R. B. Hennis

The IBM 1975 Optical Page Reader

Part I: System Design*

Abstract: The IBM 1975 Optical Page Reader, specially built for the Social Security Administration, reads over 200 fonts from quarterly employer reports printed by electric and manual typewriters, business machines, and high-speed printers. Since the SSA has no control over the means used by employers to prepare the reports, many variations in print quality are present. This paper discusses the problems involved in planning and developing a system to read these reports and summarizes the design of the specialized video signal processing circuits and the character recognition logic that are used in the system. Two companion papers treat the latter topics in more detail. Also discussed in the paper is a management information system that permitted detailed analysis of experimental data and accelerated the development process.

Introduction

Every quarter each of the approximately 3½ million employers in the U.S. sends to the Social Security Administration a report of the FICA-taxable wages earned by its employees. This report is prepared on a special form (Fig. 1) supplied by the SSA. Although the form is of standard size and the information appears in a standard format, the SSA has no control over the means used by employers to print the information on the document. The printing may be done with any typewriter or business machine normally used by the employer in conducting his business; indeed, it may be handwritten. This lack of control not only yields documents printed in a large variety of type fonts, but it also causes some documents to be submitted with distressingly poor print quality.

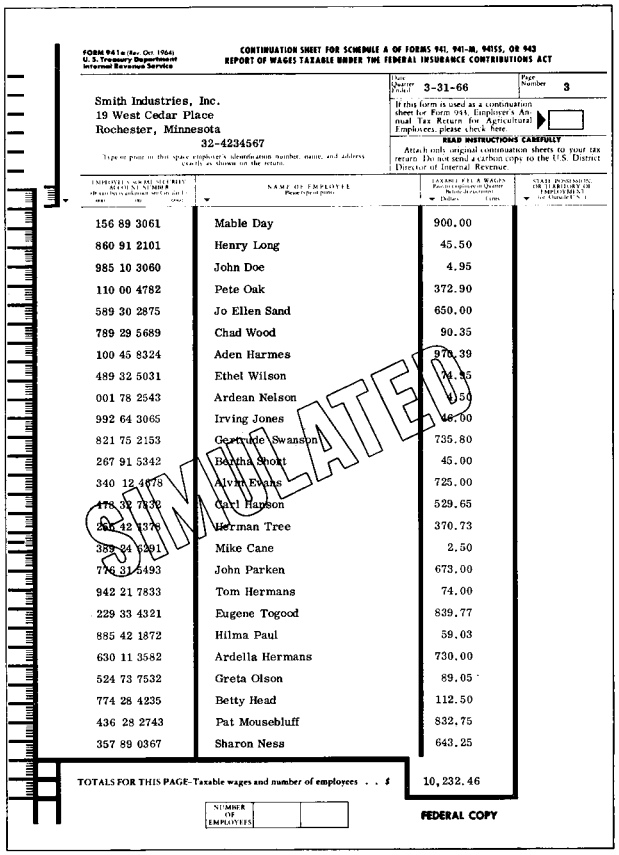
From an engineering viewpoint the SSA's desire for a machine to read a significant number of these reports represented a good vehicle for the application of a sophisticated character recognition technology. This paper describes the IBM 1975 Optical Page Reader which was developed specifically to meet the needs of the SSA. In the SSA's accounting system the incoming employer reports, representing about 80 million employee records, are separated partly by hand and partly by machine into scannable and nonscannable groups. (Each employee record is one line of type containing the person's name, social security number, and FICA-taxable wages.) The scannable group includes reports printed by electric and manual typewriters, high-speed printers, and business machines. These enter the Page Reader, are scanned, and the information is subsequently transferred to magnetic tape. The nonscannable reports are keypunched and read into a data processing system and onto magnetic tape. Both of these tapes are then used for updating employee accounts.

It is important to understand that the 1975 (Fig. 2) is not intended to read all, or even most, of the reports submitted to the SSA. In practice, the documents forwarded to the machine by the SSA employees account for more than 1/3 of all U.S. employee records. A good number of records are excluded by the SSA ground rule that employer reports will not be machine processed unless they contain three or more separate pages. Furthermore, the machine itself is used to make some of the decisions on whether or not a page is scannable. Thus, of all documents entered into the machine, perhaps as many as 35% will be ejected without being completely read.

Statistics from the fourth quarter of 1967 give a realistic picture of the machine's intended performance. A total of 33.12 million lines of printing on 1.3 million document pages were submitted to the scanner. All the characters in 16.29 million of these lines were positively recognized ("good lines"), and 3.63 million lines contained one or more unrecognizable characters. This left 12.20 million

The author is located at the Data Processing Group Headquarters in Harrison, New York.

* Some of the information in this paper was presented at the IEEE Pattern Recognition Workshop, Las Croabas, Puerto Rico, October, 1966.



FORM 941a (Rev. Oct. 1964)
U.S. Treasury Department
Internal Revenue Service

CONTINUATION SHEET FOR SCHEDULE A OF FORMS 941, 941-M, 941SS, OR 943
REPORT OF WAGES TAXABLE UNDER THE FEDERAL INSURANCE CONTRIBUTIONS ACT

Smith Industries, Inc.
19 West Cedar Place
Rochester, Minnesota
32-4234567

Date Quarter Ended 3-31-66 | Page Number 3

If this form is used as a continuation sheet for Form 944, Employer's Annual Tax Return (or Agricultural Employers, please check here ▶ ☐

READ INSTRUCTIONS CAREFULLY

Attach only original continuation sheets to your tax return. Do not send a carbon copy to the U.S. District Director of Internal Revenue.

EMPLOYEE'S SOCIAL SECURITY ACCOUNT NUMBER	NAME OF EMPLOYEE (Please type or print)	TAXABLE F.I.C.A. WAGES Paid in Quarter (Before deductions) Dollars	Cents	STATE INSURANCE OR TERRITORY OR EMPLOYMENT (See Check No. 1)
156 89 3061	Mable Day	900.00		
860 91 2101	Henry Long	45.50		
985 10 3060	John Doe	4.95		
110 00 4782	Pete Oak	372.90		
589 30 2875	Jo Ellen Sand	650.00		
789 29 5689	Chad Wood	90.35		
100 45 8324	Aden Harmes	976.39		
489 32 5031	Ethel Wilson	74.35		
001 78 2543	Ardean Nelson	9.50		
992 64 3065	Irving Jones	46.00		
821 75 2153	Gertrude Swanson	735.80		
267 91 5342	Bertha Short	45.00		
340 12 4678	Alvin Evans	725.00		
478 32 7832	Carl Hanson	529.65		
042 42 1378	Herman Tree	370.73		
389 04 6291	Mike Cane	2.50		
776 31 5493	John Parken	673.00		
942 21 7833	Tom Hermans	74.00		
229 33 4321	Eugene Togood	839.77		
985 42 1872	Hilma Paul	59.03		
630 11 3582	Ardella Hermans	730.00		
524 73 7532	Greta Olson	89.05		
774 28 4235	Betty Head	112.50		
436 28 2743	Pat Mousebluff	832.75		
357 89 0367	Sharon Ness	643.25		
TOTALS FOR THIS PAGE-Taxable wages and number of employees . . $		10,232.46		

NUMBER OF EMPLOYEES

FEDERAL COPY

SIMULATED

Figure 1 Form used by employers for submitting earnings reports to the Social Security Administration.

lines not scanned. The ratio of good lines to total scanned lines, 81.8 percent, is of course, much lower than the recognition rate that could be calculated on a per character basis.

The machine rejected 0.44 million pages as being unscannable. This decision was made if at any time during the scanning of a page, the machine identified four more lines having an unrecognizable character than good lines. Pages not intended for scanning tend to get ejected very quickly by this method since it is likely that the first four lines will contain unrecognizable characters.

The total of good lines plus lines rejected because one or more characters were unrecognizable divided by the time on the machine meter gave a net throughput of 498 lines/minute.

System planning

Before design work began on the 1975 system employer reports received by the SSA were studied to determine the variety of type fonts and the variation of print quality that the machine could be expected to handle. The reports were divided into groups according to the means used to do the printing. This distribution was useful in predicting, for example, the portion of documents that would be expected to have the relatively standard business machine type fonts.

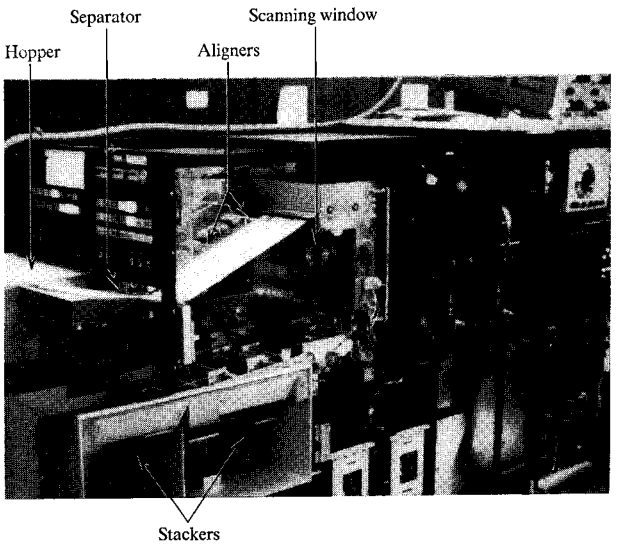


Figure 2 The IBM 1975 Optical Page Reader with covers removed.

Since the earliest statistics were extrapolated from a fairly small sample, later studies provided a more valid analysis of input document characteristics. From a study of all reports submitted during the third quarter of 1965 the following distribution was obtained. Of a total of 60.8 million lines of printing, typewritten documents containing both upper and lower case characters accounted for 24.9 million lines; typewritten documents completely in upper case added 5.3 million lines; high-speed printers such as the IBM 1403 produced 20.2 million lines; posting and billing machines produced another 3.6 million; and various other means, including handwriting, were used for 6.8 million. This distribution has remained fairly constant since then with a slight trend toward an increase in the use of high-speed printers.

In the early document studies 256 different typewriter fonts and 6 different business machine fonts were selected for a detailed analysis of character shapes. Included in this analysis were upper and lower case characters, numeric characters, and special symbols. Since it was the character shapes that were of interest here, all the characters were normalized in size and print quality. The complete variety of shapes that were identified cannot be shown here, but Fig. 3 presents as an example some of the different shapes identified for several characters.

As a result of this study, some fonts were excluded from the set of characters that the machine would be designed to recognize. The excluded fonts were those whose shapes deviated too much from the general run of observed shapes; these fonts represented less than one percent of the total input. Examples of excluded fonts are typewriter script and typewriter italic. It is not necessary to manually sort these fonts from the input, however; machine algorithms

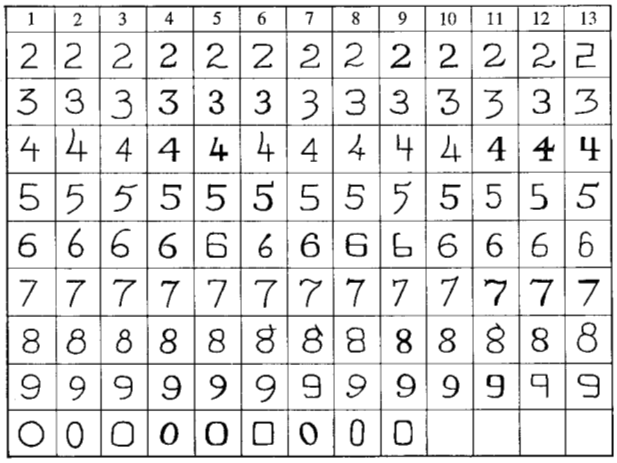


Figure 3 Examples of character shape variation.

automatically eject documents containing unrecognizable fonts. The methods used to design the recognition logic for the more than 200 fonts included in the design set are described in a companion paper by Andrews, et al.[1]

While the problem of shape variation was large, the problem caused by poor print quality was far greater. For example, on the documents studied, the ratio between the thickest and thinnest line widths was as much as 10 to 1, and the print contrast varied from barely visible to carbon black. Both of these problems occurred together with undesirable printing characteristics such as broken characters, strikeovers, individual characters printed above or below the line, and skewed lines of type. Also, in typewriter printing, more than one font per document was encountered several times, and at least one occurrence of five fonts on the same document was observed.

This variation in print quality required the design of special circuitry for enhancing the output signal from the optical scanner. A companion paper by Bartz[2] describes the adaptive video thresholder that performs this function in the 1975.

In planning the system it was necessary to consider the usual tradeoffs between throughput, scanning resolution, and scanning area. Throughput was seen as the crucial consideration. The time it takes for a document to be processed is necessarily made up of a time for mechanical movement of the document and/or the scanning apparatus and a time for electronic scanning. In the SSA application a significant part of the time would have to be spent in finding the information to be read; to keep this time as short as possible, it would be necessary to minimize the amount of mechanical motion involved in the line-finding operation.

IBM had had experience in its 1418 and 1428 optical character recognition machines with a rotating disk-type scanner in which documents were moved past the scanning optics on a rotating drum.[3] Designers of these machines were not faced with the line-finding problem, however, because the information was printed in a prescribed location on the document. Farrington Electronics had also used a disk scanner in their page reading machine. They used a line-by-line scan in which the scanner moved mechanically to look at a stationary document.[4] Since they were reading a single, stylized font, they were willing to tolerate a fairly large time for mechanical motion.

Another scanning technique is exemplified by the machines developed by the Rabinow Engineering Co. (now the Rabinow Engineering Division of the Control Data Corp.). They had been working with matrices of photo-cells, where characters were scanned on-the-fly as the document moved past the cells. In cases such as the Ryder Trucking Reading Machine, where it was necessary to read only one or two lines of type that were precisely located, this scheme could provide fast operation.[5]

Neither the photo-cell matrix nor the rotating disk type of scanner was thought to be capable of fast enough operation in reading the SSA's documents. A cathode ray tube flying spot scanner appeared to be a likely candidate since neither the scanner nor the document would have to undergo mechanical motion during the line-finding process. The Philco Corporation had used a flying spot scanner in a machine developed for the U.S. Post Office to read addresses on envelopes being sent through the mail, a problem similar in many ways to ours. They used a coarse scan for finding the address on the envelope and then shifted to a fine raster scan for operation in the recognition mode.[6]

In the design selected for the 1975 a CRT flying spot scanner is used in combination with a document handling scheme in which documents are picked from a hopper, separated by friction belts into a serial stream, passed through an aligning station, and registered at the 3 in. by 8 in. scanning window. Since the window is smaller than the total area to be scanned, the document position is incremented as required to permit all the data to be read.

System organization and function

A simplified diagram of the data flow is shown in Fig. 4. The CRT scanner, with a raster-type motion of its flying spot, is used to transform the optical image viewed on the input document into an analog signal. This signal is then transformed into a series of binary bits representing the black and white areas of the document. The bits are passed through a shift register of 663 positions to form digital representations of the scanned characters. (The shift register can be thought of as a 39-row by 17-column matrix, where the bottom cell of each column is connected to the top cell of the next column.) Measurements on the contents of the shift register by means of logic circuits detect the presence of "features" in the shapes of characters. The results of these measurements are transformed into 96-bit

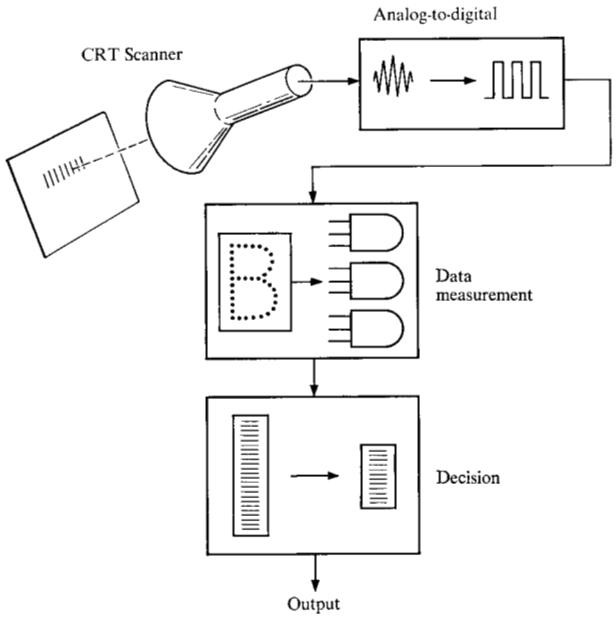


Figure 4 Data flow in Page Reader.

words called feature vectors. Finally, the feature vectors are compared with a reference set stored in a "library" to make the character recognition decisions. Thus, a 663-bit pattern is transformed to a 96-bit word by the measurements, and then to a 6-bit statement of the decision.

The system organization and function described above are conceptually the same as those of an experimental optical page reader designed at the T. J. Watson Research Center.[7] There are, however, significant differences between the experimental machine and the 1975 in the details of system design and in the scanner and recognition unit performance. One significant difference is the feature extraction logic, which was generated by automatic methods[8] in the former and by intuitive methods in the latter. The reasons for this change in design approach are based on the inability of the automatically designed logic to handle the seriously degraded printing on the documents encountered in this application.

• *Beam control*

The scanning beam has a 5-mil diameter at the half-amplitude points of its gaussian energy distribution and moves at a speed of 5000 in./sec. It scans in a raster pattern at a frequency of 25,600 scans/sec. The nominal scan height is 160 mils and the nominal distance between scans is either 5 mils or 20 mils depending on the mode of operation.

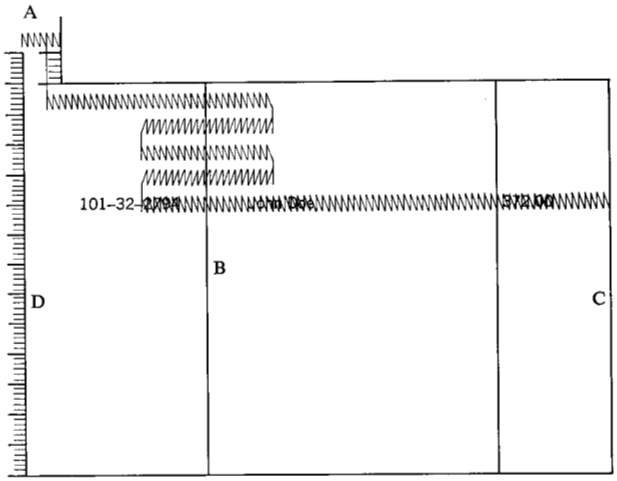
Figure 5 shows the motion of the beam as it travels over the document. The raster begins at point A, moving across the page until it hits a vertical format line which causes the beam to move down through an area from which it extracts a binary code that identifies the type of document being read. (This code is necessary because of some differences in information contained on the "blank" forms.) The raster then moves horizontally back and forth across format line B, scanning at a lower position on the page on each successive sweep. When the first line of printing is encountered, the beam centers itself on the line by a two-step centering process (fast, then slow) all the while continuing to scan until it reaches format line C. Then it moves back towards the left side of the document along the same line of printing until it reaches format line D.

On the first scan of each line (from left to right) the video signal is analyzed by circuits in the system to determine average character height, width, stroke width, pitch, and contrast. This information is used to adjust the scan height in 5-mil increments about its 160-mil nominal value so that all upper-case characters will have a digital pattern approximately 22 bits high in the shift register. Simultaneously, scan pitch is adjusted to maintain a constant aspect ratio. This normalization process allows patterns of the same shape, but varying from 75 to 115 mils in height, to appear identical in the shift register.

Recognition of characters is accomplished during the right to left scan of each line. At the rate of 25,600 scans/sec and an average of 20 scans/char., recognition occurs at 1280 char./sec. The nominal 5 mils between scans is always used on the recognition scan and on the left to right scan of the first line of printing found on the document. On the left to right scans of all succeeding lines, however, the raster spacing is increased to 20 mils. This enables greater processing speed and still provides sufficient image contrast information.

Certain conditions detected during the recognition scan can cause the beam to back up and rescan a character. For example, format control circuits can sense that a character is printed higher than the normal line of print and can direct the beam to rescan it at a higher position. Rescan of a

Figure 5 Cathode ray tube beam travel.



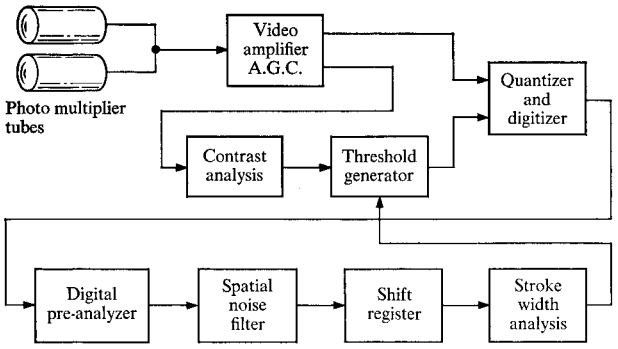


Figure 6 Video signal processing system.

character will also occur if its stroke width is sufficiently different from the average. Digitization circuits will adapt to this situation so that on the rescan the bit pattern of the character will be wider or thinner as the case demands.

• *Video system*

Two photomultiplier tubes are used in the 1975 to generate the video signal (Fig. 6). One tube is directed at the document surface to provide a signal representing the variation in reflected light from the scanning beam. The other tube is directed at the cathode ray tube surface to generate a signal representing phosphor noise. Both of these signals undergo considerable fluctuation even during the scanning of white paper containing no printing. Low frequency variations (0 to 200 Hz) are reduced with an automatic gain control system in each channel. The phosphor noise signal is subtracted from the document video signal to produce a "corrected video" signal, which serves as input to the video signal processing circuits. Although a corrected video signal determined by the ratio of the document video signal to the phosphor noise signal would be accurate over a wider range, practical difficulties in designing an adequate ratio measuring circuit led to the choice of a subtraction circuit.

The main purpose of all the signal processing on the output of the video amplifier is to establish a voltage level that can serve as the threshold dividing black from white. The variation in print quality from document to document, and even within a single document demands the use of some sort of adaptive threshold selection. In the 1975 three types of signal processing techniques are used to generate an appropriate threshold level: contrast analysis, stroke width analysis, and spatial noise filtering. Contrast analysis is the primary means of setting the threshold and the other two techniques act to refine the threshold level setting. Bartz[2] describes these methods and their implementations in detail.

The contrast analysis technique involves sampling and measuring the analog video signal, which is a representation of the variation of contrast observed by the scanner. The basic threshold setting is a linear function of the average values of the sampled signal over a predetermined area. (The threshold is always greater than a specified minimum value.) The sampled values are then digitized as "black" bits if they are greater than the threshold and as "white" bits if less. The binary data are synchronized with shift register and CRT beam movement so that each bit represents a nominal 5-mil square area on the document. The digitized signal is then passed through some circuitry where information for use in conjunction with registration, segmentation, and format control is extracted. Next, the spatial relations between the bits are examined and those considered to be "noise" in the pattern are modified. For example, a lone black bit in a field of white bits is changed to white; similarly white "holes" in the midst of a black area are filled in.

The digitized character then enters the shift register where a bit pattern in the form of the character scanned appears after an appropriate number of shifts have been made. While the character bit pattern is entering the shift register, stroke width measurements are made. On the basis of these measurements, the threshold is adjusted so that during the recognition scan as many characters as possible will be kept within desirable stroke width limits.

• *Recognition system*

An important preliminary to the recognition process is the segmentation of characters. Since the bit patterns of adjacent characters may be linked, it is necessary to determine the beginning and end of each character. This is done by an algorithm that controls the examination of the analog video signal during the left to right scan and the contents of the shift register during the right to left scan. The average distances between characters and from the start of one character to start of the next are determined from the analog signal. The segmentation decision is made by comparing shift register contents with the analog information. If two consecutive columns of white bits are found, the decision is made on that basis. If no white columns are found within a distance determined by the measurements on the analog signal, segmentation is forced. Other logical decisions are used to resolve intermediate situations.

Figure 7 is a diagram of the recognition system. After segmentation has defined a single character bit pattern, the measurement logic for a predetermined character set (i.e., upper case, upper and lower case, numbers) is activated and the existence or absence of specific conditions, or "features," in the pattern is determined. Each of these feature detectors is a combinational switching circuit which evaluates a single-output Boolean function of about 30 inputs. The inputs to the circuits are connected to specific shift register positions. After each incremental advance of a pattern through the shift register, the outputs of 96 feature detection circuits are tested and if any of them is "on," a latch is set in the corresponding position of a 96-bit meas-

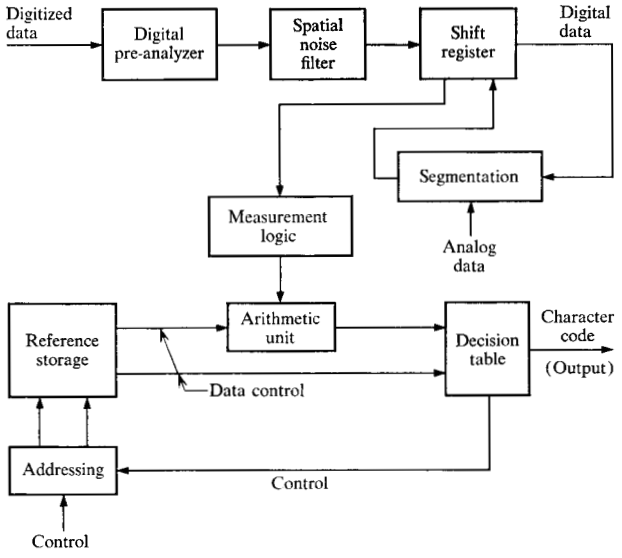


Figure 7 Recognition system.

urement buffer. When the left edge of the pattern is detected, the contents of the measurement buffer are dumped into another buffer in the decision computer. The measurement buffer is then reset and scanning of the next character continues. Thereafter, the character is represented by its 96-position binary "feature vector," each position of which indicates that the corresponding measurement was either "on" at some shift time or was always "off."

While the second character is being scanned, the 96-bit feature vector of the first character is being processed by the decision computer. The decision process involves comparing the feature vector of the input character to a set of reference vectors, and determining which reference most closely matches the input feature vector. The reference structure is ternary. For each bit of the measurement vector a statement of the desired state of the bit (either 0 or 1) or a statement that the bit is not to be considered ("don't care") is stored. Two bits of storage are used to define each position of the ternary reference. Each reference is stored in a card capacitor read only store. Within this store 192 (2×96) bits are used to store the main portion of the reference and 48 additional bits are used to store the reference's control information. Approximately 2000 references are stored and 150 to 200 of them are processed for each input character.

A mismatch counter compares the feature vector to the reference. Each time a 0,1 or 1,0 combination is found in a measurement position the counter is increased by one. If the count exceeds 15, comparison of the current reference is terminated and comparison of the next reference is begun. At the end of a given reference comparison (provided the count is less than 16) the name of the reference is stored in a scratchpad memory at a location (from 1 to 15) corresponding to the mismatch count for that reference. The final decision is made by examining the two candidates having the lowest mismatch counts and determining if the difference in count between them is sufficiently large to permit recognition.

This simple linear decision technique is only one of the many that can be used in processing the feature vectors. The system is in fact a hybrid system that may function differently for each reference vector. More specifically, each reference vector has control information associated with it to instruct the arithmetic unit and the decision table on the particular decision process to be used. For example, the control information can specify a two-level decision process in which a few character classes are first chosen by means of a select group of reference vectors, and then a larger number of reference vectors is used to decide among these classes. The two-level decision process permits a substantial increase over a single-level process in the number of reference vectors that can be compared with feature vectors at a given scanning speed.

This hybrid decision capability allows the use of a highly compromised decision technique when the decision is simple and permits the use of powerful recognition techniques when the decision is difficult. In this manner a good balance of hardware requirements against the need for high recognition rates is achieved.

The decision criteria permit the result of a recognition attempt to be reported as a "firm decision," a "good best guess," a "rough best guess," or "no decision." This is useful in evaluating the results of experiments. However, only a "firm decision" is regarded as successful recognition in the system's operation at the SSA installation—characters that could be classified in any of the other three categories are considered unrecognizable. Whenever an unrecognizable character is encountered during operation, a red mark is automatically placed in the document margin along side the line containing that character. The address of the line is identified from a binary code preprinted near the left edge of the document (see Fig. 1) and is stored on tape along with the recognition results. When the machine is through reading a stack of documents, pages containing red marks in their margins are removed for further manual processing.

Instrumentation used in system development

Management of the 1975 development project was greatly facilitated by the use of a computer-based data collection and reporting system that was developed concurrently with the Page Reader. Without this system it would have been difficult indeed to conduct experiments for testing the effectiveness of modifications to the recognition logic and signal processing circuitry. An example of the primary report produced by the instrumentation system is shown in

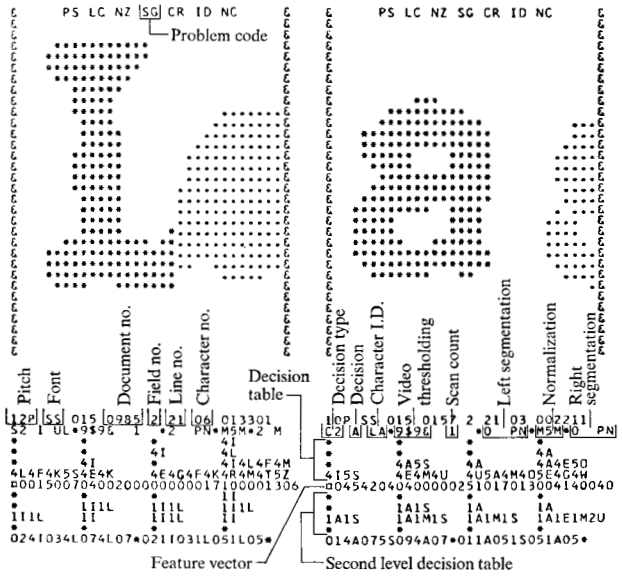


Figure 8 Report generated by instrumentation system.

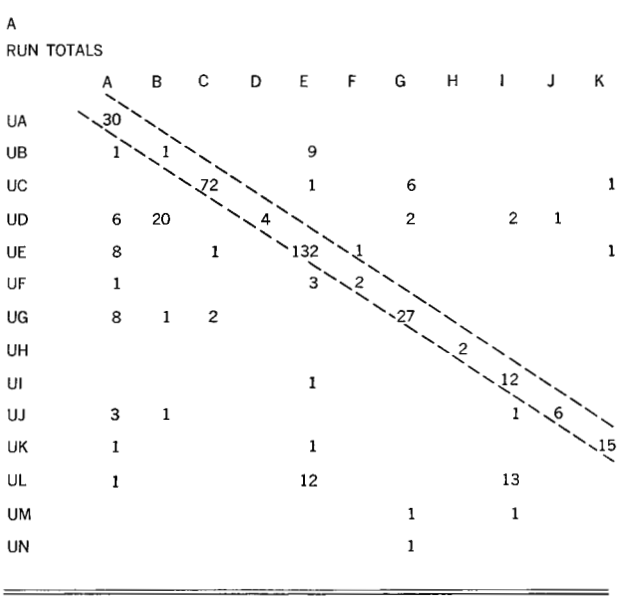


Figure 9 Portion of "best guess" table generated by instrumentation system.

Fig. 8. As each character was scanned by the laboratory model of the 1975, its bit pattern was printed out along with a great deal of information that allows one to identify important parameters involved in the recognition process all the way from the final decision back to the exact location on the document from which the character was read. This report has two main uses. The first is to allow project managers to define experimental tasks to be carried out by the appropriate design groups, and the second is to use selected information as direct input to design programs. The instrumentation system's audit of detail permitted several experiments to be performed simultaneously. For instance, engineering groups in scanning technology, recognition logic, and format control could, on one pass of documents, each determine the effects of their own design changes.

The instrumentation system also provided on request a number of summary reports of an experiment, such as: line reject rates, character recognition rates, recognition decisions as a function of scanner circuit settings, decisions as a function of segmentation circuit settings, rejects per line vs substitutions per line, character substitution table, character "best guess" table, and several others. A portion of the character "best guess" table is shown in Fig. 9. The character scanned is on the vertical axis and the "best guess" decision is on the horizontal axis. The number of times the "best guess" turned out to be correct is given by the entries that fall on the diagonal of the table. Notice that in this experiment the machine consistently guessed that the D was a B.

A chart compiled by the instrumentation system showing character substitution rate, S, vs. character reject rate, R, (Fig. 10) led to the definition of a useful "figure of merit," $\sqrt{S \times R}$, for measuring the progress of system development. After a working model of the machine had been built, improvements in recognition performance were sought by altering the design in three different basic areas: the constants in the decision equations (see Andrews et al.[1]), the reference vectors, and the measurements. It was found empirically that changing the values of the decision constants and making certain minor alterations to the feature vectors had the effect of varying system performance along a hyperbolic S vs R curve. This meant that the figure of merit remained constant. When this relationship became known, the design emphasis could most profitably be placed on making the type of change that would reduce the figure of merit. This type of change usually involved the relatively difficult process of redesigning feature measurement logic. Once the target figure of merit was reached, adjustments to the decision equations and feature vectors could be made to optimize reject and substitution rates along the new S vs. R curve. To facilitate this kind of change the reference storage and decision areas of the development hardware were implemented with magnetic core memories which could be loaded from the instrumentation system computer.

Use of contextual information for error correction

Another interesting activity made practical through the use of the instrumentation system was the creation of a name file for postprocessing of the recognition results to

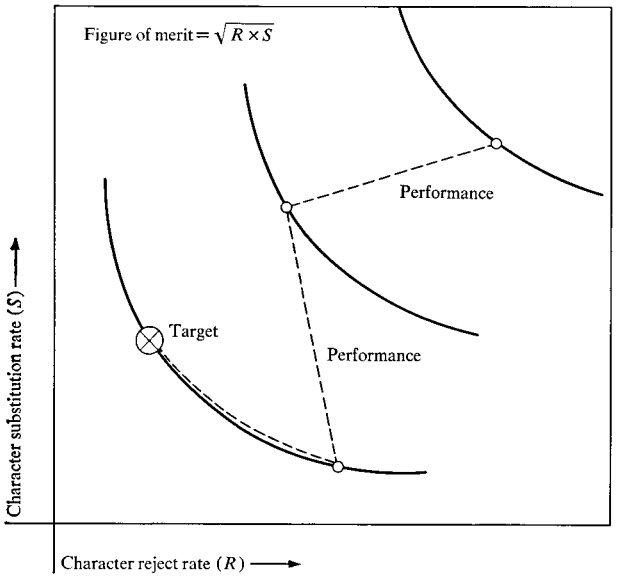


Figure 10 Progress of system development using "figure of merit" as criterion.

detect and correct errors and to correct rejects. The use of table look-up in a library containing all legitimate forms of the input data is not new, but the use of "common error" information in conjunction with table look-up does seem to be new and interesting. The postprocessing library has three sections:

I Names to be accepted as read,
II Names to be changed and then accepted,
III Names to be changed and then rejected.

Before constructing the name file a "legitimate name" list and a "common-error" name list are tabulated. The former is constructed by extracting the unique last names in the SSA's master file of about 150 million names along with their frequency of occurrence. This yields about 1 million names. The latter list is formed by taking character error information compiled by the instrumentation system and extracting the errors with the highest probabilities of occurence, e.g., "e" for "a." These are then substituted one character at a time in the names of the legitimate name list to create a common error list. Knowing the probability of error and the frequency of occurrence of the legitimate name permits the frequency of occurrence of the common error name to be calculated.

These two lists are combined to form the postprocessing name file. The section of the list to which a particular name is assigned depends on the ratio of the frequency of legitimate names having that spelling to the frequency of common error names having the same spelling, N_L/N_E. If the ratio is at least 20:1, the name is placed in Section I, accepted. A ratio between 0.05 and 1 causes the name to be changed and rejected because the number of errors corrected is not high enough in relation to the number of errors caused by the change. If the ratio is less than 0.05, however, the name is changed and accepted, that is, placed in Section II. For example, if "Janes" appears in the legitimate name file with a frequency of two ($N_L = 2$) and as an error form of "Jones" in the common error list with a frequency of 50 ($N_E = 50$), then $N_L/N_E = 0.04$ and "Janes" belongs in Section II. When "Janes" is read and found in Section II, it is changed to "Jones" and accepted. This causes two errors but corrects 50.

Since the name file compiler was designed to accept direct output from the instrumentation system, the process of creating the name file became completely automated. Experimental use of this technique resulted in significant reduction in both reject and substitution rates; however, it is not yet a part of the postprocessing method actually used at the Social Security Administration installation.

Acknowledgments

The work done on the Page Reader originated during a long period of cooperation between members of IBM's Systems Development and Research Divisions. The experimental machine built at the Research Center was used to demonstrate the feasibility of a multifont page reader. Some of the special concepts used in that machine and adopted for use in the 1975 include the method for centering the scanning beam on a line of print, the stroke width measurement technique used for adjusting threshold levels, and the decision criteria based on candidate ranking. The development efforts leading to the final design and construction of the 1975 were carried out at the SDD Laboratory in Rochester, Minn.

References

1. D. R. Andrews, A. J. Atrubin, and K. C. Hu, "The IBM 1975 Optical Page Reader, Part III: Recognition Logic Development," *IBM J. Res. Develop.* **12,** No. 5, 364–371 (1968), this issue.
2. M. R. Bartz, "The IBM 1975 Optical Page Reader, Part II: Video Thresholding System," *IBM J. Res. Develop.* **12,** No. 5, 354–363 (1968), this issue.
3. E. C. Greanias, "Some Important Factors in the Practical Utilization of Optical Character Readers," *Optical Character Recognition*, ed. G. L. Fischer, Jr., et al., Spartan Books, 1962, p. 129.
4. C. C. Heasley, Jr. and G. L. Fischer, Jr., "Some Elements of Optical Scanning," *Optical Character Recognition*, p. 15.
5. J. Rabinow, "Developments in Character Recognition Machines at Rabinow Engineering Company," *Optical Character Recognition*, p. 27.
6. J. B. Chatten and C. F. Teacher, "Character Recognition Techniques for Address Reading," *Optical Character Recognition*, p. 51.
7. H. B. Baskin, R. Bakis, R. J. Potter, J. Reines and G. L. Shelton, Jr., "System Description of a Multifont Page Reading Machine," IBM Research Report RC-1052 (1963).
8. L. A. Kamentsky and C. N. Liu, "Computer-Automated Design of Multifont Print Recognition Logic," *IBM J. Res. Develop.* **7,** 2–13 (1963).

Received May 15, 1967.