SYSTEM DESIGN - OPTICAL CHARACTER RECOGNITION

WITH WEIGHTED AREA MASKS

by

RONALD FREDERICK RUECKWALD

S.B., Massachusetts Institute of Technology

(1963)

SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

January, 1965

Signature of Author Signature redacted
Department of Mechanical Engineering, January, 1965

Certified by Signature redacted
Thesis Supervisor

Accepted by Signature redacted
Chairman, Departmental Graduate Committee

SYSTEM DESIGN - OPTICAL CHARACTER RECOGNITION
WITH WEIGHTED AREA MASKS

by

RONALD FREDERICK RUECKWALD

Submitted to the Department of Mechanical Engineering in January, 1965, in partial fulfillment of the requirements for the degree of Master of Science.

ABSTRACT

The main factors involved in a system design for a text-reading , optical character recognition machine are presented. These are developed to the point of making feasible a preliminary prototype design.

Mask generation is discussed in depth. An exact method for designing optimized masks is developed and described. In conjunction with an initial mask generation program and a development program to determine the final masks, it utilizes a tabulation of all characters in all combinations.

Formats of publications are discussed. These are divided into four categories of increasingly complex format. These categories fix the requirements for recognition machines.

A method of format analysis of the page to be machine-read is presented. It aligns the text and determines the location and content (text or non-text) of the various areas of information on the page. The results of an experiment to prove the content analysis method are presented. They indicate the potential of the proposed method.

A cursory system design or integration follows. It links the main components and operations discussed in the other chapters.

Thesis Supervisor: D. M. Baumann
Title: Assistant Professor of
Mechanical Engineering

ACKNOWLEDGEMENTS

The author is grateful to Professor Baumann, Assistant Professor of the Department of Mechanical Engineering, for inspiration and guidance during the course of this thesis investigation, and for providing the facilities used. Thanks go also to Dan Kennedy and Cord Ohlenbusch who were helpful in discussing hardware and problems encountered.

The author wishes to thank Mr. H. F. Chevalier of the M.I.T. Press who was very helpful in acquainting the author with formats of publications, and who also loaned many valuable books describing format in detail.

Many thanks go also to my good friend Yvonnë Foster who spent many difficult hours at the typewriter.

This thesis was supported in part by the U.S. Navy, Office of Naval Research, under contract Nonr - 1841(41) and sponsored by the Division of Sponsored Research of M.I.T.

TABLE OF CONTENTS

TABLE OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 The Future of Character Recognition

Human society is filling its own grave with thousands of tons of paper. As time continues, the filling accelerates. The information handling problem has been with us for a long time, since man began, but at no time has it threatened as it does now; at no time has it slowed us as it does now; at no time will it be worse than in the future.

Certainly the problem has been recognized. Computers are efficient in handling as much information as we have given them, but there the problem lies. We cannot give them enough information. The boundary between printed matter and computer or high density storage input is crossed only in a manual fashion,-an operator at a keyboard, except in a few highly specialized cases such as a typed page reader[1], a cash register slip reader[2], a specialized fixed format document reader[3], and others. The restriction is certainly at the conversion from publication to machine input. Very high speed character recognition machines must be developed to burst through this barrier. Consequently, character recognition promises to become an essential development, a development essential to our own development. It shall relieve an already overtaxing burden on information gathering and disseminating facilities.

1.2 Applications of Character Recognition

1.2.1. The Library Problem

Of primary importance to the research effort now occurring is minimizing useless duplication of effort, time, and money. So much information has been produced in the last few years that it is sometimes easier to duplicate research than to try to find a report on it, if the report ever did become generally available. This is, of course, the notorious library problem. Whole libraries of material must be converted into computer input for translating, abstracting, referencing, sorting, and storing it. The following figures for the Library of Congress for the fiscal years shown indicate the magnitude of the problem.

ITEM	1962	1963
volumes & pamphlets	12,534,351	12,752,792
manuscripts (government)	17,989,445	18,610,876
books on magnetic tape	2,864	5,506
bound volumes of newspapers	160,466	156,766
reels of microfilmed newspapers	94,058	112,320
microprint cards	150,955	166,355
microcards	53,418	63,363

The attempt to conserve space is illustrated in the conversion of bound volumes of newspapers to microfilm.

The Library of Congress is searching for new ways to prevent duplication of research effort. Its 1962 annual report notes the following.

> The scientific and technical literature searching service for a fee ($8 an hour), inaugurated late in fiscal 1961, more than proved its merit. The Library's already distinguished collections in these fields were further strengthened during the year when the Library was designated one of 12 Federal Regional Technical Report Centers to receive, on microfilm, all the unclassified technical reports collected by the Armed Services Technical Information Agency. This will add an estimated 25,000 technical reports a year, making the Library's collection of this type of material the most important public source in the country[5].

Clearly this information in a highly retrievable form is priceless, but its conversion to the form of readily accessible, high-density storage is an impossible manual task. The cost of developing a system here is negligible compared to the savings in research expense as our knowledge in printed form expands increasingly. Without such a system the surge of discovery and development of the past century threatens to be mired in its own progress.

1.2.2. Business Records

Business functions may be served well by character recognition. Costs may be reduced by smaller record storage areas possible by reading and encoding information rather than just microfilming it. Records transmitted by any method may be done faster by a reading-encoding-transmitting-receiving-decoding-printing cycle eliminating the typist and minimizing transmission. But here the recognition is the present stumbling block. To date some very special machines have been operated (as above) but these are limited to very special cases. The practicality of encoded data transmission has been shown in "dataphone" service, but all the original information must at some time be converted by a keyboard operator. The savings in time and money for keypunching in accounting and computing systems looks very favorable.

1.3 Character Recognition Methods

1.3.1 Survey

Reference 6 gives a comprehensive description of the attempted methods of character recognition. The methods described are as follows:

1. template matching
2. peephole template matching
3. coordinate description matching
4. characteristic waveform matching
5. vector crossings
6. critical feature analysis
7. curve following recognition
8. transformations of a source pattern (eg. Fourier transform[7]).

Most of these methods are innately troubled by one or more of the requirements for a general purpose, low cost machine:

(a) ability to handle adjacent characters having very small spacing;

(b) uses minimum of on-line computation;

(c) ability to handle change of fonts.

As a result, these methods may be used only in special applications. For example #4 is used to recognize specially designed characters of the American Bankers Association's font E-13B (see Figure 2.2). In this special font the particular factor precluding the method's use in a more general machine is absent. In this example (a) and (c) are unnecessary requirements of the method.

Baumann[8] has proposed a technique of weighted area scanning. This eliminates (b) and is altered in this thesis to accommodate (a).

1.3.2 Weighted Area Scanning

Baumann proposed[8] and later described[9] a method of recognition using weighted area masks to eliminate areas of a character which include low information content to increase reliability of the recognition. The method has two very important advantages over competitors.

The first advantage is that although a large, high-speed computer is requisite for designing the masks initially, the computer is not used in the recognition machine. The mask represents the computer in the actual character recognition machine. Thus, the cost of mask design, i.e. that of computation, is divided over time and the number of machines used, leaving a very small cost.

The second advantage is that horizontal scanning of characters past masks is possible, leaving lateral character alignment unnecessary. It is this advantage which has not worked well in practice. For instance, a demonstrator was built[10] to employ the weighted area mask method in recognizing the ten characters of a special 3x5 matrix font. It was found that the method worked well if the character was projected square on each mask, but did not result in adequate separation levels for actual scanning of the character past the mask. At this time Lemke and Ohlenbusch redesigned the masks by hand using a trial and error technique until the masks accomplished the required function of recognizing a scanned character. It is obvious that an improvement must be made to the mask generation program in order to allow better results for scanning. This improvement must also prevent adjacent characters impinging upon the mask from giving undesired output levels when two characters are each partly on the mask.

1.4 Thesis Objective

The objectives of this thesis are to propose a workable character recognition system for the general case of reading variable-format, variable-font, printed matter. Basic to this proposal are an improvement in the mask generation program to allow reliable outputs for scanning characters as they appear in context, an improvement in the mask generation program to minimize the number of masks required for a given alphabet, and an analysis of formats occurring in the printed matter to be read. Most of the research effort went into these three areas since the last area is basic to all general recognition systems and the first two areas are the only present blocks to the method's actually being used in a reliable, relatively low-cost, general character recognition machine for printed matter.

CHAPTER 2

MASK GENERATION

2.1 Introduction

The original mask generation method for designing masks was proposed when computer operations were far more expensive than they are now. We can now afford to use a far more extensive program to obtain an exact mask rather than using a quick but guessing approach. It has become necessary to do so for two very important reasons, (a) to allow scanning of entire words rather than single characters and (b) to increase separations of distinct character groups in order to increase reliability. The original mask method is described fully in Reference 8 and more generally in Reference 9. A very brief description follows.

The "alphabet surface" was used. This listed for each square of the mask matrix the probability of occurrance of a black square if each character was projected on the mask statically. The program removed the very high and very low probability levels, chose a number of middle levels, broke the surface at these points, and made a mask out of the portions above the split level. Then the characters were scanned across each of the masks so obtained and that one was chosen which showed the best ability to separate the characters into two or three groups.

By making approximations as this method does the separation of groups of characters becomes a very haphazard event. The magnitude of the separation is diminished beyond that required for reliable results. The method is not exact as can be shown by changing certain squares of the mask to result in better group separations. This occurs because it just happens in these cases that the borderline characters are affected as well as some non-critical characters, but the alphabet surface cannot consider these differences because the information has been lumped together. The mask can be improved by trial and error, but the potential is not great since the groups of characters may not have been chosen by the best criterion.

2.1.1 Mask Width

In order for mask output levels to be unaffected by characters adjacent to the character in question even while the character is centered on the mask, the mask width must not be greater than the width of the character plus two times the minimum spacing between characters. This prevents edges of adjacent characters from altering the output level for a particular mask-character arrangement. This requirement indicates that characters must be grouped by width and then masks designed for each group. The characters recognized in the first group need not be considered in the design of masks for the second group since preferential logic may be used to choose the correct output.

2.1.2 The Case Against Separation of Characters

Before masks are designed it must be determined if each character can be reliably separated from adjacent characters with a minimum of computation. Far simpler mask design results if characters can be separated and projected dead on each mask.

The method of transferring the characters from the page or microfilm is by scanning the line with a raster scan using high frequency in the vertical direction and relatively lower frequency in the horizontal direction. The short vertical scans must be of a width to register one to two scans per narrowest vertical stroke (hairline). A scan width of this narrowness, if scanned such that it just touches the edge of the preceeding scan, can pass between two adjacent characters without touching either one if (a) the printing is of high quality, (b) the spacing is exaggerated for clarity as in typewritten copy or in two point leaded type, and (c) there are no kerned italics.

Although separation of characters may be a better method for special applications, this is not always possible in the more general case of library material.

The reasons are that serifs often run together from poor setting, worn type, over-inking, or sliding and that kerned characters such as the "y" and "f" in Figure 2.1 on page 68 extend under or over the adjacent character, blocking any separation scan. Computational methods can solve this problem in most cases, but the amount of computation is considerable. This is to be avoided if possible. Therefore, should the masks be designable under these conditions of scanning to give good reliability, then scanning groups of characters, words, or a whole line is superior to separation of characters before attempted recognition because the computation is done in the design of the masks and is not required as part of the machine.

2.1.3 The Problem of Scanning

The stationary alphabet surface method for generation of masks is acceptable for stationary and centered inspection of each character by the masks. However, scanning is not considered by the surface. If characters vary in spacing and width, there is no simple way to determine when a character is dead on the mask without separating it from its environment. Therefore scanning the whole character past the mask is necessary. Masks chosen by operating on an alphabet surface obtained by superimposing stationary characters become meaningless because output levels for scanning are different from the dead on case except when the character is instantly dead on the mask. When that occurs is not known to the machine.

The result is that characters which are to give a low output signal may trigger the "up" level anyway when it is somewhere off-center. The design of the masks must then consider every position of a character as it is scanned across a mask in order to design a mask which will always give the correct up or down decision even though it looks at the character during its whole scan across the mask.

2.1.4 The Problem of Adjacent Characters

The problem of characters adjacent to a particular character is also a detrimental one. The concentration of character area on a mask is often greater when two halves of adjacent characters are on the mask than when one character alone is dead on the mask. Depending upon configurations of mask and characters, this combination of characters can give a higher signal on a mask than the top group of characters, thereby activating the trigger and indicating a character in the top group. This combination, of course, is not a character.

The result is that words or lines of words scanned past a row of masks designed for scanning give "up" levels at wrong times and cause errors. The mask generation program must therefore take into account all combinations of characters at all scanning positions for all spacings that the characters are likely to have.

2.1.5 Character Location

Rather than designing all masks with scanning and character combinations considered, it is better to design as few masks as possible considering these factors because of the computation involved. A character, once located, can trigger other masks to look at it. Thus, the best solution is to require that somewhere <u>each</u> character be in a high group of characters which have an output level higher than that for any combination of characters. If the trigger level for such masks is higher than any combination level, then these masks, when triggered, can trigger other masks to look at the character at the right time. The former are called <u>trigger</u> masks, the latter <u>inspect</u> masks. The trigger masks must be designed with scanning and character combinations in mind, but the inspect masks may be designed without these factors since their outputs are only sampled when the trigger mask indicated it has a character of a particular group on it.

A delay may be put on the output of each mask and adjusted so that all masks have simultaneous outputs from a particular character scanned across them serially. This simplifies the problem of when to use the output of the inspect masks. A signal from a trigger mask allows instantaneous sampling of the appropriate inspect masks by the logic set up for that purpose to determine the identity of the character.

This requirement adds a means of minimization of the number of masks. Many inspect masks may be eliminated by using the outputs of other trigger masks with the appropriate triggers added. During mask design the outputs from other masks for the character at the instant its trigger mask functions may be compared to those for other characters in order to find a unique arrangement to determine the identity of the character. This eliminates a special mask to do the job at the cost of little or no extra logic.

2.1.6 Test Example

At this point it is instructive to test these ideas in an actual case. The American Banker's Association font E-13B is filled out to a full 7x9 matrix, fourteen-character font in Figure 2.2. This is a good testing ground because the matrix is not large and there are only fourteen characters. For simplification consider combinations of characters such that the characters are separated by one column of the matrix. As a starting point the masks are obtained from the alphabet surface. The masks are then improved square by square by changing a square, checking if the results are desirable, and retaining the change if they are. A result is desirable if the group separation is increased. The output spoken of in the following is the highest level, during a scan, from a photocell situated behind the mask and intercepting all the light passing through the mask. The "top" or "unique" group of characters result in an output above a certain trigger level such that no other characters result in levels this high. It is important

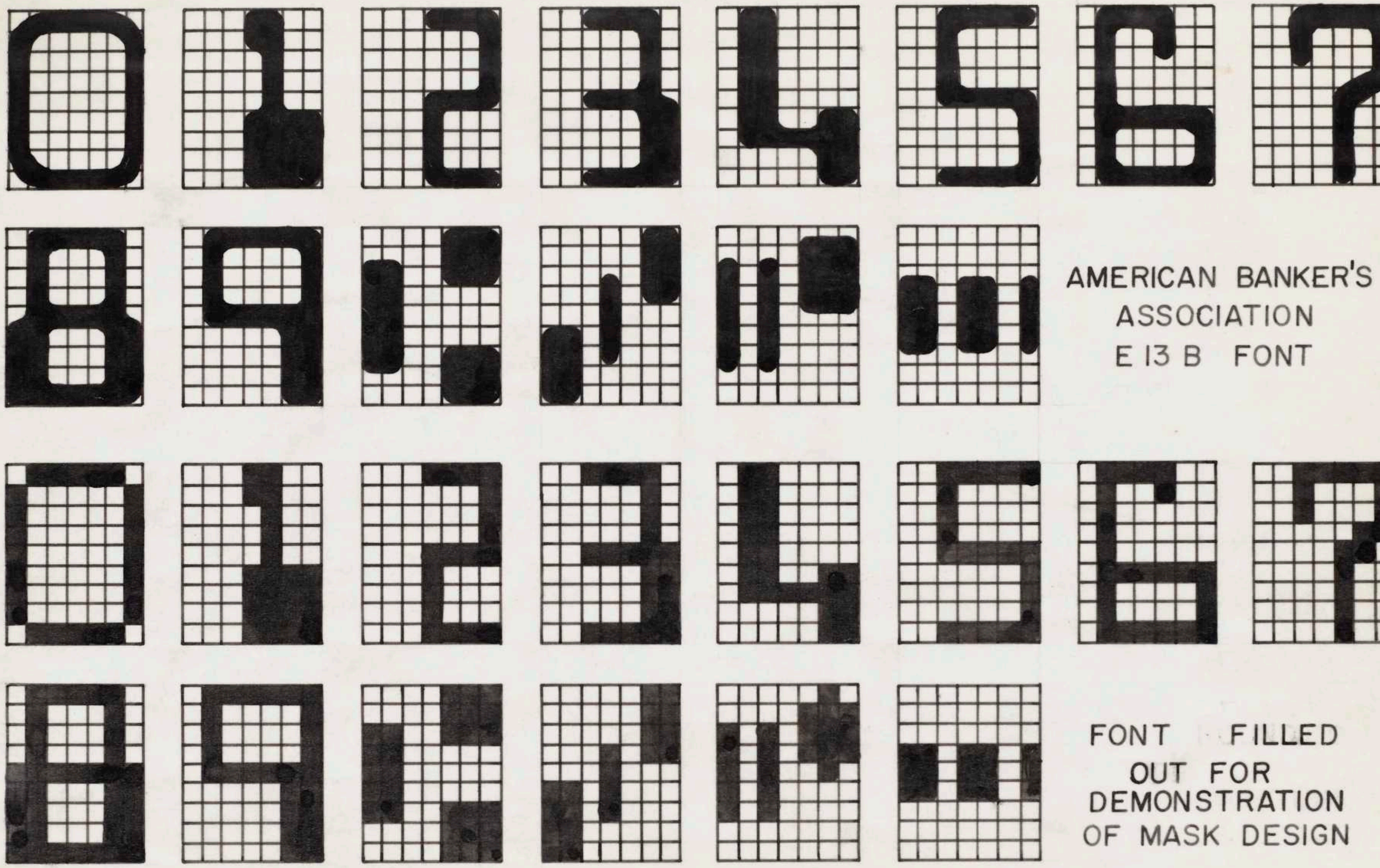


FIG. 2.2 E 13 B FONT & ADAPTATION

to note also that the negative of the font in Figure 2.2 is used in the following masks, so a high output indicates good coincidence of character with mask.

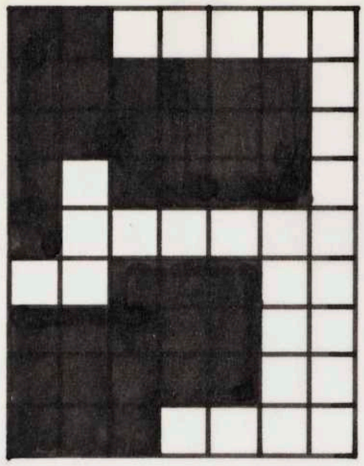
The first mask is initially from an alphabet surface of all the characters. This mask is developed in Figure 2.3. The alphabet surface is split at the 5-6 probability level and the resulting mask is shown with the highest output listed for each character scanned across the mask. The improvement program results in the "improved mask" with the output levels shown. The increase in separation of groups is finite but not spectacular.

The second mask is made from an alphabet surface of all of the characters of the middle group of Mask I. The object is to push as many characters as possible above the highest combination output. See Figure 2.4. The separation of groups is increased slightly again and two characters are maintained above the highest combination level.

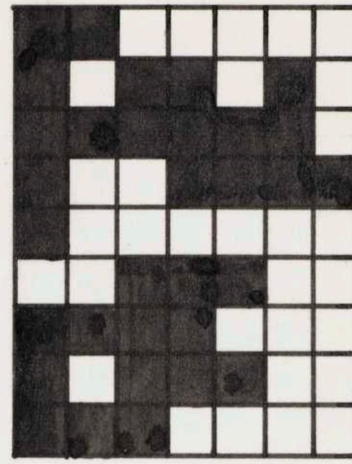
The third mask illustrates that there are further problems. It is made from an alphabet surface of the characters in the lowest group of Mask I. See Figure 2.5. The areas of the characters are so small and dissimilar that no square changing reduces the highest combinations below the chosen top or unique group. It is important to note here that any further change which reduces one combination level below seven does not reduce all of them. There seems to be a threshold below which

0	5	8	9	10	10	6
1	4	3	1	4	5	7
3	5	4	0	3	5	7
4	6	3	2	3	5	7
5	6	6	7	8	7	7
6	6	3	5	4	6	9
5	5	2	4	4	6	9
3	3	0	2	3	6	9
2	4	5	7	9	10	9

ALPHABET SURFACE



INITIAL MASK FROM 5-6 SPLIT



IMPROVED MASK

Character	Initial Mask Output	Improved Mask Output
9	23	22 •
8	23 •	25
highest combination	20 •	18
3	19	19 •
5	17	17
I!	16	17
0	15	14
2	15	13 •
1	14	15
6	14	17
4	13 •	16
7	11 •	10 •
II'	9	8
III	9	8
I'	8	7

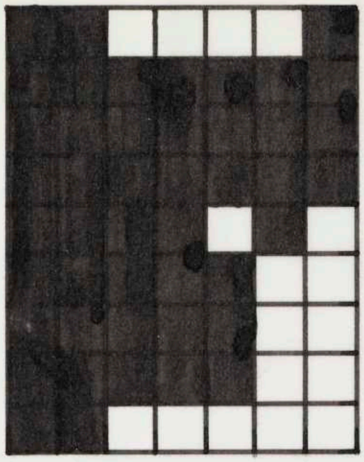
23-20 13-11	Group Separations	22-19 13-10

Source: all characters

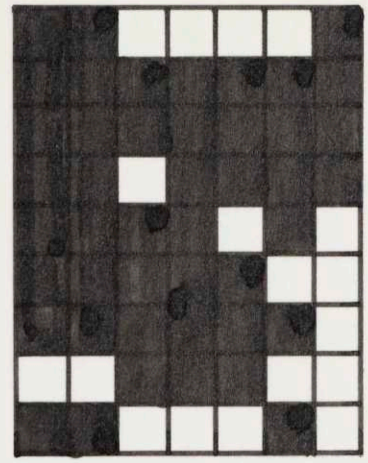
FIG. 2.3 MASK I

0	3	5	6	7	5	3
1	2	2	1	3	2	3
2	3	2	0	2	2	3
2	3	2	0	1	1	2
2	3	3	3	4	3	4
2	3	2	3	2	4	6
2	3	1	3	3	4	7
1	1	0	2	2	4	7
0	2	4	6	7	8	7

ALPHABET SURFACE



INITIAL MASK FROM 3-4 SPLIT



IMPROVED MASK

Character	Initial Mask Output	Improved Mask Output
3	17	15
5	14•	14•
highest combination	14•	13•
1	13	11•
0	12	12
6	12•	12
2	9•	8•
1:	9	7
4	9	8

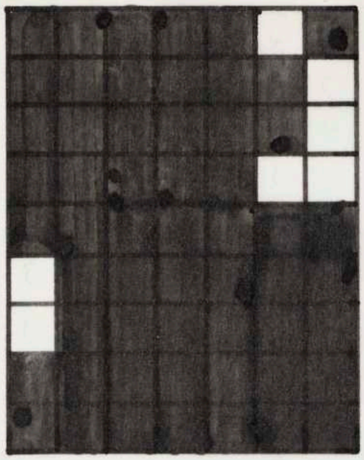
Source: middle group of mask I

14-14	Group	13-14
9-12	Separations	8-11

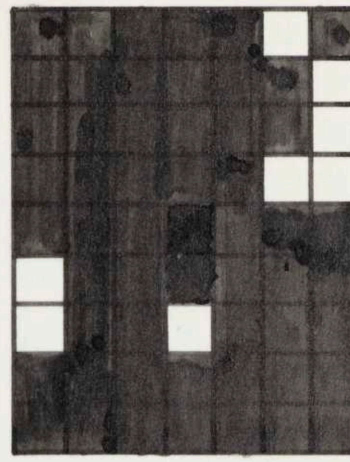
FIG. 2.4 MASK II

0	0	1	1	1	3	2
0	0	1	0	1	2	3
1	0	2	0	1	2	3
2	1	1	2	2	3	4
2	1	1	2	2	2	1
3	2	1	2	2	0	1
3	2	1	1	2	1	1
1	1	0	1	2	1	1
1	1	0	0	2	1	1

ALPHABET SURFACE



INITIAL MASK FROM 2–3 SPLIT



IMPROVED MASK

Character	Initial Mask Output	Improved Mask Output
ı'ˈ	7	8
ıı'	7·	7·
highest combination	7·	7·
7	5	5
ııı	2	2

Source: lowest group from mask I

7–7	Group Separation	7–7

Note: No further separation between groups is possible due to combination levels of 24, 28, 84, 88, ı'ˈ4, 11, 14, 18, etc.

FIG. 2.5 MASK III

the combination levels cannot be reduced. The unique group characters cannot be increased without increasing the critical combination levels.

Another important observation from this mask is that it would have been better to start with only the character "ı'ˈ" in the top group. However, the alphabet surface has no way of telling us this since both of the top two characters have the level of seven initially. This suggests a new grouping criterion is required.

2.1.7 Grouping Criteria

As evidenced by the attempt in improving Mask III above, the group of characters is not similar and therefore defies efforts to put more of them in the unique group. This means more masks are required in this system.

Since the object is to put as many characters as possible into the top or unique group it is best to choose these characters on a similarity basis. Choose the largest character (by area) of a particular width group (Section 2.1.1) and the characters which have the largest areas coincident with it as the top group. The following mask goes through the same procedure but does not consider the characters in the top group of any preceeding mask.

In order to prevent coming out with no workable mask, the top group should consist of the largest character only, then the top two, three and so on. This continues with a mask designed for each case until the highest combination level begins to interfere. The last good mask is chosen as the one minimizing the total number of masks. This means that a workable mask always results.

2.1.8 Extrapolation to Regular Fonts

These requirements have come to light as a result of problems detected while developing a mask generation program with a 3x5 matrix, ten-character font[11] and the 7x9, E-13B font used as test grounds. Do the same problems occur in a regular, sixty-two character font of a 36x36 matrix? A rough extrapolation can be made from these small fonts to the large.

The numerals of the Monotype Roman #36 font are a good bridge for extrapolation. This font is broken down into the 36x36 matrix in Reference 8. The matrix is the smallest rectangle including all of the squares used. All of the numerals are contained in a sub-matrix of 17x25 as shown in Figure 2.6. This is a good bridge for extrapolation to the full, regular font because (a) the number of characters is similar, (b) the characters are of the same kind, numerals, and (c) the ratio of total used squares to total matrix area (sum of character squares divided by the product of the number of characters and the matrix area) is about the same (note Figure 2.6).

Using these facts to indicate similarily among the three samples of numeral fonts we may estimate

	Numerals			Components of Scotch Roman font		
	3 X 5	7 X 9	17X 25	29 X 35	36 X 28	36 X 36
	Baumann's Font	E-13B Font	Monotype Scotch Roman # 36 Font			
1 Kind of Characters	Special Numerals	Special Numerals	Numerals	Lower Case	Upper Case	Full Font
2 Number of Characters	10	14	10	26	26	62
3 Smallest Character Area / Matrix Area	.33	.24	.29	.10	.17	.07
4 Largest Character Area / Matrix Area	.87	.59	.48	.22	.37	.29
5 Smallest Character Area / Largest Character Area	.38	.41	.59	.43	.45	.24
6 Total Character Area / Total Matrix Area	.67	.36	.41	.15	.25	.15
7 $\alpha = \frac{\text{Largest} - \text{Smallest}}{(\text{No. Char.}) \times (\text{Mat. Area})}$	.054	.025	.020	.005	.008	.004

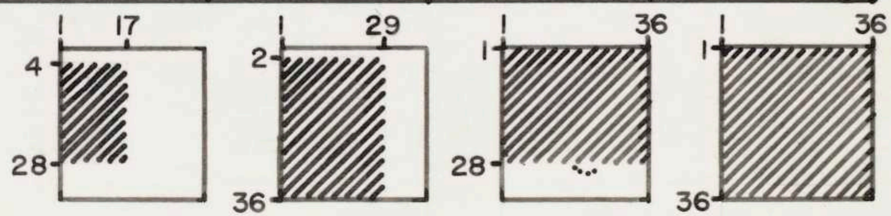


FIG. 2.6 EXTRAPOLATION DATA

the difference in effectiveness expected from any recognition technique by looking at the ratio of the smallest character area to the largest character area. This, as shown in Figure 2.6, is greater (.59) than for the two special fonts (.38 and .41) indicating more difficulty expected in separation since separation groups may be expected to be closer together. An additional factor leading in the same direction is that the regular font numerals were not designed to make recognition easy as were the first two fonts. Also, character spacing is closer in normal fonts. However, a factor which helps to push in the opposite direction is that the much smaller squares of the matrix allow excellent definition of the more continuous curvatures of the regular font. Overall, therefore, the difficulty may be expected to increase with regular numerals, but not significantly.

Now, extrapolation to the full font, letters of upper and lower case included with the numerals, may be attempted. Here it is instructive to look at Rows 1,2, and 5 of Figure 2.6. Row 5 shows that the variation of size is about the same for the 7x9 matrix font as for the upper and lower case of the regular font. However, the upper and lower case sets of characters have 2.6 times as many characters as well as having them similar, i.e. the "h" backbone of the alphabet as mentioned by Baumann[12]. This means that separation of characters of a large font is far more difficult and that separation gaps between output levels of groups of characters will be very small indeed, if existent at all. The same thing is shown more clearly in Row 7 of Figure 2.6. This alpha represents an average change of area from the smallest to largest character and is divided by the matrix to make it dimensionless. The 7x9 matrix font is seen to have an alpha number only half as large as the 3x5 font. The tendency of group separation levels to diminish increases as the font increases its number of characters.

The results of this extrapolation indicate that the mask generation technique must be more exact for the regular font than for the small-matrix fonts, as expected. Since approximate methods do not work well for the small fonts, an exact or nearly exact mask generation method is required to give reliable results, acceptable group separations, and minimum number of masks.

2.1.9 Summary of Requirements

The requirements for the mask generation method that have been determined up to this point are the following:

(a) Group characters initially by width.

(b) Group largest and most similar characters together for unique groups.

(c) Scan whole words or lines of characters.

(d) Tabulate scanning and combination results for all characters, combinations, and combination spacings.

(e) Require every character to be in a unique group above the highest combination level at some mask.

2.2 Mask Generation Method

2,2,1 Rationale

The mask generation is an exact mask optimizing method. The only disadvantage with it is the large amount of computation required. The method first categorizes all characters into "isowidth" groups, i.e. groups comprising characters of the same width. Next it takes the largest of these characters and finds the characters most similar to it. It begins with that one character as a mask and optimizes the mask to the fullest extent. Once having obtained this mask, the method continues by trying to include the next most similar character, by making the initial mask identical to the sum of the two characters on the matrix, and then by optimizing the mask. These attempts to include more similar characters continue until the output levels of scanned combinations or other characters begin interfering with the group separation levels and prevent the chosen characters from being unique to the trigger level. Thus, the maximum number of characters results as the unique group and acceptable trigger level tolerances are still maintained.

Once this is done the program begins the next mask, continuing in the same manner and no longer considering the characters already in unique groups of previous masks. This is done until every character is in a unique group at some mask. The rest of the program becomes a matter of determining each character output dead on each mask of this width group and categorizing these results to give a unique logic tree for every character. In some cases where a unique set of outputs does not result, another mask may be designed very easily without considering scanning and combinations of characters. Since this latter type of mask is easily designed for unique separation of many characters. (It is "looked at" at one instant in time when a unique group triggers.) A few masks of this type are preferable to complex logic used with outputs of the previous masks. However, the two kinds should be used so as to minimize the complexity of the machine.

At this point the program continues the same method on the next smaller width group until all are done in each group.

2.2.2 Mask Output Table

The most important part of the mask generation program is the mask output table. This table keeps a running tabulation of mask outputs for scanned characters and combinations, with varying separations, as the mask is being changed by the program. The program operates on the mask according to the levels listed in the table.

The following definitions and boundaries must be set, and following these the mask output table for the general case appears in Figure 2.7.

-Let the first character of any combination be represented by A.

-Let the second character of any combination be represented by B.

-Let the number of characters in the alphabet be N.

The character combination is therefore AB where B becomes each of N different characters for every A as it varies from 1 to N.

-Let the correlation (in squares) of scanned character A with the mask be A_p where the character A is at the scanning position p on the mask.

-Let the correlation of scanned character B with the mask be B_p where the character B is at the scanning position p on the mask.

-Define p as the number of columns between the trailing edge (right edge) of the scanned character and the similar edge (right edge) of the stationary mask where p is positive after the character has passed a dead on (p=0) position.

-Let the output (in squares of correlation of character to mask) of a given mask for a given character combination at a given position (A_{p_A} B_{p_B}) on the mask be L.

-Let the width of the mask (in matrix columns) be m.

-Let the width of the character (in matrix columns used) be n.

-Let the separation (in unused columns) between adjacent areas of the characters in a particular combination be s.

This s varies, in single matrix increments, from the smallest spacing to be found in the printed application to (m-1), the largest separation allowing both characters to have at least one column on the mask of width m. s is always positive.

Now, for a mask of width m

$$L = A_{p_A} + B_{p_B}$$

where either term is zero if its p $\geq$ m, or its p $\leq$ -n, that is, when the combination is off the mask. But,

$$p_B = p_A - (n_B + s)$$

The position of B is the position of A translated backwards (to right) the width of character B plus the spacing between the characters. Therefore,

$$L = A_{p_A} + B_{p_A - n - s}$$

Note that only combinations are scanned past the mask, but that single character levels are obtained when A is dead on the mask, i.e. $p_A = 0$

$$L_A = A_{p=0} + B_{0-n-s}$$

$$\text{but } B_{p \leq n} = 0$$

$$\text{therefore } L_A = A_{p=0}$$

Now a complete set of N^2 tables for every combination of AB can be constructed for L versus p_A and s over the full range of s.

$$s_{minimum} \leq s \leq m-1$$

The range of p_A is

$$0 \leq p_A \leq (m-1) + s_{minimum}$$

(The addition of $s_{minimum}$ allows a record of the output of every character in every position where the spacing of characters is greater th n zero.) The table follows in Figure 2.7. Note that all entries may be omitted from the table when $s > p_A$ except for the row $s_{minimum}$ where all entries are made. Note that L for the first character dead on the mask is given at $p_A=0$. This level must be maximized with respect to all others if that first character A is in the unique group of a mask. This table along with the tables for all other combinations give all the levels, making it easy to find what levels are limiting the separation of groups. Not only do these tables give the limiting levels, they also show what characters at what positions are giving these levels so that the best mask change can result. The mask output table for the general case follows on the next page.

$s \backslash p_A$	0	1	2	3
$0_{=s_{min.}}$	A_0	A_1+B_{1-n}	A_2+B_{2-n}	A_3+B_{3-n}
1		A_1	A_2+B_{1-n}	A_3+B_{2-n}
2			A_2	A_3+B_{1-n}
3				A_3
∘ ∘ ∘				
s				
∘ ∘ ∘				
m-3				
m-2				
m-1				

Combination: AB

FIG. 2.7 MA

· ·	p_A	· · ·	$m-3+s_{min.}$	$m-2+s_{min.}$	$m-1+s_{min.}$
· ·	A_p+B_{p-n}	· · ·	$A_{m-3}+B_{m-n-3}$	$A_{m-2}+B_{m-n-2}$	$A_{m-1}+B_{m-n-1}$
· ·	A_p+B_{p-n-1}	· · ·	$A_{m-3}+B_{m-n-4}$	$A_{m-2}+B_{m-n-3}$	$A_{m-1}+B_{m-n-2}$
· ·	A_p+B_{p-n-2}	· · ·	$A_{m-3}+B_{m-n-5}$	$A_{m-2}+B_{m-n-4}$	$A_{m-1}+B_{m-n-3}$
· ·	A_p+B_{p-n-3}	· · ·	$A_{m-3}+B_{m-n-6}$	$A_{m-2}+B_{m-n-5}$	$A_{m-1}+B_{m-n-4}$
· ·	· · ·	· · ·	· · ·	· · ·	· · ·
	$A_{p_A}+B_{p_A-n-s}$	· · ·	$A_{m-3}+B_{m-3-n-s}$	$A_{m-2}+B_{m-2-n-s}$	$A_{m-1}+B_{m-1-n-s}$
			· · ·	· · ·	· · ·
			A_{m-3}	$A_{m-2}+B_{1-n}$	$A_{m-1}+B_{2-n}$
				A_{m-2}	$A_{m-1}B_{1-n}$
					A_{m-1}

$$L=A_{p_A}+B_{p_A-n-s}$$

OUTPUT TABLE

2.2.3 Program

The program for generating the best weighted area masks proceeds as follows:

Step 1. Separate characters into size categories by width according to the criterion $n \leq m \leq n + 2s_{minimum}$. Start with the largest first. For each size category run the mask generation program as follows.

Step 2. Categorize similarities in each width group as follows:

(a) List areas of characters not in any previous unique group.

(b) Pick the largest of these.

(c) List the other characters in decreasing order of area common to the largest character.

(d) Choose the unique group of characters by taking the top u+1 characters where u=number of characters in the unique group for the previous run with the identical list in 2(a). (Note that this means the top character is chosen first and a mask is designed. Then the top two are chosen for the second attempt, and so on until the optimum mask is reached just before the highest combination level interferes with the separation of the unique group.)

Step 3. The initial mask is generated very simply as follows:

(a) Establish an address table for every square in the matrix.

(b) Record each unique group character in every address representing a matrix square with which the character is coincident in the static, dead-on position.

(c) Wherever there is an entry in the table of addresses in 3(b) make the mask the same as the print, i.e. clear if negative transparency is used, or opaque if black print is used. Use the former, clear, for demonstration here.

<u>Step 4.</u> Construct the N^2 mask output tables according to the algorithm $L = A_{p_A} + B_{p_A} - n - s$ as described in Section 2.2.2.

<u>Step 5.</u> Develop the mask as follows:

(a) In decreasing order, record L for all unique group characters at s=p=0, i.e. dead on mask. (These are to be maintained high relative to the nearest other level.) Let the highest be 1, the next 2, and so forth down to u.

(b) Record clear mask square which are not common to all characters having the same L as u. (This preserves squares holding up the L of the lowest characters in the unique group.)

(c) Search the mask output tables from Step 4 for the highest L below that of u.

(d) List identity and position of all combinations having the same L as 5(c).

(e) Blacken one clear mask square which is common to all positional configurations of 5(d) and which is recorded in 5(b). Only if 5(d) has no squares recorded may any square common to 5(d) alone be blackened.

(f) Reduce levels in all tables of Step 4 for every L affected by the mask change.

(g) Return to 5(b) and continue until no further changes are possible or until the minimum desired signal-to-noise ratio is reached.

(h) Print out mask and levels.

(i) Return to Step 2(d) to try to increase size of unique group if there is still an acceptable separation of unique group levels from other levels. If separation has become unacceptable, choose the best mask as the last one obtained that has acceptable group separation.

(j) Continue with following masks for this particular width group at Step 2(a). When these are finished, start at the next narrowest group at Step 1 and continue until all characters have been in a unique group.

Step 6. When all the masks are designed for unique groups, the L for each character at p=0 is found for every mask of its own and smaller width groups. These are used with the appropriate triggers as further inspection stations "looked at" when the unique group

L triggers the logic. If not enough levels are present to determine the identity of each character, additional masks for static, dead-on inspection may be easily designed.

2.2.4 Example of Program Applied

Assume that the alphabet consists of the fourteen characters of the E-13B font filled out as in Figure 2.2.

Step 1. The program categorizes the characters into groups of equal n. For the E-13B font the widths are as follows:

4 columns (minimum)	--	2 characters
5 "	--	3 "
6 "	--	3 "
7 " (maximum)	--	6 "

But s varies in the font because the seven-column matrices of each character are maintained at u constant separation ± a tolerance; but character widths vary on these constant width matrices. Therefore the s changes. This turns out in practice to vary as

$$3 \leq s \leq 7$$

but m=7 $\therefore S_{max.} = m-1 = 6$ for the program

Now $n_{min.} = 4$

$$n_{min.} + 2s_{min.} = 4 + 2 \times 3 = 10 > m - 1$$

or alternatively $(m - 1) - 2s_{min.} = 6 - 2 \times 3 = 0 < n_{min.}$

All characters may be considered in one width group since no more than two characters can ever be partially on one mask simultaneously. Therefore, for Step 2 the input is all fourteen characters of the alphabet in this special case.

Step 2. The program counts the squares of unit area of each character as in the first and second columns of Figure 2.8. The largest, "8", is chosen as the first mask.

Let us suppose that the program has recycled to this point, having obtained an optimized mask for "8" in the unique group and again for "8" and "3" with combination levels still low enough to warrant a further attempt at increasing the unique group's size. Column three of Figure 2.8 shows the areas of all characters coincident with a mask of all areas of all characters in the previous "8", "3". The next highest is 22 for "9". Therefore for this run choose "8", "3", & "9". as the unique group.

Step 3.

(a)&(b) A unique group address table is made up showing which characters of this group fall in which squares.

(c) The mask is chosen as the sum of "8", "3", and "9". This is shown in Figure 2.10. (Note: we are using the negative of the E-13B font in order to speak of "up" signals meaning highest coincidence of character to mask.) Obviously, no other mask will result in a higher signal for any of the unique group

Character	Character Area (Matrix Squares)		Area Coincident With 8&3	Area Coincident With 8,3,&9
0	24		18	21
1	23		13	13
2	18		11	15
3	23		23	23
4	24		17	17
5	21		17	18
6	25		21	21
7	17		8	12
8	37	largest	37	37
9	26		22	26
	28		19	22
	20		13	17
	21		9	12
	15		9	10

FIG. 2.8 GROUPING CRITERIA

	Character	
p	8	0
-6	5	4
-5	10	8
-4	13	4
-3	10	6
-2	13	8
-1	22	16
0	37	21
1	26	17
2	14	9
3	11	7
4	14	5
5	14	9
6	5	7

FIG. 2.9 SCAN OUTPUTS

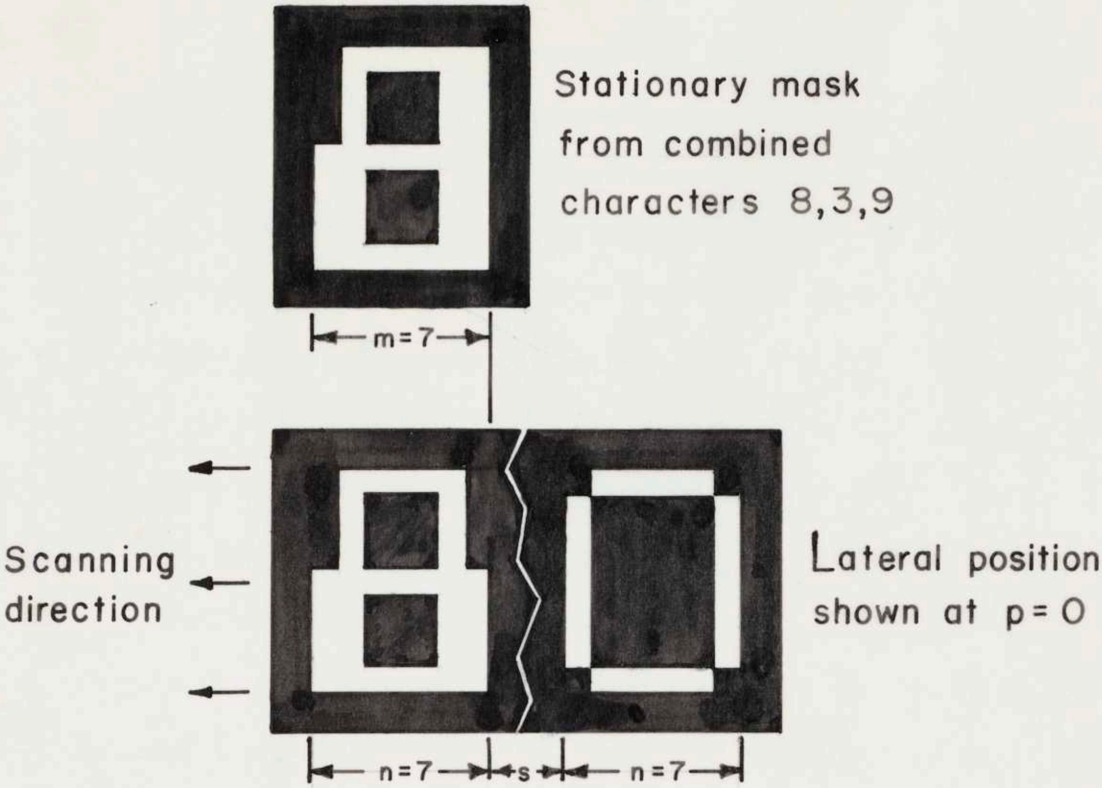


FIG. 2.10 CONFIGURATION

Combination: 80 $L = 8_p + 0_{p-7-s}$

S \ P	0	1	2	3	4	5	6	7	8	9
3	37	26	14	11	14+4= 18	14+8= 22	5+4= 9	0+6= 6	0+8= 8	0+16= 16
4					14	14+4= 18	5+13= 13	0+4= 4	0+6= 6	0+8= 8
5						14	5+4= 9	0+8= 8	0+4= 4	0+6= 6
6							5	0+4= 4	0+8= 8	0+4 4

FIG. 2.11 L FOR COMBINATION 80

characters.

Step 4.

(a) The output for all characters is recorded from the mask for every position $-(n-1) \leq p \leq m-1$. We shall continue only for the characters 8 and 0 for simplicity in demonstration. Here the range is $6 \leq p \leq 9$. See Figure 2.9 for this tabulation.

(b) The table is constructed for range of $s_{min.} \leq s \leq m-1$, i.e. $3 \leq s \leq 6$ and $0 \leq p \leq m-1+s_{min.}$, i.e. $0 \leq p \leq 9$. This allows every combination and every character to be on the mask in all positions of scan. The table is shown in Figure 2.11.

Step 5.

(a) The entry on this table which is important to keep maximum is 37. The program continues by removing a clear square on the mask which does not lower the output of the lowest unique group character but which does lower the output of the highest character or combination not in that unique group at the position in which it gives the troublesome level.

This gives an indication of the way the program works but it is a far too complicated method to continue here without a high-speed computer. However, for checking purposes, the sample was simplified to the 3x5 matrix, ten-character font where no combinations were considered. The program was tried on this and succeeded in turning out a mask identical to the best mask designed by hand for this simplified case by Lemke and Ohlenbusch[10].

2.2.5 Comments on the Program

The program considers all important factors not considered in previous methods. It is a workable solution to the problems hindering progress so far. A successful machine can be built today using it.

For future development and simplification there are a number of considerations. Masks may be designed weighting certain areas common to all fonts and minimizing effects of variable areas such as thicknesses and serifs. This allows a number of fonts to be recognized by a given set of masks.

Another improvement is to design masks of all mask sets such that the levels are all proportional between sets so that trigger levels need not be individually changed whenever a mask set is changed. If possible the levels should be the same at a mask position irrespective of the mask set used in order to simplify the electronics. The same logic arrangements should be used for each mask set as much as possible to simplify the machine.

Simplification of the program through reduction of the number of levels that must be considered results by timing characters as they move away from the position in which they are recognized. When recognized, the character width is also known. Therefore the levels of all the masks may be omitted until it is known that the following character is about to appear. This works well into the general mask output table illustrated in Figure 2.7 and should be considered for masks that are easier to design but not, in any likelihood, better than those obtained from the program described.

The amount of calculation for the program is considerable. An order of magnitude calculation gives the number of successive AND operations required for the development of one optimized mask equal to the square of the number of characters times the matrix area (in squares). For a 36x36 matrix and a one hundred character font, this results in about 10^7 operations, or 10^4 for the E-13B font.

CHAPTER 3

FORMAT CATEGORIES

Before an actual system design can be proposed for a character recognition machine to read general printed matter, a study must be made of the requirements for it as dictated by the material to be read. It has been stated that the greatest area of application for character recognition is in placing whole libraries in high density storage for subsequent library automation. This machine must therefore be capable of reading most of the printed material submitted to the library as well as that which is already in it.

A library contains the full gamut of publications from advertising broadsides to the most scientific treatises. It is clear, however, that the former is being preserved _per se_ and the latter is being preserved for reference as a building block in the furtherment of research. The largest cost saving will be a result of easing the access problem, for the latter and not the former. Thus the main interest is in automatic reading of printed documents containing highly referenced information, and only these items are considered here. These, of course, include periodic publications as well as single reports or texts on all subjects from government proceedings to nose cone ablation studies. The forms of these publications fall naturally into four distinct categories requiring increased sophistication of format analysis methods.

3.1 Simple Fixed Format

This type of publication includes reports, pamphlets and books generally of the novel type untilizing only text of two or three fonts. Simple fixed format describes publications which have every page arranged in a single block of print with paragraph indentations but rarely having spaces between paragraphs. A running headline* is the rule with exceptions in some cases where section or page headlines occur. Folios* occur at the top outside or at the bottom of the page. Page lengths are constant and breaks occur at the chapter ends.

Special complicating factors of this category are as follows:

1. Chapter headings and the first character of the paragraph may be of a large and highly stylized

---

* See definition of term in Appendix to Chapter 3, Section 3.6.

font.

2. Various but seldom occurring complications such as pictures and drawings do appear but are usually no more frequent than two to ten occurrances per volumn.
3. The first few pages preceeding the body of the text contain the title, a table of contents, a preface, foreword, and dedication which are not as simple and regular as the main body of the text.
4. Footnotes may appear very infrequently.

3.2 Complex Fixed Format

This type of publication includes the voluminous and very important sets of abstracts, abstract indexes, patent lists, and the like which use a rigid format for tens of thousands of pages. The format may be different for each title. Generally these all use very small print, are arranged in two or more columns, and have a running headline. Very few font changes are made and very few breaks occur in the columns.

Special complicating factors of this category are as follows:

1. Often a dash is used to indicate the same title or author's name listed in the preceeding statement. Similarly, as in Chemical Abstracts patent listings, recurring parts of a complete number are eliminated and only the last three

or four changing digits are given except where one of the eliminated digits changes. In that case all digits are given.

2. Frequently equations appear. Fewer diagrams appear. Graphs or tables appear very infrequently.
3. Sometimes index devices appear in the column spacings and these are in no way related to the text. For example, Chemical Abstracts uses lower case letters "a"-"h" evenly spaced down the center margin. These could make a format analyzer more complex.

3.3 Complex Continuous Format

This type of publication includes reports or texts on scientific or engineering subjects plus the body of periodicals in which advertisements have been segregated to their own special section and in which every article is continuous. An example is any I.R.E. journal. The headline and folios appear normally, similar to the simple fixed format category, but the area reserved for the main body of text differs with many complex additions. Rarely do two columns exist.

Special complicating factors of this category all appear in the main body. They are as follows:

1. Equations may occur very frequently as part of a line of text or in a separate line.
2. Pictures, graphs, and tables may occur very frequently and may be full page size and/or

may be rotated 90°.

3. Footnotes may appear frequently as may appear references at the end of each chapter.

4. Many special symbols may be used frequently.

3.4 Complex Discontinuous Format

This type of publication includes periodicals which allow intermixing of portions of articles with each other and with advertisements. Articles are usually printed in columns and may be continued in a distant part of the publication. The format may take any form and may or may not have a headline at the top of the page. Often columns are printed close together and separated by thin lines.

3.5 Conclusion

The page format analysis scanning and logic systems for the last and most complex format will obviously take care of the three simpler cases with no trouble. This is the general system. However, the bulk of the material which is referenced most often in the library is that in the third category of complex continuous format and the second category . For this reason the page format analysis method developed in Chapter 5 is for the first three categories. It is only a matter of adding logic and some storage to the method to make it work for the fourth category also. That complication is felt unnecessary at this time since the main interest in character recognition does not include that specific application. However, the method may still be termed general since the bulk of the important material in the library may be read by it.

It is wiser to start with a limited machine than a general one. For this reason a much simplified operator-monitored character recognition machine is briefly described in the Appendix to Chapter 5. Chapter 4 continues with the method for analyzing the kind of material found on the page.

3.6 Appendix - Definitions

"A running headline is a headline placed at the top of each page of a book, usually giving the main title of the work on the left-hand (verso) page and the title of the chapter or other subdivision on the right-hand (recto) page."[13]

A folio is "...a page number, usually placed at the outside of the running head at the top of the page. If placed at the bottom of the page the number is a drop folio."[14]

CHAPTER 4

FORMAT ANALYSIS

4.1 Introduction

In order to read the printed page a number of parameters must be known for guiding the scanning device. In this case a flying spot scanner is used. The raster scan covering the characters in a line of text must be placed accurately. This requires determining by preliminary scans:

(a) the type of content and its location, i.e. whether text, picture, graph, complex equation or table;

(b) the alignment of the text;

(c) the size of the text; and

(d) the spacing of the text.

Whether or not a separate scanner is used for this format analysis scanning depends entirely upon the economics of the system. For simplicity of a later system diagram, one scanner is used first in a format analysis mode and then in a read mode.

The amount of this format preserved with the text for later print-out is an important question. Is the print-out required to have the same format as the original page? Since the primary purpose of the character recognition machine is to change printed information to coded information for automatic abstracting and translation, and since the machine cannot determine the content of pictures, graphs, etc., then few format control functions are deemed necessary. This allows the printable portion of the material being read to be wholly compatible with high speed print-out machines.

However, the non-text items of a page are still encoded by different means for facsimile transmission. Each non-text item is given a reference number which is included with the text. A separate booklet of picture copy is made up as the page is being read. A copy of this booklet is transmitted by facsimile transmission or mailed if requested after the copy of the text of the article or book has been sent quickly. The text of an article is often times sufficient to satisfy the request for information.

Reference 17 describes the problems of attempting to read complex equations by any method. The conclusion is that equations accupying more than a single line of characters, similar to text, must be treated in exactly the same manner as non-text. With this and the preceeding decisions in mind, the methods for obtaining the desired format parameters, may be discussed.

4.2 Alignment

Alignment of the scan axes with the printed lines of text and columns must be done preparatory to format analysis. This is accomplished by first finding a long line of text using the content analyzer, described in a subsequent section, and then scanning down to and aligning with the base line of that line of text. It is assumed that all other lines are parallel to that one chosen, and that lines are straight over their whole length. This is true in all but the worst quality of printing.

The alignment for the whole page proceeds as follows. The scan is moved downward with small increments between scans. The output signal is inspected for significant drop-off in magnitude which may be expected when the base line is crossed because of the very different character density above and below the base line. This inspection is of a sensitivity to note a drop-off for one-tenth of the scan or more. When this drop-off occurs, the scan is rotated until there is no more drop-off. This is done by adding an increasing or decreasing voltage ramp function to the scan control voltage such that the high end of the ramp is applied to the end of the scan which is below the base line. This puts the scan above the base line again and the process is continued, but with smaller increments of scanning and of ramp voltage until the base line is reached along a large fraction of the scan. This predetermined fraction depends upon how accurate alignment is necessary and how sharp a base line is expected to be.

The ramp additions to the vertical deflection voltage are maintained for each horizontal scan during the preliminary and intensive analyses of the page. This method of changing the orientation of the scan is much faster than attempting to reorient the printed specimen itself.

4.3 Format Analysis - Horizontal Scans

Horizontal and vertical scans spaced evenly across the page determine all the information necessary for controlling the reader. The scans use the same size dot as the reading raster scan. This is smaller in diameter than the width of the thinnest stroke of the characters on the page, a fairly standard number for clarity reasons. The horizontal scans are spaced such that roughly five hundred fall on the page. This allows at least five scans through an upper case character, the number required for identification of a line of type as such and for finding spaces between lines.

4.3.1 Content Analysis

It is obvious that a character reader would have great difficulty in trying to read a picture as if it were a character. Pictures and linecuts, graphs, tables, and complex formulae require a different method for recording. These must be removed from the area of the specimen which is to be intensively scanned as text.

This implies that the machine must first determine which is text and which is not. This is not a simple thing to do for a machine which can only look at a very narrow line of data. However, there are three characteristics of a scan through text which may be exploited separately or together. They are all of a periodic nature which typifies text. These are the following:

(a) Text is by definition in parallel lines, or at least in one line with some blank space above and below. This is obvious and need not be proved, but it is not sufficient to act as the criterion for separating non-text from text; parts of various pictures or drawings may satisfy this criterion also.

(b) Text, when scanned by a flying spot scanner at a given rate, exhibits predominant or characteristic frequencies due to the periodic occurance of the vertical strokes of the characters. Similar scans of non-text, with the exception of equations, indicate that even for repetative photographs and linecuts chosen to disprove the postulate, the frequency components are not as great in the characteristic frequency range of text.

(c) Text, when scanned, results in a primarily binary output, especially when scanned through the middle of the characters. Photographs and linecuts, having generally finer lines, or dots, exhibit an output that varies more randomly than that of text where the scan dot size is on the order of half of the thinnest line in a font.

These characteristics allow identification of what type of printed matter the scan is passing through. Items (b) and (c) are not immediately obvious and are therefore proved by results obtained from an experiment using a simple flying spot scanner system assembled in the laboratory for that purpose. The experiment and results are described and discussed in the Appendix to Chapter 4 in Section 4.6.

4.3.2 Font Analysis

Determination of exactly which font is being used is almost impossible without actually recognizing the characters. Although there are many existing fonts, the number in constant use is relatively small. Either or both of size and style changes change the font, so some fonts only differ in size. This is particularly true within each publication. Therefore part of the font changes may be determined in the rough preliminary scans on a height criterion by simply noting the changes in either output frequency or height of a portion of a line of characters.

4.4 Format Analysis - Vertical Scans

Almost all of the control information necessary is obtained from the horizontal scans, but that of column separation remains. That is, the machine must note the positions of long vertical spaces which indicate column separations and separation of text from other non-text items. This information can be obtained from the horizontal scans with the aid of much computational equipment, but an easier although perhaps slightly slower solution is that of using vertical scans.

The position of a "clear" scan is recorded as information for the raster scan controls which guide the recognition. A "clear" scan is defined as that portion of a full length scan which shows no crossing of any ink for a minimum vertical distance of the height of about five normal text lines. The minimum length requirement prevents margins being determined which are really accidental vertically-aligned spaces between words. The width of the scan is on the order of half of an average character width to prevent further errors of this sort.

When these vertical spaces are known, along with the previously gained knowledge of horizontal spaces and, most important, content, a simple selective procedure will choose what columns to read and when. For example, two adjacent columns followed on a page by a column of full page width, will be read before the wide column. Should there be a column in each quadrant of the page, the order of reading is determined by the relative size of the horizontal space to the vertical space, and whether in the horizontal space there is a title (perhaps different font). Simple logic to correctly analyze the majority of cases is not difficult to envision.

4.5 Scan Sequence

The above methods are implemented as quickly as possible to avoid holding up the character recognition function of the machine. Most of these are carried on simultaneously with the exception of the alignment procedure which must necessarily come first.

The alignment process aligns the raster scans with the printed lines of text. Then the horizontal scans commence at the top of the page and work downward. During these scans the parameters of print size, line spacing (text line position), and kind of material being scanned are recorded. When the bottom of the page is reached, vertical scans begin at one side and progress to the other in discrete steps. During these scans the margin locations are recorded as control stops for the text scans. When this sequence is finished all the control information is at hand to allow direction of the text scanner and to allow coding of the non-text items.

4.6 Appendix - Content Analysis Experiment

4.6.1 Hypotheses

An experiment is included here to indicate the validity of the two following hypotheses:

(a) The output of a scan across text for a spot size less than the smallest width of a character stroke varies considerably from scans across non-text in the degree of peak-to-trough nature. Text results in variations of the output from limit to limit. Photographs and linecuts should have various grey scales in between preventing a binary signal.

(b) Text exhibits a recurring or characteristic number of vertical strokes per unit distance of scan which is different from that of photographs or linecuts. These result in a more random effect. For a given scan rate, the characteristic frequencies of the text are different from those for the latter.

4.6.2 Equipment

A flying-spot, single-line scanning system has been assembled using an oscilloscope with a very fast phosphor. This creates the flying spot which is focussed by a condensing lens onto a transparency of the specimen. A photomultiplier tube behind the transparency picks up the signal. This signal is displayed on another synchronized oscilloscope for hypothesis (a) and put through a wave analyzer and recorded for hypothesis (b). The experimental set-up is shown in Figure 4.1 and immediately following Figure 4.2 is a list of the parameters of the experiment describing in more detail the arrangement of Figure 4.1. A block diagram of the equipment connections is shown in Figure 4.3. The photomultiplier tube and output circuitry is shown in Figure 4.4.

The response time of the photomultiplier tube and of the scanning oscilloscope limits the scanning rate to 40 cm./sec. In order to supply the wave analyzer with a continuous signal the length of the scan has been adjusted optically to match the length of the photomultiplier tube aperture. The specimen covers the whole aperture.

Four samples of printed matter have been chosen for scanning specimens. They are shown in Figure 4.5.

(a) Specimen A is heavy type used for titles and includes only upper case letters as an extreme.

(b) Specimen B is light type characteristic of text material.

(c) Specimen C is a linecut chosen because of its seemingly periodic and binary nature.

(d) Specimen D is a half-tone photograph chosen as an extreme example of a photograph since its periodic nature is of the same order of magnitude as text.

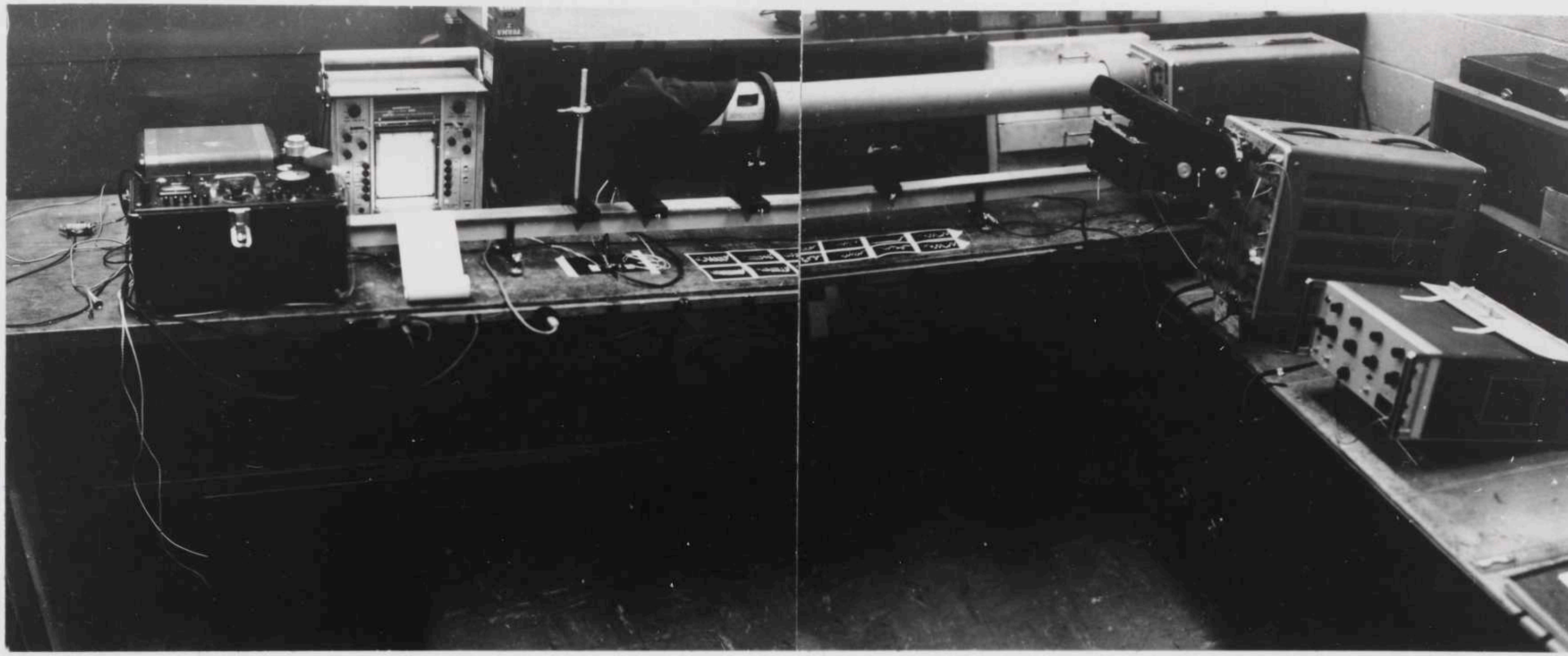


FIG. 4.1 SCAN ANALYSIS EQUIPMENT

FIG. 4.2 EXPERIMENT PARAMETERS

Scanning Oscilloscope:
- P 11 phosphor --decay time about 100 usec. for decay to 4%.
- Dot diameter < ½ mm.
- Scan length 10 cm.
- Scan speed 5 msec./cm.,i.e. 20 scans/sec.

Objective Lens:
- Focal length 7½ in.
- Diameter 2½ in.

Image Plane:
- Specimens on film positive made from high-contrast process ortho
- Dot diameter < 1/10 mm.
- Scan length 2 cm.
- Scan speed 25 msec./cm.

Photomultiplier Tube:
- RCA type 931A
- Aperture formed by black tape 2.2 cm. x 1.0 cm.

Wave Analyzer:
- General Radio type 760-A with added motor drive. One analysis of Figure 4.14 takes eight minutes.

Viewing Oscilloscope:
- Grid squares 1 cm. x 1 cm.
- Scan speed 5 msec./cm., i.e. five times magnification of length of scan at image plane.

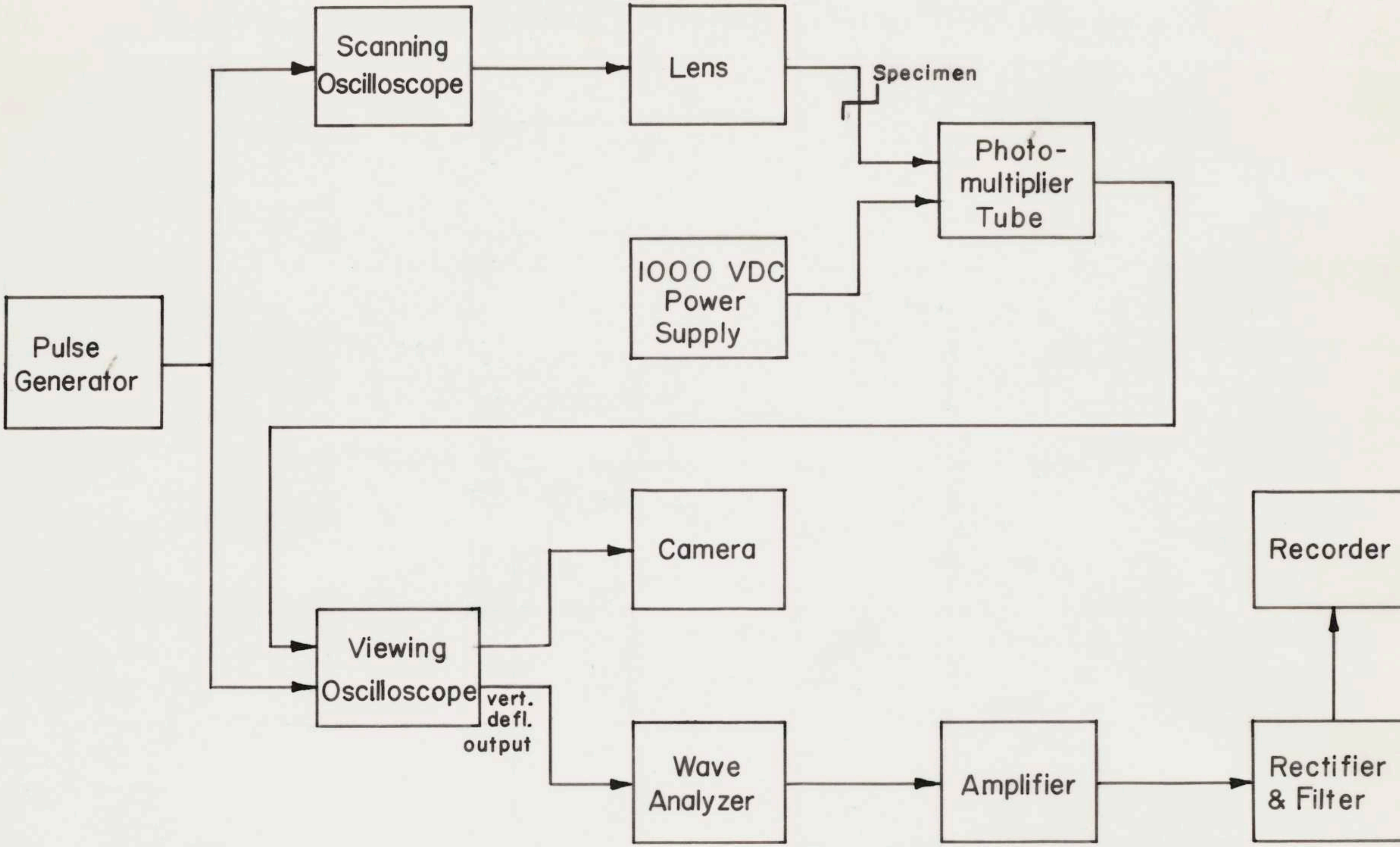


FIG. 4.3 EQUIPMENT CONNECTIONS

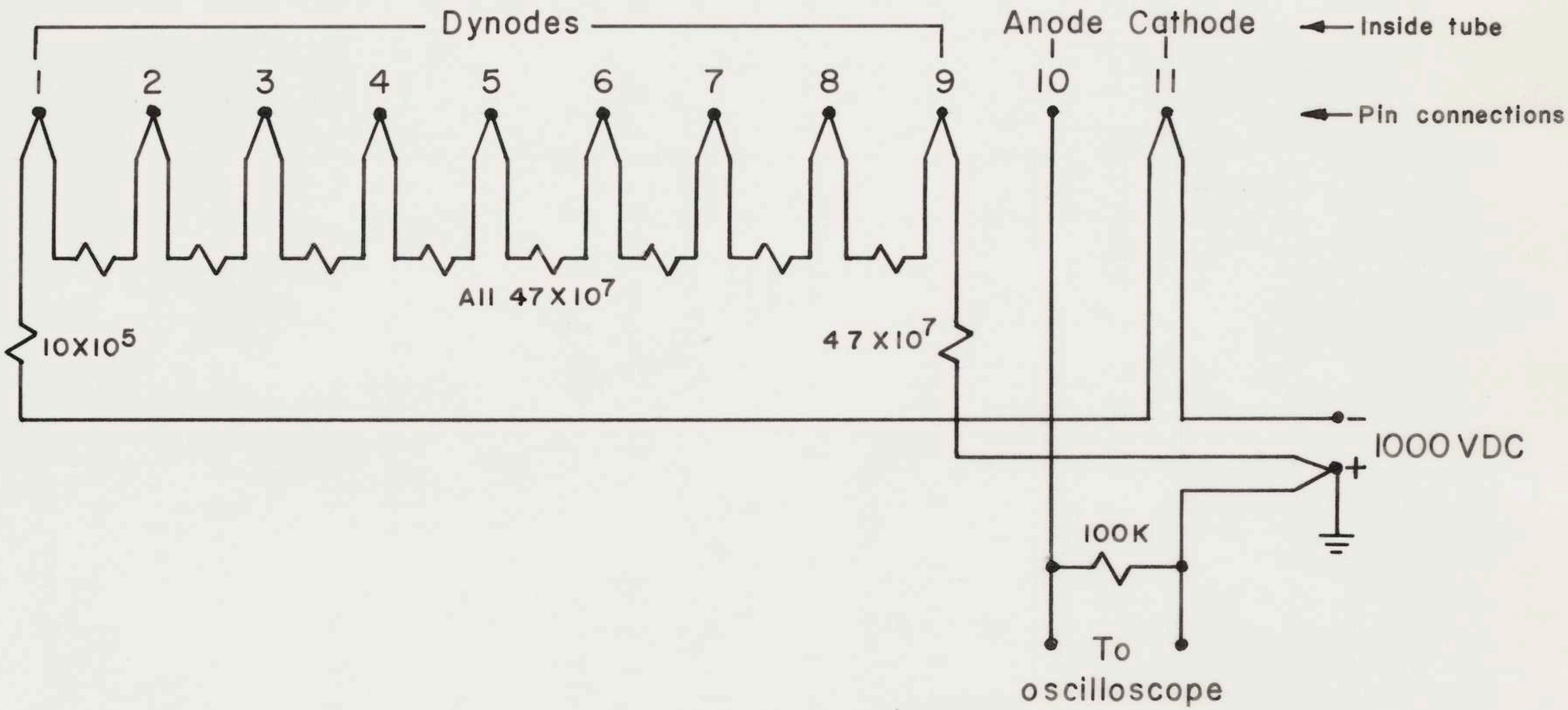


FIG. 4.4 OUTPUT CIRCUITRY

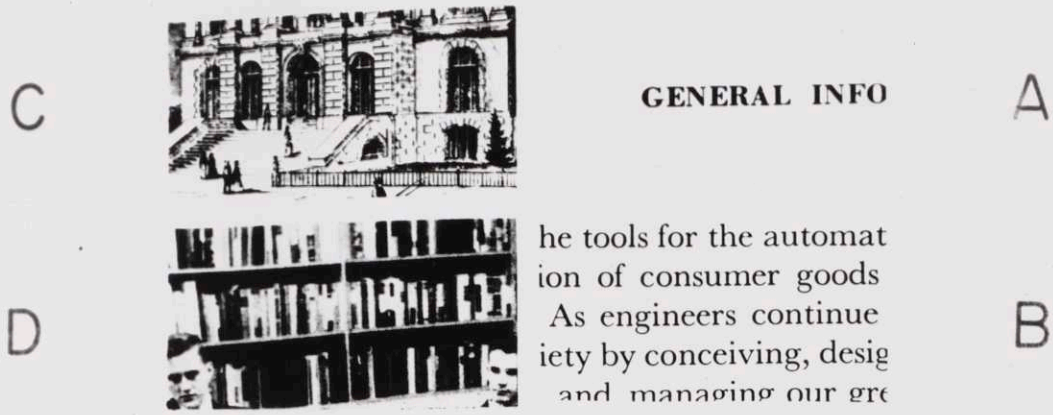
These specimens have all been chosen from the M.I.T. General Catalogue of 1963-64. The specimens are full size in Figure 4.5. The specimens are used as a positive transparency made as a contact print from the same negative from which the photograph in Figure 4.5 was made.

4.6.3 Procedure

From these four specimens nine tests can be made as follows.

Scan I Figure 4.6 shows the output of Scan I. A transparent space on the film is scanned (a) to set the scanner intensity level to just less than saturation of the photomultiplier tube (lowest grid line), and (b) to show the envelope of the results to come from subsequent scans, since the apparatus does not have perfect results. Noise, the small oscillations, and less than sharp cut off, are evident. Obviously, black interrupting the scan gives a signal in the upward direction.

Scan II The first five and a half characters of the word GENERAL from specimen A are scanned in the horizontal direction just above the center bars of the E's. The scan dot passes over the serif at the end of the center stroke and is obvious in Figure 4.7 as the narrow third peak. The next three peaks are the N and so on. Half of the A is included in the right end of the scan. It may be noted that the lower level of the signal, the



SCAN LENGTH 2 cm. ⊢——⊣

FIG. 4.5 SPECIMENS

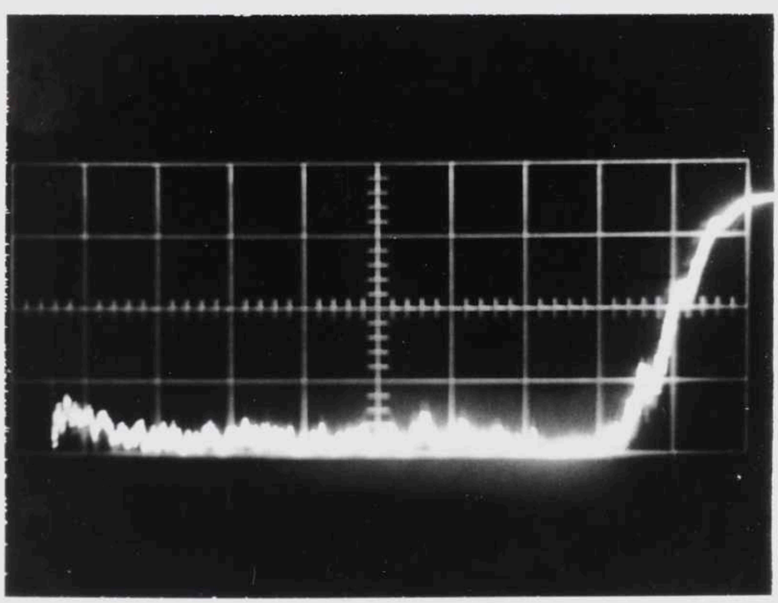


FIG. 4.6 SCAN I

zero, rises with respect to the grid. This does not compromise the experiment.

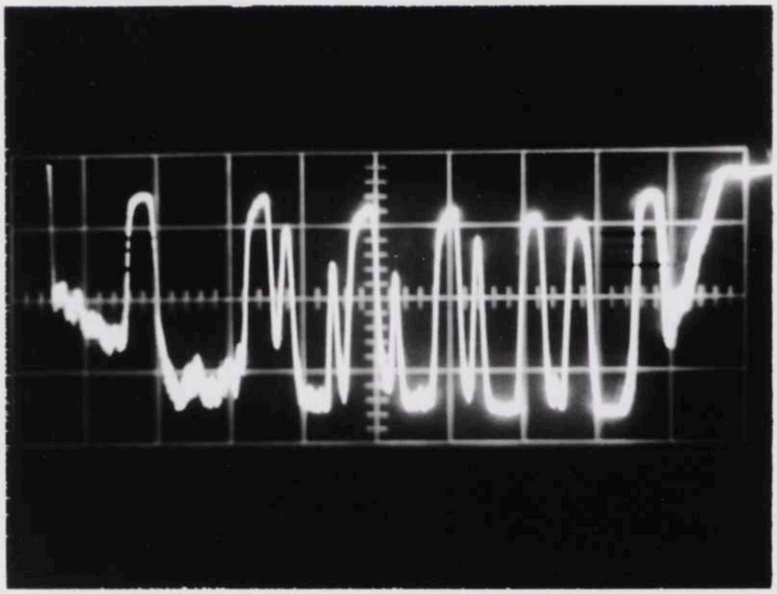
Scan III The first eight characters of the word engineers from specimen B are scanned just above the crossbars of the e's where all of the printed lines crossed are about the same width. See Figure 4.8. The first two peaks are the e. The next two peaks are the n and so on. This is noted more easily in Scan V.

Scan IV This is a second run of Scan III to test whether variations of a random nature were altering the frequency analysis. No picture is included since no change results.

Scan V This is the same specimen but scanned right along the crossbars of the e's. These crossbars are very narrow lines on the order of the scan dot diameter in width. Slight misalignment is evidenced by different levels at the e's. See Figure 4.9. The scan is more closely aligned with the first e than the pair of e's (on the right side in Figure 4.9).

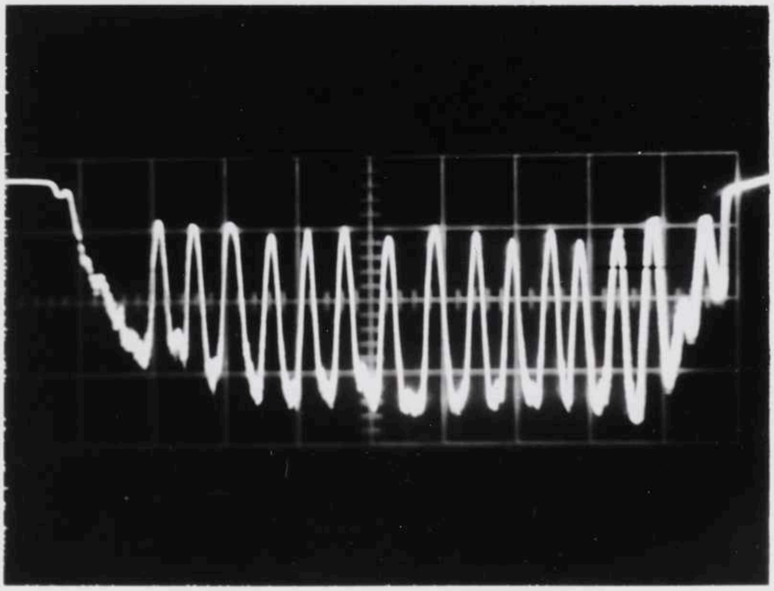
Scan VI The same specimen is scanned again, but below the crossbars of the e's. The scan just touches the tails of the e's as evidenced in Figure 4.10 by the two small peaks in the area of the latter two e's. The g shows a slowly downward sloping signal where the scan dot crosses the thin neck of the g and runs along and off the top edge of the lower loop, that is, just above the base line.

Scan VII This scan is also through specimen B but it



A

FIG. 4.7 SCAN II



B

FIG. 4.8 SCAN III

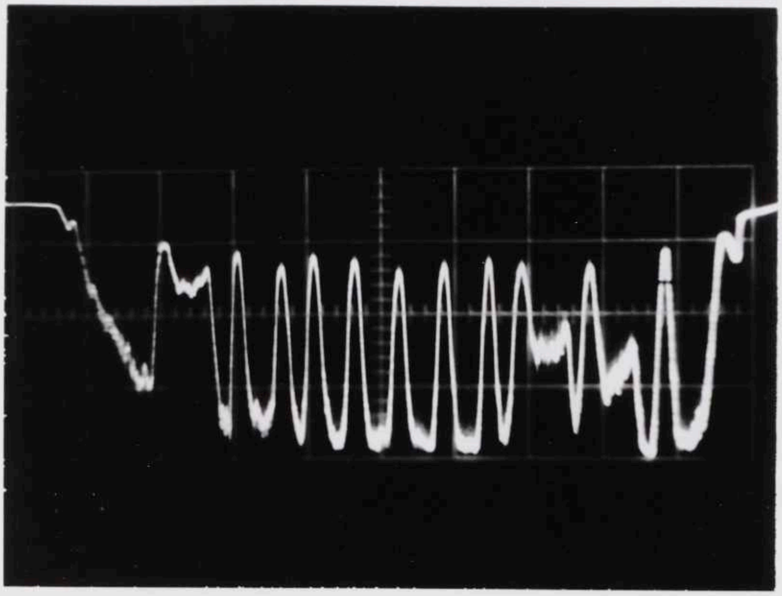


FIG. 4.9 SCAN V

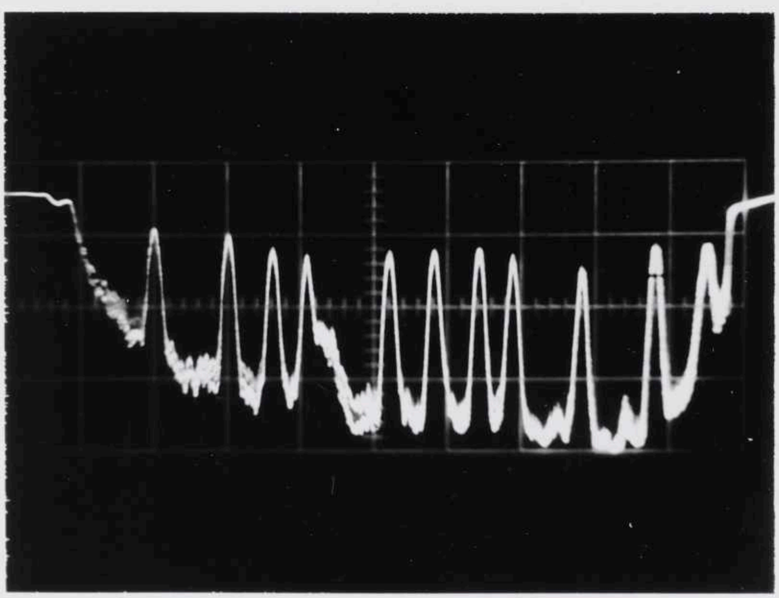
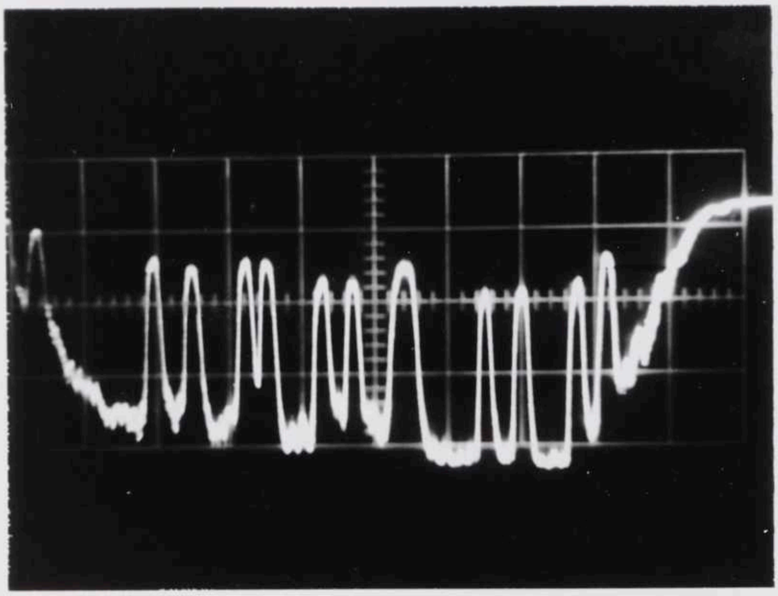


FIG. 4.10 SCAN VI

runs through the center of the words tools for. Note Figure 4.11. The first peak is the t, the next two although widely spaced, are the first o. The two peaks which are closest (third and fourth) are the adjacent portions of the two o's. Since the signal does not go to zero between these two peaks it may be concluded that (a) the scan dot size is slightly larger than the spacing between the o's and/or (b) the time of decay of the photomultiplier and/or the scan tube phosphor and/or the output circuit is (are) becoming significant as compared to the time of scan of the space involved. The former is most likely, however the results are not seriously affected. The widest peak is obviously the signal resulting from the scan crossing the middle of the s where it is diagonal.

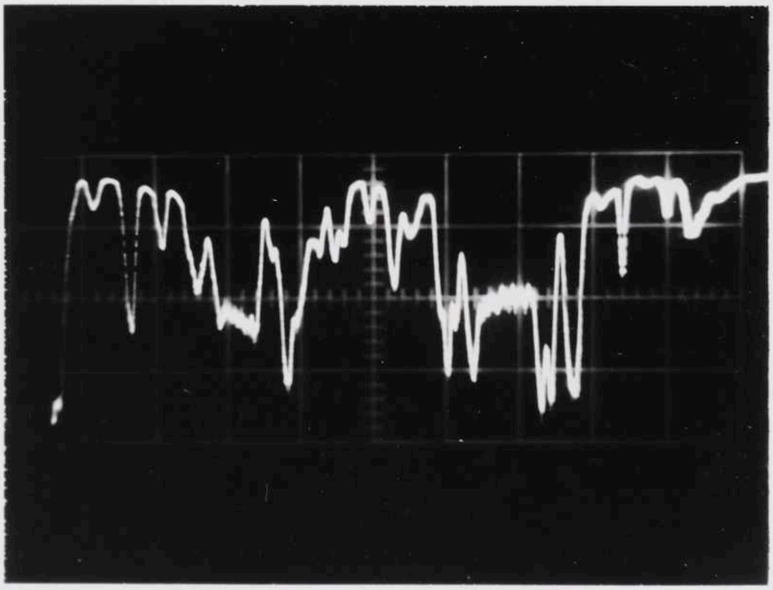
Scan VIII This scan is of the same length as the previous scans, but is directed across the three archways of specimen C, a linecut. The resulting output is pictured in Figure 4.12. The three archways are obvious as the three peaked regions in the output picture. The detail is quite noticeable high-frequency, low-amplitude regions showing up between these three peak regions are the result of scanning over five or six very fine and closely separated distinct lines (used as shading) on both sides of the middle archway.

Scan IX This scan is of specimen D, a half-tone photograph, through the center row of books and directly



B

FIG. 4.11 SCAN VII



C

FIG. 4.12 SCAN VIII

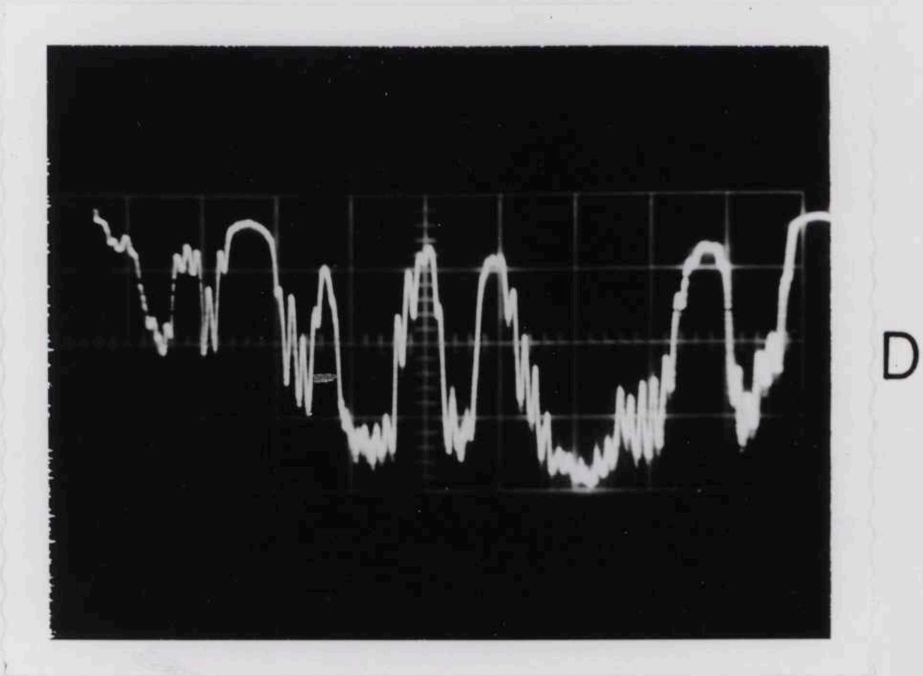


FIG. 4.13 SCAN IX

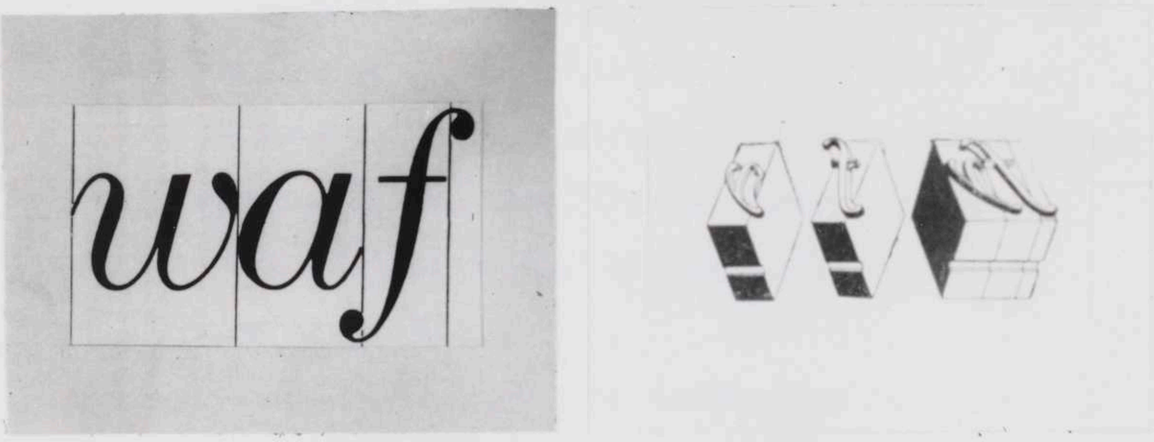


FIG. 2.1 KERNED CHARACTERS[15,16]

under the scan on specimen C. See Figure 4.13. The most interesting point to note here is the high frequency signal (Figure 4.14) superimposed on the overall signal. This is on the same order of frequency as the predominant noise frequency, but the amplitude is greater by a factor of two or three. On inspection of the specimen, the source is noted to be the occurrence of the dots making up the various tones in the half-tone photograph. This unexpected result is an indication of the great detail picked up by this crude scan.

4.6.4 Discussion of Results

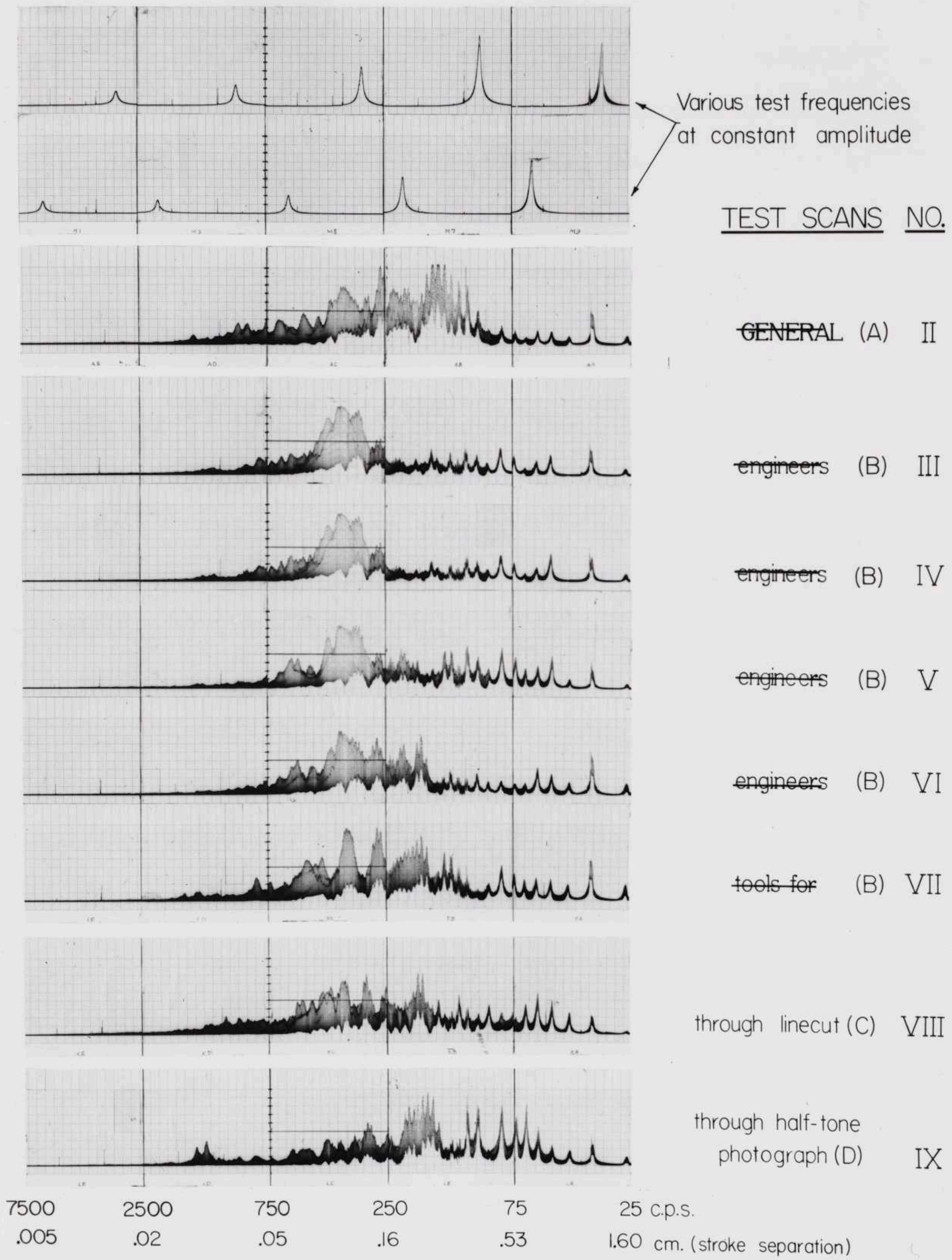
The reliability of the results is important first of all. About seventy percent of the scan length is reliable since the output envelope for no specimen in Figure 4.6 is fairly flat in the middle region. Of the ten centimeters of scan, two cm. at the right end and one cm. at the left suffer from edge effects at the aperture. Therefore, the ends of the scans are to be regarded warily.

In the same envelope photograph certain noise effects are noticeable. This is negligible, however, and is lower in magnitude and higher in frequency than the results being discussed.

Repeatability is good. No difference in frequency analysis graphs may be determined between Scans III and IV. The identical but successive scans result in the frequency vs. amplitude graphs for those scans in

Figure 4.14. A point to notice in this set of charts is that the amplitude scale varies along the frequency axis. The top two rows of graphs are of particular frequencies, all at the same amplitude. They do not appear this way in Figure 4.14. This, if remembered, is acceptable since comparisons are of interest mainly from scan to scan at a particular frequency. Thus, the results are reliable and study of the experimental results for the postulates of Section 4.6.1 may be continued.

The first postulate is that text may be differentiated from pictures and linecuts on the basis of the peak-to-trough distance of the photomultiplier output signal. Inspection of the photographs of the outputs for scanned text, Figure 4.7 to 4.11, for a linecut, Figure 4.12, and for a photograph, Figure 4.13, indicates that this postulate is justified. Although some peaks in the text scans are not at the maximum level, most are at the same level. The linecut scan, Figure 4.12, shows that even an expectedly binary linecut results in widely varied peak and trough values across the scan. The three low frequency peaks rise to the same value, but the middle frequencies do not. Similarly, the photograph scan, Figure 4.13, has peaks and troughs of varying magnitude not at all as regular as the results for text. An additional point to note is that a very short scan is enough to determine the information of



SCAN TESTS - FREQUENCY CHARACTERISTICS

FIGURE 4.14

what is being scanned.

The second postulate is that scanned text outputs display characteristic frequencies not common to similar outputs of linecuts and photographs. The samples of a linecut and photo have been chosen such that average distances between vertical strokes are similar to those of text in order to assay a worst case. Note the specimens again in Figure 4.5.

Frequency vs. amplitude for the scan tests is shown in Figure 4.14. (Remember that the response is not flat.) The frequency range of 250-750 cps. (.16-.05 cm. stroke separation) is interesting to look at first. The horizontal line is drawn across this range in each graph to aid the eye in estimating areas enclosed by the graph and the zero line. Note that Scans VIII and IX have a significantly lower area above this line than the text scans. The following is an indication of these areas where the unit of area under the signal is arbitrary and the range is 250-400 cps.

Scan	II	area of 11.3	above indicated line
	III	12.5	
	IV	12.5	same scan as III
	V	9.8	
	VI	9.0	
	VII	10.8	
	VIII	4.6	
	IX	.5	

The important part of the area is that above this "trigger" level. When integrated over the whole signal, the differences become negligible. Therefore a frequency band must be chosen with much care to approach a ten percent difference in output. A series of narrow band width filters can be used with triggers and integrators to differentiate between text scans and linecut or photograph scans.

For this size text the important frequency band is 250-400 cps. which corresponds to a .16 to .05 cm. stroke (in text) or peak (in output signal) separation. The range of longest to shortest adjacent stroke separation in text of specimen B is .15 to .05 cm. However, specimen A differs in that the range is .30 cm. to .08 cm. (Without the G appearing the larger dimension becomes .2 cm.) Thus a slightly different important frequency range is expected. It appears the component of frequency is much stronger below 250 cps. (.16 cm.) than for any other scan. See Scan II, Figure 4.14.

As expected, larger text gives a different range, and this must be included in the filter selection in order to admit large text such as that for titles and headings. But even this output signal for text exceeds the largest signal for photographs and linecuts which were chosen for their periodic nature.

4.6.5 Conclusions

Both of the postulates are therefore correct in most cases. The first may be incorrect in some cases due to lack of shading in photographs or fine lines for shading in linecuts. The second may be incorrect if the photograph or linecut happens to have a very strong periodic nature in the frequency bands characteristic of text. Thus, these may fail, but for different reasons so use of both criteria is almost infallible.

The criterion of horizontal spaces in a periodic nature above and below lines of text can indicate where text is, but there must be a running indication of text since much text does not cross the whole page making the criterion difficult to apply alone. Using the two criteria just discussed with this third allows an almost infallible running account of whether the item scanned is text or non-text.

A second benefit results from the frequency inspection or characteristic length observation. This is that the font being used is more easily estimated. Font changes with size and style of characters. The former has been determined by the line spacing inspection. The latter includes shape, width of characters, and thickness of strokes. The width of the character is easily estimated by looking at the characteristic frequency from the text analyzer. An average over a few words is sufficient. Also, an integrator of the signal gives an indication of the thickness of the strokes. Thus, a fair guide to the particular font being used is realized.

CHAPTER 5

SYSTEM DESIGN

The overall system design is partially fixed by the preceeding chapters on masks and format analysis. The methods used must be compatible with these previous commitments. In addition, the design must be the most economical without unduly sacrificing the reliability of the reader. However, there are no plans for building the reader immediately, so no detailed design of the system is attempted here. Rather, an overall idea of the direction in which design should proceed is suggested with the object being to set specifications for later design and not to fix a method in areas which are not basic to the method proposed, for example, the document handler. Figure 5.1 shows a block diagram of the system. The remainder of this chapter describes the parts of the system and how they fit together.

5.1 Sequence Controller

The sequence controller initiates operation of the

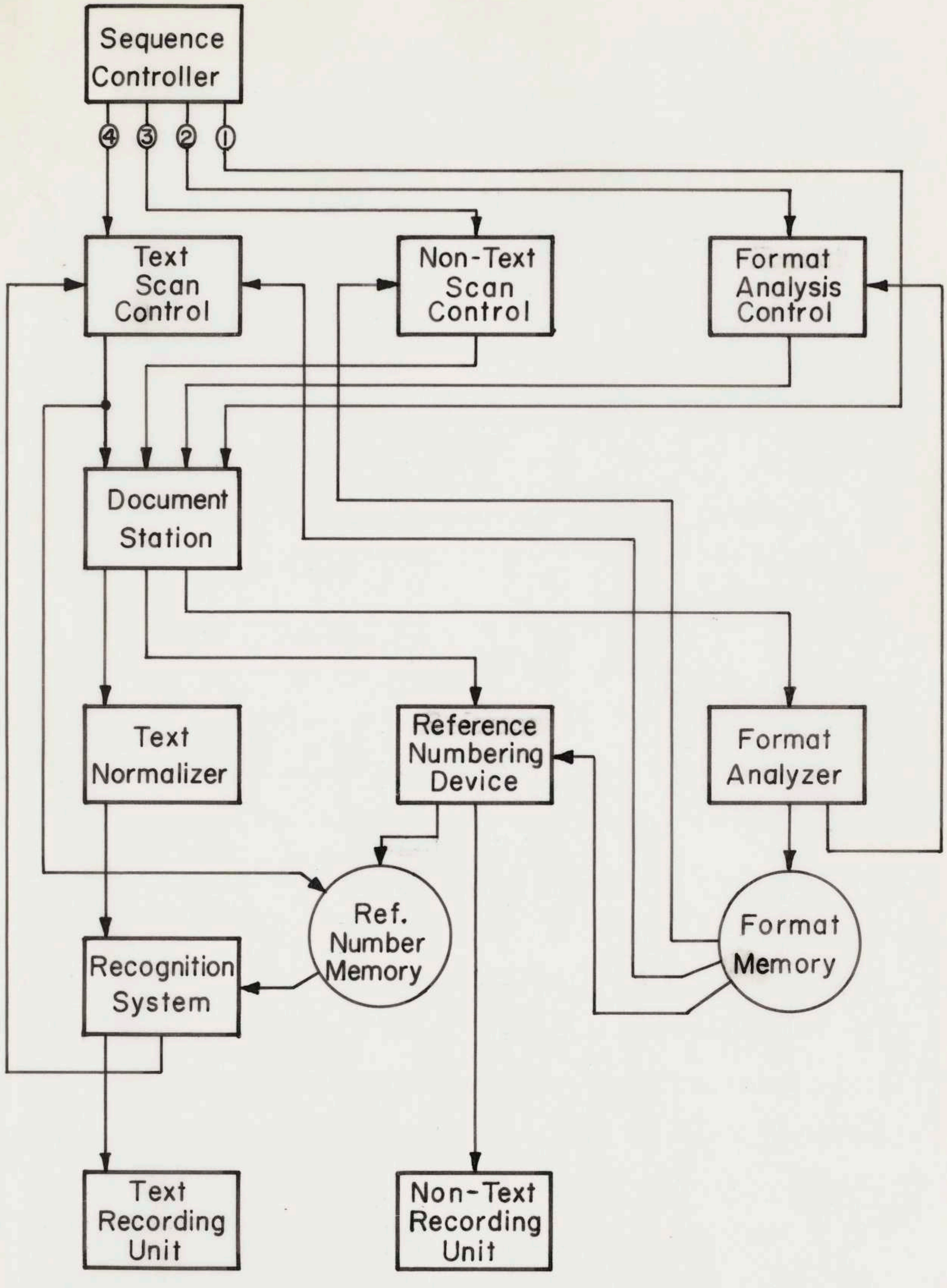


FIG. 5.1 SIMPLIFIED SYSTEM BLOCK DIAGRAM

subsystems in the proper order. First of all it signals the document transport to change pages. When this is done, the controller allows the format analyzer to store in memory the required information about the format of the page. This is discussed in detail in Section 4.5.

Then the controller initiates scanning for actual recording. It first allows the non-text scanner to scan areas known as non-text. These areas are known from the previous analysis scans. Next, the controller allows the text scanner to commence reading the document.

5.2 Document Station

The document station comprises a document transport and a document scanner. The former must have a page-turning or microfilm-indexing time as fast as possible; this is the most time-consuming part of a page-reading cycle. Microfilm-indexing is preferred to actual page-turning because of the inherent speed advantage. Most publications will be microfilmed in a large central library for preservation. The transport is shown in Figure 5.2.

The scanner-receptor (probably photomultiplier tube) configuration changes with the type of document transport to be used. The scanner is controlled by three functions: format analysis control; non-text scan control; and text scan control. A very high-speed, high-resolution scanner is required. It must be capable, with the associated optics, of scanning approximately seventy lines per widest character width, or em quad, for two

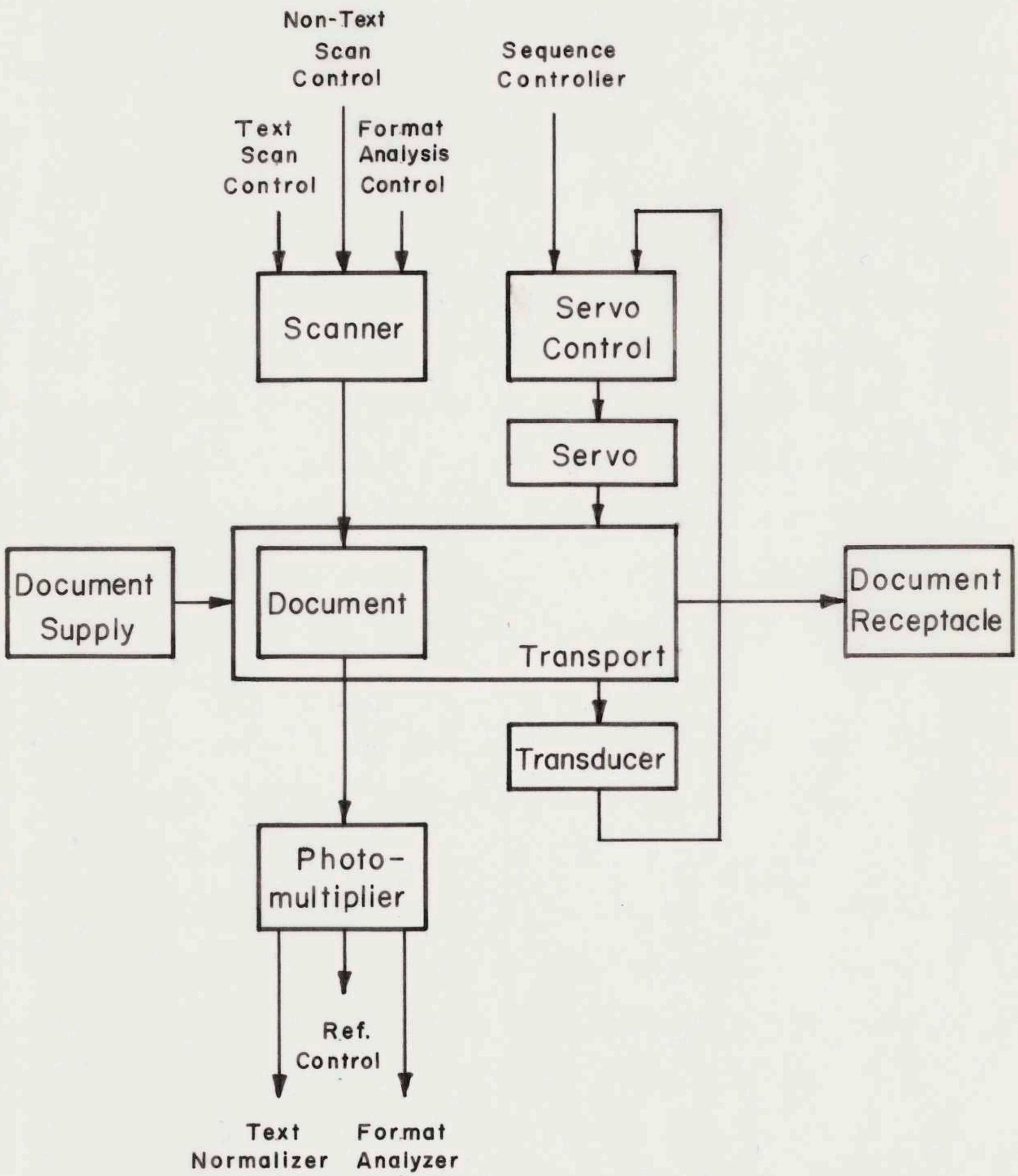


FIG. 5.2 DOCUMENT STATION
BLOCK DIAGRAM

scans per matrix column. Small-to-regular size print has approximately 150 em quads per line of text on an over-sized, $8\frac{1}{2}$-inch wide page ($7\frac{1}{2}$ inch text length). This requires $150 \times 70 \div 7\frac{1}{2} = 1,400$ scans per inch. "...Itek has prototype cathode-ray tubes currently under test and evaluation which are expected to have resolutions up to 2,000 resolvable spots per inch over the entire face of a 7-inch diameter tube."[18] Thus, in the worst case the requirements are approximately at the state of the art. Even so, a number of these tubes may be used together to reduce the resolution requirements and to allow higher speed scanning. The scanning speed is limited by the light detector used which may operate at around a megacycle. This should allow recognition at the rate of 1000 characters per second. Another limiting factor is the time required for preliminary analysis of the page and non-text recording since fast recognition does not decrease reading time significantly if preliminary analysis is slow.

5.3 Format Analyzer

The format analyzer is a subsystem which applies analysis scans to the page. This is a set of consecutive control functions for the scanner which operate according to the method worked out in Chapter 4. First a line of text is found and alignment is accomplished. Then the page is covered with spaced, horizontal scans and then with spaced, vertical scans. The output of these scans passes through various analysis operations discussed in Chapter 4, and the parameters are stored in a suitable memory device. See Figure 5.3 for a block diagram of the format analyzer.

5.4 Non-Text Scan Control

This function controls the scanner for recording of the non-text areas, i.e. pictures, graphs, complex equations, and tables. Each integral area of non-text information is referenced and that reference number is recorded in memory with its position such that it may be inserted into the recorded text later close to the appropriate text. A simple raster scan of the area is recorded on an appropriate device such as videotape or microfilm. This is not basic to the system and will not be discussed further.

5.5 Text Scan Control

This is a simple raster scan of each line of text. The scan is of high frequency in the vertical direction and lower frequency in the horizontal direction. The vertical height of the raster is known from the analysis scans as are the alignment, length, and spacing of the rasters. The object is simply to scan each character vertically without touching parts of upper or lower lines of text. The scans in the raster must be of a width small enough to allow at least one scan completely within each mask matrix square as the characters are scanned across the masks. This scan control functions

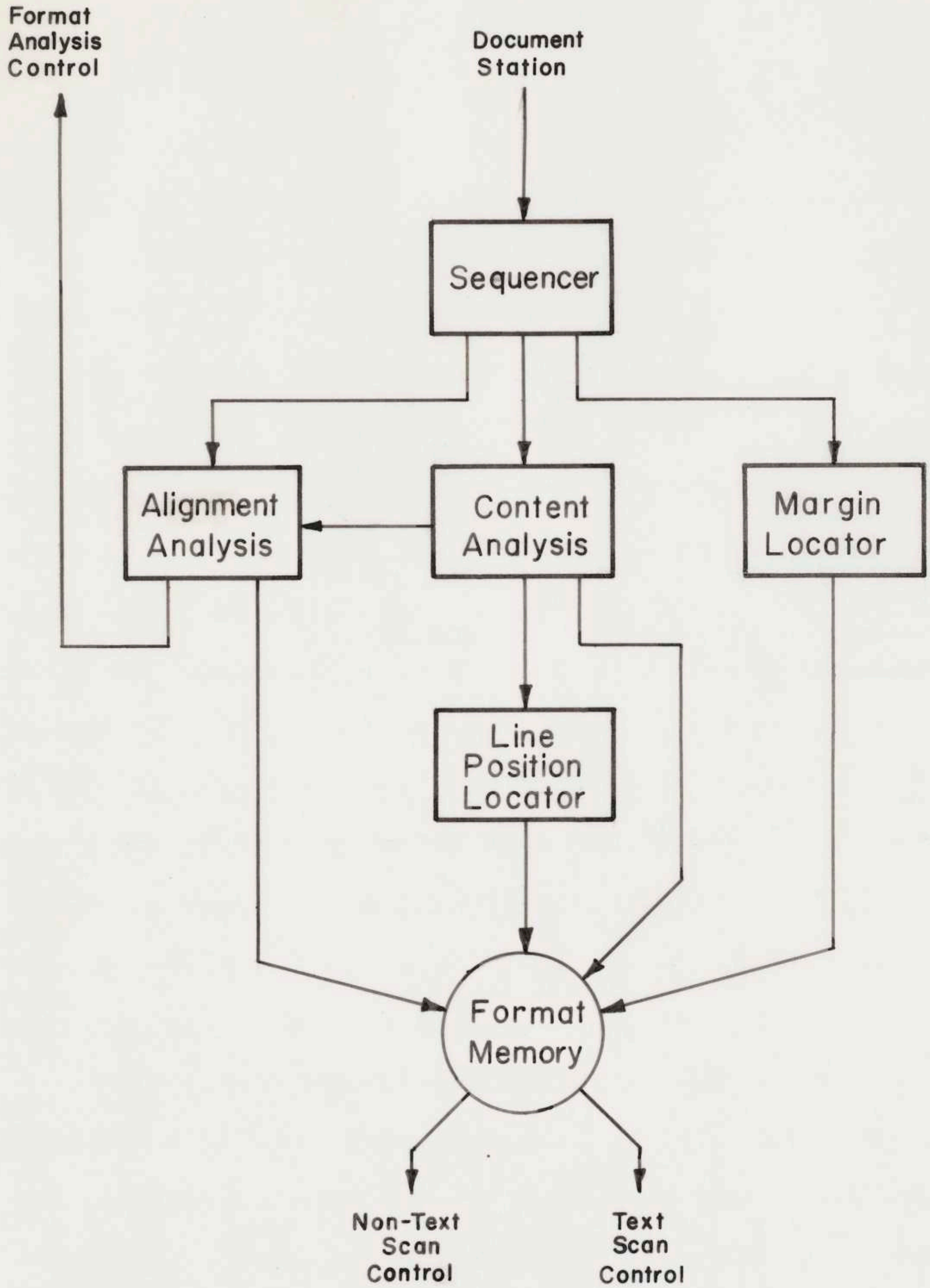


FIG. 5.3 FORMAT ANALYZER BLOCK DIAGRAM

on format information stored in memory during the analysis scans. The scan may be recycled for a line or part thereof if the recognition system indicates a mistake has occurred or the font has changed.

5.6 Text Normalizer

The output of the text scanning system discussed above is operated on by the text normalizer to reduce or enlarge the text to a given size for display and scanning past the masks. This is done to allow the use of one set of masks for all sizes of text of that particular style of characters. The other function of the normalizer is to find the base line of the scanned characters and align it more precisely with the base line position on the masks. This must be done since the location of lines was found by the format analyzer to plus or minus one-tenth of the character height. For the masks, this base line must be known within a fraction of a matrix square height. Thus, the normalizer corrects the vertical position of the characters.

Future development may make the normalizer even more important by having it remove serifs and normalize the thickness of strokes of the character. These are two means of reducing still further the number of mask sets required since many fonts differ in only these two items. Noise may also be removed from the signal at this point.

5.7 Recognition System

The requirements of the recognition system are fixed by Chapter 2 on mask design. Characters must be scanned past the row of masks. The light sensors, triggers, and logic following each mask are spelled out in Chapter 2. The logic signals for a rescan, signals a font change, or sends recognized characters to a buffer. Reference numbers for non-text are inserted at the buffer. This is shown in Figure 5.4.

The part of the recognition system preceeding the masks is less fixed. Display and image-mover devices are required for scanning the image of the line of text or portion thereof past the masks. Figure 5.4 shows a simple storage tube display with a rotating mirror for scanning. A plurality of storage tubes may be used to reduce null time due to erasing the image. This method allows a non-mechanical, quick change of masks for reading a different font. The storage tube is loaded at a different vertical position on its face such that the mirror scans the image above (or below) the previous masks. Other sets of masks are placed above and below the first set but use the same light sensors and logic.

Another concept for this display-scanning area uses a simpler scan. A moving, light-sensitive belt or disc moves past a single-line scanner which activates it. The belt or disc then passes the masks and acts as a moving film for carrying the normalized text image

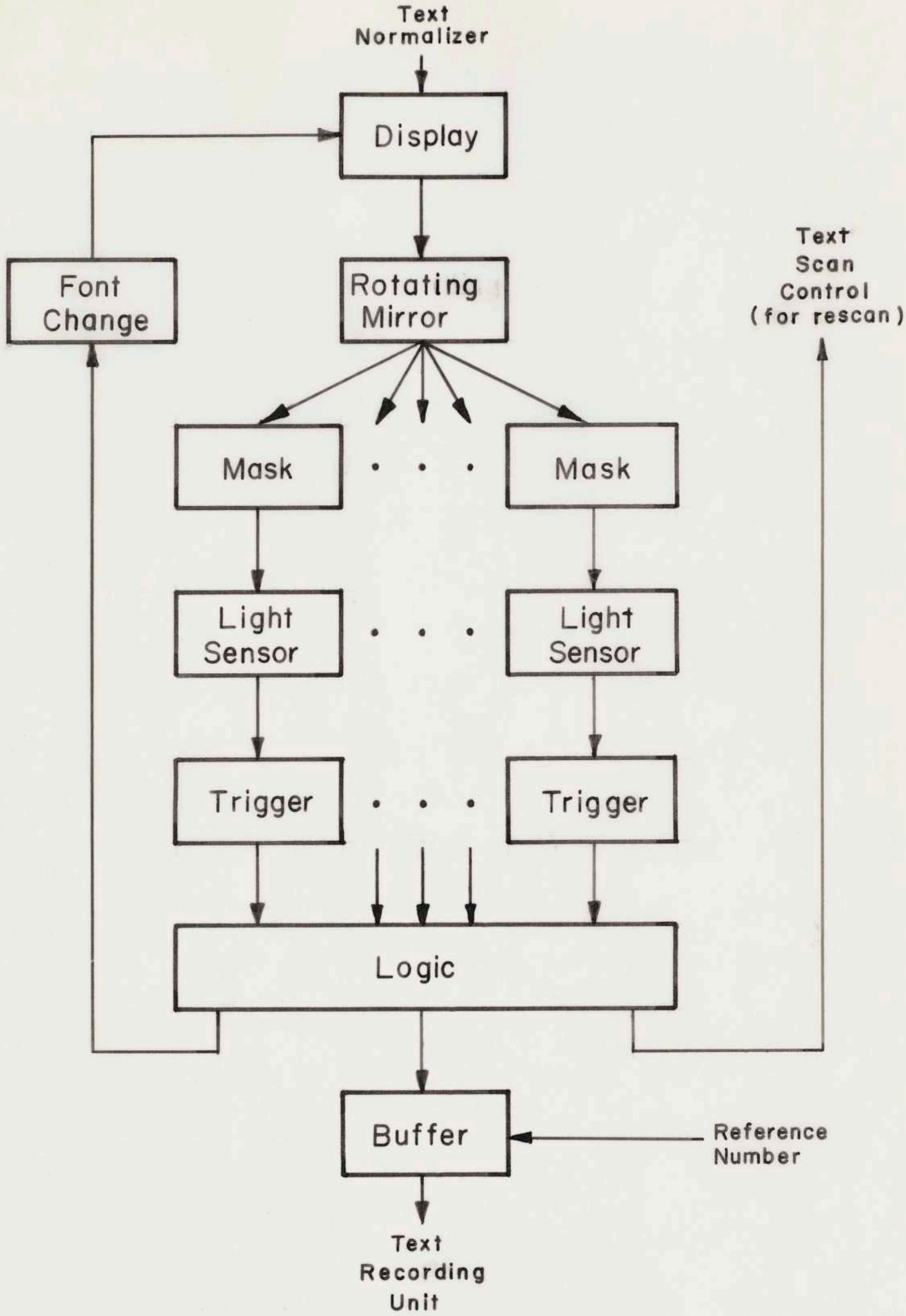


FIG. 5.4 RECOGNITION SYSTEM BLOCK DIAGRAM

across the masks.

A third concept puts masks against the face of a cathode-ray tube. The characters are moved across the tube, and therefore across the masks, without any mechanical movement or optics.

It may be noted here that in a very sophisticated recognition system a dictionary may be used to aid in picking out impossible combinations of characters and to therefore point out mistakes. This reduces the reliability required of the recognition logic. However, a dictionary requires a very large and fast memory attached to the system. In the system proposed in this chapter no dictionary is envisioned, but rather, reliable results are attempted at the outset.

5.8 Conclusion

The above proposes workable methods for an actual character recognition machine. This machine is by no means simple, but it is far simpler then machines using methods other than weighted area mask scanning. As discussed in Chapter 1, the other methods require on-line, large-scale computation which naturally increases costs.

It is felt that the solution proposed here for character recognition is now ready for actual implementation. First, masks must be designed to empirically prove the validity of the proposed method. Second, a preliminary system design with cost considerations must be done to indicate which of the exact configurations is the most useful per dollar. Third, a simplified prototype must be built to iron out the inevitable difficulties in a full scale system.

5.9 Appendix - Simple Initial Machine

The recognition system must be developed first as all other parts of the machine serve it in some manner. The following machine is proposed as a development ground for the recognition system, but at the same time, it is, when completed, a low-cost, constantly-monitored, useful machine.

The effort in design of this machine is all focussed on the recognition system. This is as involved as that in a fully automatic machine, but may be slower. However, the rest of the system is simplified by eliminating the document handler and the format analyzer by substituting in their places an operator using a display, light pen, and keyboard. These areas replaced by the operator may be added for an automatic machine when the recognition system is well on the way to an adequate state of development.

The machine functions as controlled by the operator. The operator inserts a document to be read by the scanner. An image is then displayed onto a cathode-ray tube for the operator. The operator sets controls for margins either by light pen or dial inputs. The operator then points out which areas are to be scanned as non-text and which are to be scanned as text areas. Alignment and line-finding must still be done by the machine. The machine, of course, reads characters far faster than the operator can type them. When a character is reached which does not result in recognition, the machine points out the character on the display and the operator types in the correct character. The machine then continues. Another advantage here is that the operator may point out changes of font.

This simple initial system eliminates most of the control requirements for a fully automatic machine. It could easily be used in libraries for reading simple, fixed format literature such as that described in Section 3.1 if a page turner is added. The same machine could be used for complex fixed format described in Section 3.1 if some logic operations are added. The formats described in Sections 3.3 and 3.4 require the fully automatic version of the machine when high reading rates are necessary. However, the simple machine just mentioned is an excellent place to begin. Everything is ready for that beginning.

REFERENCES

1. L. J. Mintz, K. R. Brooks, "A Typed Page Reader," *Optical Character Recognition* (Washington: Spartan Books, 1962), p. 85.

2. R. K. Gerlach, "Wide Tolerance Optical Character Recognition for Existing Printing Mechanisms," *Optical Character Recognition*, p. 93.

3. J. Rabinow, "Developments in Character Recognition Machines at Rabinow Engineering Company," *Optical Character Recognition*, p. 27.

4. "Additions to Collections and Total Contents of the Library, Fiscal Years 1962 and 1963," *Annual Report of the Library of Congress* (fiscal year ending June 30, 1963), Appendix IIB, p. 96.

5. *Annual Report of the Library of Congress* (fiscal year ending June 30, 1962), p. xvii.

6. M. E. Stevens, *Automatic Character Recognition - A State of the Art Report* (Washington: U.S. Dept. of Comm., Office of Technical Services, May 1961), pp. 32-60.

7. L. J. Cutrona, E. N. Leith, C. J. Palmero, and L. J. Porcello, "Optical Data Processing and Filtering Systems," *IRE Transactions on Information Theory*, June, 1960, pp. 386-400.

8. D. M. Baumann, *Feasibility of a Weighted Area Scanning Technique for Character Recognition* (Sc.D. Thesis, M.I.T., June, 1960).

9. D. M. Baumann, "Weighted Area Scanning Techniques for Character Recognition," *Optical Character Recognition*, p. 197

10. G. A. Lemke, *Description of Character Recognition Demonstrator*, Unpublished D.A.C.L. report, June, 1961.

11. D. M. Baumann, O.C.R., p. 199.

12. Ibid., p. 204.

13. A Manual of Style (Chicago: U. of Chicago Press, 1956), p. 253.

14. Ibid., p. 250.

15. Mathematics in Type (Richmond: William Byrd Press, 1954), p. 13.

16. Séan Jennett, The Making of Books (Pantheon Books, 1951), p. 35.

17. R. F. Rueckwald, "Problems Associated with Weighted Area Scanning Techniques for Use in Optical Character Recognition," Unpublished Bachelor's Theses, M.I.T., June, 1963.

18. "Itek Capabilities," Itek Corporation, Lexington, Mass., p. 18.