NBSIR 75-749

Preparation of Reference Data Sets for Character Recognition Research

M. Leighton Greenough & Robert M. McCabe

Optical Systems Section
Product Engineering Division
Institute for Applied Technology

June 30, 1975

Final

Technical Report to
U. S. Postal Service
Office of Postal Technology Research
Pattern Recognition and Communications Branch
Letter Agreement 74-02934, Task No. 3

Table of Contents

List of Illustrations

Introduction

This report covers the development of two reference data sets contained on magnetic tapes. These sets were derived from a common source of input, a test deck of mailpieces simulated on cards. This test deck, which was furnished by the U S Postal Service (USPS), had been developed from photographs of samples of live mail. Most of the mail was originally type-written, however a range of deliberate, controlled degradation had been introduced during preparation of the test deck.

Part 1 of the report describes the development of a reference set of character images. These Multilevel Character Image tapes were prepared for the US Postal Service to use in Optical Character Recognition (OCR) research principally for pre-processing investigation. The tapes consist primarily of digitized video signals recorded during scanning with a small grid raster placed over approximately 32 000 individual characters selected from the total of 110 000 characters on the mailpieces.

Part 2 describes the development of a tape containing the OCR recognition experience of all of the approximately 110 000 individual characters when the mailpiece cards were read by one OCR machine of advanced design. This tape was prepared in order to furnish a realistic body of information on actual substitution and rejection incidents. It was designed for usefulness in context and in other post-recognition research.

Part 1 Multilevel Character Images

1. Objective

The objective was to develop a reference data set of characters on magnetic tape for the USPS to use in research in OCR. The recorded information is a representation of the video signals generated during scanning with a grid raster placed over individual characters printed on simulated envelope mailpieces.

These tapes contain data on approximately 32 000 characters. For each character, data are recorded for a 24 x 24 point grid located on 0.006 inch (0.15mm) centers, encoded into 16 levels of equivalent reflectance.

1.1 Summary of activity

The development of the character image data set involved the following major steps:

1. Selection of material for inclusion
2. Preparation of address tapes
3. Testing of source material for OCR readability
4. Preparation of microfilm editions:
   - whole envelope faces for future address location research
   - address images only, for scanning and inclusion in the data set
5. Scanning of address films
6. Computer processing, including linearization and tape merge
7. Evaluation of output

These topics and others are covered in appropriate sections of this report. In the concluding section the results are given on a brief study of the fidelity obtained in certain critical phases of the operation. Finally an Appendix is included to document details of the spectral response calculations and of the special photographic arrangements.

The data set itself, the result of the developmental effort, is described in detail in a separate report, ref (1). This specification accompanied delivery of the data set tapes to the sponsor.

2. Selection of material for inclusion in the reference data set

One of the USPS-furnished materials for the project was an input test deck consisting of approximately 6000 mailpieces simulated on cards. Printed addresses had been applied to the cards along with individual plate (mailpiece card) numbers. It is our understanding that the address contents had been obtained by microfilming of live mail, with later substitution of artificial recipients' names. Degrees of degradation of type or print quality had been introduced in the deck by systematic voids in the strokes and by overstrikes. In some cases the addresses were printed on labels which were then attached to the cards. In addition, window coverings of glassine, as well as multi-color paper and print, were employed. Oversize characters were also fairly common in the deck as supplied. An enlargement of a typical mailpiece address is shown in Figure 1.

The specific description of mailpieces, such as color of printing ink, paper, label if used, and window type were supplied NBS by the USPS with the test deck in a contractor's report, ref (2).

In the test deck of approximately 6000 simulated addresses, many were found to be replicated in quantity. When duplicates were removed, 2100 addresses remained, containing about 110 000 characters. From this lot, it was desired to obtain approximately 32 000 representative characters.

2.1 Selection of characters

The overall selection process resulted in extracting about one-third of the envelopes from which one-third of the characters were designated, based on rules which are given below. The product of the selection activity described in this section was a computer-generated printout of those address mailpieces containing one or more characters selected for representation on the data set. The initial step was to convert the addresses containing the 110 000 characters to magnetic tape.

2.1.1 Preparation of address tapes

This magnetic tape of addresses was generated by an operator who copied the 2300 address mailpieces at a computer-connected terminal. This tape has had two uses, namely as a source of input for automated character selection and for providing context information accompanying each character represented in the data base.

2.1.2 Selection program

The address tapes generated by the techniques of 2.1.1 were processed in a separate computer operation to obtain frequency counts of the 62 characters chosen for the data set. These characters are all 52 upper and lower case letters and the ten numerals. Then the computer was instructed to select every "nth" character of each type so as to yield 500 samples of each of the 62 characters. This produced the desired results for the most part. However it left a deficiency of infrequently-encountered characters, only 26 samples of lower case j for example, even though all examples were taken. These totals were brought up to at least 100 for each character through the addition of material from typewriter font exhibits. After minor readjustment of the chosen popular characters, a list was developed of over 32 000 selected characters to be included in the data base, with individual frequencies ranging from 100 to 500, of 62 letters and numbers.

The list of selected characters was printed out from the computer in the form of a worksheet having horizontally triple-spaced characters for each line of the 2100 addresses. The selected characters were shown by computer-printed underlines. Below the address lines in the printout, three more lines were printed, blank except for underlines at the selected characters. A section of a typical printout is shown in Figure 2. The resultant printout thus became a suitable worksheet for entering notes in the underlined spaces, as described under Section 3.2.2. Two addresses were listed on a page, with the plate (mailpiece card) number appearing at the top of an address.

3. Data encoding

3.1 Mailpiece data

In the interest of future addition of information to the data set, it was deemed advisable to include all known pertinent data as part of each character record. These mailpiece data were

contained in a report from the developer of the original test deck, ref (2).

The desired information related principally to paper and ink colors, presence and type of window and the general nature of the type face. The latter had to be determined visually by the coding clerks, since it was not listed in the background report. Working from the report, an operator keypunched the indicated codes into machine tabulation cards. The format and code set of these cards were chosen to match those of the original data. Plate numbers for the mailpiece cards were also punched.

The next step was to merge these cards into the master listing represented by the address tape, using plate numbers as the common element. At this stage the master file then contained all the necessary mailpiece data. The data for any given host mailpiece was invariant for all characters derived from it.

3.2 Readability data

A separate technique was used in order to enter an indicator of readability which of course was specific for each character on a mailpiece. The general approach was to determine the readability of the selected characters and insert this information as one of the items for manual data entry during the scanning operation. The first task was that of determining the machine readability of individual characters.

3.2.1 Mailpiece runs on Postal address readers

To determine character readability, the 2100 mailpieces and 200 supplementary materials were run through three postal machines whose designs incorporated advanced forms of optical character recognition. These runs were performed at Post Offices in Boston and New York. Witnessing observers were present from NBS and the USPS. During these runs, computer printouts were generated showing summary details of reading experience for the 110 000 characters contained on the mailpieces. Although in most cases a tape was also generated as part of the printout cycle, it was believed that decoding of the tapes would have been more difficult than visual data extraction from the OCR printouts. Hence the product of the OCR runs was in the form of computer printouts showing whether a character was correctly read, rejected, improperly recognized (substituted) or disqualified for some reason (washout). On the printouts this information was shown for

the 110 000 characters fed through the postal OCR machines. There were three such printouts, with greatly different formats.

During the runs there was a degree of wear on the mailpieces, chiefly in the form of smudging of print caused by rollers in holding bins of the machinery. This was introduced noticeably in the course of the first run. Since these runs were repeated with less subsequent deterioration, and then filmed after all runs were completed, it is felt that the condition recorded on film for scanning was a reasonably good representation of that during most of the recorded OCR runs.

3.2.2 Analysis and encoding of readability

The analysis of readability began with the task of simply labelling the 2300 OCR printout sections with the plate numbers of the originating mailpieces. All three printouts were so labelled. This turned out to be surprisingly laborious but was necessitated by loss of the initial numeric order during some of the reading operations.

For the analysis, the master file was considered to be the address list worksheet as described in Section 2.1.2 - Selection program. The analyst was required to work in sequential order on the master file, hence it was sometimes necessary to search through the OCR printout to find the applicable mailpiece section. Then she was expected to determine, through examination of the OCR printout, the output character corresponding to the selected (underlined) input character. Finally she was expected to note whether correct reading, rejection, substitution or disqualification had occurred. To minimize confusion, an analyst worked exclusively with the printout from one OCR machine at a time. The analyst's responsibility was to write in an appropriate readability code in the underline boxes. The following codes were used:

C = correct ("A also used synonymously)

R = rejection

E = error, substitution

W = washout, disqualification

Review of an analyst's work was on a spot-check basis, with reduced frequency after confidence had been established.

Altogether three clerks were trained and employed in this phase of the operation.

Thus at this stage, input material had been designated for inclusion in the data set, a machine-language file had been created of applicable group characteristics such as background and print colors, and a worksheet had been prepared with the designated characters marked. This worksheet carried the information on character identification and readability, and was in effect the instruction sheet to the scanner operators.

4. Preparation of microfilm editions of input test deck

Since the NBS FOSDIC was to be employed for producing the character image data base, it was first necessary to prepare a film replica of the input test deck. Two separate versions were generated, one showing images of the entire faces of the mailpiece cards (Envelope Films) and the other containing magnified images of the address sections only (Address Films). Only the Address Films were planned for FOSDIC scanning. Both versions were produced with 16mm positive print film copies as the final product. Altogether there were:

- 11 rolls of camera negatives, Envelope Films

- 11 rolls of positive print copies, Envelope Films

- 27 rolls of camera negative, Address Films

- 27 rolls of positive print copies, Address Films

Initially it was hoped that the same type of film could be used for both applications, hence considerable effort was devoted to methods for extending the gray-scale response of microfilms. Unfortunately no one film was found suitable for both the Envelope and the Address films. As will be seen, the former required high resolution and fine grain. The latter demanded high speed and long scale rendition, which is usually accompanied by larger grain size. Thus these requirements were basically incompatible. Advice was sought from film manufacturers, and many experiments were conducted in an attempt to cover the full span from 5 to 90 per cent reflectance. Within the time available we were not able to encompass the entire range with a microfilm, finding at best a condition of saturation below 10 to 15 per cent reflectance and substantial flattening above about 80 per cent. Close control of the exposure and development steps was necessary for even these limits.

4.1 Envelope films

The Envelope Films were generated to provide input material for potential future applications such as research on OCR address location, line finding and character segmentation. To this end, the mailpiece cards were filmed, insofar as possible, under controlled conditions. These are black and white films at 12:1 reduction, made with a spectral response extending from approximately 400 to 660 nanometres. The image content includes the mailpiece card which was placed against a bordering guide, span markers for scaling, resolution charts and a gray-scale step wedge for calibration of the overall rendition of reflectance. Process controls were applied for uniformity of illumination and contrast in order to preserve as much as possible of the usable region of gray scale rendition.

The microfilming of envelopes was conducted at the Bureau of the Census, using a standard planetary (table top) camera. Mailpiece cards were placed against an L-shaped corner angle so as to minimize variations in position. Portions of the images included control elements such as size graduations, gray-scale chips and resolution charts.

The primary technical considerations on these films were:

1) Image to include whole envelope (mailpiece card)

2) High resolution

3) Low graininess

Factors of somewhat lesser importance were:

4) Gray-scale rendition

5) Spectral characteristic

The chosen negative film for the process was a high-speed type designed for rotary cameras, with which considerable experience had been gained during the 1970 Census in connection with FOSDIC processing. Other films were tried, but their slower speeds proved inadequate; when brought to low gamma by underdevelopment, excessive light was required for exposure. The selected film was used in rolls of 215 feet (65.5m) of 16mm width, packaged in the normal 100-foot size box. This film has a 2.5-mil (0.06mm) polyester base which is nearly transparent and has an antihalation coating on the back. Its spectral response extends from approximately 400 to 660 nanometres.

Details of the microfilming operation for envelope images are given in the description which follows.

For microfilming of whole envelopes (entire mailpiece cards, 4.5 x 8.5 inches) the reduction ratio was 12:1, producing an image size of .375 x .707 inch (9.5 x 18mm). When the gray scale along the top of the mailpiece card and the resolution chart samples are taken into account, the useful image area is approximately 0.5 x 0.75 inch (12 x 19mm) on the films. For the convenience of users who may wish to perform size determination or correction, two fine scratch marks were located precisely 6 inches (15.25mm) apart along the top edge of the card images. This corresponds to 0.50 inch (12.7mm) on the film.

As a check on resolution, two sections of NBS Microcopy Resolution Charts were also included in every image. It was found possible to resolve on the negative approximately 10 line pairs per millimetre as referred to the mailpiece size. On the resultant positive copies the resolution is 6 line pairs per millimetre.

Positive film copies were made of the camera negatives by a commercial laboratory. There are a total of eleven rolls of positive films, containing the whole-envelope photographs of the 6000 mailpiece cards in the test deck. With their controlled density and indicators for size, gray-scale and resolution, these films should be found satisfactory for future purposes such as address and line finding experimentation.

Other details of the microfilming and processing procedures are given in Appendix A.

4.2 Address films

These are the films containing the characters which were actually scanned. A typical image with its address content is shown in Figure 1. All 6000 mailpiece cards in the input test deck were microfilmed, producing 27 rolls of negatives. These were copied, producing the same number of positive copy films. Thus the entire test deck was transferred to microfilm. The characters selected for scanning, by the method described in Section 2.1, involved 11 of these 27 rolls.

The primary considerations on these films were:

1) Image of address portion only

2) Good gray-scale rendition

3) Spectral characteristic

Factors of somewhat less importance were considered to be:

4) High resolution

5) Low graininess

4.2.1 Film selection

From the earlier experimentation with microfilms it was recognized that these were not likely to be suitable. The spectral requirement, introduced in order to simulate practical OCR devices, necessitated limiting the spectral response to a rather narrow band. This of course resulted in considerable loss of light energy reaching the film, by a factor of approximately 5:1. As a result it was necessary to use a higher-speed film of the motion picture type. An additional reason for the selection was to reproduce the low and high reflectance regions, a problem with microfilms as described in Section 4. The material selected was one with medium speed, panchromatic spectral response, long gray-scale rendition and moderate grain. These conditions were met with a readily available film, however it was again necessary to use high illumination levels. The latter was considered preferable to employing a still higher speed but grainier film.

In order to obtain the necessary image width without sprocket hole intrusion, it was required to use negative film 35mm wide. The positive copy films of both the scanned and unscanned addresses were trimmed from 35 to 16 mm width.

4.2.2 Camera arrangement

The microfilm camera was of the standard commercial type, modified for a reduction factor of 2X by the installation of a longer focal length enlarging lens (80 mm f/5.6 stopped down to f/8). It was necessary to fabricate a new lens mount to fit the lens platform of the camera. This mount was equipped with an indexing scale for reproducible adjustment of focus.

A special work stage was constructed to hold the mailpiece cards flat and to furnish alignment guides to the operator. A photograph of the work stage is shown in Figure 3. Since the address locations varied considerably from card to card, the work stage featured a window into which the operator was expected to

place the address by moving the card. In those cases where the address lines were skewed with respect to the edges of the card, this skew was required to be shown in the photographed image. Consequently parallel guidelines were included in the stage to facilitate edge alignment. The window itself was 7/8 inch high and 3-1/2 inch wide (22 x 90 mm).

In a band across the top of the window were located 14 gray-scale reflectance chips, one resolution chart and three span marker lines. The reflectance chips covered the range from 3.7 to 85 percent, as measured with respect to barium sulfate on the 45/0 reflectometer regularly used in the paper industry, with filter #13 having a spectral centroid of 536nm. Specific values are given in Appendix B.

Illumination of the material to be photographed was provided by two reflector floodlamps operated on about 60 percent (70V) of rated voltage 300w, 120V. The adopted illumination level of 225 footcandles (2420 lux) was monitored at least hourly with a small photometer.

In photographing the test deck, each address image was preceded by an image of the plate number of the host mailpiece card. This allowed a visual check of the identification of address images.

4.2.3 Density and contrast control measures

Control of the overall input/output characteristic was effected through scanning the portion of the image containing the reflectance chips, recording the density signals for the 14 chips in the same block of data as the signals from the selected characters in the mailpiece image. This provided the basic data from which image information was calibrated in quantitative terms of reflectance. The technique for this purpose is described in Section 6.1 - Linearization and digitization. The process compensated at the same time for the non-linear, quasi-logarithmic response common with FOSDIC and other density-measuring devices.

In correcting the overall input/output characteristic to a linear function, the tacit assumption is made that at any position over a character, the area weighted proportion of reflecting material under the spot can be equated to an overall gray of proper level. In a direct document scanner the equality is inherent. However to produce the same results with photographic films, which usually show approximately logarithmic response, requires at least a reasonable match between the characteristics

of the negative and positive films. If the match is not good enough, there will be some thickening or thinning of printed lines in the subject material. Without control measures, the effect can amount to perhaps one third of the spot diameter.

The controls imposed on density and contrast took the form of upper limits on the densities of both negative and positive films. Processing was in a commercial motion picture laboratory. It was necessary to specify "pushed" development, as had been learned from test strips run previously. The group mean and range of density were determined for the processed negatives. Reference was then made to the characteristics of the copy printing equipment, previously-measured (through test strips), yielding an exposure value which was specified for the copy printer.

For the control procedure, the density on the negative was measured for the whitest chip, 85.5 per cent reflectance. It was measured on the first one or two frames on each full length roll of negative film. The density was found to range from 1.0 to 1.25, safely within the upper design limit of 1.33.

Because the microfilming took place over a period of several weeks, the exposed rolls of film were stored in a refrigerator at about 40 F (5°C). This step was taken to minimize fade of the latent image, which would cause a larger density spread in the negatives.

4.2.4 Resolution control

To verify that the necessary resolution was preserved through the photographic steps, a portion of an NBS Microscopy Resolution Chart was incorporated in the image area. This portion included the finest lines on the chart, 18 line pairs per millimetre, equivalent to 1.1-mil lines alternating with spaces of the same dimension. A visual check was maintained on these lines, which were readily resolvable on the negative. On the positive copy, the finest target, 18 line pairs per millimetre, could just be resolved. Resolution of this order is necessary to avoid degradation of the results derived from a beam width of 3 mils (0.075mm).

4.2.5 Spectral characteristic

The spectral characteristic is important in determining the response to the color of the scanned input material. The colors of both the background envelope or label, and the printing ink, affect the magnitude of signal change as the scanning beam moves

over the address. If the spectral response is "neutral", i.e. non-selective, the response to all colors except black will be weak. However most practical scanning devices exhibit a selective function, exaggerating the response to certain colors and their designations are as follows:

1)	"OCR-1"	505 to 575 nm	with max. at 530 nm
2)	"Visible"	575 to 750	" " 675
3)	"Infrared"	755 to 990	" " 850

The OCR-1 characteristic was selected as being the only one suitable for use with the available films and lenses.

The spectral characteristic prevailing during the microfilming step established the spectral response function which ultimately was assigned to the data set. Insofar as possible we attempted to match the OCR-1 characteristic, primarily through the selection of a suitable filter during the microfilming of the address test deck. For this purpose a yellow-green filter was placed in front of the camera lens. Once past the initial photographic step, the spectral characteristics of the copy film and the scanner are unimportant.

The spectral characteristic applicable to the data set was calculated from the following.

1) Measured transmission of the filter, a Wratten No. 58

2) Spectral sensitivity of the film, as supplied by the manufacturer

3) Color temperature of the illuminating lamps, (2200 K) determined from the manufacturer's data for color temperature at the actual operating voltage.

The primary influence was the filter, with a slight shift toward the longer wavelength introduced by the rapid dropoff of the illumination toward shorter wavelengths. Details of the computation are given in Appendix C.

Figure 4 shows the overall spectral characteristic as obtained by this process. It can be seen that there is a broad

peak which reaches its maximum at 550 nm, with 50% points at approximately 515 and 580 nm. This can be compared with OCR-1 response, which calls for a maximum at 530 nm and 50% points at 505 and 575 nm. The degree of this match indicates that there should be only a minor departure from design center response, slightly evident with blue and yellow inks and paper.

There is a possibility that under some conditions the data obtained from scanning with one spectral response can be converted to that of another. The apparent feasibility of this approach is covered in Section 7.5 - Feasibility of conversion to other spectral regions. Its promise was one of the major reasons for matching the spectral response represented in the data set to one of the established characteristics.

4.2.6 Positive copy film

Positive film copies were made by a commercial laboratory using conventional print film for motion pictures. Density was controlled through adjustment of the printing lamp, as determined from test strips which duplicated the measured range of density on the camera. These films have slightly less overall contrast range than the original mailpiece materials (1.1 density range on the positive compared to 1.26 on the source mailpieces). As mentioned in Section 4.2.1, the copying was carried out using 35 mm film, followed by slicing to 16 mm width. Twenty seven rolls were produced from the entire deck of 6000 mailpieces. Of these 27, 11 contained the material required for scanning.

5.0 Scanning

This section covers principally the generation of video data obtained by scanning the filmed character images. In the same operation, the readability data described in Section 3.2 above was entered manually via keyboard typing.

5.1 Description of processing equipment, FOSDIC

FOSDIC, an acronym for Film Optical Sensing Device for Input to Computers, is a computer-actuated film scanner. In the mode of operation for this data set, the scanner was used simply as a programmable microdensitometer, optically reduced so that the spot size was 0.0015 inch (0.04 mm). At this reduction the maximum attainable scanning area is 0.5 by 0.75 inch (12.7 x 18 mm) on 16 mm film. In the densitometer mode, spot positioning is programmable in increments as small as 0.0004 inch (0.01 mm or 10μm). At each point the image density can be measured and recorded

in one of 256 levels up to a maximum density of 1.6 (2.5% transmission). The general characteristic of the output is such that the digitized level is approximately proportional to the logarithm of the film transmittance.

An internal projection arrangement displays an enlarged film image on a page-size (15x) viewing screen in front of the operator. Separate observation of the scanning cathode ray tube face through windows makes it also possible to see the scanning pattern in exact registry with a 4X reproduction of the film image. These provisions for operator observation greatly facilitate the location of desired portions of the image.

The complete FOSDIC system includes, besides the scanner and the minicomputer with 4K memory, a magnetic tape unit and a low-speed paper tape and keyboard unit. For preparation of the data set, a video display device was attached, and external X and Y positional controls were added.

Other data processing equipment used at other steps in the production of the data set included a CRT/keyboard terminal and the NBS 1108 central computing system.

5.2 Scanner program

The program for actuation of the scanner and related peripheral devices included the following major routines:

a) Size calibration, wherein the amplitude of the scanning raster was ajusted to cover a square of 48 x 0.003 = 0.144 inch (3.66 mm) on a side, referred to address dimensions.

b) Gray scale calibration, which provided a quantitative basis for assigning equivalent reflectance values and which permitted linearization of overall response

c) Raster scan for location, used for positioning each character inside the raster pattern

d) Raster scan for recording, during which the scanning resulted in stored values (48 x 48 = 2304) in the computer memory.

e) Keyboard data entry, used for manual insertion of mailpiece plate number and data on character identification, readability and intrusions.

f) Magnetic tape recording, involving the transfer of video and keyboard-entered data from storage to output magnetic tape.

Each of the above items is described in more detail in the appropriately numbered paragraphs which follow.

5.2.1 Size calibration

The size of the scanning raster increment was set to 0.003 inch (0.075 mm) through the use of a special target frame photographed at the beginning of each roll of film. This target contained a ruled square of 0.72 inch (18.3 mm) on a side, corresponding to a total area of 5 x 5 or 25 rasters. The size calibration program produced visible bars which were placed at the sides of the square through manual adjustment of the horizontal and vertical-axis scale factors of the respective deflection amplifiers. It is estimated that scale factors were held to within one per cent by this technique.

5.2.2 Gray-scale calibration

This was the portion of the program which measured the densities of the individual gray-scale chips located along the top of each image. Sixteen reading points were taken on each chip so as to form a group average over the center quarter of the chip area. To take care of possible dust particles on the film, readings showing excessively great film density (apparent low reflectance of the chip) were effectively discarded. Typical readings for the darkest chip of 3.7 per cent reflectance were 10 to 15 on an overall scale of approximately 220.

5.2.3 Raster scan for location

A repetitive raster was generated to permit centering of the scan raster over the character to be read out. This locating raster had the same structure and position (48 x 48 points) as the final recording pattern. The ten-per-second repetition rate, combined with a long persistence display CRT, allowed the operator to see the exact placement of each character in the grid. Before going to the next step of recording, each character was centered by the operator in the grid, through the external XY position controls.

5.2.4 Raster scan for recording

Activation of the recording raster produced a one-time 48 x 48 grid, with memory storage of the resultant video data with 8 bits precision representing 256 gray levels. The time for the one complete raster was approximately one-fourth second.

5.2.5 Keyboard data entry

The entry of all supplemental data regarding character identification, location in the address, readability, and intrusions of adjacent characters into the scanning zone was accomplished via a keyboard while the locating raster scan was activated. Program interlocks were provided to prevent the operator from calling for the recording raster, item (4) above, until the data entry was complete.

The data entry portion of the program required the operator to type from the coding sheet, entering the desired information in a conversational mode. That is, the keyboard printer was programmed to supply the name of the next response expected of the operator, to minimize the chance of confusion. Data for the final entry, on intrusions, were obtained from observation of the video display, rather than from the coding sheet. When the intrusions field had been completed, the typing of a "@" caused the program to initiate item (4), the recording raster, followed by an automatic initiation of the data transfer to magnetic tape.

5.2.6 Magnetic tape recording

During the time of recording, no other program action was performed. Upon completion of the record, the word "MAG" was typed out, and the locating raster scan was resumed.

5.3 Scanner operation

Because of the multiple sources of input data during the keyboard data entry, it was found desirable to employ two persons in the operation. A photograph of the system in operation is shown in Figure 5. The typist controlled the sequence based on the coding sheet containing the character identification, location and readability. She conveyed verbal instructions to the FOSDIC operator, who then placed the locating raster over the desired letter, using the XY controls. When the operator had completed centering the character, she called out the intrusion codes (right, left, top, bottom), which were then entered by the typist.

After this, with the concurrence of the operator, the typist entered the "@", initiating the aforementioned data recording sequence terminating in resumption of the locating raster (3).

The throughput rate for experienced teams performing the above sequence was found to reach 200 characters per hour. Additional time was of course devoted to quality control, verification and correction re-runs.

6.0 Computer processing

The FOSDIC scanner tapes, containing the video data together with the manually-entered information on character identification, readability and intrusion were first put through a linearization process and then reduced to a 32-level representation of reflectance. The next step was to merge the video data with the mailpiece information from the original address tapes. From the resulting set of master high resolution tapes, the final tapes constituting the data set were prepared. This involved the reduction to a coarser grid and digitization to larger intervals, to meet the overall requirements of the data set. These various stages of computer processing are described below.

6.1 Linearization and digitization to 32 levels

The purpose of the linearization process was to produce an overall characteristic wherein the output indication was proportional to the input reflectance, i.e. the light reflected at each position of the scanning spot. Then the output becomes representative of a document scanner, making the data set suitable for generalized applications. The overall control provided by the linearization method was designed to make the system independent of the internal characteristics of its elements, so that the introduction of the intermediate photographic step would have negligible influence on the overall input/output relationship. Lacking this there might otherwise be an undesirable flattening of the response at very low and very high reflectances.

Linearization was accomplished as follows. Fourteen small chips having calibrated reflectances ranging from 3.7 to 85 percent were mounted to form a reflectance step wedge along the top of every address image. The arrangement of these chips is described in Section 4.2.2. The output data as recorded by the scanner included coverage of the wedge chips in the same image from which character data was generated. Thus variations in exposure due to camera shutter irregularities had no effect on the final result.

The scanner output is digital in 8 bits, allowing 256 levels, in which the region from approximately 10 to 220 was used by the 3.7% to 85% calibrated reflectance chips. In general the same range applied to the character scan signals as the beam traversed the character strokes. Tapes with these recorded signals, together with a list showing the calibrated reflectances of the chips, formed the input to the linearization program.

The first step in the program was, in effect, the construction of a curve of reflectance vs. scanner output, obtained by linear interpolation between the 14 chip signals. This relationship was computed in its entirety for each mailpiece image, then used for all computations involving the same image. A typical relationship is plotted in smoothed form in Figure 6. It can be seen that the low and high reflectance ends show some flattening, where there is less change in 256-level output for incremental changes in input reflectance. Insofar as possible, the selection of the 14 calibrating chips was made to enhance accuracy of linearization in these regions.

From the computed response function of the type shown in Figure 6, a value of reflectance was determined for the scanner signal at each point in the character scanning raster. Following this, the reflectance was looked up in a conversion table, yielding the proper one of 32 possible output levels to express the digitized reflectance. For example, reflectances greater than 6.0% but less than 9.0% were digitized to value "2". At 9.0%, transition to the value "3" took place. By this means an input range from just over 6% to just over 90% was encoded to output values from 2 to 30. While it is believed that no samples exceeded this range, there was no specific prohibition of output values below 2 or above 30.

The uncertainty in the linearization and digitization was not measured, but is estimated at about 1% plus or minus one (3%) step in output, over the desired input range of 6 to 90% reflectance. The major portion of conversion error is likely to be in the regions where straight-line interpolation departs from the smooth curve which might be drawn through the 14 calibration points.

The encoding of the digitized output indication is shown in Table 1 which follows.

TABLE 1

Encoding of digitized output into 32 levels

When input reflectance is: at least:	but less than:	Output recorded character is:	
0	3	0	(Note: Value uncertain
3	6	1	below calibration
6	9	2	limit of 6%)
9	12	3	
12	15	4	
15	18	5	
18	21	6	
21	24	7	
24	27	8	
27	30	9	
30	33	A	
33	36	B	
36	39	C	
39	42	D	
42	45	E	
45	48	F	
48	51	G	
51	54	H	
54	57	I	
57	60	J	
60	63	K	
63	66	L	
66	69	M	
69	72	N	
72	75	O	
75	78	P	
78	81	Q	
81	84	R	
84	87	S	
87	90	T	
90	93	U	(Note: Value uncertain
93	96	V	above calibration limit of 90%)

As the linearization and digitization proceeded, video records were accumulated in computer mass storage. These were filed according to mailpiece plate number, for retrieval and merging with the address tape data.

6.2 Merge with mailpiece data

Mailpiece data contained on the address tapes served for recalling the stored video records. New tapes were created with longer records combining the address and mailpiece data together with the video data. These became the eight master tapes for subsequent processing into the data set output tapes.

6.3 Reduction to 16-level output tapes

The desired detail to be represented on the data set called for reflectance digitized to 16 levels on a 24 x 24 grid. Spot size and incremental positions were both to be 0.006 inch (0.15 mm). The goal of the reduction program was to generate the final data tapes from the master tapes having 32-level data on a 48 x 48 grid, with spot size and incremental positions of 0.003 inch (0.075 mm).

Data reduction was accomplished through taking the average reflectance from clusters of four adjacent grid locations, then rounding off to 16 levels of digitized output. New magnetic tapes were created with this reduced form of data representation. With the coarser grid, the new records were 1044 bytes long rather than the previous 2772.

Other minor changes were introduced in the same processing step, such as to revise the heading data for indicated record length from 2772 to 1044 bytes, and to modify the readability encoding. The purpose of the latter was two-fold, to provide a single figure indicating readability for each character and to prohibit traceability to specific OCR machines. The encoding plan is given in Appendix D of the specifications covering the data set, showing the relationship between the input (four categories for each of three machines) and the final representation in one of four categories.

6.4 Generation of separate Setup and Test Tapes

The final processing step resulted in the extraction of a ten-percent sample of the data base characters. This operation was designed to provide an across-the-board sample of the data set characters, to be useful for system checkout. Thus the Setup tape contains approximately 3000 characters, selected as mostly readable, however some non-readable characters are also present.

The remaining approximately 29 000 characters were recorded as the four Test tapes in the data base.

6.5 Brief description of data set

The data set provides video information derived from scanning individual character images on a 24 x 24 grid. A total of approximately 32 000 characters are included, with an effective spot size of 0.006 inch (0.15 mm). Output recordings are digitized into 16 levels.

The tapes are IBM-compatible, 9-track, 800 bpi, odd parity, 8-bit EBCDIC characters. All of the information pertinent to a single character is contained in a record of 1044 bytes. Each record has two major sections of fixed length, covering its originating mailpiece and other overall information in the address data section of 316 bytes, and the specific character identification and video information in the character data section of 728 bytes. Because the address data section is included as part of each record, a compact printout of all known data regarding any desired character is facilitated.

Detailed format and other information is given in a separate publication, ref. (1).

7. Conclusions and recommendations

During the course of the work on the character image data set and subsequently in the related project (covered in Part 2) which also involved the same input material, it became possible to conduct a limited evaluation of the produced data base. Advantage was taken of the opportunity to investigate the following topics:

1) The overall input/output fidelity, that is, the relationship between true and indicated reflectance,

2) Whether the thickness of lines in characters was altered by the scanning process, as evidenced from an input/output comparison,

3) Typographic accuracy, both in the address contents and perhaps more importantly, in the identification of scanned characters.

While an exhaustive investigation was not possible, some conclusions from the above subjects are given in Section 7.1 below.

Finally to conclude Part 1 of the report, recommendations are made regarding 1) the removal of intrusions and 2) the possibility under some conditions of converting the data set so as to be representative of scanning with other spectral characteristics.

7.1 Factors affecting fidelity of input/output relationship for reflectance of lines

7.1.1 Flare and halation

Practical scanning devices often differ from their laboratory counterparts in the rendition of reflectance (gray level) of thin lines forming the character shapes. The usual evidence is that as the thickness of a dark line is reduced, it appears lighter, i.e. loses contrast. While this is readily understandable when the line is smaller than the nominal beam width, the effect may not disappear completely even though the line thickness may be many times the spot size. The problem is that the scanning spot is rarely as well concentrated as would be desired.

The closest to ideal should be a laboratory reflectometer having a well-defined aperture and carefully cleaned optical elements. The latter is most essential, for dust on lenses can result in a condition of light scattering which allows contributions from directions outside of those desired. Thus the measured reflectance over some given aperture area can be influenced by the "busy-ness" of the surrounding material. The result is that the reflectance of the background paper appears to drop as printed material is approached. Similarly when over a solid black zone, the reflectance will appear to rise near the edges.

Scanning devices which use rows of photocells or which employ moving apertures can be susceptible to the same problem. Lens flare and dust on lenses and mirrors are likely to be problems with mail processing OCR equipment, especially dust.

Scanners using a cathode ray tube have the additional problem known as halation. It results from the fact that some of the light emitted by the spot is reflected by the front glass-air interface back onto the phosphor coating. This produces a weakly-glowing disc of light surrounding the beam. It is generally about three-eighths of an inch (approximately one cm) in diameter and is fairly uniform in intensity, but with a brighter ring at its outer boundary. The total light in this disc ranges from 10 to 20 percent of the main spot. No practical method appears to eliminate this halation in cathode ray tubes. Fiber optic face plates, while effective, are prohibitively expensive.

Because the FOSDIC scanner used in the preparation of the data base involves a cathode ray tube, it was deemed desirable to evaluate the magnitude of the flare and halation effect present in the data base. Such an evaluation consisted in comparing the

laboratory reflectometer measurements with the output data for exactly the same position on specific characters. From the video printouts used for quality control, it was readily possible to identify sufficient material to give a reasonable estimate of the overall input/output fidelity. Since the reflectometer had an aperture of 0.005 x 0.035 inch (0.12 x 0.9 mm), it was necessary to derive an average reflectance from 24 output data points at 0.003 inch spacing. Because the most precise readings of data were available at the scanner's output in 256 levels, these data were used for comparison. Readings were averaged, converted to reflectance according to the procedure in Section 6.1 using readings on the gray scale step wedge, and then compared with the reflectometer measurements for the same area. Because the information sought was for purely geometric factors, comparisons were made primarily on black and gray lines spanning a wide range of reflectance. Portions of characters were selected for a range of line thickness and a minimum of nearby printed material.

The results of such a comparison are shown in Figure 7. The dashed line at 45 dgreees is the calibration function, the mathematically-forced agreement on the gray scale wedge. Points have been plotted for a variety of printed lines ranging from solid black to no black, i.e. the unprinted background of the mailpiece card. The unbroken line in Figure 7 represents an approximate mean of the plotted points. It can be seen that the indicated reflectance is slightly low (5 or 6 percent) at the high end, with an opposite effect at the low end. Near zero "true" reflectance, the output appears to be about 13 percent too high. Both of these effects are to be expected, as was mentioned.

The slope of the unbroken line is less than unity, indicating a loss in contrast in the overall process. This slope appears to be in the order of .75 to .80. Thus the recorded scanner signals in the data set show a light-to-dark excursion which is 75 to 80 percent of that which can be obtained from a laboratory reflectometer. This loss can often be considerably higher for practical scanning devices.

It may be worth noting that many of the points which are intermediate between the low and high extremes of reflectance were derived from mailpiece cards having glassine windows. Earlier work on window materials had shown that the rise in low end reflectance, and decrease in high end, was to be expected. Measurements on window materials are covered in an NBS report of 1972, ref (3).

The source of the apparent thickening was believed to lie not in the photographic resolution, but rather in a more complex

effect brought about by residual nonlinearity in the photographic steps. Photographic resolution causes can be disqualified since the loss in resolution was slight. As was mentioned in Section 4.2.4, the 1.1-mil lines and spaces (18 line pairs per mm) could just be resolved, meaning that the thickening did not exceed 1.1 mil.

The observed thickening of lines is instead believed attributable to the residual non-linearity shown in Figure 6, before application of the linearization process. In this process, line reflectance responses were compared to those from the large-area step wedge responses. It is known that thickness changes can be introduced if the system uncorrected response (that shown in Figure 6) departs from linearity. Such a condition can arise for example when the uncorrected response for 50 percent reflectance deviates from midway between that for 10 and 90 percent reflectance. Then the scanning spot position for equal signal will not be exactly at the half-on, half-off location at the edges of the lines. Thus the apparent line width can depart slightly from the true value.

7.1.2 Uniformity of measurement over scanned address area

An important characteristic in a scanning microdensitometer is uniformity of optical efficiency over the total scanned area. For example if the system efficiency were found to show a 10-percent reduction in sensitivity at the edges compared to the center, measured reflectances would have the same degree of uncertainty. Under this condition the indicated reflectance on a known uniform white card might be 90 percent at the center, but only 81 percent at the edges of the scanned area. Such a variation therefore causes a position-dependent uncertainty in indicated reflectance value.

In the arrangement employed for scanning to create the data set, reflectance values for character signals were calibrated through reference to those derived from the gray scale wedge along the top of the scanning area. Because the characters were located throughout the scanning area, a possible variability was introduced. Additionally, non-uniform lighting of the area viewed by the microfilm camera could generate the same effect. For these reasons the overall system uniformity was measured.

The test method consisted of microfilming a test card of 85 percent reflectance, then scanning at six sample positions over the total scannable area. The variation was found to be approximately ± 1.7 percent. It must be noted however that scanner signals in the white and near-white region are effectively

expanded by the linearization procedure. Hence the above measured sensitivity variation indicates that the output values in the data set may have an uncertainty of as much as ±2.2 percent due to positional location in the address. The uncertainty in terms of percentage reflectance is greatest in the representation of the white paper stock and least for black ink.

7.2 Thickening or thinning of lines

The fidelity of rendition of line width was checked by comparing measured line thicknesses with the scanned data. As was the case with gray levels, the scanner output data furnished the most precise indicator of system performance, since it resolved 256 levels of reflectance signal. Comparisons were carried out on lines of different widths and gray levels. Because of the work entailed, these tests were conducted on only a limited basis.

Line thickness on input material was measured with a small scaling magnifier, using the average of edge roughness as the criterion to determine the thickness. At the output, line edges were defined as those points having a threshold reflectance midway between the smoothed lowest value over the printed line and highest values on the paper away from the line. This 50-percent criterion was chosen for its independence of the scanner spot width. Using this criterion, line segments were analyzed in the printouts of scanner data used for quality control.

The conclusion from these measurements showed a slight tendency toward apparent thickening of lines. Very thin or gray lines (9 to 10 thousandths of an inch (0.25 mm) 50% reflectance) were increased by about one thousandth. Thicker, blacker lines (20 thousandths (0.5 mm) 20% reflectance) were found to show an apparent increase of 3 to 4 thousandths by the threshold criterion.

7.3 Typographic accuracy

In each character record there is a reproduction of the source address on the mailpiece from which it was derived. This material was added principally to provide context. While normal care was used in the transcription of the address, a few examples of typographic errors in the approximately 2200 mailpiece cards have come to our attention. These errors should have no relation to possible errors in character identification.

The accuracy of character identification, one of the critical parameters in a data base, was promoted by verbal confirmation between the operators during character scanning. Subsequent

verification was conducted on a sampled basis, by checking one character per mailpiece. While no errors have been discovered so far in this process, it cannot be guaranteed that errors of character misidentification are completely absent.

7.4 Intrusion removal

In the preparation of the data set, no attempt was made to eliminate intrusions from portions of adjacent characters when the scanning raster of 0.144 inch square included such material. However as an aid for researchers who wish to deal with isolated character images, an intrusion code was included in the record for each character. These codes were inserted by the machine operator based upon visual observation of an enlarged video presentation. Details of the coding are included in the accompanying specification sheet, ref. (1). The actual implementation of intrusion removal is left to the user.

7.5 Feasibility of conversion to other spectral regions

Conversion to other spectral responses should be feasible under some conditions. The key would be the availability of simultaneous reflectance measurements in two spectral regions, at the same position on the address. It may be noted that spectral conversion involves amplitude scaling but not positional shift. For example it appears necessary to measure only two values near the extreme conditions, such as with the scanning spot over background paper for one and over solid inking for the other. The reflectance values for the two spectral regions may then be plotted on rectangular coordinates, such as "OCR-1 vs. Visible" reflectance. Two points are thereby established. Since at other positions over a character image, there is an area averaging of paper and ink contributions under the measuring spot, a linear relationship is reasonable to expect. Measurements bear out this assumption to at least a first-order approximation. Thus it should be possible to convert the signal value at each point in the character image via a mathematical scaling during computer processing. The result would be a data set representative of the new spectral response.

As was mentioned, empirical readings of reflectance are required for the basic and converted spectral regions. The feasibility examination was considerably assisted by the availability of a prototype reflectometer developed commercially under USPS sponsorship. Correlation between reflectances measured over a full range of spot/ink coverage showed that the straight-line interpolation resulted in agreement within about 5 percent

reflectance. Conversion was checked going from OCR-1 to Visible and to Infrared.

While conversion thus appears possible when two colors, a paper and an ink, are involved, it cannot work when additional colors are present. The basic assumption is made that the spot is at any point receiving some proportionate sum of paper and ink in the reflected light. The locus for converting values is determined by the two sets of originally measured values. However when an additional background color is introduced, as was the case occasionally in the test deck, there is no means for conveying the necessary information about the relative proportions of the two background colors. Hence, it should be recognized that the proposed technique would be applicable only to simple combinations of one background and one ink color.

Part 2 OCR Character Recognition Results

8. Objective

The project objective was to generate a set of data on magnetic tape, containing information on AOCR (Advanced Optical Character Reader) recognition results on 110 000 characters from a 2100-piece test deck of simulated mail. The desired data include whether 1) each character was correctly read, 2) rejected as unreadable, 3) if substituted, specific detail and 4) if expanded into multiple characters, specific details. This classification by individual recognition results was in contrast to the analysis performed in Part 1, where the interest was in whether or not the character was readable. Part 2 was instead concerned with supplying the raw data from which confusion matrices may be developed.

Determination of the apparent causes of the misreadings observed was not within the scope of the project.

8.1 Summary of activity

The development of the OCR character recognition data set involved the following major steps:

1) Selection of material

2) Evaluation of recognition results

3) Conversion to machine language

4) Computer processing, formatting

5) Preparation of documentation including specifications describing the resultant data set, ref (4)

9. Selection of material

Input material was selected from the same test deck of simulated mailpieces supplied by the USPS and used for the character image data set reported on in Part 1.

The source material was typewritten in the majority, but with an appreciable quantity of printed oversized characters. Prior to reproduction, many of the typewritten addresses were systematically degraded through half-tone line screens. Overstrikes, obtained by typing two characters in the same location, were also introduced. The intent of the test deck was to span the range from readily readable to

plainly unreadable characters. As mentioned, these steps in the production of the test deck had been performed earlier by a commercial supplier.

10. Readability evaluation

This initial step, common to both data sets, was designed to obtain a realistic evaluation of each character's recognition experience. For this purpose, the mailpiece cards were put through three large mailsorting OCR machines. The results from one of these, Advanced OCR, (AOCR) was used for the project described herein. As each card was processed by the AOCR, information was stored regarding the machine's recognition. Subsequent printout furnished a hard copy version of the recognition experience of some 110 000 characters. The complex nature of the printout, plus the human analysis required for character segmentation, dictated that the desired information could best be obtained through visual analysis of the AOCR printout.

10.1 Generation of printout work sheet

A sample work sheet is shown in Figure 8. The work sheet was prepared by printing out the contents of the address tapes described in Section 2.1.1 of Part 1. The tape dump program was formatted so that two addresses were printed on a page, the text was spread out horizontally and underlines were placed under each character in the address. As can be seen, the analyst's notes on character recognition results were entered in the underlined spaces. This technique was chosen to minimize the possibility of misplaced notation. There were approximately 1100 work sheet pages produced to contain the listed addresses.

10.2 Analysis of AOCR printout

This activity was the major phase of the project. It was recognized that the AOCR printout contained all that could be learned about character readability. A guiding rule in the analysis and transcription was to preserve and carry along to the next step as much information as possible about each character. This can best be explained by first considering certain details of the AOCR machine operation.

In the AOCR system, recognition of an input character is performed simultaneously in two modes, alphabetic and numeric. The results of both are shown in the data set, that is, the alphabetic and the numeric choices are both included. In many cases recognition was not accomplished and the machine rejected the character for alphabetic and/ or numeric recognition. In

other cases there were misrecognitions, which were regarded as errors by substitution. Finally there were also occasions when no output character was indicated, because the reader did not detect the input character. This was called a washout. The most common washouts were caused by the loss of a whole line, brought about by extraneous printed material near the genuine address. This resulted in the absence of a basis for decision.

A sample section of the AOCR printout is shown in Figure 9. There are two lines of principal interest, labelled ALPHA and NUMERIC. These contain the reader's decision on the best choice of character identification, from alphabetic and numeric repertories respectively. Under the ALPHA and NUMERIC lines there are two pairs of lines; these represent the internal machine symbol for the printed ALPHA or NUMERIC entries. These lower lines were useful when the ALPHA or NUMERIC lines contained an asterisk indicating rejection. While the equipment manufacturer indicated that the internal codes could be of only limited use, it was felt worthwhile to preserve the detail contained in these codes.

Besides these lines of data on the AOCR printout there are several others containing information primarily on the number of characters in each line. These latter lines were not needed in the analysis.

In the analysis and its concomitant encoding, the work sheet formed the master list. That is, it was considered necessary to report the treatment given every character on the work sheet, but no more. Material not included in the address and, consequently, not appearing in the typed copy of address files would frequently be found by the AOCR reader; such characters were ignored in the analysis. This rule was set up so that the final data set could be generated directly from the original typed address files.

In the interest of preserving word separation in the data base, spaces and punctuation as typed were retained in the set.

The analysis itself was straightforward when there were no discrepancies in character count, i.e. none missing or expanded. The analyst was required to note on the work sheet, at the appropriate AL or NU line and character, whether the character was correctly read or not, using a check mark for the former. In cases of substitutions, the replacement character was written in. When an asterisk appeared in the ALPHA line, the analyst was directed to write the internal code, such as 12, 13, 14 or 15. In the NUMERIC line, most of the printed entries were asterisks;

these were left blank in the work sheets to save labor. Various other labor saving techniques were introduced, such as special marking to indicate that entire words were correct.

About one-third of the addresses contained expanded characters, with two and three apparent characters being indicated on the AOCR printout for one genuine input. Almost invariably these were caused by oversize figures on the mailpiece cards. Such expanded characters were segmented by the analyst as appeared appropriate, sometimes taking into account the indicated numeric interpretation. All of these data were noted on the work sheet with bounding lines as needed. It was found that segmentation of expanded characters was quite laborious, as was to be expected. The availability of the original mailpiece cards was helpful at this stage.

Other aberrations such as dropped characters (washout) were noted on the work sheet by selected codes.

10.3 Transcription to machine language

A keyboard typing system was employed to convert the data from the worksheet into machine language. Insofar as possible, steps were taken to minimize the amount of typing. The total character content was approximately 110 000 characters; thus typing two characters for each input (ALPHA and NUMERIC) meant typing 220 000 characters, or the equivalent of a manuscript of 44 000 words. The adopted technique was to use a remote terminal device attached to a small computer.

The remote terminal included a keyboard and a cathode ray tube for display. As the initial step the address tape was fed into the computer. Addresses were formatted so that each complete address was displayed, with a pointer underlining one character. The purpose of the pointer was to indicate to the typist the particular input character for which she was expected to type in the ALPHA and NUMERIC entries from the worksheet. After the worksheet data had been typed for an input character, a Transmit button caused the typed AL and NU entries to be joined to the specific input character in the address. Additionally the pointer moved on to the next character.

In the interest of reducing entry time, the underlining pointer was programmed to jump over punctuation and spaces in the original address contents. This eliminated the need for a response from the typist for the blank remainders of lines in which the address contained less than the maximum of 38 characters.

Keying conventions were developed for all of the special conditions to be reported. These are listed in Table 2 below.

TABLE 2

Typing Codes for Transcription of Recognition Results

Reported character on AOCR Printout	Written character on Worksheet	Typed character at CRT
ALPHA line		
Alpha (correct)	✓	. (Period)
Alpha (incorrect)	Same alpha character as on AOCR printout	Same as left
*(12,13,14 or 15)	12,13,14 or 15	" # $ % (See note below)
Space (40)	40	Space
None	Δ	- (Minus)
NUMERIC line		
Numeral (correct)	✓	. (Period)
Numeral (incorrect)	Same numeric character as on AOCR printout	Same as left
*(10)	Blank	Space
Space (40)	40	Space
None	Δ	- (Minus)

Note = The characters ", #, $ and % were obtained by typing Shift 2, Shift 3, Shift $ and Shift 5 according to the information on the AOCR Printout.

When multiple output characters appeared in the AOCR printout for one input character, each ALPHA and NUMERIC pair was typed according to the above table. Thus if two output characters were generated for one input, two ALPHA

and NUMERIC pairs were typed before the message was transmitted. Three or even four pairs were permitted, however in these cases the information beyond two pairs was truncated at a later stage.

11. Computer processing and formatting

Processing of the typed data resulted in formatted records on magnetic tape, one record of 720 characters for each mailpiece card. The precise layout of these records is covered in a separate specification which is included with the data set, ref (4). In brief, each record contains a Heading followed by a Recognition Results section. The latter is a series of six-character groups for each input character contained in the address tapes. The first character in the group is a reproduction of the original input from the mailpiece, the second through the fifth are the AOCR reading data, and the final character is a status flag or code to report on errors, rejects, joined characters etc. In each record there are 3 (lines) x 38 (input characters per line) x 6 (output characters for one input), or 684 characters of data on recognition results. The remaining, initial 36 characters, include the heading data.

In the computer processing the action was to generate the six character groups, one from each input character contained on the address tapes. The recognition results were inserted into the second through fifth byte positions, substantially as typed. The major exception was for the typed period, which was the abbreviation for correct reading. In the processing the period was replaced with the original alphabetic or numeric character. Finally the sixth byte, the status code, was generated automatically by the program based upon logical comparisons of the data in the first five byte positions.

This status code was employed to facilitate rapid search through the tape with minimum programming effort by users. Thus it is possible to extract only those characters having substitution errors, for example, by searching on the basis of an "E" in the last position. Other possibilities of status code usage are indicated in the summary tape description, reference (4).

The total number of records is greater than 110 000 by more than a factor of two. In the interest of complete context preservation and uniform file length, the data set also includes the residual space

characters following the address characters in each line. Thus the data set contains approximately 240 000 actual records, of which 110 000 are the useful address characters, including embedded spaces and punctuation between words.

12. Acknowledgment

The authors wish to acknowledge the contribution of Miss Harriet A. Baker for her painstaking analysis of the character recognition results which form the reference data set in Part 2. Mention is due also of Mrs. Betsey Heacock's careful transcription of the data into the computer.

13. References:

1) "Summary tape description, multilevel character data tapes for computer evaluation of optical character recognition techniques," June 1974, prepared by NBS to accompany data set tapes.

2) Optical Character Reader Test Decks, Final Report, Control Data Corp Rockville, Md. March, 1972 for USPS, Washington, D C Vol. 1 & Vol. 2 Contract RER 143-71.

3) "Laboratory evaluation of effects of envelope windows on bar-code reading" NBS Report No. 10823, March 15, 1972.

4) "Summary tape description, OCR character recognition results, data tape for evaluation of context processing" June 1975, prepared by NBS to accompany data set tape.

APPENDIX

A. Notes on microfilming and processing of Envelope Films

A.1 Illumination

The illumination during microfilming employed conventional incandescent lighting which was checked for uniformity within a few percent. In order to provide as much tolerance as possible for the overexposure-underdevelopment technique for enhancing the gray scale range, the illumination was raised by a factor of 5 to 7 over normal levels. At this high intensity, it was found desirable for the operators to wear dark glasses. Operators were also changed frequently during the approximately 40 hours of filming to reduce fatigue.

A.2 Density and contrast control

The techniques used for extending gray-scale rendition were the standard ones, namely choice of one of the fastest microfilms, followed by overexposure and matching underdevelopment. By these means the gamma (contrast slope) of the film was reduced from its usual 4 to 1.5. While almost identical speed and contrast characteristics were obtained on test strips with diluted normal developer and with less active continuous-tone solutions, we preferred the latter as less prone to rapid exhaustion.

The design center for development control called for a density on the negative of 1.25 for the image of the 85 percent reflectance chip. Under these conditions, good rendition from white to near-black reflectances was found. At the extreme black end, the transparency of the resultant negative made it difficult to see chips whose reflectance was less than 10 per cent.

The films were processed at the Bureau of the Census, with a developer normally used for low contrast motion picture film. Processing took place in a small laboratory unit designed for microfilm work, with the result that the development time was considerably less than normal for the solution employed. No replenisher was used; rather the developer was discarded after two rolls had been processed. Exhaustion was checked with pre-exposed test strips which showed negligible fade as full rolls were developed.

To produce the required positive copies, the rolls of negatives were printed onto conventional microcopy film by a commercial laboratory. Optimum exposure in the printer was worked out through submission of test strips, after which the entire batch was printed. There was a slight increase in contrast produced by the characteristic of the print film. As a result, it can be noted on the films that the

high reflectance chips, 80 percent and above, are difficult to distinguish individually. However this was considered an inevitable consequence for keeping the maximum density (1.2 nominal) within the useful region for film scanning devices.

Approximate linearity exists for the reflectance range from 15 to 75 per cent, allowing the user a wide region for the application of thresholding techniques.

B. Calibrated values of test reflectance chips

Envelope and Address films

Chip No.	Measured Reflectance at 536nm (in percent compared to barium sulfate)
1	4.0
2	5.8
3	12.0
4	17.4
5	24.1
6	29.0
7	33.9
8	41.9
9	48.7
10	56.0
11	68.0
12	74.1
13	81.8
14	85.5

C. Computation of Spectral Output

Wavelength nm	Log (film sensitivity)*	Film sensitivity*	Source relative radiance (color T=2200 K)	Filter transmittance (Wratten No. 58)	System spectral output (relative)	System spectral output (normalized)
488	-0.06	0.871	.356	0	0	0
495	-0.1	0.794	.400	.035	.011	.03
500	-0.1	0.794	.434	.10	.034	.09
510	-0.04	0.912	.508	.29	.134	.35
520	0.02	1.05	.590	.42	.260	.69
530	0.06	1.15	.680	.44	.344	.91
540	0.10	1.26	.778	.38	.373	.98
550	0.14	1.38	.885	.31	.379	1.00
560	0.16	1.45	1.000	.24	.348	.92
570	0.16	1.45	1.123	.17	.277	.73
580	0.15	1.41	1.255	.10	.177	.47
590	0.14	1.38	1.395	.045	.087	.23
600	0.12	1.32	1.542	.015	.031	.08
610	0.10	1.26	1.698	.004	.009	.02
616	0.10	1.26	1.795	0	0	0

*Sensitivity = reciprocal of exposure (ergs/cm) as published by the manufacturer required to produce specified density, in this case 1.00, above gross fog.

FIGURE 5 SCANNER IN

OCRA Test Deck
Plate No. C1154995

Income Tax Bureau
Box 3122 General Post Office
New York, New York 10001
U.S.A.

FIGURE 1 ENLARGEMENT OF TYPICAL SECTIONS OF ADDRESS FILMS

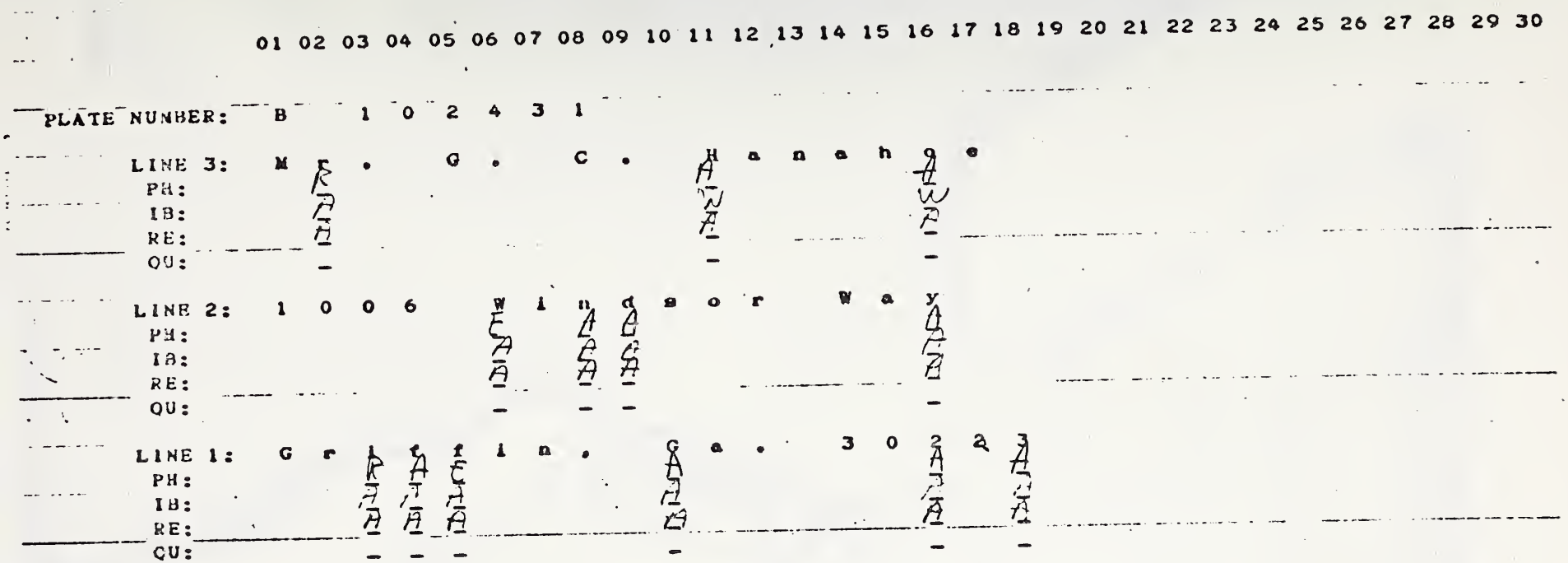

FIGURE 2 PORTION OF COMPUTER-GENERATED WORK SHEET SHOWING DESIGNATION OF SELECTED CHARACTERS

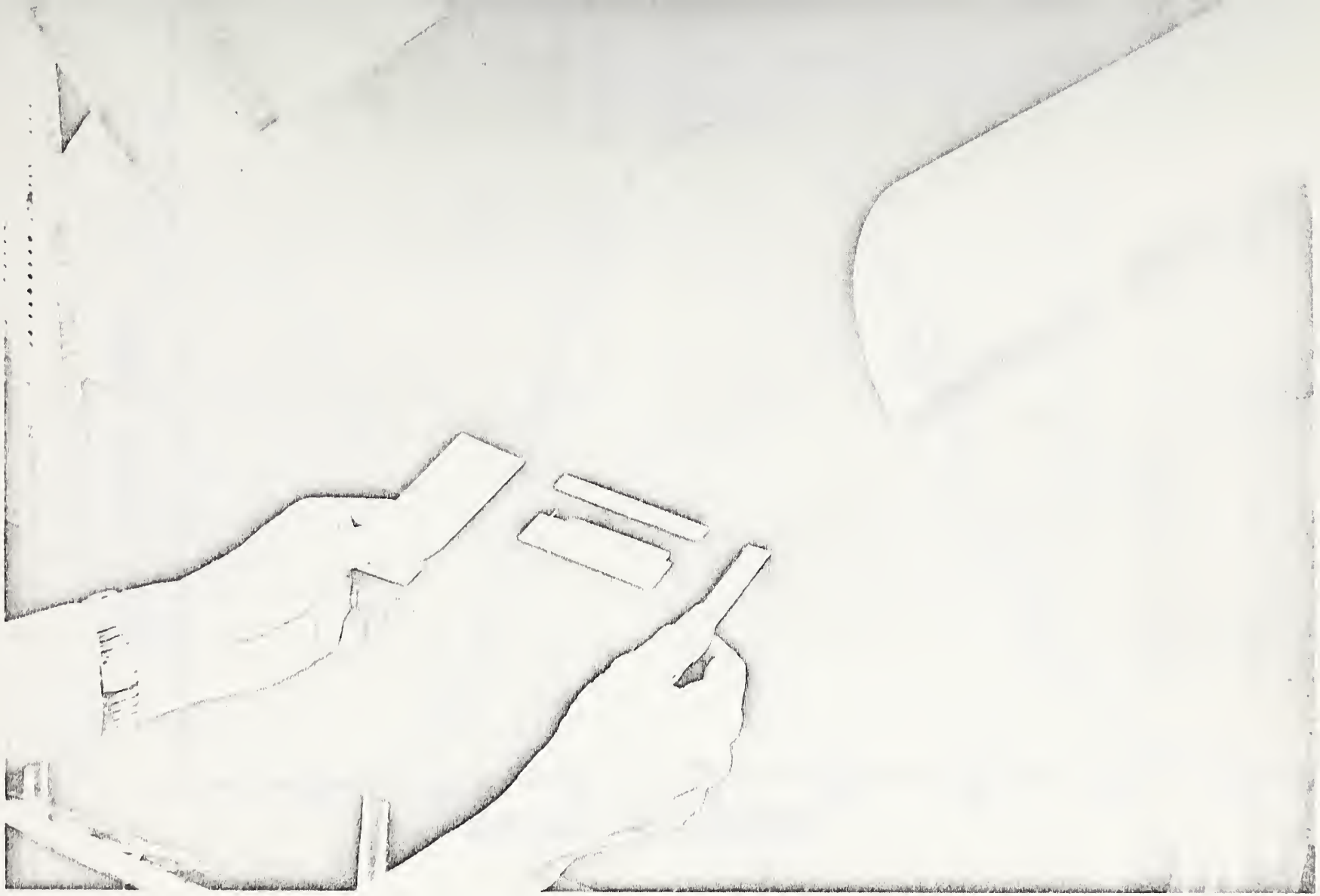

FIGURE 3 WORKHOLDER FOR MICROFILMING MAILPIECE CARDS

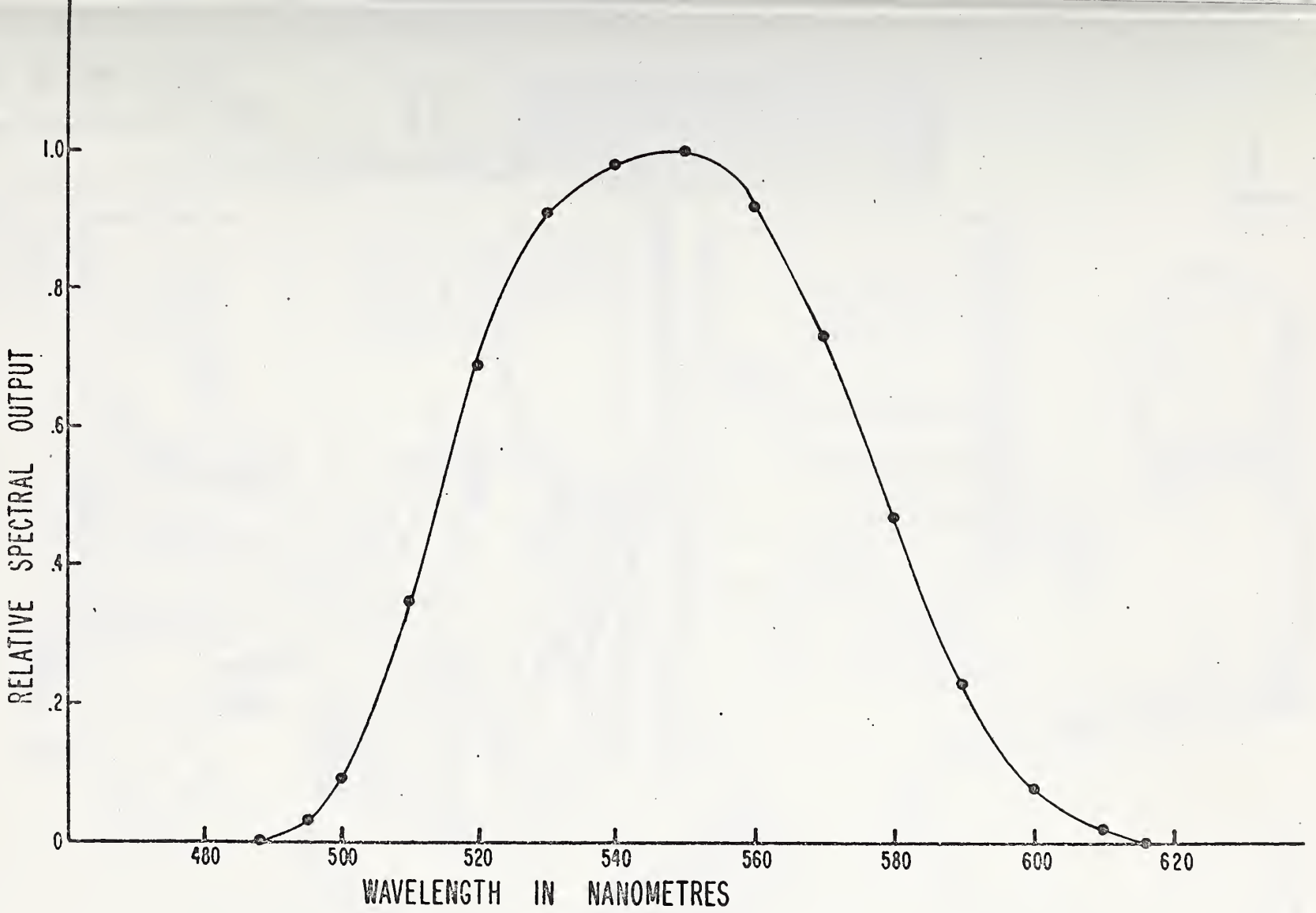

FIGURE 4 SPECTRAL CHARACTERISTIC OF DATA SET

FIGURE 4 SPECTRAL CHARACTERISTIC OF DATA SET

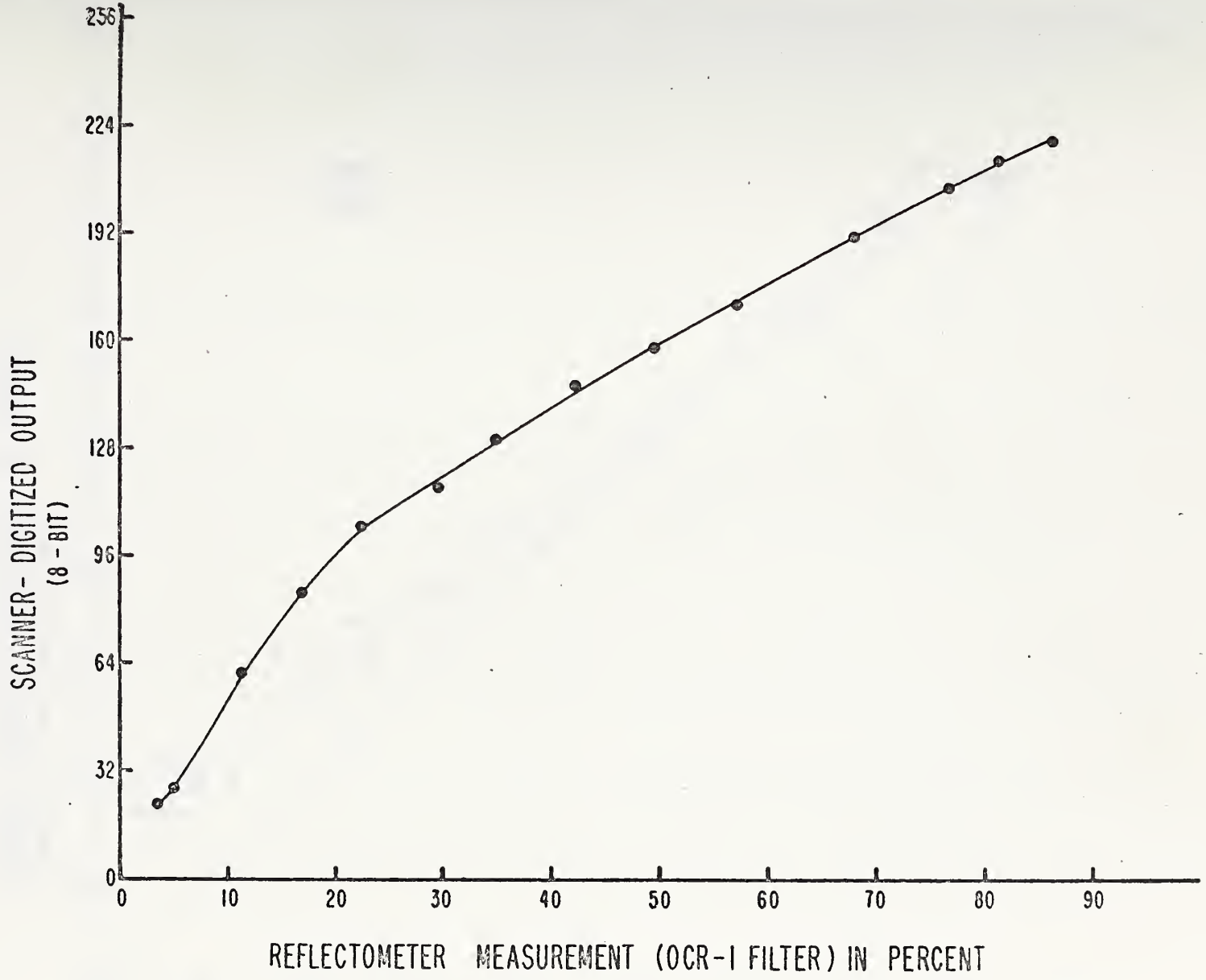

FIGURE 6 SCANNER OUTPUT AS A FUNCTION OF TEST CHIP REFLECTANCE

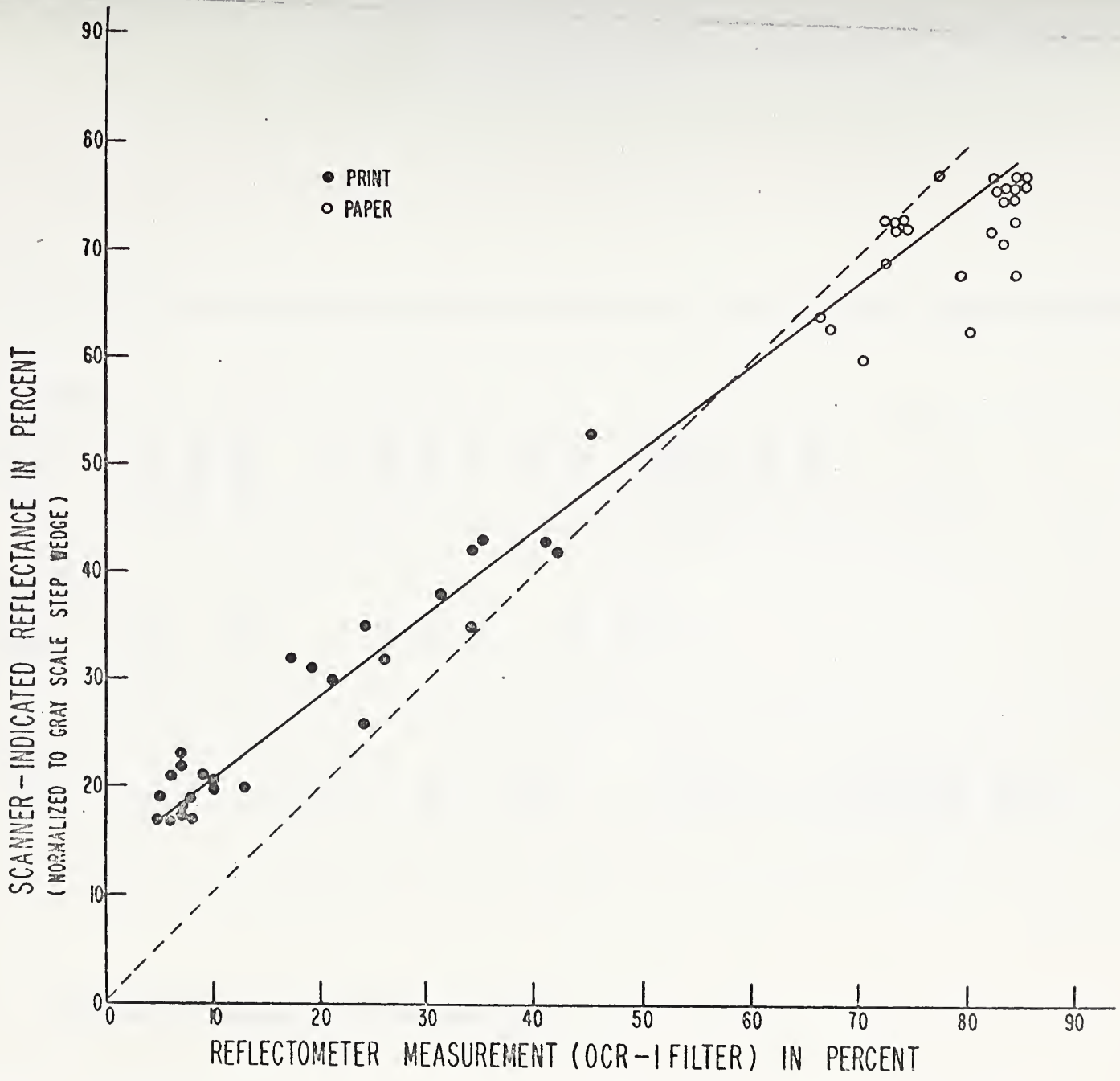

FIGURE 7 OBSERVED RELATION BETWEEN MEASURED AND SCANNER-INDICATED REFLECTANCE, AS DERIVED FROM SEGMENTS (LINES) OF CHARACTERS

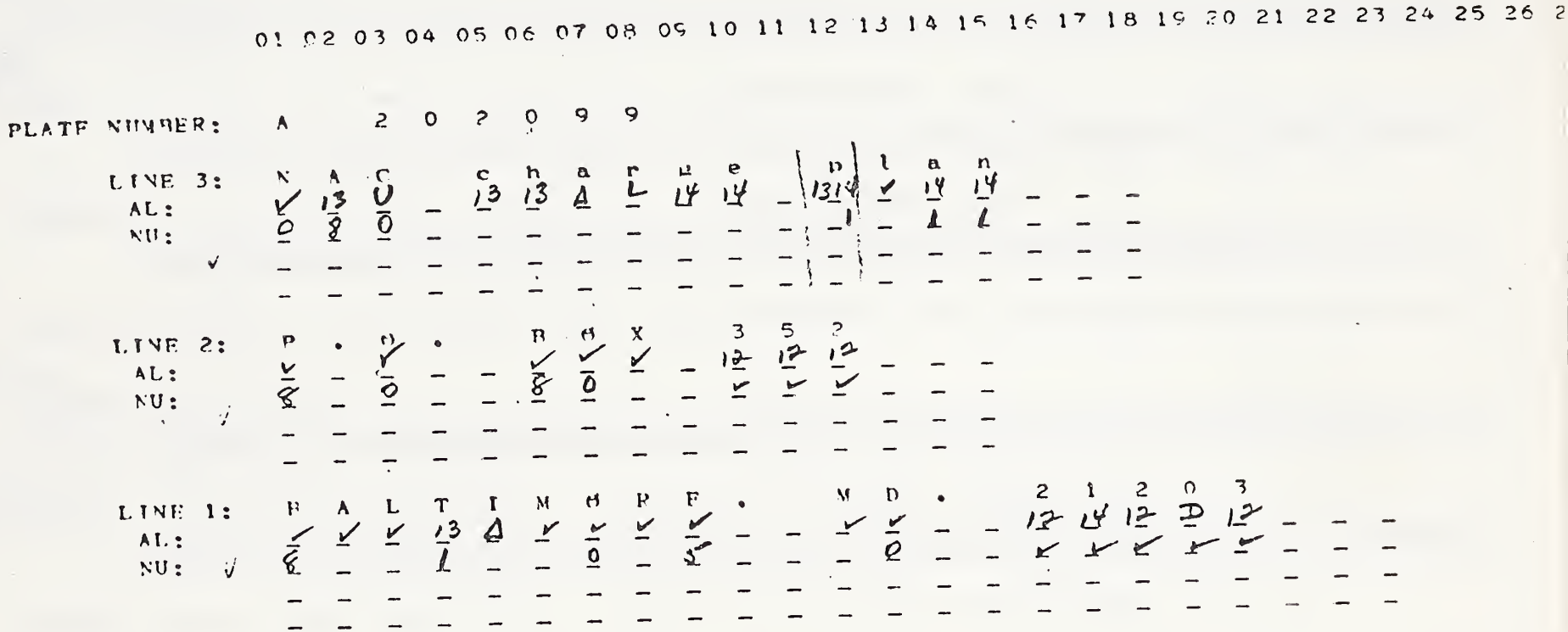

FIGURE 8 PORTION OF COMPUTER GENERATED WORKSHEET

```
DOC NO 00088  BIN 127

CONTROL 030000001111DF4000000000000000000001F04410010000000000000000000001111223032200000000000000
        C9A413332369110300AA000000000000000010042100500000F500000000000006A038F5E10077000000003000000

ALPHA   MAC , N*U **L** **L** , PO BOX *** , BAL*MORE MD ***0* <----

        DCC424D1F41101141101142400D4CDE4111424CCD1DDDC4DC4111D1AF0000000000000000000000000000000000000
        413360534033322034344060760267022206021334695044024220F0000000000000000000000000000000000000

NUMERIC **0 , 080 ***** *1*11 , 80 80* 352 , 8**1*0*5 *0 21203 <----

        11E424FFF41111141F1FF424FF4FF14FFF424F11F1F1F41F4FFFFFAF0000000000000000000000000000000000003
        000060080000000001011060800800035206080010005000021203DF0000000000000000000000000000000000000

SORT CODE 010C007F  SORTD 00037C6C  MTU1  FILEAC 1709C5FF  FLAGS 00004188882200000000000000000000

DOC NO 00089  BIN 127

CONTROL 040000000111000400000000000000000001000444100100000100000000000000112222340033000000000000000
        C7833044301590403BA0000000000000003000012000100005500000000000004B0E0260700066000000000000003

ALPHA   NAC CHARGE PLAN NCORPORATED , P O BOX W*** , BALT*MORE MARYLAND *L*W

        DCC4CCCDCC4DDCD4DCDDDDDCECC424D4D4CDE4E111424CCDE1DDDC4DCDEDCDC41D1EAF0000000000000000000000000
        513038197507315053697691354060706026706222060213344695041983154023260F0000000000000000000000000

NUMERIC 040 08436* 8*** 100*80*1*50 , 8 0 80* *352 , 8***1*08* **8****0 2120
```

FIGURE 9 SAMPLE SECTION OF AOCR PRINTOUT SHOWING CHARACTER RECOGNITION RESULTS

FORM NBS-114A (REV. 7-73)

U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET	1. PUBLICATION OR REPORT NO. NBSIR 75-746	2. Gov't Accession No.	3. Recipient's Accession No.

4. TITLE AND SUBTITLE

Preparation of Reference Data Sets for Character Recognition Research

5. Publication Date

6. Performing Organization Code

7. AUTHOR(S)

M. Leighton Greenough & Robert M. McCabe

8. Performing Organ. Report No.

9. PERFORMING ORGANIZATION NAME AND ADDRESS

NATIONAL BUREAU OF STANDARDS
DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20234

10. Project/Task/Work Unit No.

4460492

11. Contract/Grant No.

Letter Agreement 74-02934

12. Sponsoring Organization Name and Complete Address (Street, City, State, ZIP)

U. S. Postal Service
11711 Parklawn Drive
Rockville, Maryland 20852

13. Type of Report & Period Covered

Final Report

14. Sponsoring Agency Code

15. SUPPLEMENTARY NOTES

16. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)

A reference data set contained on magnetic tape has been generated for research in optical character recognition and related fields. The data set contains video data in 16 levels for each point in a 24 x 24 scanning grid, applied to approximately 30 000 characters. The input material, on some 2200 simulated address mailpieces, includes a range of character quality from excellent to poor.

The data set was prepared by first microfilming the address fields and printing positive transparencies. Then preselected characters were scanned in the NBS FOSDIC. Through the use of calibrated gray scales which were filmed and scanned along with the addresses, corrections were applied to assure linearity of overall response.

A second reference data set was also prepared from the same input material of simulated mailpieces. Through analysis of the recognition results printed out during runs on one address reader, details of identification on both alphabetic and numeric recognition modes were noted. These were encoded into six-character groups on the data tapes, formatted to provide one such group for each input character in the original addresses. Recognition results are shown for approximately 110 000 input characters.

17. KEY WORDS (six to twelve entries; alphabetical order; capitalize only the first letter of the first key word unless a proper name; separated by semicolons)

Character recognition; context research; data bases; optical character recognition; pattern recognition; reading machines

18. AVAILABILITY

[X] Unlimited

[ ] For Official Distribution. Do Not Release to NTIS

[ ] Order From Sup. of Doc., U.S. Government Printing Office Washington, D.C. 20402, SD Cat. No. C13

[X] Order From National Technical Information Service (NTIS) Springfield, Virginia 22151

19. SECURITY CLASS (THIS REPORT)	21. NO. OF PAGES
UNCLASSIFIED	52
20. SECURITY CLASS (THIS PAGE)	22. Price
UNCLASSIFIED	$ 4.25

USCOMM-DC [illegible]