NATIONAL BUREAU OF STANDARDS REPORT

3804

A Study of Print Reading Systems Leading to a Proposed Reader for Typewritten Material

By

Herbert D. Cook

NBS

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

U. S. DEPARTMENT OF COMMERCE
Sinclair Weeks, *Secretary*

NATIONAL BUREAU OF STANDARDS
A. V. Astin, *Director*

THE NATIONAL BUREAU OF STANDARDS

The scope of activities of the National Bureau of Standards is suggested in the following listing of the divisions and sections engaged in technical work. In general, each section is engaged in specialized research, development, and engineering in the field indicated by its title. A brief description of the activities, and of the resultant reports and publications, appears on the inside of the back cover of this report.

Electricity. Resistance and Reactance Measurements. Electrical Instruments. Magnetic Measurements. Electrochemistry.

Optics and Metrology. Photometry and Colorimetry. Optical Instruments. Photographic Technology. Length. Engineering Metrology.

Heat and Power. Temperature Measurements. Thermodynamics. Cryogenic Physics. Engines and Lubrication. Engine Fuels. Cryogenic Engineering.

Atomic and Radiation Physics. Spectroscopy. Radiometry. Mass Spectrometry. Solid State Physics. Electron Physics. Atomic Physics. Neutron Measurements. Infrared Spectroscopy. Nuclear Physics. Radioactivity. X-Ray. Betatron. Nucleonic Instrumentation. Radiological Equipment. Atomic Energy Commission Radiation Instruments Branch.

Chemistry. Organic Coatings. Surface Chemistry. Organic Chemistry. Analytical Chemistry. Inorganic Chemistry. Electrodeposition. Gas Chemistry. Physical Chemistry. Thermochemistry. Spectrochemistry. Pure Substances.

Mechanics. Sound. Mechanical Instruments. Fluid Mechanics. Engineering Mechanics. Mass and Scale. Capacity, Density, and Fluid Meters. Combustion Control.

Organic and Fibrous Materials. Rubber. Textiles. Paper. Leather. Testing and Specifications. Polymer Structure. Organic Plastics. Dental Research.

Metallurgy. Thermal Metallurgy. Chemical Metallurgy. Mechanical Metallurgy. Corrosion.

Mineral Products. Porcelain and Pottery. Glass. Refractories. Enameled Metals. Concreting Materials. Constitution and Microstructure.

Building Technology. Structural Engineering. Fire Protection. Heating and Air Conditioning. Floor, Roof, and Wall Coverings. Codes and Specifications.

Applied Mathematics. Numerical Analysis. Computation. Statistical Engineering.

Electronics. Engineering Electronics. Electron Tubes. Electronic Computers. Electronic Instrumentation. Process Technology.

Radio Propagation. Upper Atmosphere Research. Ionospheric Research. Regular Propagation Services. Frequency Utilization Research. Tropospheric Propagation Research. High Frequency Standards. Microwave Standards.

●Office of Basic Instrumentation ●Office of Weights and Measures.

NATIONAL BUREAU OF STANDARDS REPORT

NBS PROJECT | **NBS REPORT**

0107-20-5515 | November 26, 1954 | 3804

A Study of Print Reading Systems Leading to a Proposed Reader for Typewritten Material

By

Herbert D. Cook

Electronic Instrumentation Section
Electricity and Electronics Division

Developed through

Data Processing Systems Division

for

Department of the Navy

Bureau of Supplies and Accounts

BuSandA 56782

NBS

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

1 INTRODUCTION

The purpose of this study is to investigate the principles of reading print in order to establish a suitable design for a reader for typewritten material. This will form the basis for a proposed project to develope a prototype model of such a reader.

As a background for the study a number of existing print reading systems were analyzed with respect to methods of operation and also practical considerations such as speed, accuracy and suitability for the application. Analysis of these systems is not presented in this report because of the difficulty of obtaining complete details of the many developments now under way in this field. It is believed, however, that a sufficiently wide cross section of the field has been examined to formulate reliable conclusions with respect to the suitability of the methods. In this report a brief analysis of the basic principles of print reading is given and also a description of the design of a proposed reading system.

The proposed reader is expected to read at the rate of at least 10 characters per second.

1.1 Characteristics of Typewritten Material

The special problems encountered in reading typewritten material arise from the nature of the typewriter itself. It is essentially a portable device, constructed with a minimum of precision, for writing what is sometimes little more than barely legible print. Most typewritten material is readable with processes of which the eye is capable, but which would be difficult to build into a machine reader. The eye sees characters largely by their outlines, taking less account of the total black area of the printed material. A machine reader must rely on both the area and the shape, and this leads to the special difficulties.

First, the shape of the characters is not uniform because of uneven distribution of ink transfer in the printing process. This is due in part to the texture of the ribbon which is usually a woven fabric, giving a superimposed pattern of threads on the character and in the space between and surrounding the type face.

Also the roughness of the paper surface prevents an even deposit of ink over the surface. As a result, the characters are generally imperfect especially at the edges, although any part of the character may have missing portions. This difficulty must be taken care of by the reader so that each character can be identified in spite of defects, unless the defects cause it to be changed so as to be indistinguishable from some other character. Some improvement in quality may be obtained by the use of a paper tape ribbon that is now commercially available.

Also, the position and spacing of the characters is not uniform. Often the characters are so close as to leave no space between them. This rules out any system that utilizes this space in the register or recognition processes.

The shape and size of the characters are not uniform between different typewriters due to the large number of makes of typewriters and the variety of type faces available. Unless a program of standardization of type face is undertaken, which appears impractical at present, a reader for material gathered from many sources must be capable of reading all the various fonts. No reader has yet been built that will do this without a change in the reader for each font being read. If a large amount of machine reading is contemplated, the proposal for font standardization as set forth by Rabinow[1] deserves study.

1.2 Register Problem

Other problems also arise that are basic to all print reading systems. These are the problems of register and speed. Obtaining register is the most difficult practical problem in any system. By register is meant the relative geometrical positions of two images or shapes. This is without reference to a fixed coordinate system but only with reference to the relationship between the images themselves. Two identical images are said to be in register when all points of one coincide spatially with the corresponding points of the other.

In a recognition system the comparison is between the image presented to the system and a reproduction of the image stored in the system. In this case register does not require that the images be superimposed actually, but only in a relative manner depending on the manner in which the image is stored. It is convenient to think of register by superposition as taking place by an inverse process of projection of a stored image back on the image being read.

1 Report on standardization of the 5x7 font, Diamond Ordnance and Fuse Laboratories Report No. TR-39

The material is generally read one character at a time, each of which is placed in turn in the field of view of the optical pick-up head of the reader. Accurate register of each character is required in order to achieve recognition. The accuracy required is too high for mechanical positioning because of the inaccurate manner in which the material is written on the paper.

In order to locate the characters properly, several steps are necessary. First the paper must be inserted in approximate position and orientation to put the appropriate line of print in view of the reader. Even if this is done accurately, there is generally sufficient variation along the line of print to require an additional process of searching for each character. The accuracy of register that must be attained is high in order to make it possible to distinguish among the 40 or more characters that may appear. It is estimated that the tolerance from exact register for ordinary print to be distinguishable by machine cannot be greater than about one thirtieth of the overall width of the widest character. Better reading accuracy would be obtained in any system with a tolerance of less than this value.

In the various reading systems now being developed a number of different methods are used to attain register. Some require that there be a clear space surrounding each character, and utilize this in the searching process. When it is impractical to require that such a space exist, as in typewritten material, these systems cannot be used. In this case a system must use other methods to achieve register. This may take the form of relying on the recognition process as an indicator for register while searching blindly in the area where the character may be found.

When recognition is used in register sensing, a complete test is made of many possible positions throughout the area in which each character may exist. This is usually done in a geometric pattern of tries with several in the vertical direction, which is already positioned closely, and with a continuously moving pattern in the horizontal direction, in order to cover all possible positions. Since register must be obtained within a distance of not more than one thirtieth of the character spacing, a large number of tries are necessary for each character. There are ways of reducing the number of tries such as taking advantage of the knowledge gained from the position of one character to help find the next one, and some increase in speed may be obtained by such methods.

1.3 Speed Problem

The large number of individual operations required for comparing all of the stored characters with the field of view throughout the area where the characters are likely to be found is an indication of the large amount of information processed during the reading operation.

Consider a system in which the field of view is sampled at about 1000 points, as in a 32 line raster and 32 point resolution along each line. Without knowledge of the horizontal register, and with the vertical register known within a tolerance of one fourth the character height, there is required a total of 8x32 or 256 individual rasters to find each character. It is assumed that 8 complete rasters are made at each horizontal position, each consecutive raster being moved one line vertically. Since there are 1000 points to sample in each field, this amounts to 256,000 points per character. If 50 characters are in the vocabulary and all must be tried in sequence, there are 12,800,000 points to be sampled per recognition. This must be multiplied by the number of characters per second to find the number of point samples per second.

This rate is too high to be achieved with any single existing electronic device alone. In order to achieve a practical reading speed, as many operations as possible must be done simultaneously, or in parallel. Every practical existing system has some portion operating in a parallel manner. An example of this is the comparison of the scanned input with all of the stored characters simultaneously. Other methods for increasing the reading speed utilize some means for sensing misregister in order to reduce the searching required for recognition.

1.4 Basic Operations

Although the methods used in reading print differ in many respects, there are certain necessary operations basic to all systems. First, there must be conversion of the image on the paper to the medium used in the device, usually an electrical signal to be handled by electronic means. Next, there must be stored in the device in some manner the complete set of characters to be recognized. Then there must be a comparison between the stored character information and that derived from the printed image. Beyond this, there is a selection process to indicate which of the characters is recognized, and a coding of the selection that is appropriate to the computer or other device into which the reader operates. These operations are not immediately apparent in every system because of the many combinations in which they may appear, but are essential to the operation of any system.

It may be seen that since a variety of methods can accomplish each basic operation, if used in their possible combinations, a very large number of different reading machines might be built all of which would operate successfully. This is apparent from the large number of memory systems that have been used in computers that could serve in a reader. A practical reader must, however, use components that are practical in themselves and that work well in combination.

1.5 System Classification

The various systems for reading print can be classified roughly by the manner in which the image is analyzed. Recognition is always by comparison with stored information that includes all of the possible characters, and in order to make the comparison, the image as viewed by the optical reading head must be converted to the same type of information that is available in the storage. The classification suggested here is in terms of the dimensions of the converted image.

The three categories will be called (A) Optical matching, (B) Scan line analysis, (C) Image point analysis. In "optical matching" (A) the complete image in two-dimensional form is compared directly with two-dimensional images stored in the reader. In (B) the image is scanned with a continuously moving spot, forming a line pattern over the image field and producing a time varying signal. The signal is compared with a similar signal produced by the storage mechanism. In (C) the image is analyzed point by point over the entire field according to a specific pattern, either in sequence or in combination. The values obtained from examining each point are then processed in a digital manner to determine the nature of the image. Given the values from a sufficiently large number of points, a digital computer is capable of recognizing the characters that appear in the field of view of the reader.

A system using the "optical matching" principle (A) requires some means for making the image of the field of view coincide with that of each of the characters in the vocabulary. Since this is an optical-mechanical process, the speed of such a system is generally limited.

With the scan line analysis (B), less information is derived from the image than for (A) but the information does maintain continuous dimensional characteristics in time that allow accurate comparisons. Further, the derived information is compatible with established continuous signal techniques in electronics.

In image point analysis (C), some accuracy is lost in the loss of the interpolation possibilities that exist in continuous signals. As a result more image points are required to reproduce an image with the same accuracy. This means in turn that more sampling of the field is required. The techniques of analysis are those common to digital computers.

A number of readers may be briefly mentioned as examples of the combination of principles.

An early reader for the blind (RCA Laboratories) uses a method of scanning and makes a comparison with the character storage with respect to the number of intersections the scan makes in passing over different parts of the character. Only a small amount of information about the character is utilized.

More information is stored in an early reader by Sheppard in which the character being read is compared with stored parts of characters by direct optical means. Combinations of parts determine the identity of the character.

A later reader by Sheppard scans the character by image points as in (C) above. The necessary computing is done at the same time that the scanning progresses.

The reader built by Rabinow is in the category of (A) with a direct optical comparison between read and stored images. Rabinow's chief contribution is in the comparison process in which the two superimposed images are scanned with a raster, searching over the field sequentially for unmatched image points.

The reader demonstrator built by Greenough[2] is a more rapid model of Rabinow's system with electronic scanning and control circuitry. Intermittent mechanical motion of the sequential vocabulary storage also is a factor in the increased speed.

Although all of the readers are capable of reading they require much improvement in speed and accuracy to be practical. It is believed that this improvement can be achieved with methods that are available in the present state of electronics. One of the important factors is that there be a large amount of information stored in the vocabulary storage of the reader, and that it be available very quickly on demand. Also important is that a suitable method be used for attaining accurate register as has already been mentioned. It is believed that the above considerations point to method (B) of system classification as most suitable for a practical reader of typewritten material, largely because of the high degree of dimensional accuracy that can be obtained with a minimum of information processing.

2 NBS Report No. 3634 "Technical Details of Print Reader Demonstrator", M. L. Greenough and C. C. Gordon

2 THE PROPOSED SYSTEM

On the basis of the previous considerations and on experimental work done while this study was in progress a reading system was designed incorporating methods that are presently achievable in electronic practice. It is of a type characterized by comparison of electrical signals derived from scanning as described in the classification of (B) in the previous section. The comparisons are made against all characters in parallel as is required for a practical reading speed. The process is entirely electronic except for motion of the paper through the reader. The block diagram of the system is shown in Figure 1.

A description of the proposed reader in terms of its functional operation follows.

2.1 Paper-moving Equipment

This is the only mechanical part of the reader. It consists of a carriage, yet to be designed, for holding the document in position beneath the optical pickup head. A continuous relative motion between the paper and the pickup head moves the reading point horizontally along the line of print. The line to be read is chosen by the operator in accordance with the proposed application, as some choice is necessary as to which material on the document is of value. Automatic machinery can be designed for paper insertion but is complicated unless similar automatic feed were used in printing. For typewritten material it is doubtful whether the value of such equipment would outweigh its complexity.

2.2 Light Source

The field of view of the reader is illuminated by a light source consisting of a projection type lamp and condensing lenses to focus the light on the required paper. The paper, illuminated at about 20,000 foot candles as required, can be viewed by the operator through an appropriate dark glass filter.

2.3 Optical Pickup Head

The illuminated field is seen by an optical system consisting of a lens and an "image dissector" converter tube. The image dissector converts the optical image to electrical signals in accordance with an applied scan pattern. This tube was chosen for its ability to operate with an open light source rather than in darkness, and also, since it is a non-storage type of converter, its ability to operate with the irregular scan patterns that are necessary in the reading process. The image dissector is of a type used in commercial closed-circuit television.

2.4 Scanning of Pickup

The image is scanned in a raster similar to television scanning, covering the field in a linear grid pattern. This produces a time-varying signal that is compared with similar signals generated in the vocabulary storage. A 32-line raster is considered suitable for recognition of the typewritten characters.

2.5 Storage of the Character Vocabulary Images

The information required for recognition of the forty or more characters in the vocabulary is stored in the form of masks of all of the images. These correspond in shape to the original type face and are changeable for different varieties of type. Each mask is scanned by an individual scanner consisting of a small cathode ray tube for a light source and a phototube pickup. These generate signals in the same manner as the pickup head.

2.6 Scanning of the Storage Masks

The scan pattern of the mask scanners is identical to that of the optical pickup head and thus the generated signals are of the same type. The signal derived from scanning any mask is identical to that from the pickup only when the corresponding character appears in register before the reading head.

2.7 Unitizing

The signals generated from both sources (pickup and storage) vary in amplitude depending on the density of ink, darkness of paper and for other reasons. Since only the shape of the character is of interest, signals corresponding to white and black are more desirable. The variable signals are therefore given values of either one or zero by a circuit that makes a decision as to whether the instantaneous value of the signal is above or below a prescribed level. This makes all signals uniform in order that comparison may be made only on the basis of shape and area.

2.8 Mismatch Indicating

The unitized signals from the pickup are compared for identity with the signals from all of the storage scanners simultaneously. For each stored character a comparison circuit indicates for each part of the scanned field the disagreement with the image dissector signal. The disagreement or mismatch is zero only for the case of a character in register, compared with the corresponding mask scanner. All other comparison circuits indicate some mismatch as does also the correct character when register is not proper.

2.9 Integrating

An integrator for each mask adds up all the integrated mismatch every time the field is scanned for a comparison. A low value is an indication of correspondence and a small amount of mismatch.

2.10 Comparison of Integrated Values

Comparing the values indicated by the integrators associated with each mask, if any one is sufficiently small, it is assumed that a character has been recognized.

2.11 Selection of Character

The integrator giving a value of mismatch less than the threshold value is that designating the character being read. If more than one indicate less than the required minimum value, the one indicating least is taken as the designating integrator.

2.12 Encoding for Computer Input

When a character is identified a code signal is generated that is appropriate for the input to the computer.

2.13 Output to the Computer

The coded output signal goes either directly to a computer or to some intermediate medium such as magnetic tape or teletype tape.

2.14 Checking

With even the most perfect reader there is a possibility of error due to imperfect print or defects in the paper surface. Accordingly, it is necessary to include some ability to make checks on the accuracy of reading. There are several forms this may take, some of which may reject doubtful reading and some of which may ask for re-reading the material, taking cognizance of significant portions of the characters in view of the information gained on first reading. Other methods anticipate the possible characters being read and sensitize only those in order to make a more certain identification. As an example it is possible to sensitize only the numeric characters when no alphabetic characters are expected. This would be done automatically in a programmed sequence of operations.

Other rejections of material may come after reading because of logical inconsistencies. It is possible to arrange the material when written to assist in such checks by such means as putting a known number of characters in each line etc.

With respect to the over-all accuracy of the reader, it can be assumed that if material is rejected for inconsistency, whether due to faulty printing or inability of the machine to read, it is not necessarily to be considered a machine error.

3 EXPERIMENTAL WORK

In order to check the validity of the foregoing proposal, the various portions of the proposed reader were tried experimentally. A complete scanning system was constructed including an image dissector tube and its associated deflection amplifiers. A mask scanner with a single character was operated in conjunction with this in order to develope the comparison circuits. This effort resulted in a number of system improvements that are important in the solution of the reading problem.

3.1 Automatic Threshold Bias Control

In typewritten material there is encountered a wide variation in the density of the printing ink and the reflectivity of the paper surface. In scanning such material the signals vary as a result of these differences. Unless special provision is made, the effect is to change the apparent size and shape of the characters. This effect is of such a magnitude occasionally as to prevent the recognition circuits from operating.

A solution of this problem is found in the use of a circuit that automatically controls the threshold bias with respect to which the signals are unitized. This threshold value should properly be midway between the values of signal corresponding to the ink density and that of the paper surface. The circuit as developed makes the adjustment automatically whenever a character appears in the field of view of the reader. The effect of using the circuit is equivalent to improving the optical resolution of the image and also the resolution as affected by the bandwidth of the system and persistence effects of the cathode ray scanners.

3.2 Two-Image Mask

This is a means for deriving two separate images from each mask scanner of the vocabulary storage. The two images are used for the operations of recognition and register control. Register control is best achieved by using a complete character image that is identical in shape to the type face, as accurate positional information can be derived with its use. The other image is used in the recognition process. It is thinner and is designed to fit entirely within the boundaries of the complete character image. Recognition by the use of this second image is tolerant of slight amounts of residual misregister and tilt of the character, and this results in greater accuracy.

The two images are produced by scanning both the normal black mask and a semi-transparent mask superimposed on it. When this combination is scanned in the normal manner, the information of both appears in the signal. These are separated by establishing two separate threshold bias levels at which two Schmitt trigger circuits operate, each of which then responds to one of the images.

3.3 Image Register Sensing

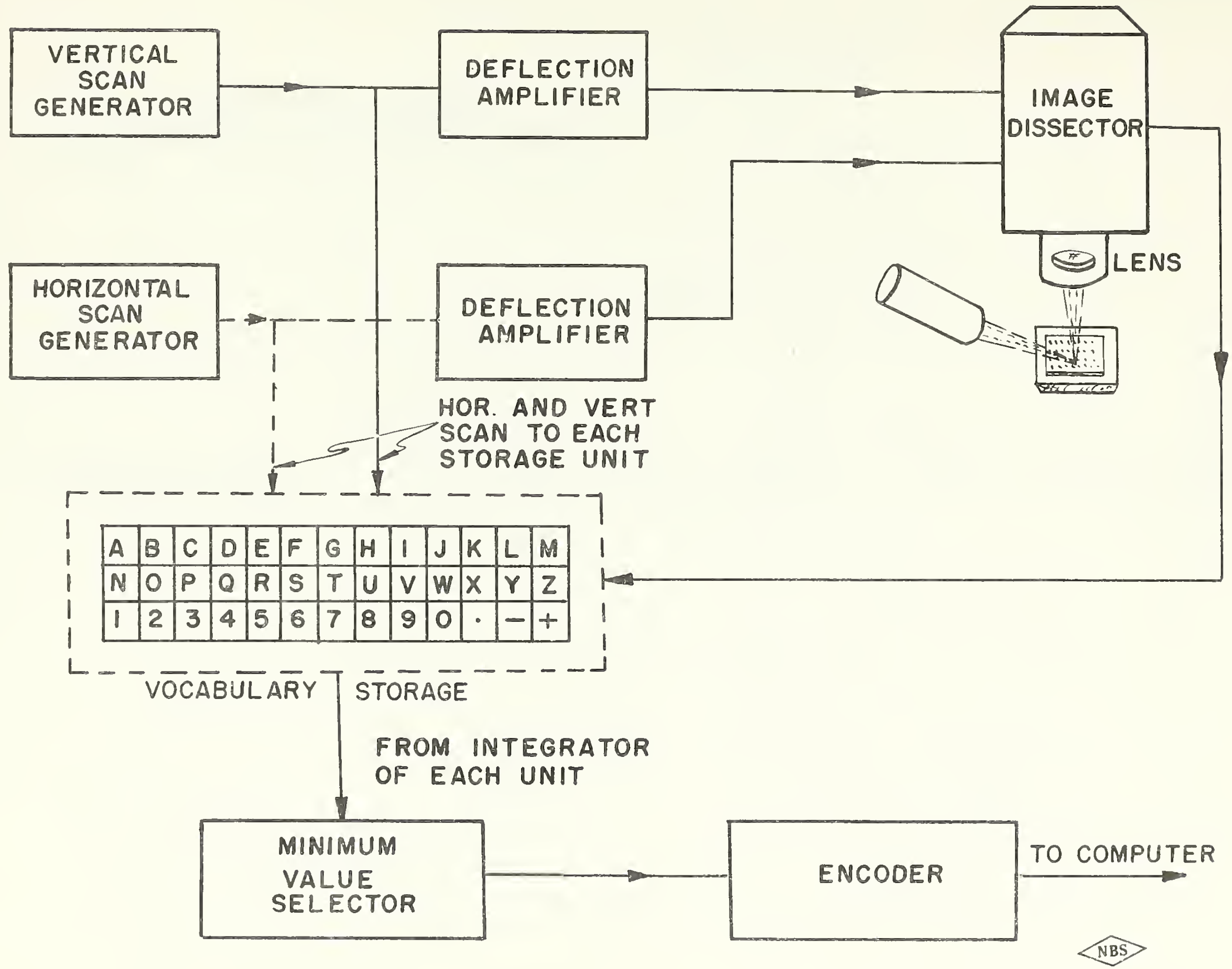
The number of operations required for searching to obtain register can be substantially reduced if some knowledge can be gained as to where the searching might best be done. Various devices external to the reading head such as photocells servoed to the paper carriage or to the position controls of the reading scan can be used, but some searching is generally still required to achieve accurate register. Other methods using information obtained internally by the use of the normal reading scanner with non-uniform scan patterns have been used, but also are limited in the accuracy with which they ultimately achieve register, especially if there is no clear space horizontally between characters.

A method of sensing the direction of misregister has been devised in which information is derived from the whole area scanned during the reading process. From the information derived from comparing the stored image and the pickup image, a positive or negative voltage is obtained from the comparison circuits depending on the direction of misregister. This is done for both the vertical and horizontal directions. When the direction voltages are applied to the positioning controls of the scanning tubes, a complete servo loop is formed that moves the images into register. It is not necessary to allow a definite space between characters as the information is derived from the same scan field as the information used in the recognition process, and this can be made to cover only the area occupied by the character.

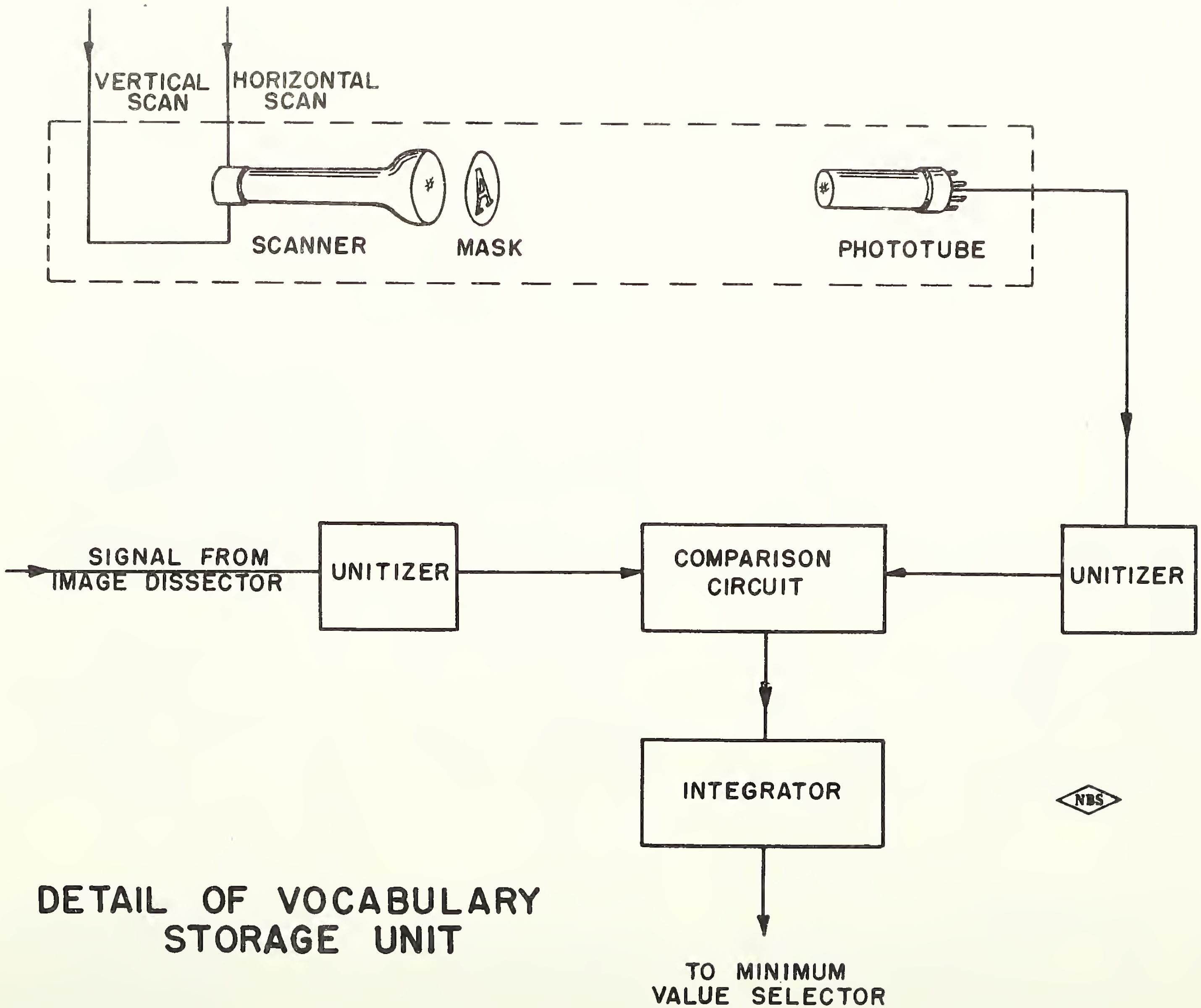
This registering scheme and the other circuits described were built into the experimental system, and were found to operate as expected. In addition to testing these circuits, the construction of the experimental equipment was of value in the solution of other problems incidental to the development of a prototype model of reader.

4 SUMMARY

The considerations upon which the design of the proposed reader are based were derived from analysis of the problem of reading print, and information with respect to known existing readers and their development. The operation of these readers has demonstrated that a successful reader can be built, although there are problems associated with their speed and accuracy. One of the greatest problems has been that of register of the material with respect to the reader, and this can be considered the main one in the way of increasing reading speed. It is believed that the system chosen here has a high probability of success in solving these problems when fully developed.



BLOCK DIAGRAM OF PROPOSED READER



DETAIL OF VOCABULARY STORAGE UNIT