PRINTED LANGUAGE TO MACHINE CODE TRANSLATION

by

HENRY D'ANGELO

B. S., College of the City of New York, 1955

A THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Electrical Engineering

KANSAS STATE COLLEGE
OF AGRICULTURE AND APPLIED SCIENCE

1957

TABLE OF CONTENTS

INTRODUCTION

The ability of man to store knowledge in the form of writings and prints is a major factor responsible for the advanced state of our present-day civilization. Unfortunately, in developing our present-day writings and prints our forefathers did not have any of the modern computers and business machines in mind. If they had, our written material would be of a form that could easily be recognized by simple machines as well as by the human reader.

Most business machines and computers have the ability to process data at enormous rates--once the data is put into a form that is acceptable to the machine. Getting data into usable form usually requires a human operator to translate printed data into a pattern of holes punched into cards or into a pattern on a magnetic tape. The concerted effort of many human translators is required to keep one of these machines in operation for the percentage of time necessary to justify the machine's cost. A device that could perform this translating at a rate consistent with the rate at which the machines process the data would markedly increase the usefulness and the possibilities of the machine. Such a device would also be especially valuable in performing mechanized translations from one language to another, and from the printed page to an audible message (for the blind).

The problems in recognizing printed material are many and not at all clear. The purposes of this paper will be to determine just what these difficulties are and to describe a new system that will perform the printed language to machine code translation. This paper will be limited to the recognition of individual letters ("character recognition") of one type of print rather than the more complicated problem involved in the recognition of many types of prints or words and phrases.

ESSENTIALS OF CHARACTER RECOGNITION

General

Before the art of character recognition can become a science it will be necessary to determine what it is that makes one character appear different from other characters in an alphabet. One may look at any character in an alphabet and without too much difficulty point out hundreds of differences between it and any other of the characters. Yet if the character is produced with so much error as to seemingly destroy most of these differences, it may still be possible to distinguish this character from the rest of the characters in the alphabet. This points up the fact that each character of an alphabet, in most types of prints and writings, carries many more distinguishing features than are actually necessary to merely distinguish it from the other characters in the alphabet. In fact, each character may be distinguished from almost any other character in any other alphabet, or for that matter it may be distinguished from any other pattern produced. Then since a character can be distinguished from an infinite number[1] of patterns, each character can possibly carry an infinite amount of information.[2]

Implications of Character Recognition

A person engaged in reading makes no attempt to distinguish one character from all the other characters possible, but he merely attempts to distinguish it from the other characters in the given alphabet. In general he assumes that the next character which he will look at is a member of the given alphabet and then he tries to decide which letter of the alphabet the character is supposed to represent. If for some reason it could be expected that the next character he will look at could be any one of an infinite number of characters, with the occurrence of each one having equal probability, then he would be required to have an infinite amount of information in order to determine which of the infinite set of characters this particular one is. The time required to make this decision would, of course, be infinite.

[1] If a finite character area is considered to be quantized, the number of patterns possible is not infinite. The number of patterns, however, is still very large.

[2] If a character must be chosen from an infinite number of possibilities, then the character carries infinite information. According to Shannon (26),

$$H = -\log_2 p \text{ bits}$$

where p is the probability of a character occurring. If $p = 0$ (infinite number of possibilities or zero probability), $H = \infty$.

An Analogous Situation

As an example approaching the aforementioned situation, one could imagine having all the necessary statistics on every grain of sand on a certain beach so that any particular grain of sand could be distinguished from any other. Then further imagine being given a particular grain of sand from this beach and being asked to find the set of statistics, from those given, that correspond to this grain of sand. It seems clear that even if all the information were filed in the most efficient way possible, the task would not be a small one. It also seems clear that the size of the task would increase as the number of grains on the beach increased.

Now, on the other hand, suppose that instead of being given statistics on each grain of sand, 27 sets of limits were given such that any grain of sand would have to fall within one of these sets of limits. It is also known that the probability is high, for this particular beach, that most of the grains of sand will fall within 26 of the given sets of limits, while the probability of falling within the 27th set of limits is extremely small. It can be seen that the problem of classifying a grain of sand into one of these sets of limits is a much more feasible undertaking than the one of completely distinguishing a grain of sand from all of the rest.

This latter example, though farfetched, resembles in many ways the problem involved in character recognition. The beach may be thought of as being analogous to a book. Once the language in which the book is written is determined, the number of letters in the alphabet can usually also be determined. If it is assumed that this particular book is written in a language using only 26 characters, then 27 sets of limits can be defined. The 27th set of limits is added for characters not belonging to the other 26 sets of defined characters of the alphabet because of errors or other reasons. It is usually expected that the probability is very small that a character will fall within the 27th set of limits.

Character Limits

A cell[1] description rather than a detailed description is used to recognize characters. The task of finding the limits of the cells is not, a priori, obvious. It could be hoped that a true reading machine would itself determine cell limits. At present this is too much of a demand on machines and they must depend on previously determined cell limits.

For the sake of devising a character recognition system it will be necessary to define cell limits. These defined limits, not being human limits, would then cause the system to make errors that would not be made by a human reader. The converse may also occur but the probability of this seems to be very small.

The limits that have been discussed are actually very broad because the distinguishing features that should be observed from

[1] The region between a set of limits is defined as a cell.

a particular character to distinguish it should be a minimum to allow the least complication and the highest speed in recognizing characters. For the most efficient operation, any reading device, including the human reader, should make use of this minimum detail to extract the total information from each character.

Conclusions

The following seemingly obvious but important observations may be made:

(a) In recognizing a character, the amount of indecision increases as the number of characters in the alphabet increases, and likewise the amount of time required to perform the recognition also increases. (This corresponds to the result stated by Shannon (26) in information theory which says that the amount of information carried by a message increases as the number of possible messages increases.)

(b) In recognizing a character, the amount of indecision increases as the limits on each letter become more stringent, and likewise the amount of time required to perform the recognition also increases. (This corresponds to the result stated by Shannon (26) in information theory which says that as the probability of a message decreases, the information carried by that message increases. By making more stringent limits, the probability of error, or the 27th set of characters, increases, and therefore the probability of all other letters decreases, the total probability being unity.)

The time statement made in both of the above generalizations corresponds to the result in information theory which says that for a given channel the rate at which information can be transmitted is finite.[1]

INVESTIGATION OF THE HUMAN READER

General

It seems that in reading a person makes use of character outlines to recognize a character. When it is considered that these character outlines have sufficient detail to allow a person to distinguish between an infinite number of characters, there would be a tendency to assume that the human reader obtains much more information from a character than the minimum amount necessary to merely perform the recognition. If this were true, the implication would be that the information capacity, or the correct-decision-making rate, of the human reader is very much greater than it has been estimated to be. Unfortunately, our correct-decision-making mechanisms are not as fast as might appear. Rather the human memory, which at times seems to be quite poor, is largely responsible for the human reader's rapid reading rate.

[1] This implies finite bandwidth and power sources. Shannon (26) has derived equations for channel capacity in terms of power and bandwidth.

"Information" Gained in Reading

Consider the definition of information which Shannon (26) is responsible for:

$$H = \log_2 \left(\frac{\text{probability of event after message is received}}{\text{probability of event before message is received}} \right)$$

From this expression it can be seen that any information that we have about a character before we look at it reduces the amount of information that we receive by looking at it.

A person reading a text usually does not get any more information from the 100th reading of a character than that which was obtained from the 99th reading; i.e., after a few readings, a person has usually absorbed all the information about the details of a particular character that he is ever going to absorb. From then on the only effort is directed into extracting the few bits of information necessary for distinguishing a particular character from the rest of expected characters in the material being read. No attempt is made by the reader to distinguish each character from the infinity of possible characters since the reader assumes that in most cases the printer is trying to print a character of the given alphabet.

Effect of Odd Characters and Errors

Assume, for simplicity, that all of the characters of a particular font of N expected characters occur with equal frequency. Then if no odd characters are expected to occur, the information carried by each character would be

$$H = \log_2 N$$

Shannon (26) gives

$$H = - \sum_{i=1}^{n} p \log_2 p$$

as the average information per message, where the possible messages have different probability. In this application a message is a character.

If it can be expected that all characters occur with equal probability X and that odd characters occur with the probability P, then the average information per character is

$$H = -N X \log_2 X -p \log_2 P$$

Assuming, very generously, that an odd character occurs with a probability equal to any of the other characters, the average information per character is

$$H = -N X \log_2 X -X \log_2 X$$

which reduces to

$$H = -(N + 1) X \log_2 X$$

Note, however, that since the sum of the probabilities of all the characters expected is unity, that

$$(N + 1) X = 1$$

or $$(N + 1) = \frac{1}{X}$$

Therefore $$H = -\log_2 \frac{1}{(N + 1)}$$

If N is large

$$(N + 1) \doteq N$$

and

$$H = -\log_2 \frac{1}{N}$$

and the average information per character is again given by

$$H = \log_2 N$$

The results here seem to be as expected since if in some reading material there is a high frequency of unexpected symbols, the reading rate goes down. The unexpected symbols carry very much information and, according to Pierce and Karlin (24), the rate at which the human reader can assimilate information is limited and fairly constant. Conversely, an occasional odd symbol hardly affects the reading rate.

Conclusions

The memory of the human reader enables him to remember a sufficient amount of detail of the character necessary for recognition. It is thus possible for him to look at a character and receive information to allow him to distinguish a character from other characters and yet receive no information about the details of the character itself. This helps the human reader to make use of his maximum information capacity.

The memory also enables him to remember something about the probabilities of the various characters. This makes it possible for the reader to adjust his reading rate to the information content of the reading. This again helps the reader to make use of his maximum information capacity.

DESIRABLE FEATURES OF A READING SYSTEM

General

If it is desired to construct a reading device that will be able to read the various prints and writings as effectively as does a human reader, it will be necessary to build two very essential features into this device. The device should have first a large memory, and, second, a variable scanning rate.

Variable Scanning

A variable scanning rate is necessary to take full advantage of the maximum information assimilating capacity of the particular device. The scanning rate is naturally high for reading material having low information content ("quick reading") and is low for reading material having a high information content ("slow reading"). It shall be noted that this is a very difficult feature to realize. It would require that each character be coded with the same number of bits as the information that is carried by that character. The scanning rate would then have to be made inversely proportional to the number of bits carried by the character.

A reading device that does not have a variable scanning rate must scan at a lower rate than is actually possible for most reading material. This lower rate will then enable the system to recognize the high information content material that it will encounter.

Memory

In the memory it would be necessary to put much of the information about the various styles of prints and writings that most human readers have acquired through many years of reading experience. This would be a large problem in more ways than one. It is obvious that with the present constructional limitations, the physical size of the memory unit would be prohibitive because of the large amounts of information that would have to be stored in it. But more basic than this is the fact that it is not at all clear what should be placed into this memory unit. If, for example, a reader should be asked to reproduce the small letter form of "G" from a certain type print which he has read often, it is very probable that he would not do it with any degree of accuracy; it probably would not be done in the same way by different persons. What, then, should be placed into the memory?

Versatility and accuracy will have to be sacrificed if the system does not have a large memory. The system will be able to recognize the many varied types of prints and the various distortions and errors possible in printing only if the possibility for recognition of each of these particular variations is built into the system's memory.

In order to keep the size of the machine memory to a minimum, it will be necessary to have the sensing device sense the minimum number of bits of information necessary for recognition from each character. This also keeps the construction of the sensing device as simple as possible.

LIMITATIONS OF CHARACTER RECOGNITION

It should be noted that so far nothing has been said about the fact that in reading a person may not even look at any one character but will observe letter combinations, or the word as a whole, or possibly even groups of words. Recalling all of these various combinations of letters and words plus the many characters implies an even larger memory, larger by no smaller factor than the memory which has already been termed prohibitive. For the sake of simplicity, and of achieving some practical results, the possibilities along these lines will not be discussed any further, important and significant as the possibilities seem to be.

PRESENT METHODS OF CHARACTER RECOGNITION AND THEIR LIMITATIONS

General

It would be interesting to investigate some character sensing methods that have been realized or suggested and the principles that each of them utilizes to achieve recognition. Their advantages and disadvantages can be determined in the light of the preceding discussion. The present methods to be discussed will then serve as a reference for comparison for the new method to be introduced.

It should be noted that all of the methods to be described will be in no way comparable to a human reader as far as versatility in reading the various types of reading materials is concerned. However, the one major advantage that all of these systems will have over the human reader is the tremendously higher speeds at which they can or will be able to operate.

The need for high speeds in character recognition is felt mostly by the business machines people. They would surely be willing to print their data in a special way if the human translator of the printed language to a machine code could be replaced by a high-speed character recognition system. For a short period this limitation would require human translators to translate from any given print to the type print that could be recognized by the system; hence no saving would be realized. But in time most of the printing done would be of the type that the machine could use and the human translators would be eliminated.

Rectangular Mosaic of Photocells

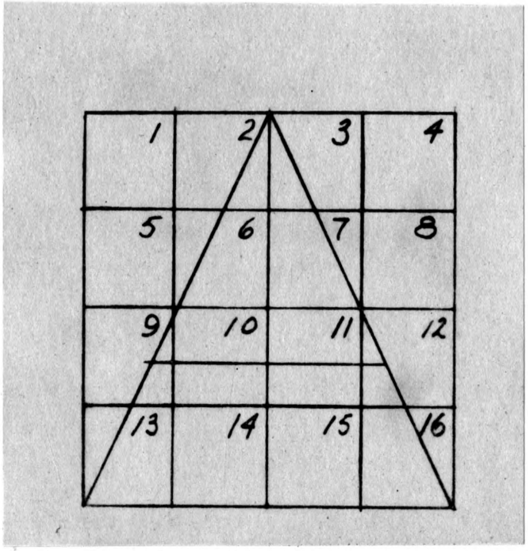
This method seems to be the most intuitive of all to be described. To achieve character recognition each character area is divided into a rectangular array of smaller areas as shown in Plate I, Fig. 1. The light that is reflected from each one of these smaller areas will energize a particular photocell. If a character is occupying a character area some of the photocells will not be energized since the character itself is black and

EXPLANATION OF PLATE I

Fig. 1. Schematic of rectangular mosaic of photocells receiving the image of the letter A. The coding for the letter A is given at the bottom. ("1" indicates the character image crosses a photocell; "0" indicates the character image does not cross a photocell.)

Fig. 2. Schematic showing an example of a closed path and an incomplete path that an automatic pantograph tracer would follow.

PLATE I



A = 0110, 0110, 1111, 1001

Fig. 1.

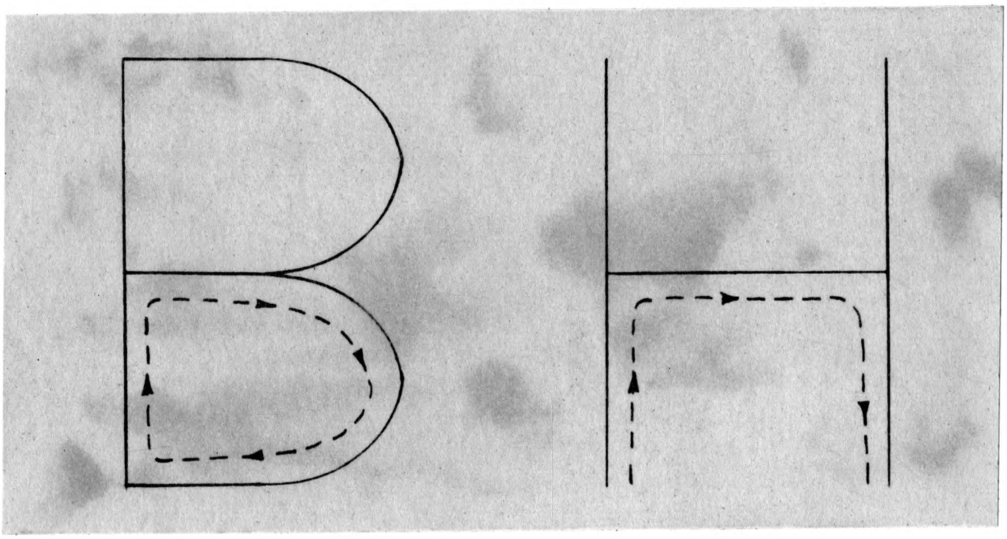


Fig. 2.

reflects no light. If the number of smaller areas is large enough, the combination of energized and unenergized photocells will be unique for each character of a particular type print.

The following discussion by Loeb (19) will make clear the major disadvantages of a device using this principle.

> Quantize the problem by dividing the area into a finite number of small squares. Squares are black where there is drawing, white where there is not. If N is the number of squares, then the number of possible drawings is 2^N. If all these patterns are assumed equally probable, then the entropy per pattern is $H = N$. The entropy of a simple message will be less if it is known beforehand that the message will be simple (that is, we may know that only five per cent of the total number of squares will be black).[1] It is therefore worth while with a view to saving bandwidth or power transmission line to develop methods that are more suitable for simple drawings. A facsimile device transmits the total amount of information on a given page. It expends as much effort in areas having little information content as in areas having high information content. The machine must answer the question, "Is there a black dot in this point?", even when there is no such possibility.

As an illustration, consider the sketch showing a four by four, or 16-cell, mosaic. It is possible for this mosaic to be used for a 37-character alphabet (26 letters, 10 numbers, and a period), and still have each character defined uniquely, but anything less than this, such as a three by three, or nine-cell, mosaic, which can distinguish between 2^9 or 512 patterns, will give the same photocell combinations for some of the different characters as commonly recognized. If it is considered that it is possible to distinguish between 2^{16} possible patterns with 16

[1] Loeb (19) gives a proof on page 318 that shows that the entropy of a simple message is less if it is known beforehand that the message is simple.

photocells, a large degree of overdesign is indicated. However, since a system of this type takes more information from the character than is actually necessary for recognition (system is called redundant), error-correcting schemes can be built into it, whereas any scheme that extracts only the minimum amount of information necessary for recognition can tolerate no errors.

In a device of this type the scanning mechanism would have to scan in jumps of one character width and information could be registered only when the photocell mosaic corresponded exactly with the character area. In order to achieve this, either the characters will have to be printed in predetermined areas and then all of these areas will have to be checked for presence of characters, or each character will have to carry additional information about its location. In this latter case a servomechanism is implied since the sensing device will need information about where it is going to scan which will depend on the position that it is presently in.

Pantograph or Letter-tracing Device

In this scheme the outline of each character is traced at some fixed speed. This motion is then broken up into its "x" or "y" variations against time. These variations will be unique for each character, and hence recognition will be possible.

Loeb (19) stated that a major advantage and disadvantage of a pantograph-type device are that:

> In a pantograph device there is no scanning and the

> receiving pencil has no possibility of indicating any points that are not spotted by the transmitting one. Yet this device is able to give more information than the amount that is strictly required. The speed at which the curve is traced is also transmitted. High speeds would require wide frequency characteristic channels.

Another major disadvantage seems to be in the design of the tracing device itself. It is, of course, assumed that the tracing will be done automatically by some sort of a servo arrangement. Such a system has been described by R. A. Fairthorne in Loeb's (19) paper but has not been constructed because of prohibitive cost. A quick glance at some of the characters will show that there are many closed paths that a tracing device could follow, and therefore never reach the end of the character. Another possibility that is apparent is that the tracing device be shunted at certain character junctions so that it will never trace out the character completely (Plate I, Fig. 2). These difficulties could be avoided by stylizing the alphabet with this problem in mind, but it seems that if the alphabet must be stylized other sensing devices requiring simpler mechanisms may be used instead.

Scanning Light Beam or Beams

In this scheme a light beam sweeps across a character area and a photocell senses the light beam crossing the character. The direction of sweep, the number of light beams, whether variable or constant sweep velocity be used, and the way information is used to achieve recognition, is in no way necessarily unique in this scheme. Several systems of this type will be described.

Parallel Light Beams. Zworykin (34) described as follows a system which he had constructed:

> In order to recognize letters, the scanner of the device divides the line of type into a number of horizontal bands as shown in Plate II, Fig. 1, each of which it explores with a spot of light. The light reflected from each band is converted into an electrical signal by a phototube. Counting circuits interconnected with the scanner count the number of times each light spot encounters part of a letter as it traverses its zone of line of type recording the total number of black areas per letter per zone. From this information other circuits in the device recognize the letter or letters scanned.
>
> To avoid the bulk of eight light sources and phototubes, a combination of flying-spot and time-division techniques is used.

"Stopping-spot" Process (Variable Speed Sweep). Loeb (19) suggests:

> Scan with a "flying spot". The spot speed will be very high so long as no black point is found. When one is found, the spot will stop and a pulse code signal will be transmitted which specifies how much time has passed since the last encounter of the traced line and the spot will not resume its movement until complete transmission of the code signal has occurred. The total pulses will not contain more information than the necessary coding of the curve (character). Also no limitation as to the per cent of black spots in the area will have to be imposed.[1]

Proportional Parts System. This system has been constructed by the International Business Machines Corporation (Greanias, 12). Serial scanning is performed from top to bottom and right to left across the characters with a fine spot of light along closely spaced vertical lines. The light signals reflected from

[1] Loc. cit.

EXPLANATION OF PLATE II

Fig. 1. Illustration of scanning in parallel-light beams' scheme.

Fig. 2. Examples of sets of characters that would give the same machine code in the parallel-light beams' scanning scheme.

PLATE II

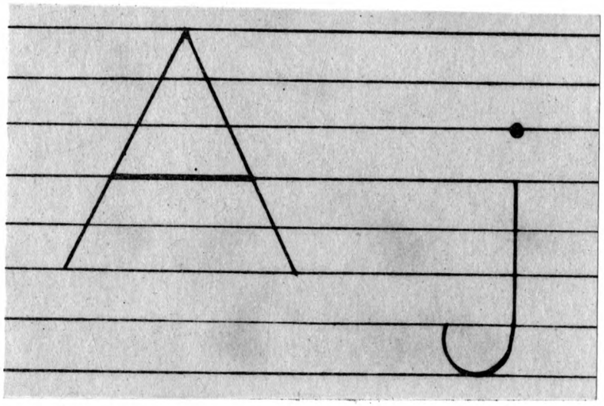


Fig. 1.

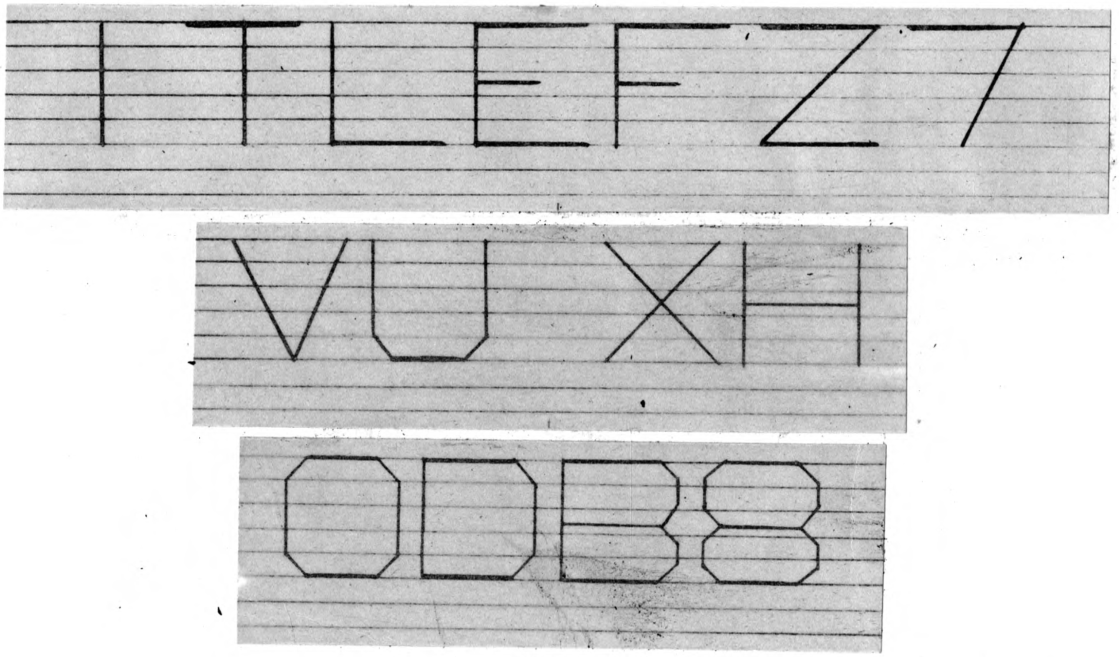


Fig. 2.

the paper are first quantized into black and white cells. The binary data from each vertical scan are then coded into three decimal digits which designate the number, position, and size of the black areas in the scan. The first digit relates to the number and proportions of black areas detected. The second digit indicates the change in altitude of the top of the character relative to a selected earlier scan. The third digit designates the distance between extreme black areas in the scan. Individual characters can be represented by a succession of coded scan data, and the basic recognition criteria for each character can be defined as a sequence of these codes.

All three of these methods have the disadvantage of having to sense many more details of a character than should be required for recognition. This extra amount of detail which has to be processed makes it necessary for the sensing mechanism to become more complicated. The recognition circuitry will also have to be more complicated since the extra detail will necessitate the use of a longer machine code than is necessary. In the case of binary coding, the number of digits in the character coded with the largest binary number will determine the least amount of apparatus necessary.

The "stopping spot" and the "proportional parts" systems would require mechanisms which are complex and expensive. Whereas the parallel light beams seem fairly simple to achieve, there are characters which cannot be distinguished from one another by this system. Some examples are shown in Plate II, Fig. 2. In order to make these characters distinguishable, major steps would have to be taken in stylizing the alphabet.

The Scanning Slit. Kelner (15) has constructed a device that performs character recognition. The principle of this scheme is to scan each character area with a slit of light rather than a beam of light. The slit should be long enough to cover the length of the character area while being as thin as possible. The reflection from this slit of light will strike a photocell. (Plate III, Fig. 1.) As this slit scans the character area, the amount of light being reflected from the character area varies (Plate III, Fig. 2). It is possible to obtain unique variations from the various characters to be recognized. The major difficulty with this scheme seems to be in transforming the many variations of the photocell into a form that will be useful to a computer. Also some stylizing will have to be done for characters such as are shown in Plate III, Fig. 3, where the variations would not be essentially different. The system that has been constructed can only recognize the ten characters of our decimal system. As far as the sensing device is concerned, this system would seem to have the advantage of being capable of being operated at very high speeds.

PROPOSED SYSTEM

Discussion

The sensing device and the method of recognition in this proposed system are not essentially different from those used in

EXPLANATION OF PLATE III

Fig. 1. Illustration of slit of light moving across a character in the scanning slit scheme.

Fig. 2. Response of phototube against distance scanned for the letter E shown in Fig. 1 of this plate.

Fig. 3. Examples of sets of characters that would give similar responses in the scanning slit scheme.

PLATE III

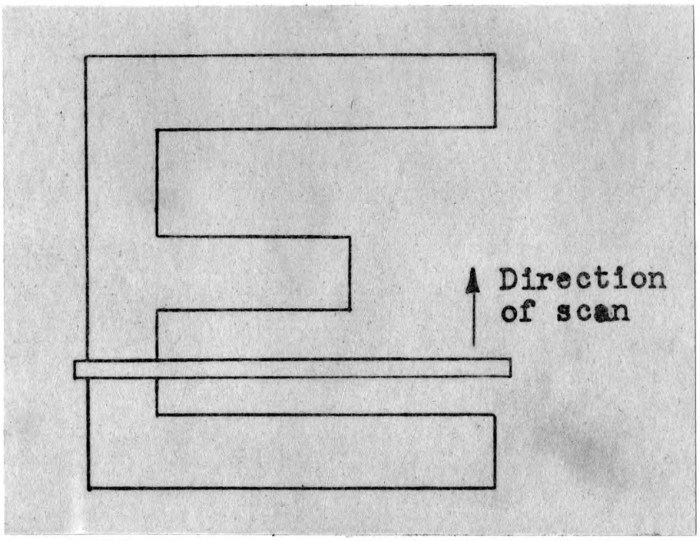


Fig. 1.

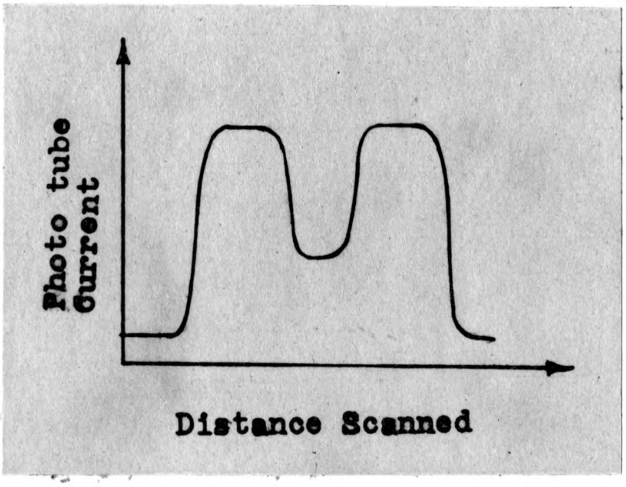


Fig. 2.

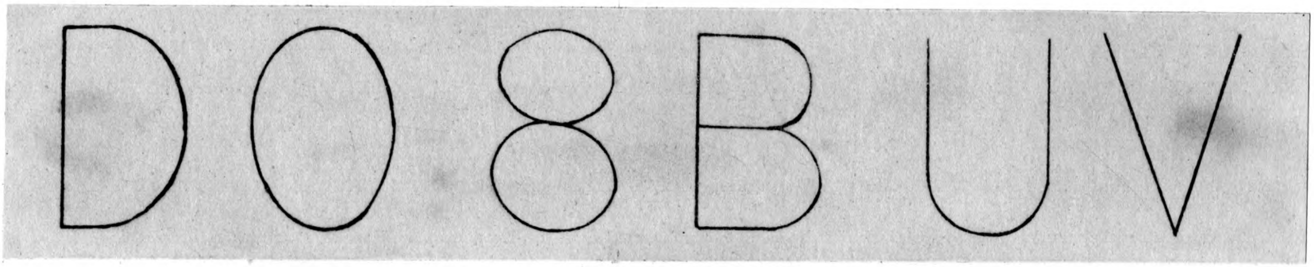


Fig. 3.

the system utilizing the rectangular mosaic. The essential difference lies in the idea of making use of as little character detail as possible for recognition. This makes the number of photocells used and their locations in the character area critical.

The proposal is to locate the minimum number of photocells necessary for recognition in such a way that the combination of excited and unexcited photocells is unique for each character.

The amount of information carried by each character in an alphabet of N letters, each letter having equal probability of occurrence, is

$$H = \log_2 N \text{ bits}$$

The amount of information in bits is also equal to the minimum number of photocells required for recognition.

The question now comes up as to whether the location of this minimum number of photocells is necessarily unique. A simple illustration will show that the locations are not unique.

Consider the four-letter alphabet shown in Plate IV, Fig. 1. Here the number of photocells required for recognition is the same as the information per character, which is

$$H = \log_2 4 = 2 \text{ bits}$$

or two photocells are required. Let it be stipulated that all characters lie in character areas of the same dimensions. Let the two photocells be located in the character area as shown in Plate IV, Fig. 2. Plate IV, Fig. 3, shows the location of the photocells with respect to each of the characters, and Plate IV, Fig. 4, shows that the combinations of excited and unexcited

EXPLANATION OF PLATE IV

Fig. 1. The four characters of a four-letter alphabet.

Fig. 2. Location of phototubes in a character area. Character area is defined by the dotted line.

Fig. 3. Location of phototubes with respect to each character of the four-letter alphabet defined in Fig. 1 of this plate.

Fig. 4. A chart showing the unique combination of excited and unexcited photocells for each of the characters.

PLATE IV

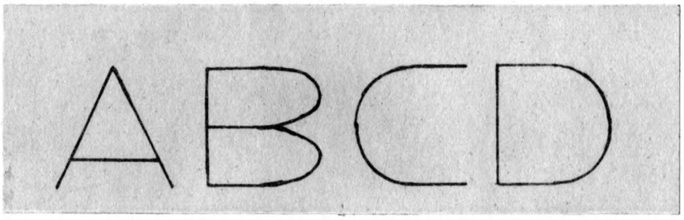


Fig. 1.

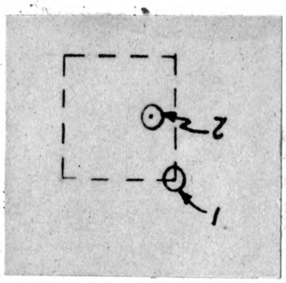


Fig. 2.

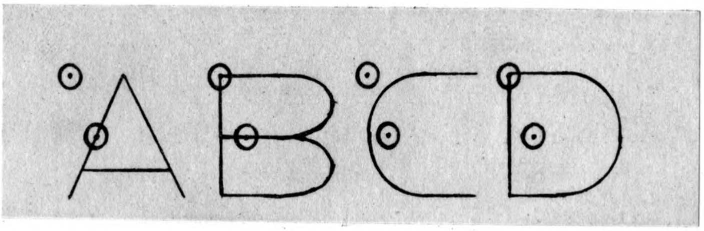
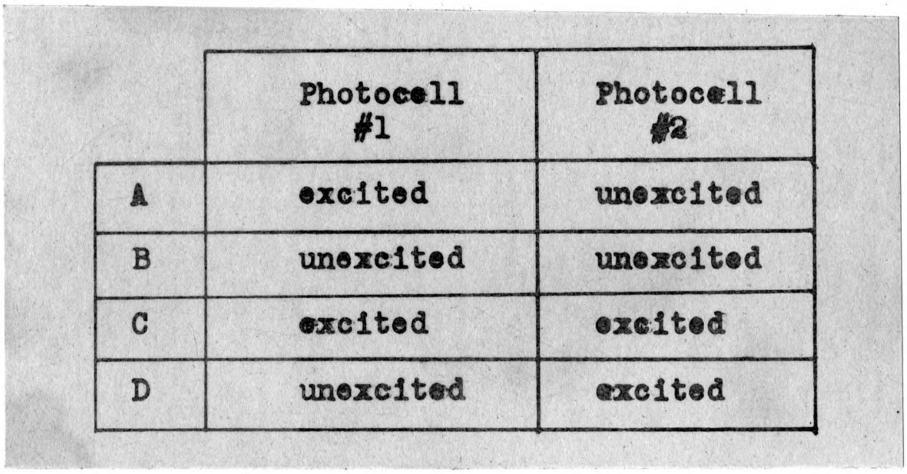


Fig. 3.



	Photocell #1	Photocell #2
A	excited	unexcited
B	unexcited	unexcited
C	excited	excited
D	unexcited	excited

Fig. 4.

photocells are different for each character.

Now let the location of the photocells be changed to the positions shown in Plate V, Fig. 1. Plate V, Fig. 2, shows the location of the photocells with respect to each of the characters, and Plate V, Fig. 3, shows that the combinations of excited and unexcited photocells is again different for each character.

This shows that the ways of choosing the minimum amount of distinguishing features is not necessarily unique.

The determination of the location of the photocells is essentially a matter of trial and error. However, due to the fact that the locations are not unique, the effort involved in performing the trial-and-error process is not great. There are many location possibilities for each photocell; therefore the probability of guessing the locations correctly is fairly high.

Details

In order to determine the location of these photocells, all the characters of the particular alphabet being used should be superimposed on one character area[1], as shown in Plate VI. As a

[1] If a bold print alphabet is superimposed, the result will undoubtedly be a meaningless black smudge. In this case each letter will have to be trimmed; that is, each letter must be drawn with thin lines having the same form as the heavy print. Once the photocell locations are determined with the thin line drawings, these same locations can be used for the bold print as long as the locations of the photocells are not too close together. If the thin line drawings bisect the heavy print, then the photocells should be no closer to each other than one heavy line thickness, nor should a photocell be within one heavy line thickness of a character which it is not to detect.

EXPLANATION OF PLATE V

Fig. 1. Location of phototubes in a character area. Character area is defined by the dotted line. Note these locations are different from those shown in Plate IV, Fig. 2.

Fig. 2. Location of phototubes with respect to each character of the four-letter alphabet defined in Fig. 1 of Plate IV.

Fig. 3. A chart showing the unique combinations of excited and unexcited photocells for each of the characters.

PLATE V

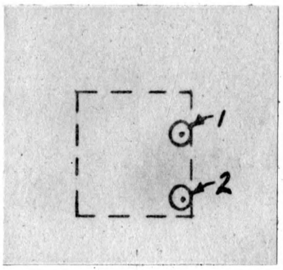


Fig. 1.

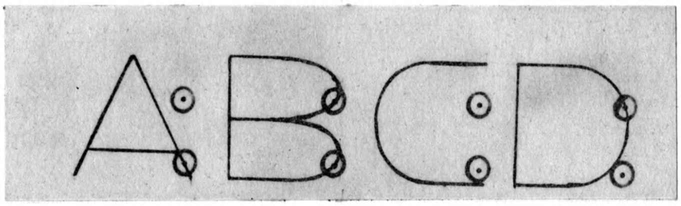
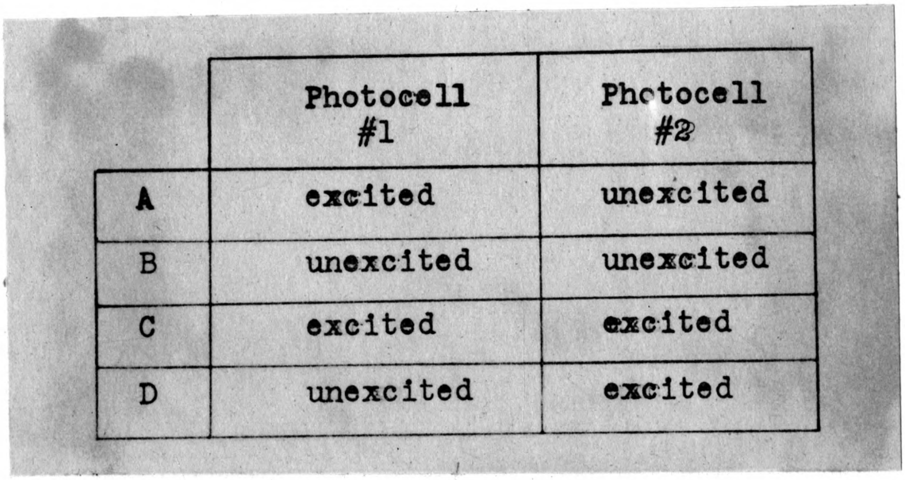


Fig. 2.



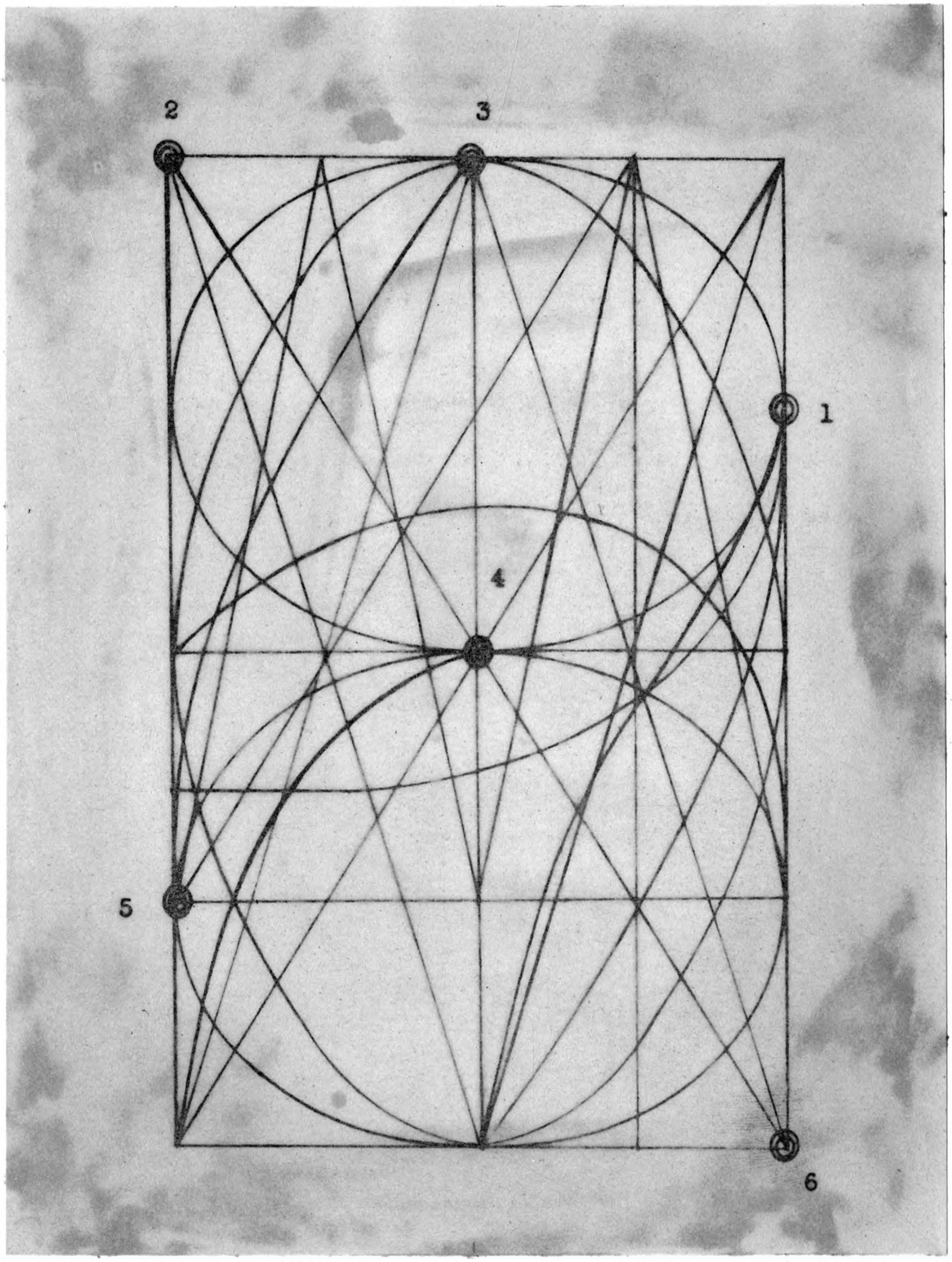
	Photocell #1	Photocell #2
A	excited	unexcited
B	unexcited	unexcited
C	excited	excited
D	unexcited	excited

Fig. 3.

EXPLANATION OF PLATE VI

The location of six phototubes is shown on character area which has all 37 characters superimposed upon it.

PLATE VI



first trial, the position of the photocells can be completely arbitrary; however, points where many of the characters intersect each other will probably give the best results. Each of the photocells then can be numbered so that it corresponds to a binary digit of a binary number that will represent a character. When a character crosses a point occupied by a photocell, the photocell will be unexcited and the digit represented by that photocell will be called "1". If the character does not cross that point, the digit will be called "0". A chart can now be made as shown in Plate VIII, where each character will have its corresponding binary number written next to it. If all these binary numbers are unique, the job is done. If some of the binary numbers are the same, then a new location must be found for some of the photocells. If it is found that it is not possible to locate these photocells so that a different binary number corresponds to each character, it will be necessary to have some of the photocells correspond to areas rather than points.

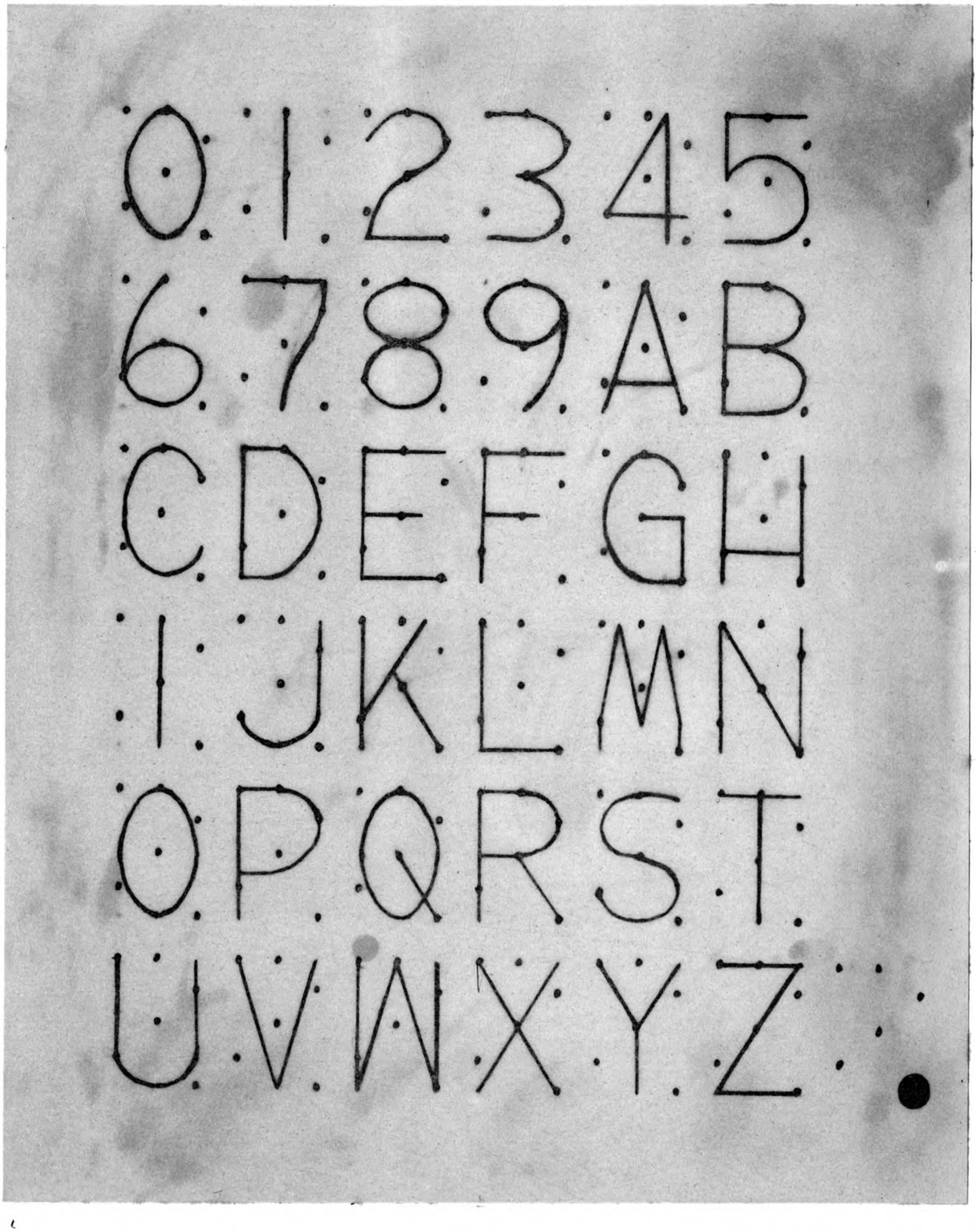
The location of six[1] photocells (Plate VI) for the recognition of a 37-character alphabet (Plate VII) has been achieved. A chart giving the binary number for each character is shown in Plate VIII. This alphabet does not correspond to any particular type of print or any combinations of prints; in fact, some of the characters have been stylized to facilitate the location of photocells. The photocells here are all focused on points. It will be noticed, however, that none of the characters will offer

[1] $H = \log_2 37 = 5.2$. Therefore six photocells are required.

EXPLANATION OF PLATE VII

The 37 characters of a 37-letter alphabet.

PLATE VII



EXPLANATION OF PLATE VIII

A chart showing the binary number representing each character.

PLATE VIII

0	001000
1	001100
2	101101
3	111100
4	000010
5	011000
6	000110
7	111000
8	101110
9	101100
A	001001
B	111110
C	101000
D	011010
E	011111
F	011110
G	101011
H	110011
I	001100
J	100010
K	010111
L	010011
M	000011
N	110111
O	001000
P	111010
Q	001101
R	111111
S	001110
T	011100
U	110010
V	010000
W	111011
X	010101
Y	010100
Z	011101
.	000001

the reader any difficulty as far as quick recognition is concerned. The possibility is not denied that someone with more artistic talents may have achieved a more beautiful looking alphabet using the same photocell locations. No attempt was made to distinguish between the letter "O" and the number "0", or between the letter "I" and the number "1". Plate IX shows a schematic of the recognition circuit, and Plate X shows a block diagram of the whole system.

The major advantages of this system are the small amount of apparatus required and the simplicity of the recognition circuits used. Scanning would have to be done in steps of one character width and would have to be done over all possible character areas unless some additional information is put into the character.

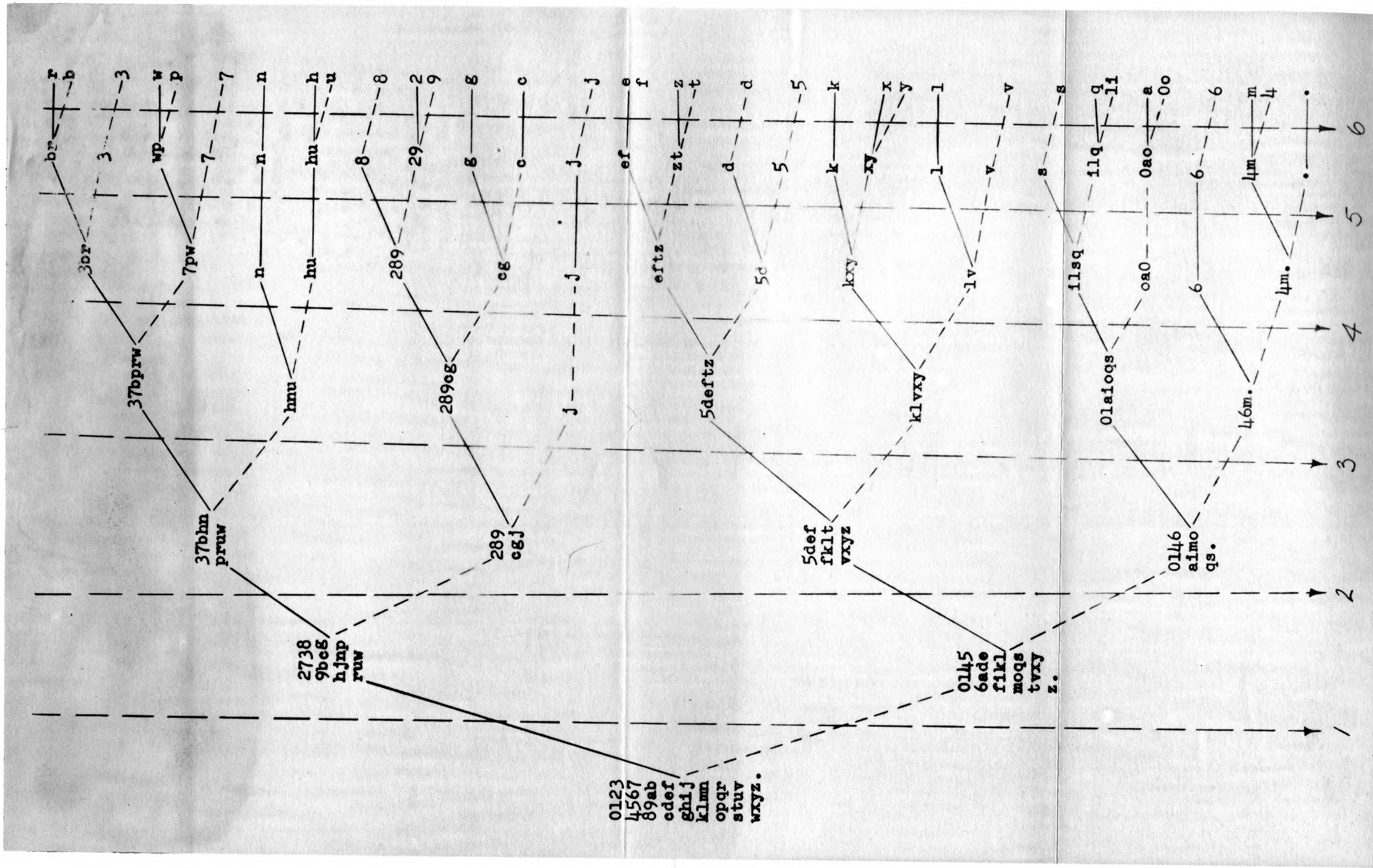
If uniform scanning is used, the time consumed for each character will be at least equal to the time required for the coder (see Plate X) to transmit the code of the character having the longest code. However, it is possible to achieve a variable scanning rate in this system. The coder can be constructed so that the length of the code for each character is roughly inversely proportional to the probability of the occurrence of that particular character. The coder could then be made to trigger the scanning mechanism after a character is coded. In this way variable scanning is achieved in a fairly simple manner.

Certain type errors in printing will cause errors in recognition. If it is desired to reduce these errors, redundancy may be added to the system in the form of additional photocells. If full advantage is to be taken of these additional photocells, it

EXPLANATION OF PLATE IX

Standard representation of a recognition circuit. Each long-dashed line can be the "flip-flop" which is excited by the phototube whose number is shown.

PLATE IX


0123 4567 89ab cdef ghij klmn opqr stuv wxyz.
2738 9bcg hjnp ruw
0145 6ade fjkl moqs tvxy z.
37bhn pruw
289 cgj
5def fklt vxyz
0146 aimo qs.
37bprw
hnu
289og
j
5deftz
klvxy
01aioqs
46m.
3br
7pw
n
hu
289
cg
j
eftz
5d
kxy
lv
11sq
oa0
6
4m.
br
3
wp
7
n
hu
8
29
g
c
j
ef
zt
d
5
k
xy
1
v
s
11q
0ao
6
4m
.
r
b
3
w
p
7
n
h
u
8
2
9
g
c
j
e
f
z
t
d
5
k
x
y
1
v
s
q
11
a
0o
6
m
4
.
1
2
3
4
5
6

EXPLANATION OF PLATE X

Block diagram of the printed language to machine code translator.

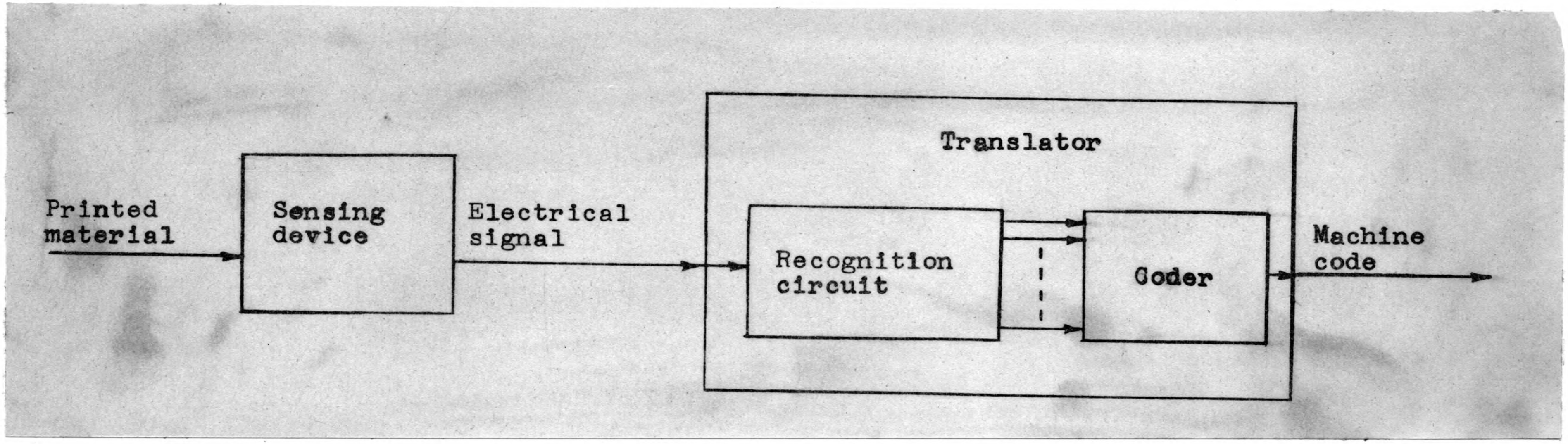
Block description:

Sensing device - The particular arrangement of the phototube and the scanning mechanism.

Recognition circuit - Shown in Plate IX.

Coder - A device that will give a digital signal when excited.

PLATE X


Printed material
Sensing device
Electrical signal
Translator
Recognition circuit
Coder
Machine code

would be necessary to know the type of errors that will be expected for the particular characters and to place the photocells so that recognition is still possible. The number of additional photocells will, of course, depend on how much it is worth to reduce certain types of errors.

Possibility of Using This Device as Part of an Ideal Character Recognition System

The possibility of these phototubes finding a suitable location automatically should be investigated. If this could be achieved, then a single system would be able to translate various types of print into machine codes. A system of this type would not require a large memory since it would learn about a new type of print by having a few samples of the print fed into it. The system would then relocate its phototubes so that it could recognize the new print.

However, in the process of relocating its phototubes, the system is actually forgetting about the type of print that it was previously set to recognize. In other words, while this system is in the process of translating it will be able to recognize only one type of print. Any changes in print will cause the system to make errors unless the system is instructed to relocate its phototubes. A system of this type will not be useful where the type of print being read changes with a high frequency since the process of relocating the phototubes will require samples of print for relocating the phototubes before recognition can begin.

The following features will be difficult to build into the system:

1. Having the system recognize new types of print as opposed to errors in old print so that the system will know when to start relocating its phototubes.

2. Having the system choose the proper samples so that one of each letter will be selected for sampling. Hence the phototubes will find their proper locations for recognition of all characters of the alphabet.

3. Having the system make the proper connections for each new type of print so that one particular output terminal of the system represents the same letter of the alphabet for all types of print. (It is very probable that different binary numbers will represent the same letters in different type prints; therefore the same letter would cause different output terminals to be excited unless some new connections were made when the phototubes were relocated.

It may be necessary for the sake of obtaining an immediate practical solution to have the above mentioned functions performed by a human operator until automatic means are devised.

It may be argued that all of this complication and expense can be avoided by using a rectangular mosaic type system with a large number of photocell elements. A fine mesh mosaic of this type will allow the recognition of many types of print. But it should be remembered that in order to perform the recognition, information about the various types of print will have to be known and placed into the device beforehand. On the other hand, a system that would be able to relocate its phototubes after a few exposures to a print would be able to recognize prints which were nonexistent at the time the system was constructed.

A method for relocating the phototubes will be described. The assumption will be made that a sample of each character of the print to be recognized is available to the system in the desired positions.

It will be recalled that in the trial-and-error process of determining these phototube locations, as a first trial, points were sought in the character area where many of the characters crossed. It will not always be possible to superimpose all of the characters in one character area to determine these points. Bold print will blacken the character area completely and in most cases it will be impossible to make any distinctions at all. When a human operator is going to determine the phototube locations, this difficulty may be overcome (footnote 1, page 30). However, if the locations are to be found automatically, the described scheme may not be used.

It is suggested that each of the possible characters of the alphabet to be read be scanned simultaneously and completely by separate beams of light. In the alphabet that has been used, 37 beams would be required. The reflections of each of the scanning light beams would then be sensed by separate photocells. The output of these photocells would then be put into a summing amplifier. The output of this summing amplifier could be expressed in terms of the number of crossings made by all the characters of the alphabet at a particular point of the character area.

The exact point of the character area would be a function of time, considering zero time to be the instant of time that the scanning is begun. A large output would indicate few characters crossing a particular point on the character area, while small outputs would indicate a large number of character crossings. The most desirable points for ultimate phototube locations would be those points having about one-half of the characters going through them.

Since this is a trial-and-error process, it is quite possible that the locations so found will not provide a different binary number for each character of the alphabet. This probability can be made small by using more than the minimum number of phototubes required for recognition. A block diagram for this automatic relocating is given in Plate XI.

EXPLANATION OF PLATE XI

Block diagram showing how automatic location of phototubes is achieved.

Block description:

Scanning mechanism - Produces mechanical motion so that each character may be scanned from left to right and top to bottom.

Light sources - Beams of light to be driven by the scanning mechanism. There should be as many light sources as characters.

Sampled printed characters - A sample of each character possible should be made available to the system.

Phototubes - A separate phototube should be available for each light beam.

Summing amplifier - Combines the outputs of the phototubes to give an output proportional to their sum.

Biasing voltage - Bias should be made so that zero voltage would represent half of the characters crossing a particular point.

Absolute value - All negative voltages made positive voltages of same magnitude.

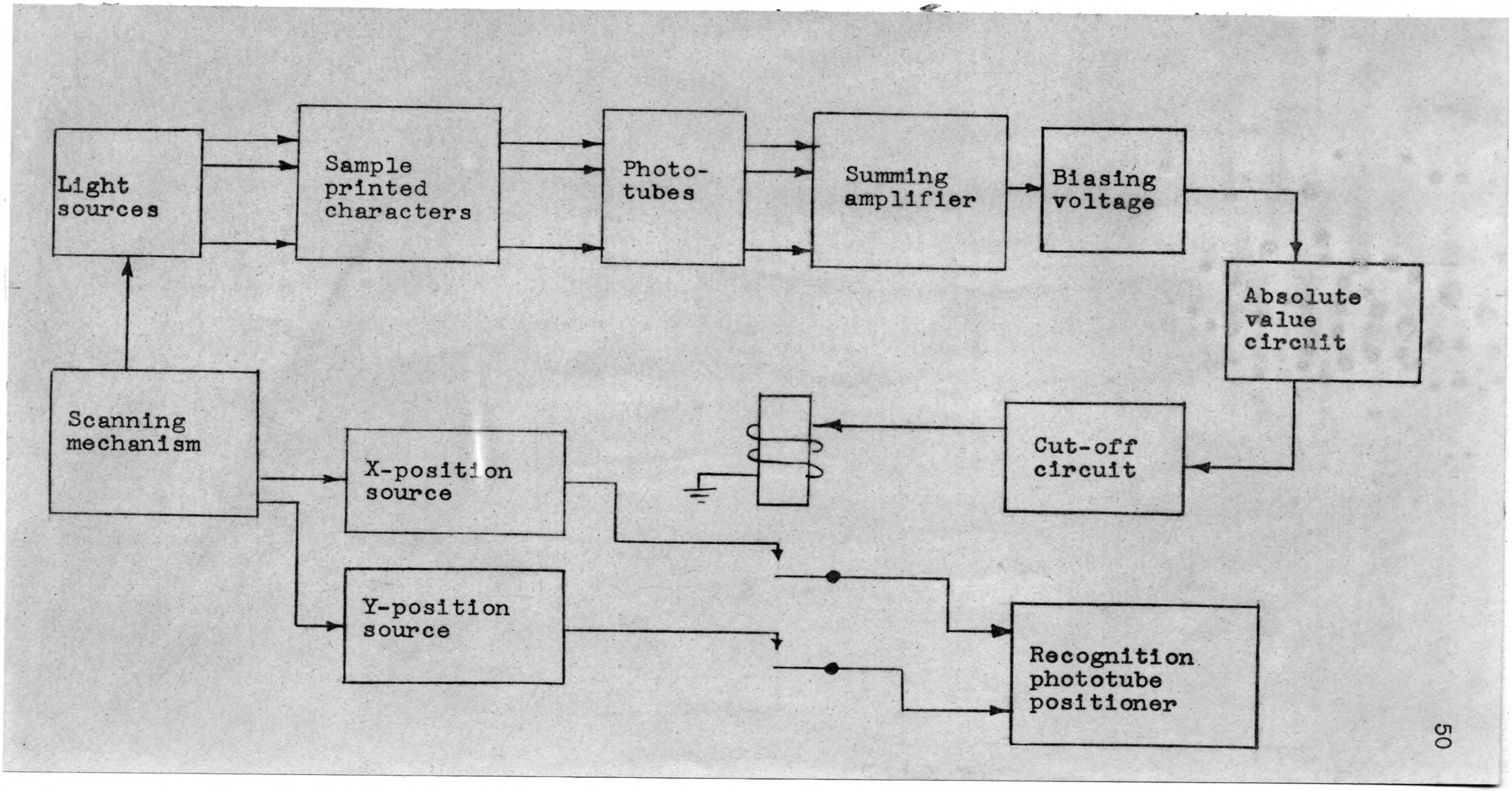
Cut-off circuit - Would give an output for a range of voltages slightly above zero. The actual value would depend on the number of character crossings that a point would be required to have for a phototube to be located there.

Y-position source - A voltage proportional to the number of scans from the top of the character area.

X-position source - A voltage proportional to the travel of each individual scan.

Recognition phototube positioner - A driving mechanism that will position the recognition phototubes (as opposed to the scanning phototubes) according to the outputs of X- and Y-position sources.

PLATE XI


Light sources
Sample printed characters
Photo-tubes
Summing amplifier
Biasing voltage
Absolute value circuit
Scanning mechanism
X-position source
Y-position source
Cut-off circuit
Recognition phototube positioner

ACKNOWLEDGMENTS

The writer sincerely appreciates the aid and guidance so generously given by the two major instructors, Dr. C. A. Halijak and Professor Joe E. Ward, Jr., in the preparation of this paper.

The writer is also deeply grateful for the suggestions and ideas given to him by Mr. Ugo Gagliardi, Professor P. G. Kirmser, and Mr. Isaac Wakabayashi.

BIBLIOGRAPHY

1. Adrian, E. E. "The Physical Background of Perception." *The Waynflete Lectures*. Oxford: Clarendon Press, 1946.

2. Adrian, E. D. "Sensory Integration." *The Sherrington Lectures I*. England: University of Liverpool, 1949.

3. Bar-Hillel, Y. "An Examination of Information Theory." *Philosophy of Science* 22 (April, 1955) 86-105.

4. Bar-Hillel, Y. "Linguistic Problems Connected with Machine Translations." *British Journal of Philosophical Science*. 20 (July, 1953) 217-225.

5. Bar-Hillel, Y. "The Present State of Research on Mechanical Translation." *American Documentation*, 2(April, 1952).

6. Bar-Hillel, Y. "Can Translation Be Mechanical?" *American Scientist*. 42 (April, 1954) 248-260.

7. Brandt, H. J. *The Psychology of Seeing*. New York: The Philosophical Library, 1945.

8. Cherry, Colin. "A History of the Theory of Information." *Transactions of the Institute of Radio Engineers*. PGIT-1 (February, 1953) 37.

9. Cherry, Colin. *On Human Communication*. New York: Technology Press of M.I.T. and John Wiley and Sons, 1957.

10. Diringer, D. *The Alphabet*. London: Hutchinsons Ltd., 1948.

11. Goldman, Stanford. *Information Theory*. New York: Prentice-Hall, 1953.

12. Greanias, E. C., and others. "Design of Logic for Recognition of Printed Characters by Simulation." *I.B.M. Journal of Research and Development*, January, 1957, 1:8-18.

13. Hebb, C. *The Organization of Behavior*. New York: John Wiley and Sons, 1949.

14. Hollander, G. L. "Bibliography on Data Storage and Recording." *American Institute of Electrical Engineers Transactions*. 73 (March, 1954) 49-58.

15. Kelner, R. C., and M. H. Glauberman. "Magnetic Shift-Register Correlator." *Electronics*. 29 (August, 1956) 172-175.

16. Klemmer, E. T., and P. F. Muller, Jr. "The Rate of Handling Information--Key-pressing Responses to Light Patterns." *Human Factors Oper. Research Lab. Memo Report.* Washington: Air Research and Development Commission, Bolling Air Force Base, 34 (March, 1953).

17. Kris, Ernst. *Psychoanalytic Explorations in Art.* New York: International Universities Press Inc., 1952.

18. Locke, W. N. and A. D. Boothe. *Machine Translation of Languages.* The Technology Press of Massachusetts Institute of Technology, 1955.

19. Loeb, J. "Communication Theory of Transmission of Simple Drawings." 317-323. *Communication Theory*, edited by Jackson Willis. New York: Academic Press Inc., 1953.

20. MacKay, D. M. "On Comparing the Brain with Machines." *American Scientist*, 42 (April, 1954) 261.

21. *Mechanical Translation* (A bibliography), Vol. 1, No. 1. Massachusetts: Massachusetts Institute of Technology, Room 14-N307.

22. Newman, E. B., and L. S. Gerstman. "A New Method for Analyzing Printed English." *Journal of Experimental Psychology*, 44 (1952), 114-125.

23. Oswald, Jr., V. A., and S. L. Fletcher. "Proposals for the Mechanical Resolution of German Syntax Patterns." *Modern Language Forum.* 36 (1951) 81-104.

24. Pierce, J. R., and J. E. Karlin. "Reading Rates and the Information Rate of a Human Channel." *The Bell System Technical Journal*, 36 (March, 1957) 497-516.

25. "Research on Reading Machines for the Blind." *National Academy of Science, Common Sensory Devices.* Haskins Lab., 1947. 131 p.

26. Shannon, C. E., and W. Weaver. *The Mathematical Theory of Communication.* Urbana: University of Illinois Press, 1949.

27. Shepart, David H. "The Analyzing Reader." A paper presented before a meeting of the *Association of Computing Machinery*, Cambridge, Mass.

28. Stumpers, F. L. "Information Theory Bibliography." *Research Lab. of Electronics of Massachusetts Institute of Technology.* February 2, 1953.

29. "Typed Figures Translated into Computer Code." Engineering. 183 (Friday, March 15, 1957).

30. Walter, G. "The Functioning of Electrical Rhythms in the Brain." Journal of Mental Science. 96 (1950) 1.

31. Weiner, N. Cybernetics. New York: John Wiley. 1948.

32. Wolfe, J. M. A First Course in Cryptanalysis. New York: Brooklyn College Press. 1943.

33. Young, J. Z. "Doubt and Certainty in Science." Reith Lectures. London: Oxford University Press, 1951.

34. Zworykin, V. K., and others. "Letter Reading Machine." Electronics. 22 (June, 1949) 80-86.

PRINTED LANGUAGE TO MACHINE CODE TRANSLATION

by

HENRY D'ANGELO

B. S., College of the City of New York, 1955

AN ABSTRACT OF
A THESIS

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Electrical Engineering

KANSAS STATE COLLEGE
OF AGRICULTURE AND APPLIED SCIENCE

1957

This paper deals with the problem of finding what is required to achieve true character recognition, and then introduces a system which achieves a limited form of character recognition.

In trying to determine what essential features a system would be required to have, the human reader was investigated. Shannon's principles of information theory were applied throughout and the results that were obtained have strong intuitive justification.

It was found that a true character recognition system that takes full advantage of its information capacity requires a large memory unit and a variable rate of scan which depends on the information content of the material. It was also found that the information to be placed into the memory is not readily available, nor can it easily be found, even if it were physically and economically feasible to construct the large memory unit. At present it is not feasible. Either limiting the size of the memory unit or the lack of a variable scanning rate will limit the usefulness of the system. A small memory will limit the number of characters that can be recognized, while not having a variable scanning rate will reduce the possible reading rate of a given system.

The system that is introduced has a very limited memory, and is therefore quite limited. It can distinguish between 37 characters (26 letters, 10 numbers, and a period). The discriminating way in which this system samples the characters to be recognized marks the major deviation of this system from others which have been constructed or suggested. This deviation makes it possible to code the printed character into a machine code using the least amount of apparatus. A variable scanning rate may or may not be used, depending on desired expense. The chief advantage of this system is its simplicity and the way it makes the best possible use of the memory available. Trial and error are required in the design stages, but due to the many possibilities the task is fairly simple. An automatic method of performing the trial-and-error design is described.