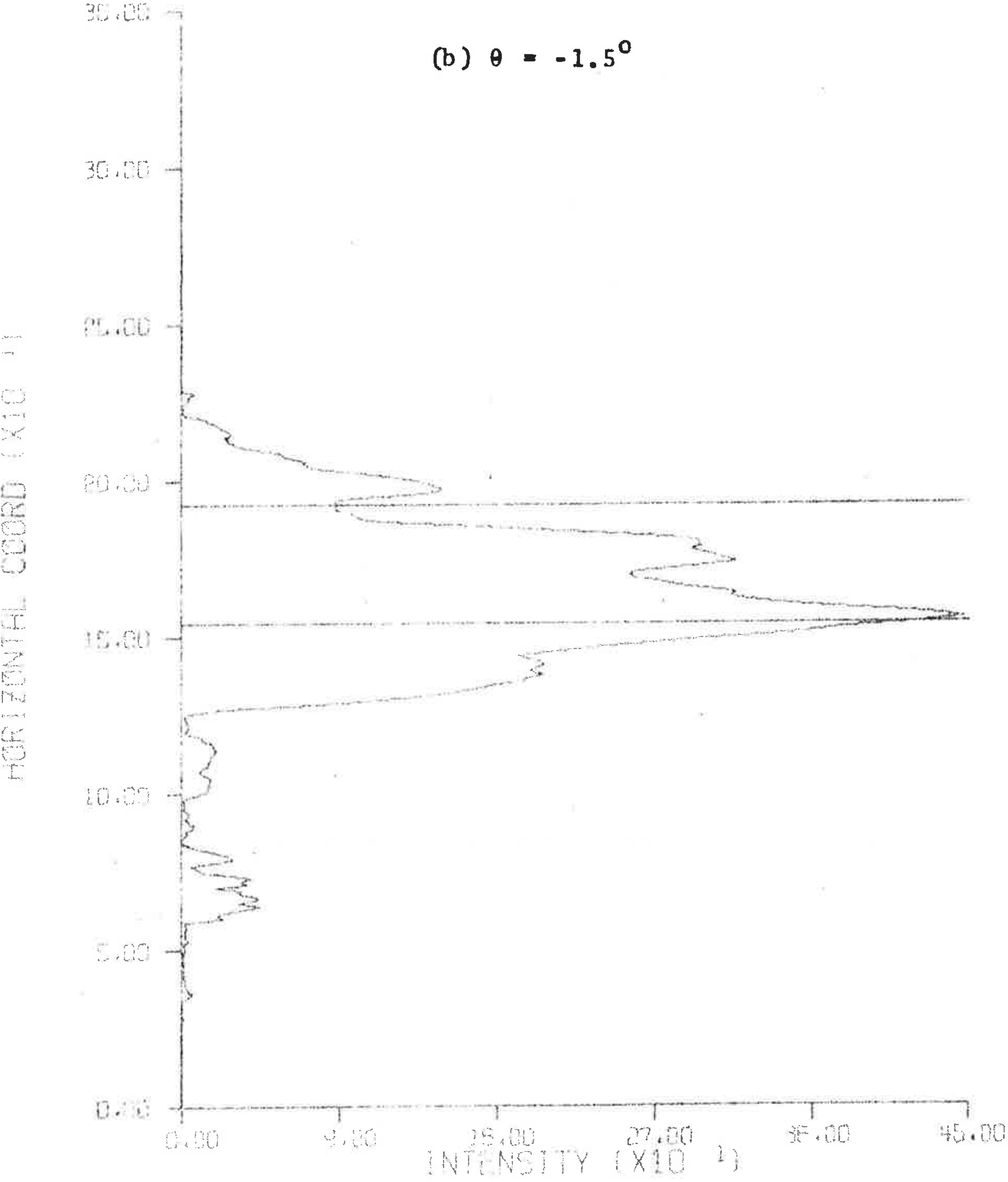

(e) n=16
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
HORIZONTAL COORD (X10 1)
0.00
9.00
18.00
27.00
36.00
45.00
INTENSITY (X10 1)

Figure A.2. The effect of θ on an intensity plot for a single line of print with n=0



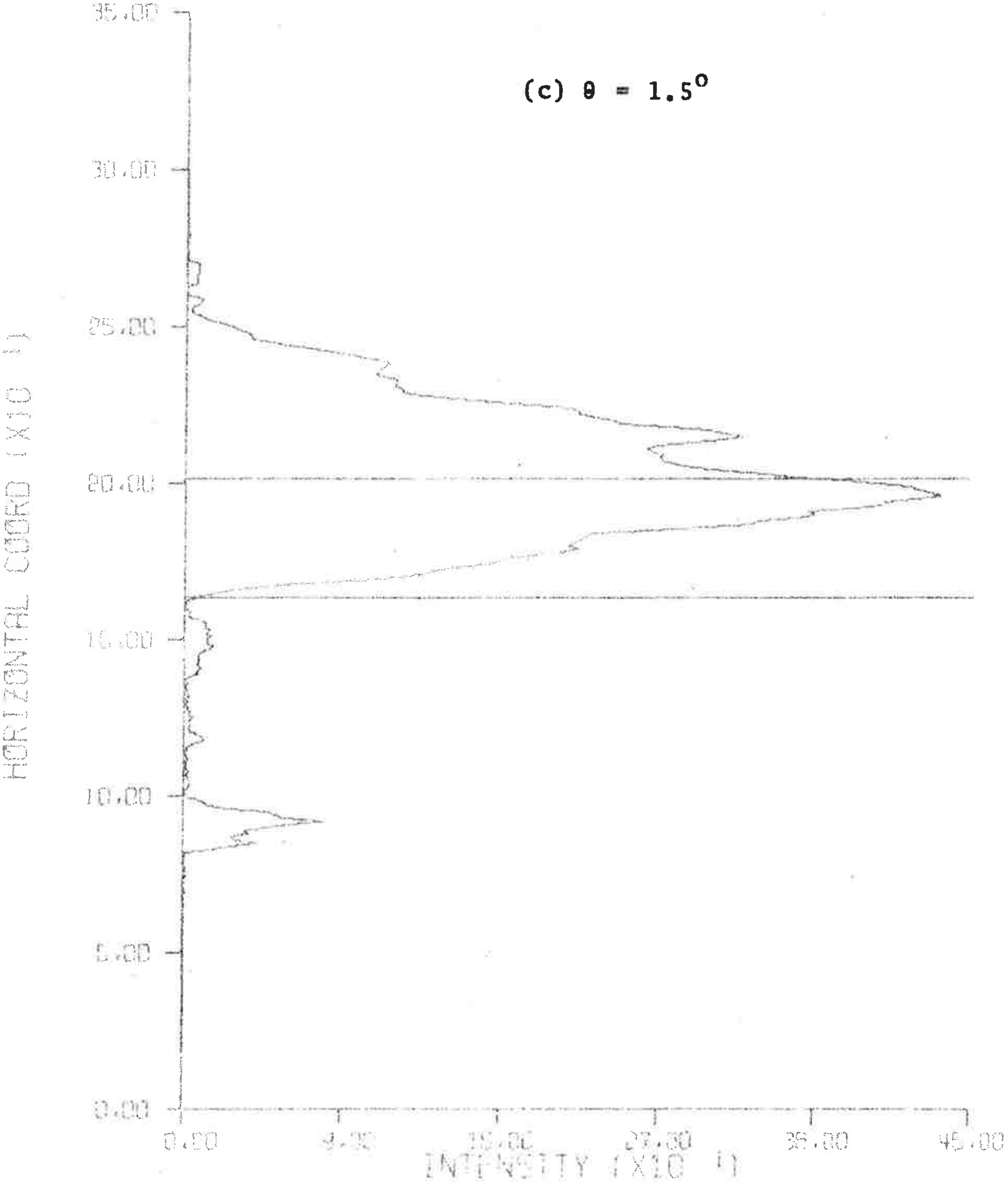
(a) $\theta$ = -2°

35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
HORIZONTAL COORD (X10 1)
0.00
7.00
14.00
21.00
28.00
35.00
INTENSITY (X10 1)

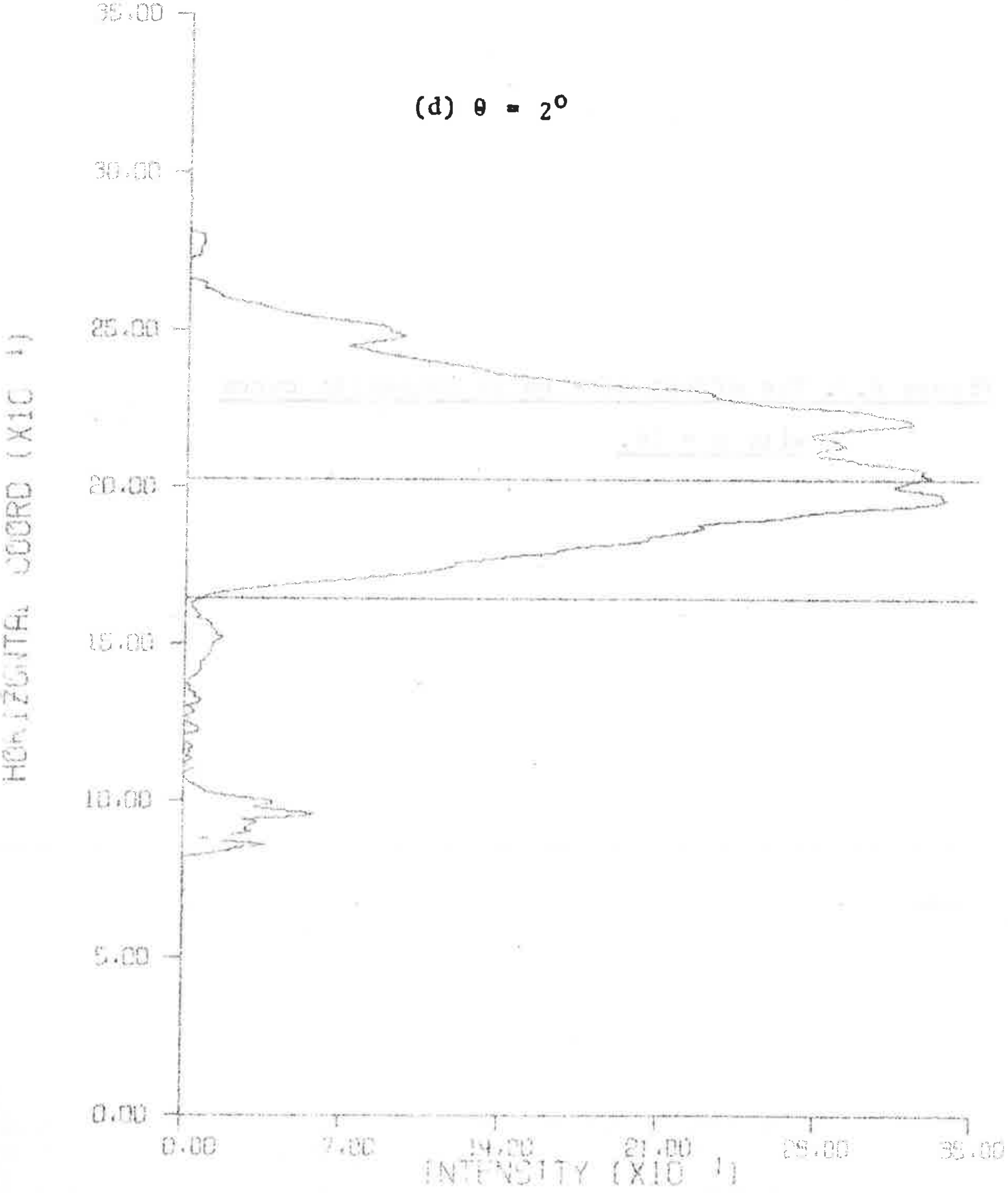


(b) θ = -1.5°

35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
HORIZONTAL COORD (X10 1)
0.00
9.00
18.00
27.00
36.00
45.00
INTENSITY (X10 1)


35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
HORIZONTAL COORD (X10 1)
0.00
9.00
18.00
27.00
36.00
45.00
INTENSITY (X10 1)

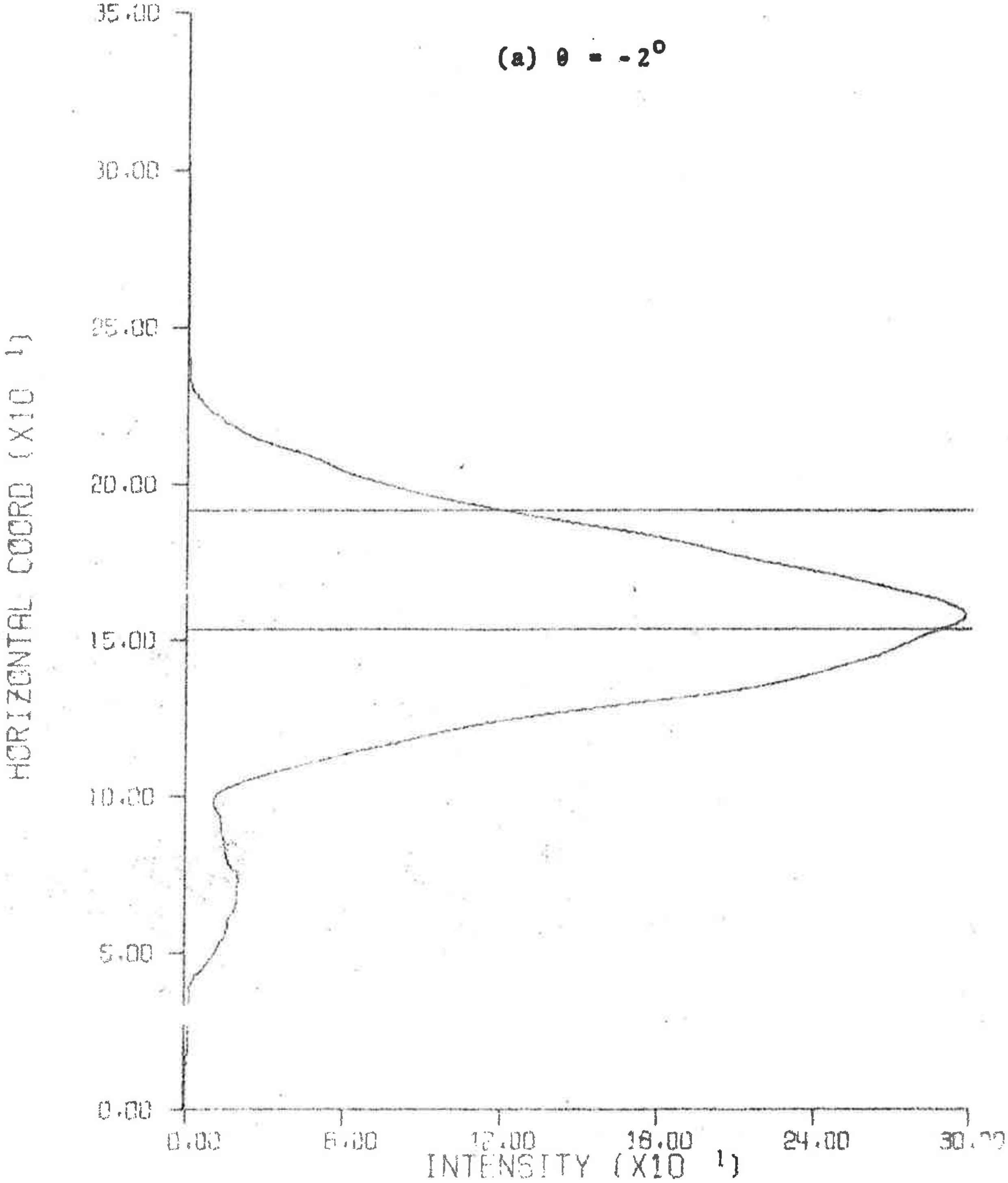
(c) $\theta = 1.5^{o}$

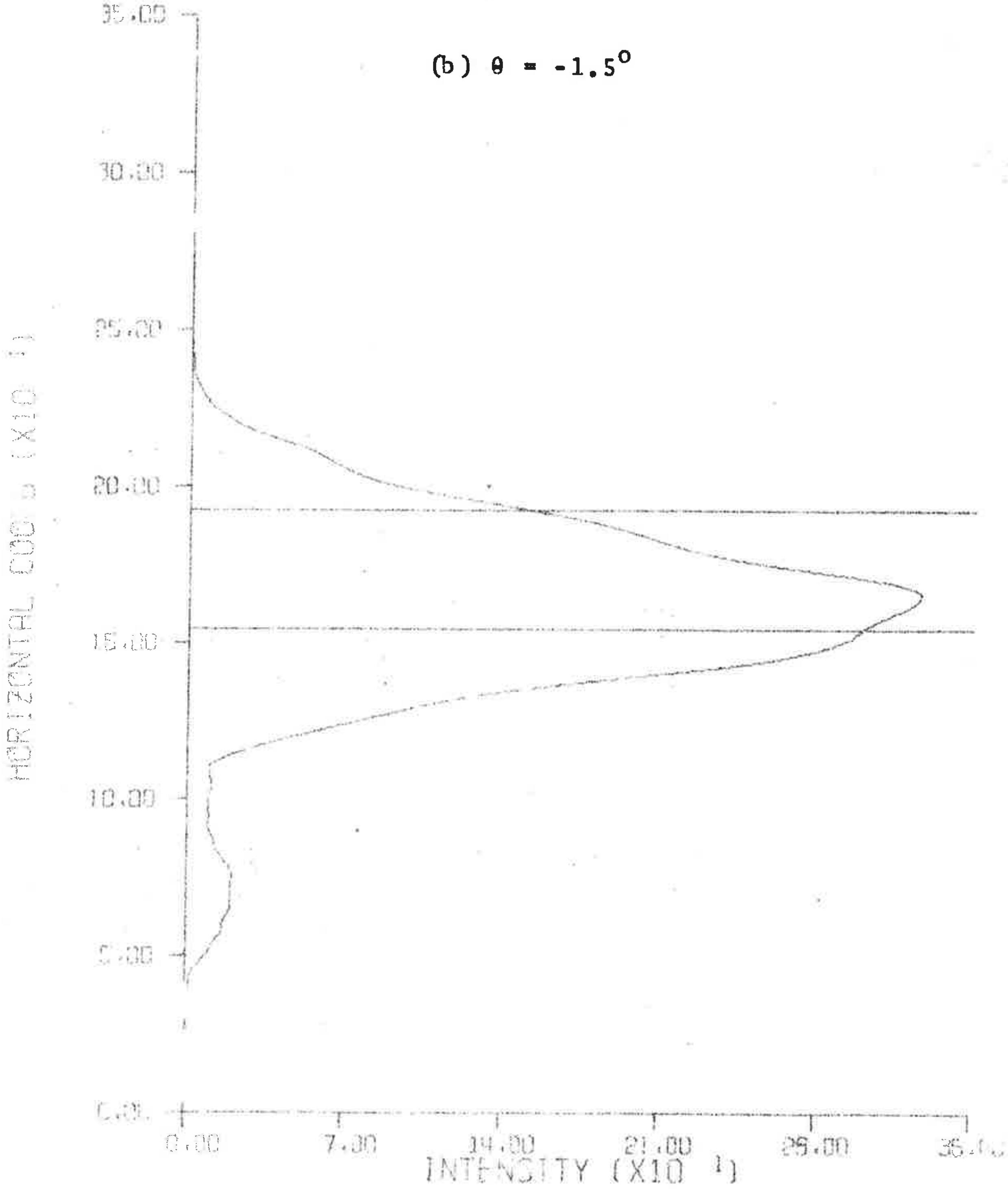


(d) θ = 2°

35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
HORIZONTAL COORD (X10 1)
0.00
7.00
14.00
21.00
28.00
35.00
INTENSITY (X10 1)

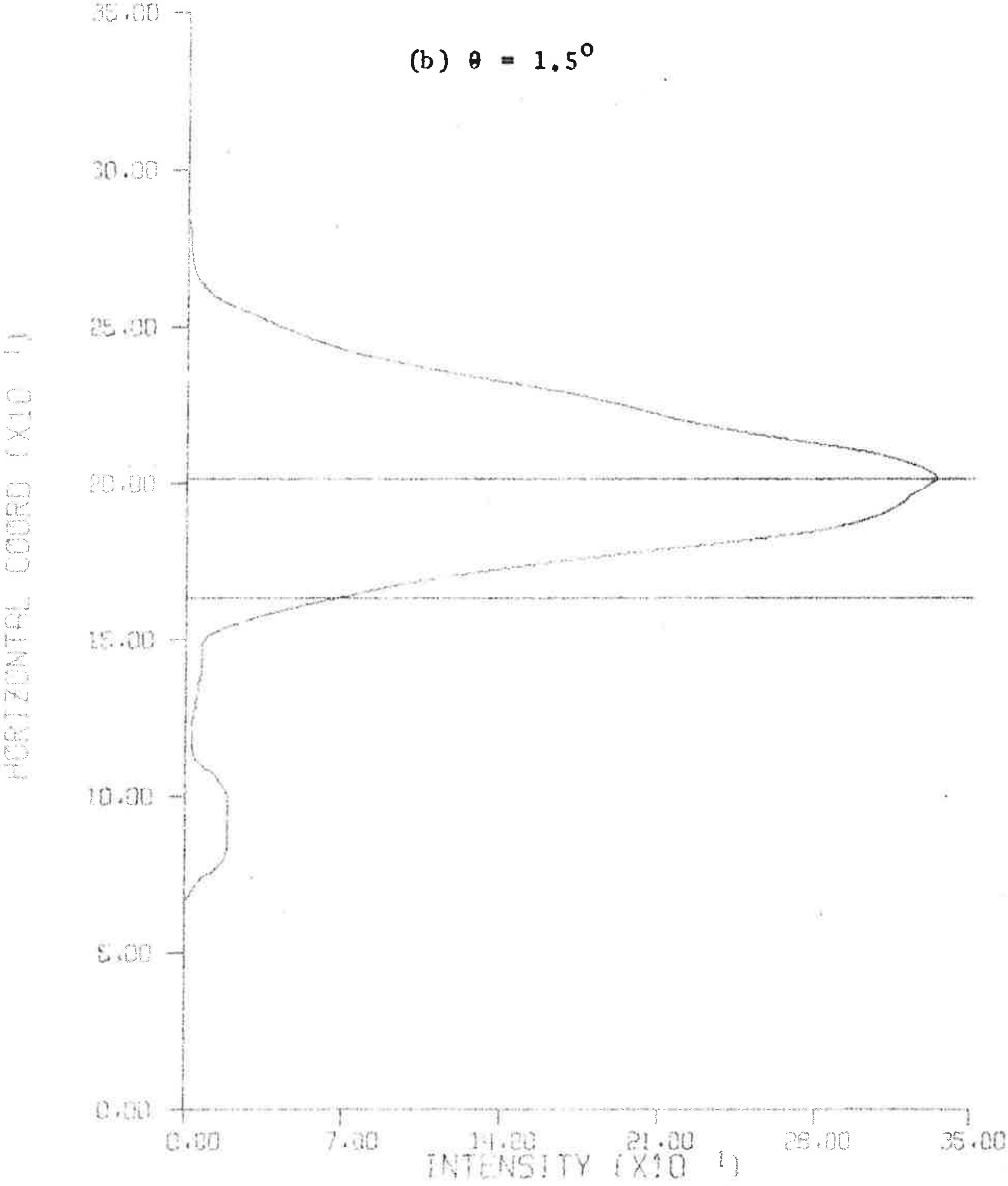
Figure A.3. The effect of $\theta$ on an intensity curve with n = 16.

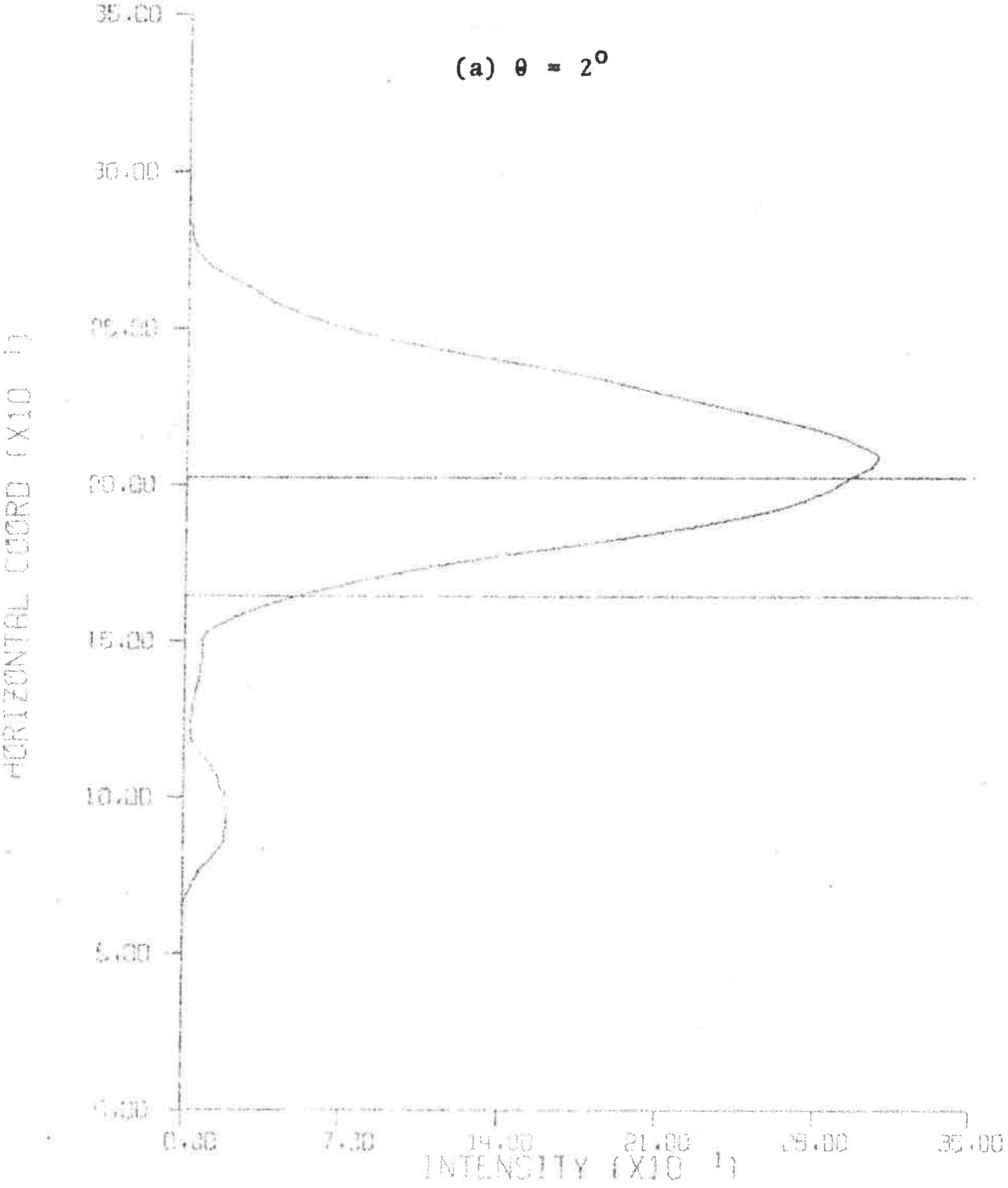

(a) θ = -2°
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
HORIZONTAL COORD (X10 1)
0.00
6.00
12.00
18.00
24.00
30.00
INTENSITY (X10 1)



(b) θ = -1.5°

35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
HORIZONTAL COORD (X10 1)
0.00
7.00
14.00
21.00
28.00
35.00
INTENSITY (X10 1)


(b) θ = 1.5°
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
HORIZONTAL COORD (X10 1)
0.00
7.00
14.00
21.00
28.00
35.00
INTENSITY (X10 1)


(a) θ = 2°
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
HORIZONTAL COORD (X10 1)
0.00
7.00
14.00
21.00
28.00
35.00
INTENSITY (X10 1)

# APPENDIX B

In the following pages the moments of all the symbols of the Thai Alphabet are presented from orders 3 to 7 inclusive in the following format.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 30 | 21 | 12 | 03 | | | | |
| 40 | 31 | 22 | 13 | 04 | | | |
| 50 | 41 | 32 | 23 | 14 | 05 | | |
| 60 | 51 | 42 | 33 | 24 | 15 | 06 | |
| 70 | 61 | 52 | 43 | 34 | 25 | 16 | 07 |

where the numbering is representative of the moment, e.g. 30 $m_{30}$

## TABLE B.1. MOMENTS FOR MIDDLE LINE CHARACTERS

### CHARACTER 1

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .023 | -.269 | .006 | -.223 | | | | |
| 1.238 | .002 | .897 | .015 | 1.728 | | | |
| .038 | -.400 | .002 | -.691 | .024 | -.910 | | |
| 1.658 | -.005 | 1.090 | .005 | 1.584 | .043 | 3.648 | |
| .040 | -.544 | -.006 | -.968 | .002 | -1.820 | .056 | -3.004 |

### CHARACTER 2

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .280 | .290 | -.063 | -.470 | | | | |
| 1.901 | .176 | .801 | .134 | 1.933 | | | |
| 1.011 | 1.052 | .209 | .112 | -.214 | -1.871 | | |
| 4.553 | .875 | 1.453 | .398 | 1.179 | .463 | 4.849 | |
| 3.354 | 3,265 | 1.084 | .978 | .211 | -.260 | -.743 | -6.477 |

### CHARACTER 3

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .086 | -.219 | .063 | -.096 | | | | |
| 1.446 | -.120 | .967 | -.039 | 2.108 | | | |
| .201 | -.369 | .251 | -.666 | .173 | -.558 | | |
| 2.341 | -.330 | 1.316 | -.288 | 1.979 | -.147 | 5.369 | |
| .367 | -.604 | .581 | -1.120 | .608 | -2,090 | .486 | -2.477 |

### CHARACTER 4

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.544 | .375 | .154 | .054 | | | | |
| 2.329 | -.055 | .723 | -.159 | 2.087 | | | |
| -2.697 | 1.061 | 0.186 | .657 | .294 | -.194 | | |
| 7.523 | -.808 | 1.601 | -.294 | 1.155 | -.494 | 5.231 | |
| -12.113 | 3.713 | -1.441 | 1.732 | -.264 | 1.105 | .759 | -1.730 |

### CHARACTER 5

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .430 | .278 | -.084 | -.036 | | | | |
| 1.938 | .307 | .947 | .044 | 1.477 | | | |
| 1.605 | 1.112 | .299 | ..293 | -.110 | -.093 | | |
| 4.751 | 1.415 | 1.868 | .549 | 1.372 | .115 | 2.517 | |
| 5.393 | 3.730 | 1.622 | 1.411 | .502 | .382 | -.159 | -.236 |

TABLE B.1. (CONTINUED)

CHARACTER 6

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -1.403 | -.060 | .067 | -.043 | | | | |
| 4.879 | .402 | .519 | -.190 | 1.611 | | | |
| -13.833 | -1.099 | -.205 | .113 | .213 | -.321 | | |
| 44.922 | 3.721 | .876 | -.107 | .655 | -.482 | 3.131 | |
| *45.849 | -11.383 | -1.370 | .113 | -.061 | .130 | .612 | -1.274 |

CHARACTER 7

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.114 | .015 | .170 | .263 | | | | |
| 1.488 | -.099 | .726 | .090 | 1.837 | | | |
| -.527 | .116 | .238 | .004 | .281 | 1.075 | | |
| 2.777 | -.370 | .882 | .035 | 1.116 | .214 | 4.011 | |
| -1.853 | .363 | .350 | -.030 | .403 | .152 | .509 | 3.407 |

CHARACTER 8

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .017 | .364 | .178 | -.328 | | | | |
| 1.711 | .045 | .890 | .239 | 1.959 | | | |
| -.226 | 1.013 | .434 | .446 | .545 | -1.169 | | |
| 3.607 | .139 | 1.499 | .623 | 1.634 | .804 | 4.802 | |
| -1.199 | 2.544 | .943 | 1.472 | 1.333 | .963 | 1.578 | -3.637 |

CHARACTER 9

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .039 | .436 | .236 | -.467 | | | | |
| 1.827 | .134 | .860 | .267 | 2.161 | | | |
| -.031 | 1.198 | .604 | .569 | .679 | -1.862 | | |
| 4.047 | .551 | 1.658 | .846 | 1.653 | .884 | 6.089 | |
| -.358 | 3.139 | 1.556 | 1.949 | 1.772 | 1.272 | 1.923 | -6.690 |

CHARACTER 10

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .062 | -.078 | .035 | .195 | | | | |
| 1.376 | .021 | .780 | .002 | 1.671 | | | |
| .043 | -.226 | .134 | -.029 | .028 | .711 | | |
| 2.176 | .058 | .953 | -.009 | 1.156 | .041 | 3.235 | |
| -.198 | -.507 | .245 | -.196 | .167 | .059 | .030 | 2.094 |

TABLE B.1. (CONTINUED)

CHARACTER 11

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .043 | -.055 | .040 | .215 | | | | |
| 1.390 | .042 | .765 | .003 | 1.658 | | | |
| -.011 | -.186 | .119 | .024 | .043 | .799 | | |
| 2.228 | .106 | .929 | .030 | 1.115 | .034 | 3.197 | |
| -.338 | -.443 | .208 | -.113 | .141 | .192 | .075 | 2.367 |

CHARACTER 12

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .145 | .064 | -.020 | .315 | | | | |
| 1.673 | .204 | .941 | -.014 | 1.681 | | | |
| .485 | .231 | .225 | .331 | -.045 | 1.141 | | |
| 3.027 | .620 | 1.548 | .391 | 1.543 | -.030 | 3.473 | |
| 1.285 | .653 | .719 | .713 | .552 | 1.037 | -.080 | 3.435 |

CHARACTER 13

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.113 | .219 | .147 | -.020 | | | | |
| 1.861 | .268 | .720 | -.092 | 1.826 | | | |
| -.318 | .512 | .274 | .420 | .372 | -.208 | | |
| 3.849 | .880 | 1.258 | .346 | 1.094 | -.255 | 3.967 | |
| -.855 | 1.146 | .704 | 1.014 | .718 | .786 | .911 | -.898 |

CHARACTER 14

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.506 | .056 | -.157 | .059 | | | | |
| 2.011 | -.011 | 1.154 | .089 | 1.594 | | | |
| -2.201 | .182 | -.772 | .138 | -.235 | .335 | | |
| 5.622 | -.100 | 2.588 | .182 | 1.915 | .279 | 2.900 | |
| -8.510 | .739 | -2.969 | .351 | -1.074 | .489 | -.223 | 1.106 |

CHARACTER 15

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .117 | .199 | .036 | -.620 | | | | |
| 1.657 | .097 | .930 | .137 | 2.200 | | | |
| .269 | .726 | .187 | .254 | -.001 | -2.769 | | |
| 3.318 | .342 | 1.304 | .419 | 1.865 | .457 | 6.663 | |
| .542 | 1.931 | .570 | .354 | .131 | -1.760 | -.338 | -11.041 |

## TABLE B.1. (CONTINUED)

### CHARACTER 16

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .321 | .022 | .095 | .127 | | | | |
| 1.866 | .225 | .956 | .022 | 1.787 | | | |
| 1.164 | .232 | .496 | .129 | .211 | .561 | | |
| 3.986 | .819 | 1.824 | .488 | 1.676 | .090 | 3.781 | |
| 3.428 | .990 | 1.564 | .588 | 1.063 | .507 | .504 | 1.883 |

### CHARACTER 17

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.001 | -.008 | .004 | .343 | | | | |
| 1.600 | .203 | .931 | .013 | 1.702 | | | |
| -.025 | .021 | .160 | .269 | .031 | 1.262 | | |
| 2.761 | .562 | 1.463 | .414 | 1.515 | .062 | 3.610 | |
| -.094 | .065 | .401 | .472 | .528 | 1.012 | .155 | 3.867 |

### CHARACTER 18

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .140 | -.120 | .066 | .055 | | | | |
| 1.605 | -.164 | .910 | -.015 | 1.961 | | | |
| .316 | -.299 | .364 | -.311 | .173 | .165 | | |
| 2.907 | -.456 | 1.322 | -.365 | 1.657 | -.082 | 4.530 | |
| .571 | -.685 | .879 | -.723 | .811 | -.868 | .473 | .339 |

### CHARACTER 19

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .148 | -.123 | .059 | .038 | | | | |
| 1.608 | -.157 | .919 | -.014 | 1.958 | | | |
| .346 | -.302 | .354 | -.314 | .162 | .100 | | |
| 2.919 | -.435 | 1.324 | -.355 | 1.704 | -.073 | 4.528 | |
| .658 | -.676 | .859 | -.733 | .794 | -.860 | .464 | .128 |

### CHARACTER 20

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .227 | -.039 | -.025 | .091 | | | | |
| 1.362 | -.135 | .759 | -.004 | 1.594 | | | |
| .559 | -.108 | .173 | -.084 | -.038 | .349 | | |
| 2.052 | -.331 | .982 | -.204 | 1.014 | -.030 | 2.957 | |
| 1.137 | -.237 | .430 | -.219 | .256 | -.176 | -.073 | 1.054 |

TABLE B.1. (CONTINUED)

CHARACTER 21

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.009 | -.021 | .190 | .013 | | | | |
| 1.281 | .127 | .777 | -.203 | 1.796 | | | |
| -.062 | .028 | .176 | -.101 | .644 | -.208 | | |
| 1.834 | .349 | .880 | -.038 | 1.192 | -.798 | 4.096 | |
| -.172 | .097 | .170 | -.008 | .538 | 0.479 | 1.961 | -1.442 |

CHARACTER 22

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .183 | .050 | -.021 | -.418 | | | | |
| 1.391 | .047 | 1.082 | .052 | 1.977 | | | |
| .442 | .288 | .008 | -.371 | -.059 | -1.847 | | |
| 2.253 | .145 | 1.423 | .195 | 2.135 | .207 | 5.166 | |
| .912 | .835 | .074 | -.168 | -.146 | -1.831 | -.242 | -6.925 |

CHARACTER 23

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .056 | .025 | -.053 | -.099 | | | | |
| 1.484 | -.014 | .856 | .046 | 1.675 | | | |
| .070 | .235 | -.129 | .002 | -.060 | -.369 | | |
| 2.687 | -.125 | 1.282 | .033 | 1.282 | .185 | 3.265 | |
| -.123 | .891 | -.358 | .288 | -.173 | -.082 | -.044 | -1.103 |

CHARACTER 24

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .190 | .189 | -.060 | .188 | | | | |
| 1.390 | .311 | 1.046 | -.002 | 1.569 | | | |
| .557 | .475 | .184 | .459 | -.110 | .657 | | |
| 2.262 | .853 | 1.617 | .533 | 1.732 | .013 | 2.920 | |
| 1.362 | 1.141 | .724 | 1.011 | .396 | 1.136 | -.173 | 1.873 |

CHARACTER 25

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .277 | .240 | 0.081 | -.031 | | | | |
| 1.349 | .232 | .923 | .004 | 1.574 | | | |
| .728 | .586 | .166 | .396 | -.140 | -.062 | | |
| 2.130 | .650 | 1.314 | .351 | 1.389 | .031 | 2.809 | |
| 1.622 | 1.319 | .637 | .938 | .271 | .743 | -.236 | -.059 |

TABLE B.1. (CONTINUED)

CHARACTER 26

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.187 | .210 | .247 | .057 | | | | |
| 1.320 | -.245 | .806 | .305 | 1.901 | | | |
| -.672 | .456 | .085 | .322 | .851 | .597 | | |
| 2.188 | -.729 | .957 | .054 | 1.314 | 1.193 | 4.624 | |
| -1.944 | 1.075 | -.252 | .469 | .608 | .959 | 2.691 | 3.108 |

CHARACTER 27

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .145 | .031 | -.008 | .045 | | | | |
| 1.313 | .111 | .852 | -.004 | 1.542 | | | |
| .340 | .099 | .169 | .086 | .007 | .172 | | |
| 1.876 | .269 | 1.068 | .169 | 1.274 | -.006 | 2.734 | |
| .673 | .219 | .383 | .217 | .325 | .224 | .053 | .502 |

CHARACTER 28

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.118 | -.053 | .231 | .178 | | | | |
| 1.355 | -.240 | .774 | .248 | 1.884 | | | |
| -.340 | -.121 | .177 | -.035 | .773 | 1.136 | | |
| 2.117 | -.600 | .923 | -.154 | 1.224 | 1.048 | 4.791 | |
| -.792 | -.156 | .168 | -.250 | .588 | .272 | 2.510 | 5.169 |

CHARACTER 29

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .026 | .024 | -.107 | -.344 | | | | |
| 1.443 | .115 | 1.152 | .064 | 2.006 | | | |
| .077 | .303 | -.230 | -.338 | -.267 | -1.611 | | |
| 2.507 | .316 | 1.779 | .418 | 2.418 | .328 | 5.298 | |
| .152 | 1.035 | -.378 | -.083 | -.692 | -1.784 | -.789 | -6.419 |

CHARACTER 30

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.177 | .075 | .264 | -.006 | | | | |
| 1.518 | -.301 | 1.087 | .376 | 2.364 | | | |
| -.856 | .402 | .222 | .012 | 1.149 | .467 | | |
| 3.090 | -1.054 | 1.520 | .013 | 2.506 | 1.801 | 7.598 | |
| -3.207 | 1.450 | -.124 | .157 | 1.346 | .315 | 4.658 | 3.809 |

TABLE B.1. (CONTINUED)

CHARACTER 31

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .306 | -.296 | -.076 | .062 | | | | |
| 1.387 | -.227 | .894 | -.007 | 1.606 | | | |
| .821 | -.710 | .211 | -.470 | -.133 | .197 | | |
| 2.289 | -.685 | 1.320 | -.347 | 1.329 | -.045 | 2.956 | |
| 1.880 | -1.600 | .768 | -1.108 | .341 | -.845 | -.227 | .497 |

CHARACTER 32

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .276 | .180 | -.133 | .174 | | | | |
| 1.423 | .170 | 1.076 | 0.052 | 1.559 | | | |
| .642 | .474 | .050 | .439 | -.268 | .595 | | |
| 2.348 | .481 | 1.615 | .217 | 1.788 | -.149 | 2.856 | |
| 1.301 | 1.103 | .321 | .970 | .026 | 1.071 | -.531 | 1.661 |

CHARACTER 33

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .239 | .067 | -.021 | 0.240 | | | | |
| 1.499 | .112 | .812 | .069 | 1.648 | | | |
| .651 | .307 | .154 | -.042 | 0.039 | 0.830 | | |
| 2.558 | .334 | 1.163 | .268 | 1.148 | .206 | 3.201 | |
| 1.493 | .856 | .495 | .249 | .210 | -.261 | -.112 | -2.327 |

CHARACTER 34

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.432 | .306 | 0.328 | -.172 | | | | |
| 2.243 | -.192 | 1.004 | .151 | 1.450 | | | |
| -2.093 | 1.369 | -.804 | .176 | -.601 | -.556 | | |
| 7.450 | -1.432 | 2.410 | .027 | 1.378 | .485 | 2.420 | |
| -9.828 | 6.125 | -2.610 | 1.243 | -1.162 | -.098 | -1.128 | -1.450 |

CHARACTER 35

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .014 | -.108 | -.108 | .363 | | | | |
| 1.470 | -.123 | .854 | -.156 | 2.070 | | | |
| -.064 | -.147 | -.081 | .035 | -.437 | 1.634 | | |
| 2.466 | -.390 | 1.246 | -.411 | 1.496 | -.724 | 5.316 | |
| -.368 | -.103 | -.213 | .041 | -.416 | .511 | -1.507 | 5.953 |

TABLE B.1. (CONTINUED)

CHARACTER 36

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.840 | .420 | -.372 | .156 | | | | |
| 3.021 | -.786 | 1.119 | -.151 | 1.471 | | | |
| -6.222 | 2.665 | -1.490 | .707 | -.752 | .446 | | |
| 17.954 | -6.646 | 3.875 | -1.340 | 1.654 | -.654 | -.420 | 2.456 |
| -48.043 | 19.161 | -8.559 | 3.828 | -2.386 | 1.244 | -1.450 | 1.036 |

CHARACTER 37

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .139 | -.280 | .208 | .033 | | | | |
| 1.544 | -.303 | .904 | .042 | 2.209 | | | |
| .388 | -.590 | .478 | -.670 | .681 | .166 | | |
| 2.735 | -.812 | 1.356 | -.644 | 1.907 | .342 | 6.179 | |
| .871 | -1.237 | 1.130 | -1.430 | 1.266 | -1.782 | 2.308 | 1.098 |

CHARACTER 38

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .147 | .295 | -.174 | -.061 | | | | |
| 1.554 | .221 | 1.035 | 0.012 | 1.970 | | | |
| .547 | .747 | -.024 | .578 | -.364 | -.256 | | |
| 2.875 | .689 | 1.689 | .329 | 1.885 | -.015 | 4.518 | |
| 1.634 | 1.803 | .358 | 1.421 | 0.004 | 1.330 | 0.802 | -.839 |

CHARACTER 39

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.199 | -.219 | -.037 | .295 | | | | |
| 1.604 | -.111 | .821 | .019 | 2.008 | | | |
| -.727 | -.337 | -.169 | -.268 | -.002 | 1.369 | | |
| 3.124 | -.359 | 1.281 | -.251 | 1.285 | .240 | 5.206 | |
| -2.252 | -.377 | -.515 | -.446 | -.228 | -.357 | .475 | 5.622 |

CHARACTER 40

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .014 | .022 | -.016 | 0.535 | | | | |
| 1.406 | -.118 | 1.111 | .057 | 2.147 | | | |
| -.114 | .237 | -.146 | -.548 | -.035 | -2.432 | | |
| 2.376 | -.393 | 1.487 | .019 | 2.387 | .252 | 6.283 | |
| -.648 | .826 | -.439 | -.420 | -.330 | -2.617 | -.200 | -9.709 |

TABLE B.1. (CONTINUED)

CHARACTER 41

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.280 | -.125 | .268 | -.115 | | | | |
| 1.781 | -.397 | .896 | .144 | 2.261 | | | |
| -1.398 | .043 | .346 | -.583 | .795 | -.554 | | |
| 4.481 | -1.324 | 1.288 | -.452 | 2.001 | .397 | 6.411 | |
| -5.796 | .957 | .351 | -.938 | 1.283 | -1.858 | 2.422 | -2.351 |

CHARACTER 42

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .038 | -.021 | .004 | .043 | | | | |
| 1.463 | -.015 | .794 | 0.039 | 1.951 | | | |
| .041 | .024 | -.045 | .026 | -.051 | .311 | | |
| 2.396 | -.094 | 1.097 | -.108 | 1.211 | -.138 | 4.421 | |
| -.033 | .174 | -.177 | .129 | 0.169 | .144 | 0.193 | 1.256 |

CHARACTER 43

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .007 | -.009 | .108 | -.185 | | | | |
| 1.573 | -.050 | .901 | -.011 | 1.871 | | | |
| -,054 | -.013 | .153 | -.027 | .287 | -.701 | | |
| 2.807 | -.175 | 1.347 | -.065 | 1.476 | -.002 | 4.116 | |
| -.281 | .009 | .224 | .046 | .437 | -.121 | .769 | -2.175 |

CHARACTER 44

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .336 | .146 | -.008 | .047 | | | | |
| 2.510 | .237 | .980 | -.003 | 1.605 | | | |
| 2.243 | .916 | .430 | .328 | -.045 | .126 | | |
| 8.659 | 1.705 | 2.674 | .553 | 1.369 | -.023 | 3.055 | |
| 12.106 | 4.999 | 2.988 | 1.916 | .780 | .639 | -.145 | .330 |

CHARACTER 45

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.989 | .623 | 0.366 | 0.396 | | | | |
| 4.145 | 0.969 | .837 | .188 | 1.758 | | | |
| -9.477 | 3.478 | -1.129 | .364 | -.838 | -1.576 | | |
| 30.686 | -8.465 | 3.272 | -.786 | 1.082 | .872 | 4.025 | |
| -88.581 | 26.336 | -7.732 | 2.720 | -1.333 | -.179 | -2.155 | -5.218 |

TABLE B.1. (CONTINUED)

CHARACTER 46

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .548 | -.579 | .241 | .548 | | | | |
| 2.757 | -.469 | .739 | .248 | 1.962 | | | |
| 3.318 | -2.077 | .527 | -.275 | .684 | 2.224 | | |
| 10.670 | -2.710 | 1.890 | -.202 | .970 | 1.038 | 5.246 | |
| 17.903 | -8.486 | 2.334 | -1.442 | .687 | .333 | 2.211 | 7.966 |

CHARACTER 47

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .038 | -.095 | .089 | .532 | | | | |
| 1.509 | -.036 | .955 | .094 | 1.945 | | | |
| .303 | -.476 | .264 | .398 | .249 | 2.155 | | |
| 2.913 | -.272 | 1.296 | .213 | 1.778 | .387 | 5.127 | |
| 1.298 | -1.484 | .617 | .132 | .678 | 1.847 | .801 | 7.669 |

CHARACTER 48

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .188 | .672 | .069 | -.065 | | | | |
| 3.483 | .259 | 1.263 | .054 | 1.404 | | | |
| 2.427 | 3.620 | .640 | .878 | .226 | -.188 | | |
| 18.517 | 3.503 | 4.755 | .770 | 1.761 | .178 | 2.192 | |
| 23.664 | 21.583 | 5.388 | 5.097 | 1.379 | 1.222 | .492 | -.452 |

CHARACTER 49

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.222 | .512 | -.076 | -.046 | | | | |
| 2.285 | -.242 | .982 | .039 | 1.366 | | | |
| -1.479 | 1.762 | -.210 | .470 | -.098 | -.133 | | |
| 7.227 | -1.407 | 2.153 | -.141 | 1.152 | .157 | 2.085 | |
| -7.749 | 6.334 | -1.179 | 1.659 | -.159 | .484 | -.143 | -.317 |

CHARACTER 50

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.131 | .220 | -.008 | .069 | | | | |
| 2.192 | -.529 | 1.286 | -.060 | 1.399 | | | |
| -.814 | .873 | -.324 | .515 | -.031 | .211 | | |
| 6.458 | -2.557 | 3.121 | -.844 | 1.893 | -.090 | 2.173 | |
| 04.152 | 3.574 | -1.820 | 1.887 | -.688 | 1.071 | -.064 | .506 |

TABLE B.1. (CONTINUED)

CHARACTER 51

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -,305 | .526 | -.214 | -.678 | | | | |
| 2.075 | -.205 | .598 | .341 | 2.087 | | | |
| -1.492 | 1.272 | -.185 | .193 | -.738 | -2.839 | | |
| 5.439 | -.871 | .994 | .050 | .801 | 1.427 | 6.189 | |
| -5.870 | 3.320 | -.533 | .680 | -.279 | -.454 | -2.808 | -10.703 |

CHARACTER 52

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.382 | .543 | -.208 | -.609 | | | | |
| 2.070 | -.337 | .618 | .358 | 1.992 | | | |
| -1.764 | 1.385 | -.326 | .205 | -.549 | -2.491 | | |
| 5.544 | -1.385 | 1.129 | -.070 | .781 | 1.383 | 5.542 | |
| -6.901 | 3.909 | -1.088 | .790 | -.374 | -.414 | -2.404 | -9.043 |

CHARACTER 53

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .118 | -.035 | -.013 | .235 | | | | |
| 1.747 | .061 | .703 | -.046 | 1.951 | | | |
| .474 | -.069 | -.003 | -.015 | -.039 | .905 | | |
| 3.573 | .201 | .918 | .018 | 1.055 | -.222 | 4.518 | |
| 1.542 | -.145 | .027 | -.073 | -.030 | .031 | -.099 | 2.985 |

CHARACTER 54

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .170 | -.007 | -.281 | -.010 | | | | |
| 1.578 | .226 | .962 | -.291 | 1.995 | | | |
| .597 | .012 | -.304 | .807 | -1.008 | .487 | | |
| 2.943 | .765 | 1.392 | -.050 | 1.880 | -1.343 | 5.142 | |
| 1.711 | .187 | -.310 | .022 | -1.222 | .809 | -3.475 | 3.351 |

CHARACTER 55

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .110 | -.108 | -.018 | .015 | | | | |
| 1.675 | -.132 | .847 | -.007 | 1.678 | | | |
| .377 | -.291 | .122 | -.201 | -.027 | .047 | | |
| 3.076 | -.410 | 1.289 | -.233 | 1.309 | -.307 | 3.264 | |
| 1.011 | -.684 | .434 | -.559 | .239 | -.400 | -.035 | .118 |

TABLE B.1. (CONTINUED)

CHARACTER 56

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.012 | -.040 | .489 | .373 | | | | |
| 1.887 | -.299 | 1.096 | .443 | 1.970 | | | |
| .252 | -.363 | .954 | .382 | 1.551 | 1.771 | | |
| 4.528 | -1.352 | 2.081 | .239 | 2.380 | 2.124 | 5.235 | |
| 1.747 | -1.725 | 2.384 | .003 | 2.726 | 2.284 | 5.176 | 7.198 |

CHARACTER 57

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .148 | -.090 | .314 | .045 | | | | |
| 1.931 | -.009 | 1.128 | .078 | 1.632 | | | |
| .899 | -.258 | 1.032 | -.012 | .767 | .187 | | |
| 4.781 | -.209 | 2.544 | .175 | 2.040 | .315 | 3.067 | |
| 3.979 | -.878 | 3.392 | -.053 | 2.404 | .309 | 1.772 | .647 |

CHARACTER 58

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .406 | .095 | -.355 | .185 | | | | |
| 1.808 | .082 | .847 | -.295 | 1.867 | | | |
| 1.487 | .167 | -.359 | .404 | -1.104 | .866 | | |
| 4.009 | .314 | 1.147 | -.320 | 1.618 | -1.322 | 4.566 | |
| 4.567 | .379 | -.307 | .549 | -1.334 | 1.565 | -3.565 | 3.807 |

CHARACTER 59

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.255 | .020 | .027 | -.154 | | | | |
| 1.511 | -.012 | .859 | .183 | 1.752 | | | |
| -.778 | .168 | .003 | -.002 | .168 | -.505 | | |
| 2.736 | -.122 | 1.068 | .209 | 1.442 | .670 | 3.779 | |
| -2.159 | .489 | -.148 | .209 | .264 | -.007 | .633 | -1.349 |

CHARACTER 60

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.030 | -.189 | .264 | .341 | | | | |
| 1.645 | -.243 | .936 | .287 | 2.040 | | | |
| .057 | -.497 | .363 | .027 | 1.010 | 1.627 | | |
| 3.239 | -.832 | 1.370 | .039 | 1.847 | 1.513 | 5.555 | |
| .521 | -1.252 | .732 | -.372 | 1.179 | 1.047 | 3.796 | 6.924 |

TABLE B.1. (CONTINUED)

CHARACTER 61

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .010 | .039 | .182 | -.030 | | | | |
| 1.865 | -.145 | .996 | -.023 | 1.612 | | | |
| ,125 | -.035 | .676 | .010 | .458 | -.129 | | |
| 4,114 | -.619 | 2.030 | -.263 | 1.699 | -.024 | 3.096 | |
| .714 | -.473 | 1.997 | -.284 | 1.675 | -.021 | 1.140 | -.415 |

CHARACTER 62

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.011 | -.000 | .006 | -.000 | | | | |
| 1.373 | .000 | .579 | .000 | 1.676 | | | |
| -.050 | -.000 | .016 | -.000 | .008 | -.000 | | |
| 2.100 | .000 | .618 | .000 | .667 | .000 | 3,260 | |
| -.136 | -.000 | .010 | -.000 | .032 | 0.000 | .003 | -.000 |

CHARACTER 63

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .034 | -.000 | .769 | -.000 | | | | |
| 2.013 | .000 | 1.159 | -.000 | 1.772 | | | |
| .187 | .000 | 1.512 | -.000 | 1.829 | -.000 | | |
| 5.014 | .000 | 2.516 | .000 | 2.545 | -.000 | 3.737 | |
| .821 | .000 | 3.804 | .000 | 3.717 | -.000 | 4.336 | -.000 |

CHARACTER 64

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.034 | -.000 | -.769 | -.000 | | | | |
| 2.-13 | -.000 | 1.159 | .000 | 1.772 | | | |
| -.187 | -.000 | -1.512 | -.000 | -1.829 | -.000 | | |
| 5.014 | -.000 | 2.516 | -.000 | 2.545 | .000 | 3.737 | |
| -.821 | .000 | -3.804 | .000 | -3.717 | .000 | -4.336 | -.000 |

CHARACTER 65

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .000 | 0.000 | .000 | 0.000 | | | | |
| 1.770 | 0.000 | 1.000 | 0.000 | 1.500 | | | |
| .000 | 0.000 | .000 | 0.000 | .000 | 0.000 | | |
| 3.667 | 0.000 | 1.770 | 0.000 | 1.500 | 0.000 | 2.250 | |
| .000 | 0.000 | .000 | 0.0-0 | .000 | 0.000 | 0.000 | 0.000 |

TABLE B.1. (CONTINUED)

CHARACTER 66

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .000 | -.000 | .000 | -.000 | | | | |
| 1.926 | 0.000 | .654 | 0.000 | 1.926 | | | |
| .000 | 0.000 | .000 | 0.000 | .000 | 0.000 | | |
| 4.421 | -.000 | .942 | 0.000 | .942 | 0.000 | 4.421 | |
| .000 | 0.000 | .000 | -.000 | .000 | -.000 | .000 | -.000 |

TABLE B.2. MOMENTS FOR UPPER AREA

CHARACTER 1.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.019 | .385 | .021 | -.485 | | | | |
| 2.092 | -.111 | .533 | .092 | 2.174 | | | |
| -.442 | .997 | .021 | .303 | -.059 | -2.469 | | |
| 5.568 | -.530 | .952 | .021 | .608 | .409 | 6.746 | |
| -2.695 | 2.912 | -.203 | .747 | .100 | .221 | -.595 | -11.064 |

CHARACTER 2.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.452 | .399 | .631 | -.052 | | | | |
| 2.046 | .052 | .937 | .379 | 2.708 | | | |
| -2.132 | 1.036 | .744 | 2.450 | .313 | | | |
| 5.855 | -.278 | 1.478 | 1.171 | 2.876 | 2.029 | 9.990 | |
| -8.855 | 2.805 | .957 | 2.016 | 3.342 | 3.695 | 10.066 | 3.387 |

CHARACTER 3.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.697 | .133 | .580 | -.084 | | | | |
| 2.343 | -.075 | .669 | .023 | 2.491 | | | |
| -3.424 | .459 | .451 | .107 | 1.813 | -.358 | | |
| 8.197 | -.496 | .793 | .159 | 1.670 | -.123 | 7.933 | |
| -15.660 | 1.632 | .305 | .361 | 11.577 | .119 | 6.123 | -1.833 |

CHARACTER 4

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.616 | .130 | .544 | -.096 | | | | |
| 2.343 | -.028 | .622 | -.052 | 2.541 | | | |
| -3.036 | .470 | .408 | -.007 | 1.738 | -.536 | | |
| 7.880 | -.314 | .728 | .027 | 1.521 | -.532 | 8.175 | |
| -13.790 | 1.586 | .286 | .180 | 1.348 | -.465 | 5.990 | -3.061 |

CHARACTER 5

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .640 | -.217 | .047 | .037 | | | | |
| 2.821 | -.208 | .519 | .086 | 1.762 | | | |
| 4.509 | -.684 | .108 | -.162 | .122 | .203 | | |
| 12.773 | -.943 | .691 | -.038 | .640 | .242 | 3.559 | |
| 27.528 | -2.367 | .456 | -.353 | .097 | -.173 | .305 | .771 |

TABLE B.2. (CONTINUED)

CHARACTER 6.

```
  .000  -.000   .000  -.000
 1.616   .000   .515   .000  1.526
  .000  -.000   .000  -.000   .000  -.000
 3.030   .000   .560   .000   .548   .000  2.668
  .000  -.000   .000   .000   .000  -.000   .000   .000
```

CHARACTER 7.

```
  .000  -.000   .000  -.000
 1.000  -.000  1.000  -.000  1.776
  .000  -.000   .000  -.000   .000  -.000
 1.000  -.000  1.000  -.000  1.776  -.000  3.703
  .000  -.000   .000  -.000   .000  -.000   .000  -.000
```

CHARACTER 8

```
  .473   .426  -.176  -.235
 1.922   .185   .842   .156  1.710
 1.542  1.158   .118   .450  0.432  0.858
 4.533   .739  1.471   .331  1.264   .566  3.568
 4.569  3.025   .707  1.293   .042   .502 -1.131 -2.695
```

CHARACTER 9

```
  .117  -.175  -.139  -.019
 1.627   .045   .762  -.016  1.749
  .529  -.196  -.173  -.321  -.286  -.091
 3.257   .206   .955  -.043  1.193  -.037  3.650
 1.941  -.093  -.178  -.443  -.411  -.664  -.600  -.349
```

CHARACTER 10

```
  .000  -.000   .000  -.000
 3.235  -.000   .056  -.000  3.235
  .000  -.000   .000  -.000   .000  -.000
12.530  -.000   .092  -.000   .092  -.000 12.530
  .000  -.000   .000  -.000   .000  -.000   .000  -.000
```

TABLE B.2. (CONTINUED)

CHARACTER 11

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.164 | -.231 | .373 | .617 | | | | |
| 1.676 | -.180 | .791 | .515 | 2.599 | | | |
| 0.472 | -.539 | .311 | .068 | 1.621 | 3.769 | | |
| 3,406 | -.537 | 1.058 | .134 | 1.841 | 3.062 | 10.545 | |
| -1.243 | -1.155 | .448 | -.434 | 1.249 | 1.919 | 7.807 | 21.022 |

CHARACTER 12

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .070 | -.398 | .144 | .552 | | | | |
| 2.121 | -.218 | .518 | .133 | 2.212 | | | |
| .535 | -.991 | .194 | -.233 | .316 | 2.506 | | |
| 5.630 | -.888 | .887 | -.069 | .643 | .438 | 6.527 | |
| 2,651 | -2.745 | .616 | -.686 | .222 | .018 | .872 | 10.052 |

SYMBOL 13

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.643 | -.094 | .699 | .729 | | | | |
| 2.190 | -.170 | .668 | .394 | 3.010 | | | |
| -2.978 | -.198 | .555 | .102 | 2.172 | 4.373 | | |
| 7.061 | -.390 | .667 | -.106 | 1.742 | 2.567 | 12.521 | |
| -12.669 | -.242 | .555 | -.306 | 1.507 | 1.362 | 8.590 | 23.772 |

SYMBOL 14

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .121 | .021 | .396 | .164 | | | | |
| 2.182 | .558 | 1.065 | -.138 | 1.761 | | | |
| .285 | .425 | 1.091 | .302 | .936 | .439 | | |
| 6.320 | 2.460 | 2.492 | .774 | 1.971 | -.374 | 3.739 | |
| .334 | 1.982 | 3.198 | 1.708 | 2.552 | .744 | 2.250 | 1.009 |

SYMBOL 15

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.092 | .112 | .268 | -.079 | | | | |
| 1.874 | .253 | .767 | -.236 | 1.851 | | | |
| -.189 | .384 | .337 | .031 | .717 | -.488 | | |
| 4.648 | 1.111 | 1.254 | .032 | 1.181 | -.871 | 4.305 | |
| -.051 | 1.272 | .759 | .377 | .751 | -.280 | 2.030 | -2.194 |

TABLE B.2. (CONTINUED)

SYMBOL 16

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.403 | -.185 | .365 | .253 | | | | |
| 2.494 | .221 | .785 | -.127 | 1.828 | | | |
| -2.263 | -.453 | .382 | -.112 | .888 | .957 | | |
| 8.685 | 1.448 | 1.453 | -.190 | 1.133 | -.162 | 4.486 | |
| -11.838 | -1.833 | .499 | -.050 | .892 | -.297 | 2.311 | 3.736 |

SYMBOL 17

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.748 | -.483 | .420 | .573 | | | | |
| 2.952 | .069 | .611 | -.091 | 2.488 | | | |
| -4.568 | -1.411 | .401 | -.471 | 1.119 | 3.110 | | |
| 12.781 | 1.162 | 1.147 | -.575 | 1.026 | .330 | 8.975 | |
| -26.117 | -5.362 | .294 | -1.147 | 1.080 | -.665 | 3.530 | 15.437 |

SYMBOL 18

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.798 | -.332 | .393 | 1.023 | | | | |
| 2.6-3 | .003 | .398 | .077 | 3.208 | | | |
| -4.124 | -.768 | .266 | -.223 | .995 | 5.718 | | |
| 10.098 | .546 | .533 | -.297 | .634 | .884 | 14.272 | |
| -20.283 | -2.422 | .212 | -.491 | .554 | -.113 | 3.371 | 30.390 |

SYMBOL 19

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.081 | -.029 | .178 | .237 | | | | |
| 2.047 | .520 | .904 | -.030 | 1.569 | | | |
| -.803 | -.068 | .447 | .338 | .470 | .693 | | |
| 5.776 | 2.096 | 1.812 | .735 | 1.418 | -.075 | 2.986 | |
| -4.656 | -.647 | .939 | 1.027 | 1.286 | .962 | 1.061 | 1.740 |

SYMBOL 20

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.199 | .068 | .081 | -.000 | | | | |
| 1.831 | .227 | .766 | -.109 | 1.522 | | | |
| -.802 | .084 | -.037 | .111 | .270 | -.084 | | |
| 4.425 | .998 | 1.270 | .149 | .921 | -.392 | 2.842 | |
| -2.986 | -.168 | -.262 | .277 | .152 | .125 | .699 | -.470 |

TABLE B.2. (CONTINUED)

SYMBOL 21

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.489 | -.227 | .122 | .405 | | | | |
| 2.423 | .310 | .681 | -.102 | 1.683 | | | |
| -2.849 | -.712 | -.079 | .043 | .336 | 1.290 | | |
| 8.515 | 1.728 | 1.301 | -.009 | .833 | -.161 | 3.675 | |
| -14.771 | -3.367 | -.844 | .098 | .180 | .195 | .785 | 3.979 |

SYMBOL 22

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.801 | -.534 | .199 | .589 | | | | |
| 3.024 | .336 | .549 | -.122 | 2.126 | | | |
| -5.211 | -1.712 | .021 | -.381 | .507 | 2.620 | | |
| 14.120 | 2.518 | 1.202 | -.204 | .627 | -.062 | 6.627 | |
| -31.396 | -7.631 | -1.033 | -.941 | .443 | -.441 | 1.341 | 11.058 |

SYMBOL 23

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.854 | -.382 | .399 | .945 | | | | |
| 2.893 | .041 | .455 | .031 | 3.274 | | | |
| -4.872 | -.993 | .294 | -.303 | 1.087 | 5.722 | | |
| 12.500 | .872 | .704 | -.383 | .777 | .789 | 15.156 | |
| -26.560 | -3.581 | .190 | -.681 | .730 | -.294 | 3.807 | 32.486 |

SYMBOL 24

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.184 | -.014 | .149 | .294 | | | | |
| 2.156 | .561 | .933 | -.020 | 1.685 | | | |
| -1.261 | -.107 | .401 | .425 | .453 | .909 | | |
| 6.628 | 2.371 | 1,950 | .832 | 1.573 | -,045 | 3.470 | |
| -6.977 | -1,104 | .811 | 1.193 | 1.378 | 1.276 | 1.112 | 2.465 |

SYMBOL 25

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.296 | .083 | .048 | .052 | | | | |
| 1.923 | .240 | .791 | -.097 | 1.604 | | | |
| -1.191 | .082 | -.117 | .176 | .220 | .060 | | |
| 5.007 | 1.099 | 1,394 | .169 | .994 | -.378 | 3.165 | |
| -4.579 | -.370 | -.500 | .364 | .073 | .281 | .623 | -.120 |

TABLE B.2. (CONTINUED)

SYMBOL 26

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.602 | -.222 | .090 | .447 | | | | |
| 2.623 | .354 | .708 | -.117 | 1.827 | | | |
| -3.578 | -.803 | -.162 | .079 | .303 | 1.529 | | |
| 10.266 | 2.080 | 1.469 | -.011 | .927 | -.211 | 4.354 | |
| -19.690 | -4.282 | -1.226 | .160 | .105 | .299 | .779 | 5.092 |

SYMBOL 27

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.876 | -.547 | .193 | .629 | | | | |
| 3.310 | .384 | .578 | -.132 | 2.308 | | | |
| -6.098 | -1.917 | -.001 | -.409 | .536 | 3.014 | | |
| 17.006 | 3.049 | 1.372 | -.214 | .703 | -.083 | 7.846 | |
| -39.879 | -9.408 | -1.326 | -1.093 | .495 | -.501 | 1.532 | 13.774 |

SYMBOL 28

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.742 | -.327 | .257 | .952 | | | | |
| 2.795 | .145 | .345 | -.133 | 3.370 | | | |
| -4.267 | -.791 | .148 | -.279 | .558 | 5.983 | | |
| 11.366 | 1.043 | .499 | -.243 | .510 | -.204 | 16.178 | |
| -22.807 | -2.951 | -.094 | -.428 | .433 | -.425 | 1.316 | 35.295 |

SYMBOL 29

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.224 | -.104 | .098 | .334 | | | | |
| 2.263 | .491 | .879 | -.066 | 1.733 | | | |
| -1.690 | -.571 | .196 | .358 | .332 | 1.063 | | |
| 7.430 | 2.098 | 1.669 | .589 | 1.477 | -.184 | 3.700 | |
| -9.833 | -3.053 | -.080 | .618 | .921 | 1.197 | .828 | 2.988 |

SYMBOL 30

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.268 | .013 | .017 | .086 | | | | |
| 1.999 | .275 | .791 | -.117 | 1.627 | | | |
| -1.284 | -.239 | -.173 | .168 | .161 | .165 | | |
| 5.558 | 1.294 | 1.404 | .144 | 1.017 | -.430 | 3.283 | |
| -5.902 | -1.657 | -.725 | .228 | -.044 | .336 | .511 | .184 |

TABLE B.2. (CONTINUED)

SYMBOL 31

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.569 | -.260 | .051 | .501 | | | | |
| 2.700 | .358 | .646 | -.160 | 1.917 | | | |
| -3.467 | -.968 | -.173 | .061 | .163 | 1.782 | | |
| 10.660 | 2.056 | 1.289 | .015 | .853 | -.424 | 4.845 | |
| -19.549 | -4.841 | -1.194 | .029 | .072 | .299 | .290 | 6.186 |

SYMBOL 32

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.820 | -.512 | .119 | .682 | | | | |
| 3.251 | .404 | .519 | -.238 | 2.428 | | | |
| -5.663 | -1.756 | -.044 | -.356 | .195 | 3.380 | | |
| 16.020 | 2.874 | 1.192 | -.140 | .613 | -.726 | 8.764 | |
| -36.131 | -8.445 | -1.300 | -.896 | .374 | -.437 | -.036 | 16.146 |

SYMBOL 33

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.033 | -.471 | .019 | .578 | | | | |
| 2.128 | -.010 | .411 | .002 | 2.358 | | | |
| -.121 | -1.026 | .013 | -.327 | .030 | 2.961 | | |
| 5.289 | -.007 | .734 | -.015 | .385 | .006 | 7.622 | |
| -.420 | -2.550 | .021 | -.584 | .020 | -.267 | .048 | 13.003 |

SYMBOL 34

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .286 | .043 | .111 | -.079 | | | | |
| 1.820 | .331 | 1.077 | -.078 | 1.698 | | | |
| .991 | .302 | .563 | .164 | .211 | -.344 | | |
| 4.029 | 1.178 | 2.098 | .391 | 1.657 | -.198 | 3.349 | |
| 3.080 | 1.241 | 1.729 | .746 | .968 | .354 | .444 | -1.112 |

SYMBOL 35

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .130 | .106 | -.029 | -.300 | | | | |
| 1.687 | .117 | .799 | -.081 | 1.844 | | | |
| .594 | .437 | .006 | -.102 | -.070 | -1.167 | | |
| 3.563 | .527 | 1.244 | -.020 | 1.133 | -.216 | 4.141 | |
| 2.300 | 1.479 | .241 | .129 | -.085 | -.442 | -.112 | -3.799 |

TABLE B.2. (CONTINUED)

SYMBOL 36

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.016 | -.190 | .006 | .083 | | | | |
| 2.168 | .001 | .851 | .013 | 1.651 | | | |
| -.091 | -.223 | -.025 | -.176 | .024 | .370 | | |
| 5.636 | .015 | 1.826 | .006 | .941 | .047 | 3.405 | |
| -.383 | -.016 | -.114 | -.095 | -.023 | -.303 | .079 | 1.605 |

SYMBOL 37

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -.327 | -.601 | .044 | .296 | | | | |
| 2.513 | .207 | .695 | -.101 | 2.090 | | | |
| -1.860 | -1.813 | -.049 | -.576 | .065 | 1.619 | | |
| 8.072 | 1.175 | 1.660 | -.012 | .781 | -.264 | 5.833 | |
| -8.675 | -6.018 | -.686 | -1.515 | .084 | -.679 | -.003 | 7.183 |

SYMBOL 38

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .199 | -.038 | .559 | .557 | | | | |
| 1.854 | -.280 | .783 | .502 | 2.531 | | | |
| .907 | -.347 | .632 | .329 | 1.738 | 3.308 | | |
| 4.369 | -.995 | 1.037 | .152 | 1.584 | 2.525 | 9.233 | |
| 3.544 | -1.300 | 1.049 | .010 | 1.364 | 1.800 | 6.418 | 16.775 |

SYMBOL 39

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .603 | -.254 | .242 | .163 | | | | |
| 2.603 | -.217 | 1.040 | -.091 | 1.604 | | | |
| 3.345 | -1.247 | 1.065 | -.036 | .452 | .491 | | |
| 9.814 | -2.111 | 2.587 | -.163 | 1.567 | -.183 | 3.001 | |
| 17.729 | -6.477 | 4.014 | -.684 | 1.667 | .329 | .848 | 1.309 |

SYMBOL 40

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .454 | -.111 | .142 | -.076 | | | | |
| 2.132 | -.171 | .945 | -.177 | 1.578 | | | |
| 2.250 | -.661 | .646 | -.221 | .276 | -.309 | | |
| 6.460 | -1.204 | 2.029 | -.466 | 1.269 | -.477 | 2.944 | |
| 10.129 | -3.291 | 2.555 | -.842 | .889 | -.454 | .561 | -.931 |

TABLE B.2. (CONTINUED)

SYMBOL 41

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .446 | -.373 | .245 | .317 | | | | |
| 2.671 | -.390 | .802 | -.078 | 1.691 | | | |
| 3.027 | -1.461 | .754 | -.225 | .389 | 1.059 | | |
| 10.125 | -2.388 | 1.976 | -.410 | .977 | -.152 | 3.529 | |
| 17.122 | -6.622 | 2.965 | -.957 | .938 | -.158 | .613 | 3.341 |

SYMBOL 42

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .000 | -.000 | .000 | -.000 | | | | |
| 1.221 | .000 | 1.000 | 0.000 | 1.776 | | | |
| .000 | -.000 | .000 | -.000 | .000 | -.000 | | |
| 1.664 | 0.000 | 1.221 | .000 | 1.776 | 0.000 | 3.703 | |
| .000 | -.000 | .000 | -.000 | .000 | -.000 | .000 | -.000 |

## TABLE B.3. MOMENTS FOR LOWER AREA CHARACTERS

### CHARACTER 1.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .208 | .499 | .190 | -.551 | | | | |
| 2.357 | ,168 | .763 | -.225 | 2.031 | | | |
| 1.130 | 1.603 | .295 | .232 | .604 | -2.295 | | |
| 7.185 | .806 | 1.597 | -.041 | 1.147 | -.944 | 5.558 | |
| 5.227 | 5.249 | .743 | 1.172 | .647 | -.389 | 2.046 | -8.447 |

### CHARACTER 2

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| .233 | .244 | -.051 | .081 | | | | |
| 1.357 | .211 | .945 | .010 | 1.552 | | | |
| .609 | .560 | .186 | .460 | -.065 | .344 | | |
| 2.132 | .582 | 1.357 | .371 | 1.468 | .052 | 2.765 | |
| 1.349 | 1.226 | .615 | .977 | .386 | .980 | -.059 | 1.088 |

## APPENDIX C.

Table C.1. Comparison of Recognition Rates for the Middle, Upper and Lower Areas with Variance, Scatter Ratio, Natural and Random Ordering using 2,4,....,30 Feature elements for test data with 10 percent noise level.

(a) The Middle Area

| No. of elements used for recognition | Percent Recognised Correctly | | | |
|---|---|---|---|---|
| | Variance | Scatter Ratio | Natural | Random |
| 2 | 65.30 | 65.30 | 47.65 | 41.36 |
| 4 | 89.24 | 71.82 | 85.30 | 62.88 |
| 6 | 92.50 | 77.65 | 92.73 | 74.70 |
| 8 | 91.67 | 87.88 | 94.15 | 80.91 |
| 10 | 93.33 | 91.59 | 83.18 | 86.21 |
| 12 | 93.18 | 88.94 | 89.39 | 77.65 |
| 14 | 92.73 | 90.76 | 91.21 | 83.33 |
| 16 | 94.32 | 92.42 | 89.92 | 85.68 |
| 18 | 93.18 | 91.82 | 91.67 | 86.29 |
| 20 | 92.50 | 91.74 | 92.20 | 86.89 |
| 22 | 92.20 | 84.77 | 92.12 | 87.27 |
| 24 | 91.14 | 86.74 | 87.20 | 82.27 |
| 26 | 91.52 | 87.80 | 87.42 | 87.58 |
| 28 | 87.27 | 88.79 | 88.41 | 88.94 |
| 30 | 86.21 | 86.21 | 86.21 | 86.21 |

Table C.1. (continued)

(b) The Upper Area

| No. of elements used for recognition | Percent Recognised Correctly | | | |
|---|---|---|---|---|
| | Variance | Scatter Ratio | Natural | Random |
| 2 | 72.02 | 72.02 | 67.02 | 44.05 |
| 4 | 89.17 | 80.71 | 84.40 | 62.98 |
| 6 | 87.14 | 87.02 | 89.05 | 73.10 |
| 8 | 89.64 | 78.21 | 91.67 | 78.21 |
| 10 | 89.64 | 80.71 | 76.55 | 81.43 |
| 12 | 92.86 | 86.43 | 83.10 | 65.60 |
| 14 | 92.14 | 87.38 | 86.31 | 74.29 |
| 16 | 93.69 | 87.50 | 78.45 | 77.26 |
| 18 | 92.50 | 87.14 | 80.24 | 77.62 |
| 20 | 91.55 | 78.45 | 81.07 | 78.45 |
| 22 | 91.19 | 76.90 | 83.10 | 78.69 |
| 24 | 90.12 | 79.88 | 76.19 | 78.81 |
| 26 | 89.40 | 80.00 | 77.14 | 79.05 |
| 28 | 87.02 | 79.88 | 77.86 | 79.76 |
| 30 | 80.24 | 80.24 | 80.24 | 80.24 |

(c) The Lower Area

Recognition was found to be 100 percent correct for the test data with this noise level irrespective of the element ordering.

## APPENDIX D.

### Table D.1. Roots and Vectors of $|S-\lambda I|=0$

(a) Middle Area Samples.

| $k^j$ | $a_{1k}$ | $a_{2k}$ | $a_{3k}$ |
|---|---|---|---|
| 1 | -1.137E-01 | 2.116E-02 | 1.336E-01 |
| 2 | 1.703E-01 | 1.316E-02 | -1.282E-01 |
| 3 | -8.208E-02 | 1.948E-01 | 1.578E-02 |
| 4 | -8.315E-02 | -2.962E-03 | 1.624E-01 |
| 5 | 3.363E-01 | 6.2.6E-02 | 9.992E-02 |
| 6 | -4.040E-02 | -6.751E-02 | 5.083E-02 |
| 7 | 3.959E-02 | 6.598E-02 | 1.723E-01 |
| 8 | 6.377E-03 | 1.276E-01 | 1.213E-01 |
| 9 | -7.207E-02 | 5.216E-02 | -8.550E-02 |
| 10 | 7.054E-01 | 6.293E-02 | -3.027E-01 |
| 11 | -1.956E-01 | 3.743E-01 | 1.024E-01 |
| 12 | 1.825E-01 | 3.045E-02 | -4.269E-02 |
| 13 | -2.047E-01 | 5.769E-01 | 2.410E-02 |
| 14 | 4.585E-01 | 2.738E-01 | 5.895E-01 |
| 15 | -1.155E-02 | 1.085E-01 | -7.604E-02 |
| 16 | -1.395E-02 | 2.169E-01 | 4.236E-01 |
| 17 | 2.166E-02 | 5.639E-01 | -4.874E-01 |
| Roots ($\lambda_j$) | 1.230E+00 | 6.384E-01 | 3.966E-01 |

| $k^j$ | $a_{4k}$ | $a_{5k}$ | $a_{6k}$ |
|---|---|---|---|
| 1 | 3.104E-01 | -1.114E-03 | -2.011E-01 |
| 2 | 1.178E-01 | -6.463E-02 | 2.208E-02 |
| 3 | -1.214E-02 | -1.611E-01 | 1.334E-01 |
| 4 | -1.001E-01 | -1.006E-01 | -9.788E-02 |
| 5 | -6.102E-01 | -2.890E-01 | -2.775E-01 |
| 6 | 2.382E-01 | -1.834E-01 | -2.756E-01 |
| 7 | 9.340E-02 | 1.235E-01 | 1.285E-02 |
| 8 | -1.944E-02 | 1.065E-01 | -7.798E-02 |
| 9 | -8.581E-03 | 1.316E-01 | 1.359E-01 |
| 10 | 3.722E-01 | -1.180E-01 | 3.398E-01 |
| 11 | 1.948E-01 | -4.072E-01 | 1.077E-02 |
| 12 | 1.149E-01 | -2.452E-01 | -1.890E-01 |
| 13 | -6.990E-02 | -3.398E-01 | 3.483E-01 |
| 14 | -6.468E-02 | 1.197E-01 | -1.398E-01 |
| 15 | 3.905E-01 | -1.232E-01 | -5.916E-01 |
| 16 | 2.500E-01 | 4.609E-01 | 1.425E-01 |
| 17 | -1.661E-01 | 4.488E-01 | -3.027E-01 |
| Roots ($\lambda_j$) | 3.092E-01 | 2.711E-01 | 1.442E-01 |

Table D.I. (Cont.)

| $k^j$ | $a_{7k}$ | $a_{8k}$ | $a_{9k}$ |
|---|---|---|---|
| 1 | -2.063E-01 | -4.654E-01 | -3.651E-01 |
| 2 | 3.005E-02 | 9.475E-02 | -7.429E-02 |
| 3 | 3.319E-02 | 3.434E-02 | 3.905E-02 |
| 4 | -4.641E-01 | 3.065E-01 | -2.064E-01 |
| 5 | 4.207E-01 | -2.518E-02 | -2.012E-01 |
| 6 | 2.292E-01 | 1.475E-01 | 1.709E-01 |
| 7 | -2.575E-02 | 5.154E-02 | 9.150E-02 |
| 8 | -2.794E-02 | -4.183E-02 | -5.895E-03 |
| 9 | 3.442E-01 | 8.121E-02 | -5.800E-01 |
| 10 | 2.950E-02 | -7.473E-02 | -2.273E-02 |
| 11 | 7.797E-02 | -3.574E-01 | -2.866E-01 |
| 12 | -3.366E-01 | 5.433E-01 | -3.686E-01 |
| 13 | -9.490E-03 | 2.121E-01 | 2.676E-01 |
| 14 | -2.254E-01 | -1.889E-01 | 1.540E-01 |
| 15 | 2.700E-01 | 1.385E-01 | 2.458E-01 |
| 16 | 3.510E-01 | 3.399E-01 | -1.650E-01 |
| 17 | -1.529E-01 | -4.502E-02 | -3.191E-02 |
| Roots ($\lambda_j$) | 1.002E-01 | 7.625E-02 | 4.268E-02 |

| $k^j$ | $a_{10k}$ | $a_{11k}$ | $a_{12k}$ |
|---|---|---|---|
| 1 | 1.056E-01 | 4.652E-02 | 6.411E-01 |
| 2 | -2.876E-01 | 1.405E-01 | 1.094E-02 |
| 3 | 1.386E-02 | 1.077E-01 | -6.000E-02 |
| 4 | 4.981E-01 | -5.001E-01 | -1.465E-01 |
| 5 | -3.161E-02 | -2.127E-01 | 2.652E-01 |
| 6 | 5.516E-02 | -6.670E-02 | -3.093E-02 |
| 7 | -1.053E-01 | -9.490E-02 | 7.922E-02 |
| 8 | -1.885E-03 | 8.905E-03 | -4.965E-02 |
| 9 | 4.632E-01 | 4.307E-01 | -2.082E-01 |
| 10 | 2.159E-01 | -2.573E-01 | 4.117E-02 |
| 11 | -2.891E-01 | -2.542E-01 | -4.673E-01 |
| 12 | -3.703E-01 | 2.734E-01 | 6.644E-02 |
| 13 | 1.865E-01 | 1.568E-01 | 3.695E-01 |
| 14 | 1.168E-01 | 3.592E-01 | -2.393E-01 |
| 15 | 2.743E-01 | 6.999E-02 | -4.346E-02 |
| 16 | -1.830E-01 | -3.112E-01 | 1.406E-01 |
| 17 | -6.158E-02 | -1.060E-01 | -2.163E-02 |
| Roots ($\lambda_j$) | 2.068E-02 | 1,267E-02 | 8.102E-03 |

Table D.1 (Cont.)

| $k^j$ | $a_{13k}$ | $a_{14k}$ | $a_{15k}$ |
|---|---|---|---|
| 1 | -4.583E-02 | -4.102E-02 | -6.569E-02 |
| 2 | -4.208E-02 | -7.971E-01 | -3.656E-01 |
| 3 | 1.512E-01 | -1.130E-01 | -1.432E-01 |
| 4 | 3.539E-03 | -2.321E-01 | -1.052E-01 |
| 5 | 4.590E-02 | -3.513E-02 | 3.366E-02 |
| 6 | -8.026E-01 | 2.711E-03 | -5.568E-02 |
| 7 | 9.797E-04 | -4.191E-01 | 8.405E-01 |
| 8 | 2.160E-01 | -9.831E-02 | -1.418E-01 |
| 9 | -1.061E-01 | -8.430E-02 | 1.535E-01 |
| 10 | -1.078E-02 | 1.127E-01 | 4.478E-02 |
| 11 | 1.747E-02 | 4.599E-02 | 7.232E-02 |
| 12 | 2.680E-02 | 2.719E-01 | 1.235E-01 |
| 13 | -5.562E-02 | 3.073E-03 | 6.057E-03 |
| 14 | -9.377E-02 | 1.457E-02 | -9.330E-02 |
| 15 | 4.511E-01 | -1.685E-02 | 2.707E-02 |
| 16 | 7.737E-02 | 1.106E-01 | -2.087E-01 |
| 17 | -2.158E-01 | 3.394E-02 | 2.355E-02 |
| Roots ($\lambda_j$) | 3.014E-03 | 9.496E-04 | 7.767E-04 |

| $k^j$ | $a_{16k}$ | $a_{17k}$ |
|---|---|---|
| 1 | -81359E-02 | 4.999E-03 |
| 2 | 1.727E-01 | -1.311E-01 |
| 3 | -9.105E-01 | -8.311E-02 |
| 4 | 6.213E-03 | -2.133E-02 |
| 5 | -2.877E-02 | 1.577E-02 |
| 6 | -1.478E-01 | 2.076E-01 |
| 7 | -1.091E-01 | 99.030E-02 |
| 8 | -1.869E-02 | 9.323E-01 |
| 9 | 4.280E-02 | 1.431E-02 |
| 10 | -4.684E-02 | 1.958E-02 |
| 11 | 1.306E-01 | -9.202E-03 |
| 12 | -3.562E-02 | 5.629E-02 |
| 13 | 2.563E-01 | 4.084E-02 |
| 14 | 3.628E-02 | -1.103E-02 |
| 15 | 7.926E-02 | -1.391E-01 |
| 16 | -6.173E-03 | -1.333E-02 |
| 17 | -5.998E-02 | -1.719E-01 |
| Roots ($\lambda_j$) | 4.577E-04 | 2.990E-04 |

Table D.1. (cont.)

(b) Upper Area Samples.

| $k^j$ | $a_{1k}$ | $a_{2k}$ | $a_{3k}$ |
|---|---|---|---|
| 1 | 1.971E-01 | -2.406E-01 | 4.657E-02 |
| 2 | 2.071E-01 | 3.538E-02 | -1.059E-01 |
| 3 | -4.169E-03 | 3.138E-01 | 9.841E-02 |
| 4 | -2.431E-01 | 1.401E-01 | 8.734E-02 |
| 5 | -3.792E-01 | 8.438E-02 | 1.153E-01 |
| 6 | 3.056E-03 | -6.505E-02 | 4.985E-02 |
| 7 | 1.368E-01 | -4.897E-02 | 1.937E-01 |
| 8 | 4.274E-02 | 1.048E-01 | -3.178E-02 |
| 9 | -2.254E-01 | 4.419E-01 | -2.344E-01 |
| 10 | 6.344E-01 | 1.559E-01 | -4.310E-01 |
| 11 | 8.357E-02 | 1.593E-01 | 2.813E-01 |
| 12 | 2.081E-01 | 7.471E-02 | 7.391E-04 |
| 13 | 4.992E-02 | 7.022E-01 | 2.544E-01 |
| 14 | 1.659E-01 | -2.326E-01 | 6.337E-01 |
| 15 | 2.091E-01 | 3.158E-02 | 4.761E-02 |
| 16 | 3.428E-01 | 2.282E-01 | 3.558E-01 |
| Roots ($\lambda_j$) | 1.115E-00 | 7.408E-01 | 4.612E-01 |

| $k^j$ | $a_{4k}$ | $a_{5k}$ | $a_{6k}$ |
|---|---|---|---|
| 1 | 4.926E-01 | -4.439E-01 | 2.941E-01 |
| 2 | -9.245E-02 | -1.140E-01 | -7.076E-02 |
| 3 | 2.910E-02 | -3.692E-02 | -1.246E-01 |
| 4 | 2.613E-02 | 2.058E-01 | 3.914E-01 |
| 5 | -4.615E-01 | -6.016E-01 | -8.913E-02 |
| 6 | -4.514E-01 | 1.891E-01 | 1.025E-01 |
| 7 | 2.741E-02 | 1.142E-01 | -2.632E-02 |
| 8 | 1.493E-01 | 1.404E-02 | 2.022E-01 |
| 9 | 6.679E-02 | -1.804E-01 | 4.530E-01 |
| 10 | -2.144E-01 | -2.382E-01 | -1.369E-01 |
| 11 | 1.949E-01 | -3.835E-01 | 6.048E-02 |
| 12 | -2.125E-01 | -7.104E-02 | 2.620E-01 |
| 13 | 6.371E-02 | 4.322E-02 | -3.053E-01 |
| 14 | -1.914E-01 | -1.662E-01 | 1.487E-02 |
| 15 | -3.565E-01 | 7.485E-02 | 5.222E-01 |
| 16 | 7.852E-02 | 2.425E-01 | 1.320E-01 |
| Roots ($\lambda_j$) | 1.944E-01 | 1.354E-01 | 7.840E-02 |

Table D.1. (Cont).

| $k^j$ | $a_{7k}$ | $a_{8k}$ | $a_{9k}$ |
|---|---|---|---|
| 1 | -2.462E-01 | 6.455E-02 | 1.139E-01 |
| 2 | 5.022E-03 | -1.220E-02 | -1.975E-01 |
| 3 | -5.281E-02 | -1.481E-01 | 9.845E-02 |
| 4 | 9.720E-04 | -5.600E-01 | -2.254E-01 |
| 5 | -9.391E-02 | 2.912E-01 | -1.311E-01 |
| 6 | 8.797E-02 | -1.355E-01 | 3.881E-01 |
| 7 | 1.279E-01 | 1.512E-01 | -9.142E-02 |
| 8 | -2.283E-01 | 7.236E-02 | -3.523E-01 |
| 9 | 5.353E-01 | 1.146E-01 | -1.330E-02 |
| 10 | 2.060E-01 | -2.400E-01 | -3.295E-02 |
| 11 | 1.117E-01 | -3.038E-01 | 4.145E-01 |
| 12 | -2.907E-01 | -1.027E-01 | -4.569E-01 |
| 13 | -3.555E-01 | -7.624E-02 | 5.151E-02 |
| 14 | 2.959E-01 | -2.206E-01 | -2.493E-01 |
| 15 | -3.739E-01 | 1.458E-01 | 3.706E-01 |
| 16 | 2.679E-01 | 5.497E-01 | -5.712E-02 |
| Roots ($\lambda_j$) | 5.209E-02 | 3.577E-02 | 2.286E-02 |

| $k^j$ | $a_{10k}$ | $a_{11k}$ | $a_{12k}$ |
|---|---|---|---|
| 1 | 3.742E-02 | -1.826E-01 | -4.802E-01 |
| 2 | 1.080E-01 | 2.839E-01 | -1.143E-01 |
| 3 | 2.171E-02 | 4.618E-02 | 3.854E-02 |
| 4 | 3.171E-01 | -3.907E-01 | 4.279E-02 |
| 5 | 2.189E-01 | -2.944E-01 | 3.541E-02 |
| 6 | 1.479E-01 | -7.031E-02 | -6.045E-01 |
| 7 | 6.391E-02 | -9.955E-02 | -7.814E-02 |
| 8 | -2.512E-01 | -3.121E-01 | 1.244E-01 |
| 9 | -2.815E-01 | 1.946E-01 | -1.729E-01 |
| 10 | -6.212E-03 | -3.631E-01 | 6.954E-02 |
| 11 | 2.917E-01 | 2.186E-01 | 3.688E-01 |
| 12 | 2.806E-01 | 5.359E-01 | -1.143E-01 |
| 13 | -2.227E-01 | -3.490E-02 | -2.820E-01 |
| 14 | -4.674E-01 | 1.451E-02 | -3.868E-02 |
| 15 | -2.799E-01 | -1.362E-02 | 3.131E-01 |
| 16 | 3.993E-01 | -1.616E-01 | 4.209E-02 |
| Roots ($\lambda_j$) | 1.486E-02 | 8.249E-03 | 4.899E-03 |

Table D.1. (Cont).

| $k^j$ | $a_{13k}$ | $a_{14k}$ | $a_{15k}$ |
|---|---|---|---|
| 1 | -1.597E-01 | 4.417E-02 | -2.197E-02 |
| 2 | 1.167E-01 | 5.273E-01 | 6.569E-01 |
| 3 | 3.289E-02 | 4.091E-01 | -5.494E-01 |
| 4 | -2.314E-01 | 8.697E-02 | 1.830E-01 |
| 5 | -2.373E-02 | 5.941E-02 | -2.111E-02 |
| 6 | 4.022E-01 | -9.681E-02 | -2.734E-03 |
| 7 | 9.385E-02 | 6.229E-01 | -2.999E-01 |
| 8 | 7.469E-01 | -5,038E-02 | -2.859E-03 |
| 9 | -2.214E-02 | 5.389E-02 | -6.468E-02 |
| 10 | -9.085E-02 | -1.007E-01 | -9.263E-02 |
| 11 | 3.326E-01 | -9.404E-02 | 9.942E-02 |
| 12 | -1.267E-02 | -2.176E-01 | -3.018E-01 |
| 13 | -1.295E-01 | -8.343E-02 | 1.435E-01 |
| 14 | -8.085E-02 | -9.892E-02 | 5.092E-02 |
| 15 | -1.762E-01 | 1.940E-01 | 5.967E-02 |
| 16 | -5.718E-02 | -1.344E-01 | 4.163E-02 |
| Roots ($\lambda_j$) | 4.864E-03 | 7.429E-04 | 4.802E-02 |

| $k^j$ | $a_{16k}$ |
|---|---|
| 1 | -6.219E-02 |
| 2 | -2.246E-01 |
| 3 | -6.465E-01 |
| 4 | -5.787E-02 |
| 5 | 4.243E-02 |
| 6 | -5.644E-02 |
| 7 | 6.286E-01 |
| 8 | -6.987E-02 |
| 9 | 4.796E-02 |
| 10 | 3.377E-02 |
| 11 | 1.579E-01 |
| 12 | 1.042E-01 |
| 13 | 1.762E-01 |
| 14 | -1.221E-01 |
| 15 | 2.032E-02 |
| 16 | -1.838E-01 |
| Roots ($\lambda_j$) | 1.814E-04 |

# Bibliography

[1] Reifler, E. (1967): Chinese-English Machine Translation, its Lexiographic and Linguistic Problems," Machine Translation, Booth, A.D. (ed), North Holland Publishing Co. Amsterdam.

[2] Booth, A.D., Bandwood, L. and Cleave, J.P. (1958): Mechanical Resolution of Linguistic Problems, Butterworth Scientific Publications, London.

[3] Perry, P.G. and Strehlow, T.J. (1969): "A Preliminary Investigation into an Automated Process for Translating Written Thai into English", T.R. available Comp. Sc. Dept., University of Adelaide.

[4] Allison, G.H. (1962): Modern Thai, Nibondh and Co. Ltd., Thailand (Second Edition).

[5] Haas, M.R. (1964): Thai-English Student's Dictionary, Stanford, Stanford Univ. Press.

[6] Jessup, A.M. and Wallace, C.S. (1968): "A Cheap Graphic Input Device", Australian Comp. Journal, Vol. 1,2,95-96.

[7] Lee, K.C. and Perry, P.G. (1969): "Computer Assisted Instruction", Proc. of Fourth Aust. Comp. Conf., 401-406.

[8] Fitzmaurice, A. (1962): "Reading Russian Scientific Literature", Optical Char. Recog., Spartan Books, McGregor and Werner, Inc., Washington 12, D.C. 61-72.

[9] Hammans, B. (1969): "The Design of a Page Scanner for Character Recognition", Marconi Review, Vol. XXXII, 72, 31-48.

[10] Mason, C.J.W. and McFall, S.H. (1968): "Some Experience with the Method of Potential Functions", IEE Conf. Publication 42,69-76.

[11] Unger, S.H. (1959): "Pattern Detection and Recognition", Proc. IRE, Vol. 47, 10, 1737-1752.

[12] Sherman, H. (1959): "A Quasi-Topological Method for the Recognition of Line Patterns", Proc. International. Conf. on Inf. Proc. UNESCO, 227-238.

[13] Deutsch, E.S. (1968): "Preprocessing for Character Recognition", IEE Conf. Publication 42, 179-185.

[14] Dineen, G.P. (1955): "Programming Pattern Recognition", Western Joint Comp. Conf., 94-100.

[15] Alcorn, T.M. and Hogger, C.W. (1969): "Preprocessing of Data for Character Recognition", Marconi Review, Vol. XXXII, 17, 61-81.

[16] Bomba, J.S. (1959): "Alpha-Numeric Character Recognition using Local Operations", Proc. of Eastern Joint Comp. Conf., 218-224.

[17] Baeb Rian G. Gai (The Thai ABC). Printed by Prachachang, 1966.

[18] Chow, C.K. (1957): "Optimal Character Recognition System using Decision Functions", IRE Trans on Elec. Comp. Vol. EC-6,4, 247-254.

[19] Highleyman, W.H. (1961): "A note on Optimum Pattern Recognition Systems", IRE Trans. on Elec. Comp., Vol. EC-10, 2, 287-288.

[20] Flores, I. (1958): "An Optimum Character Recognition System using Decision Functions", IRE Trans. on Elec. Comp., Vol. EC-7,2,180.

[21] Chu, J.T. (1965): "Optimal Decision Functions for Computer Character Recognition", J. of Assoc. Comp. Mach., Vol. 12, 213.

[22] Chow, C.K. (1959): "Comments on Optimum Character Recognition Systems using Decision Functions", IRE Trans. on Elec. Comp. Vol. EC-8,2,230.

[23] Sebestyen, G. (1962): Decision-Making Processes in Pattern Recognition, New York: Macmillan, 42.

[24] Horwitz, L.P. and Shelton, G.L. (1961): "Pattern Recognition using Autocorrelation", Proc. IRE, Vol. 49,1, 121-128.

[25] Clowes, M.B. (1962): "The Use of Multiple Autocorrelation in Character Recognition", Optical Char. Recog., Spartan Books, McGregor and Werner, Inc. Washington 12, D.C., 305-318.

[26] Clowes, M.B. and Parks, J.R. (1961): "A New Technique in Automatic Character Recognition", Brit. Comp. Journal, Vol. 4,1, 121-128.

[27] Minneman,M.J. (1966): "Handwritten Character Recognition Employing Topology, Cross-Correlation and Decision Theory", IEEE Trans. on Systems Sc. and Cybernetics, Vol. SSC-2, 2, 86-96.

[28] Parks, J.R., Elliot, J.R. and Gowin, G. (1968): "Simulation of an Alpha-Numeric Character Recognition System for Unsegmented Low Quality Print", IEE Conf. Publication 42, 95-105.

[29] Greanias, E.C., Meagher, P.F., Norman, R.J., and Essinger, P. (1963): "The Recognition of Numerals by Contour Analysis", IBM Journal of Res. and Devel., Vol. 7,1, 14-21.

[30] Guiliano, V.E., Jones, P.E., Kimball, G.E., Meyer, R.F. and Stein, B.A. (1961): "Automatic Pattern Recognition by a Gestalt Method", Inf. and Control, Vol.4, 332-345.

[31] Alt, F.M. (1962): "Digital Pattern Recognition by Moments", Optical Char. Recog., Spartan Books, McGregor and Werner, Inc., Washington 12, D.C. 153-179.

[32] Hu, M. (1961): "Pattern Recognition by Moment Invariants", Proc. IRE, Vol. 49,9, 1428.

[33] Kullback, S. (1959): Information Theory and Statistics, New York. Wiley, p 195.

[34] Marill, T. and Green, D.M. (1963): "On the Effectiveness of Receptors in Recognition Systems", IEEE Trans. on Inf. Theory, Vol. IT-9,1, 11-17.

[35] Lewis, P.M. (1962): "The Characteristic Selection Problem in a Recognition System", IRE Trans. on Inf. Theory, Vol. IT-8,2, 171-178.

[36] Kullback, S. and Leibler, R.A. (1951): "On Information and Sufficiency", Annals of Math. Stat., Vol. 22,1, 79-86.

[37] Kamensky, L.A. and Liu, C.N. (1963): "Computer-Automated Design of Multifont Print Recognition Logic", IBM Journal of Res. and Devel., Vol.7,1, 2-13.

[38] Morrison, D.F. (1967): Multivariate Statistical Methods, McGraw and Hill, 221-258.

[39] Marill, T. and Green, D.M. (1960):"Statistical Recognition Function and the Design of Pattern Recognisers." IRE Trans. on Elec. Comp. Vol. EC-9,4,472-477.

[40] Highleyman, W.H. (1962): "Linear Decision Functions with Application to Character Recognition", Optical Char. Recog., Spartan Books, McGregor and Werner, Inc., Washington 12 D.C., 249-285.

[41) Parzen, E. (1962): "On the Estimation of a Probability Density Function and Mode".Annals Math. Stat., Vol. 33,3, 1065-1076.

[42] Murthy, V.K. (1965): "Non-Parametric Estimation of Multivariate Densities with Applications", Proc. Int. Symp. on Multivariate Anal., Dayton, Ohio, 43-56.

[43] Cacoullous, T. (1966): "Estimation of a Multivariate Density", Annals of Inst. Stat. Math., Vol.18,2, 179-189.

[44] Specht, D.F. (1967): "Generation of Polynomial Discriminant Functions for Pattern Recognition", IEEE Trans. on Elec. Comp., Vol. EC-16,3, 308-319.

[45] Specht, D.F. (1967): "Vector-cardiographic Diagnosis using the Polynomial Discriminant Method of Pattern Recognition", IEEE Trans. on Bio-Medical Eng., Vol.BME-14,2,90-95.

[46] Edwards, A.W. and Chambers, R.L. (1964): "Can A Priori Probabilities Help in Character Recognition", Journal of ACM, Vol. 11,4, 465-470.