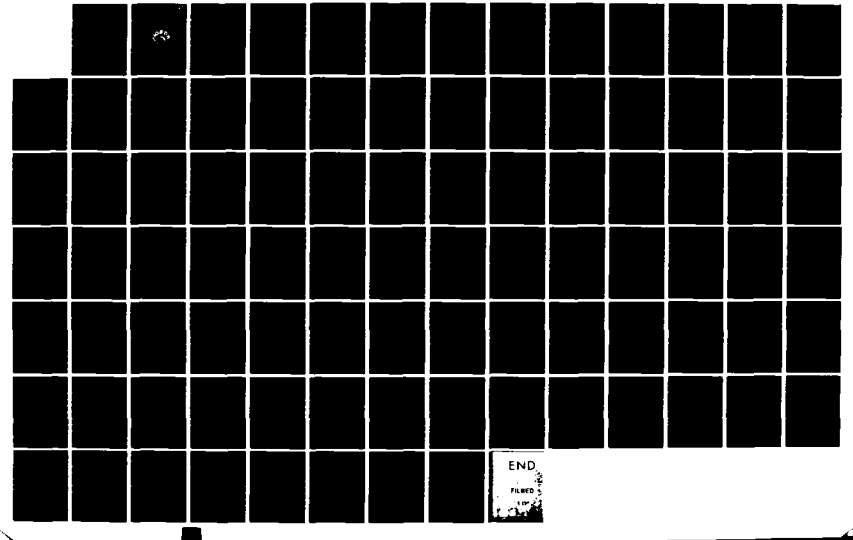
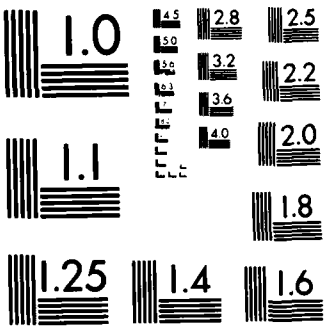
AD-A131 341 AN OVERVIEW OF OPTICAL CHARACTER RECOGNITION (OCR) TECHNOLOGY AND TECHNIQUES(U) NAVAL OCEAN RESEARCH AND DEVELOPMENT ACTIVITY NSTL STATION MS. 1/1

UNCLASSIFIED L K GRONMEYER ET AL. JUN 78 NORDA-TN-217 F/G 9/2 NL





NORDA Technical Note 217

Naval Ocean Research
and Development Activity
NSTL Station, Mississippi 39529

An Overview of Optical Character Recognition (OCR) Technology and Techniques

L.K. Gronmeyer B.W. Ruffin
Ocean Science and Technology Laboratory
Mapping, Charting, and Geodesy
Development Group

M.A. Lybanon P.L. Neely
S.E. Pierce, Jr.
Computer Sciences Corporation
NSTL Station, Mississippi

Prepared for Defense Mapping
Agency, HQSST

June 1978

The following individuals participated in writing this report:

NAVAL OCEAN RESEARCH AND DEVELOPMENT ACTIVITY

Larry K. Gronmeyer
Byron W. Ruffin

COMPUTER SCIENCES CORPORATION

Matthew Lybanon
Patrick L. Neely
Stephen E. Pierce, Jr.

ABSTRACT

This report presents the results of a survey of 1978 Optical Character Recognition (OCR) technology conducted by NORDA Code 302, the Mapping, Charting, and Geodesy Development Group. The Systems Engineering Branch, Engineering and Science Services Laboratory (ESSL), National Space Technology Laboratories was contracted for a major portion of this effort. The survey was required by the Defense Mapping Agency (DMA) as a prelude to continuation of DMA funded OCR system development efforts within NORDA. Three principal areas of OCR technology development were reviewed:

- Government applications of OCR.
- Commercial OCR products.
- Software and basic research.

This document also contains an extensive bibliography and discussion of selected papers.

ACKNOWLEDGMENTS

This NORDA Technical Note is a reprint of a contractor-prepared report printed in June, 1978. The work was sponsored by DMA under Program Element 63701B, "Optical Character Recognition System Development," and was performed by NORDA Code 302, the Mapping, Charting, and Geodesy Development Group. The DMA Program Manager was Mr. Robert Penny. The following individuals participated in the project: Larry K. Gronmeyer and Byron W. Ruffing, of NORDA; Matthew Lybanon, Patrick L. Neely, and Stephen E. Pierce, Jr., of Computer Sciences Corporation, under a NASA technical work request.

AN OVERVIEW OF OPTICAL CHARACTER RECOGNITION (OCR) TECHNOLOGY AND TECHNIQUES

JUNE 7, 1978

Prepared For

DEFENSE MAPPING AGENCY

By

NAVAL OCEAN RESEARCH AND DEVELOPMENT ACTIVITY

And

COMPUTER SCIENCES CORPORATION
SYSTEMS ENGINEERING BRANCH
ENGINEERING AND SCIENCE SERVICES LABORATORY
NATIONAL SPACE TECHNOLOGY LABORATORIES
NSTL STATION, MISSISSIPPI 39529

TABLE OF CONTENTS

I. INTRODUCTION

This report presents the results of a survey of current Optical Character Recognition (OCR) technology conducted by the Systems Engineering Branch, Engineering and Science Services Laboratory (ESSL), National Space Technology Laboratories for the Naval Ocean Research and Development Activity (NORDA). The survey was requested by the Defense Mapping Agency (DMA) as a prelude to continuation of DMA funded OCR system development efforts within NORDA. Three principal areas of OCR technology development were reviewed:

- Government applications of OCR.
- Commercial OCR products.
- Software and basic research.

OCR application requirements within DMA exhibit features which, in many instances, markedly distinguish them from typical OCR applications found in commercial, business and governmental areas. In particular, the free format of the character data to be read, such as the smooth sheet oceanographic sounding problem being investigated by NORDA, departs radically from the constrained data input presented on prescribed forms or cards which normally are required with existing OCR systems. The survey was concerned with identifying details of systems, techniques or research that could be related directly to DMA applications, in particular smooth sheet digitization.

The survey of government-sponsored OCR work was conducted to identify any ongoing development activities or prior experience that could be incorporated in a solution to DMA's requirements. Information was obtained through the Defense Documentation Center and through field surveys of Federal, State and local government OCR projects.

Commercial OCR systems were reviewed to determine their applicability and/or adaptability to the DMA problems. It should be emphasized that this portion of the survey was concerned with complete systems and not individual components that might be configured into a suitable laboratory development system for DMA OCR applications research.

In gathering material for the review of software and research activities, it was found that a few other surveys of the OCR field have been published in recent years. Two

such articles, by Harmon [38]* and Ullmann [89], include tutorials on methods.

Harmon's article is problem-oriented; it discusses the general topic of OCR, then concentrates on the recognition of handprint and script, then further particularizes to a discussion of decision-making methods. Since Harmon's review covers the period up to 1972 it is an excellent supplement to the present report, which mostly concentrates on later work.

The Ullmann survey is more technique-oriented. After a review of scanners and other hardware, several preprocessing techniques are discussed. Then some recognition methods are covered, including a few used in commercial systems. This is followed by a discussion of the recognition of distorted characters (e.g. handprint, which deviates from a fixed font), and some special topics.

Both articles couple their discussions of methods with references to the literature. Aside from these, the annual surveys on image processing by Rosenfeld, which are narrated bibliographies, contain sections on the character recognition literature. The three most recent ones as of this writing are [65, 66, 67].

Some other surveys of special types have also been published. Baty [14] wrote a nontechnical article describing the current state of the field, several applications and projecting future developments. Himmel [39] presented four case studies in handprint OCR. Kegel [45] made a survey of OCR users who read numeric handprint.

The section on software and research gives a representative view of recent activity in the field. The majority of the references are concerned with the recognition of handprint, since that (if cursive script is included) is the area where most of the remaining unsolved problems lie. Not all of the work dealt with complete systems for recognizing characters. Some of the research was concerned with specific techniques, and some related topics were also covered. There were a few applications to particular problems, such as postal address reading.

A summary of conclusions reached from the survey results is presented at the end of the report.

*Numbers in brackets refer to the list of references in Appendix E.

II. GOVERNMENT APPLICATIONS OF OCR

An extensive literature search was conducted in the field of optical character recognition through the Defense Documentation Center (DDC). The following products were received: (1) Report Bibliography, (2) Work Unit Summaries, and (3) Independent Research and Development Reports. In addition, a National Technical Information Service (NTIS) search was performed. The results of the NTIS search and the DDC Report Bibliography are presented together. Summaries of field surveys are also provided.

A. WORK UNIT SUMMARIES

Work unit summaries were received on the activities of three military facilities. ECOM Communication/ADP Lab, of Ft. Monmouth, New Jersey, was to devise new techniques for automated data reduction that would be applied to the reading of handprinted material. They developed a handprint reader system but came to the conclusion that handprinted readers are only practical when they are tailored to an individual's writing style. This effort was completed in 1972.

The Marine Corps Development Command, Quantico, Virginia, was to interface a ruggedized version of a commercial grade OCR reader with the Automated Message Entry System (AMES). The effort is still in progress as of 1977.

Rome Air Development Center (RADC), Griffiss Air Force Base, New York, has had four significant OCR projects. One was to evaluate the existing optical character recognition capabilities of a Russian typeset print reader. This was done by Information International, Inc. A similar effort is to be tried by Logos Development Corporation. Another project was initiated to develop a capability to input handprinted alphanumerics by character recognition techniques for a variety of records and documents. The conclusion was that there should possibly be a family of OCR readers, each tailored to a particular set of variables and specific applications. This effort was completed in 1977. The final effort was to write an AGARD report entitled "Optical Character Recognition and its Application to Documentation - A State of the Art Review." This document addressed optical character recognition of machine print only. The conclusion was that there is no present method for

the automatic detection, scanning and processing of graphics. This effort was completed in 1977.

B. INDEPENDENT RESEARCH AND DEVELOPMENT REPORTS

The Independent Research and Development Reports identified five companies involved in OCR developmental work under government contract. Control Data Corporation had an effort in 1975 to develop an improved set of numeric handprint and symbol character recognition algorithms which could recognize a large range of character shapes without a training set. The approach chosen was to extract basic features such as line beginnings, ends, splits, joints, and a class measurement feature related to line slope and flat information. In 1974, a prototype handprint unit was built and proven satisfactory. This effort was terminated in 1975.

E-Systems, Inc. put forth an effort to develop a single font numeric reader. The result was a single font OCR with a sixteen-character capacity. This effort was terminated in 1976.

The Ford Motor Company received a research contract in 1973 to develop a means of reading postal mail by OCR techniques, converting the information to a bar code and imprinting the code on the envelopes. A printer/reader would then be inserted between the OCR and the Letter Sorting Machine (LSM). The reading process would be simplified if the information was in digitized code. The project is continuing.

The Singer Company was involved in an effort to develop a low cost, hand-held OCR reader that reads a 16 character field, and to develop a wand-based system for reading the Universal Product Code (bar code).

Stanford Research Institute developed a means of producing scientific and technical publications using OCR as an input device.

System Development Corporation had an effort to develop software which would recognize handprinted input. Their result was an average recognition rate of 88 percent. This was completed in 1975.

C. FIELD SURVEYS

1. City of Baltimore, Maryland

The city of Baltimore has one IBM #1288 OCR reader which they use to prepare property tax bills, water meter bills, income tax forms, and to control food stamps. They use it strictly for machine print in a "turn-around-document" mode. This is the same reader the State of Maryland currently uses for their income tax scanning tasks.

2. State of Maryland Income Tax Division

The Income Tax Division currently scans Maryland withholding form #506 from each employer on an IBM-1288 OCR device located in Baltimore. The forms are prepared using constrained handprint methods of "filling-in-the-box." The acceptance rate is 85%, but this includes the preprinted data (OCR-A font) as well as the handprint. The personnel estimate that most of the errors are rejects; there are little substitutions.

They currently have an RFP out for the scanning of income tax documents utilizing OCR-A font and constrained numeric handprint. Additionally, the reader must be able to read handprint applied in lead pencil. All scanned data will be output to magnetic tape. The equipment must be able to halt when a character is rejected and allow the operator to manually key it in via a CRT terminal. The equipment must also be able to continuously scan and send all rejected documents to a special receiver for later manual processing.

3. National Bureau of Standards (NBS)

Discussions were held with Mr. Jacob Rabinow of NBS concerning various techniques he used for optical character recognition, such as the Crosswalk Technique, and Watchbird Technique. Mr. Rabinow could provide no documentation since these techniques all are patents. He suggested that thinning the character after acquiring it was an improper technique; instead, if thinning was necessary, it should be a part of the acquisition system. He does not believe in feature selection for OCR because of the increasing number of features needed as the character data sets become larger.

Mr. Rabinow indicated that the Social Security Administration and Internal Revenue Service are studying the feasibility of using OCR to read 26 machine fonts. He also said the standard commercial reading error rate (machine print) is 1 character in 100,000. It was suggested that Mr. Art Hamburgen of IBM in Rochester, Minnesota, be contacted. It was

also mentioned that Recognition Equipment, Inc., (REI) is using a Retina device (an analog array, with all gray scales) for Sears Roebuck scanning activities.

4. U.S. Postal Service

The Postal Service has completed a data base for use in experiments with OCR handwritten algorithms. This data base was prepared at the Post Office Research Lab in Washington, D.C., and at the Air Force Weapons Lab at Kirtland AFB, Albuquerque, New Mexico. The purpose of the data base is to express in measurable terms the problem of free (unconstrained) handwritten numerals.

This data base consists of approximately 9500 handwritten Zip Codes (or approximately 47,500 numerics) obtained from 14 dead letter offices across the country. The Zip Codes are photographically reduced (2:1) on 35 mm film and then scanned using a 1.615 micro-programmable CRT scanner (Information International, Inc., Model PF4-3). The scanner output provides 16 reflectance levels (shades of gray) on a 64X224 grid for each Zip Code. This represents a resolution of 14,336 individual gray levels. This data is output to tape.

At the completion of the scanning, the original material is evaluated at the Postal Lab for the following criteria: texture and primary color of the paper; ink color; pen type; stroke width and uniformity; character-to-character height variation; Zip Code skew; character slant; distortion of character. Each Zip Code is encoded with identifying data using a Threshold T-500 speech recognizer. A tape containing this information is prepared.

The tape containing the encoded information is then sent to Kirtland AFB and is merged with the scanned data tape. The resulting data base is known as Scan/Descriptor Tapes. These tapes are 800 bpi IBM-compatible 9-track tapes with odd parity. The descriptors are recorded in 8-bit EBCDIC characters for descriptors; the scan data are recorded in double-packed 4-bit binary format. The data can then be accessed by various criteria (i.e., all 5's, numbers that slant left, etc.) for testing in various algorithms. This data base is now complete.

Originally, three contractors were under consideration for associated hardware and software development. These were Philco Ford, IBM, and Recognition Equipment, Inc. Only REI remains as a viable contractor.

Currently, the Postal Service is making a data base of constrained handprinted numerics. The input is a box-type document for the date and Zip Code. It must be noted that these data bases are for input to a recognition technique(s) only; the Postal Service is not developing algorithms for handprint recognition. Only an estimated 15% of the mail is handprinted.

The Postal Service estimates that 85% of the mail is now machine printed (typed). To support this type of mail, the Postal Service has 20 OCR-1 machine-print readers placed in various cities in the country. These read at a rate of 12 letters (pieces of mail) per second, and they scan the last line only (Zip Code). The Postal Service has one OCR-2 reader in Boston, Massachusetts, which also reads 12 letters per second, but scans all lines of the address. In New York City, the Postal Service has an IBM Advanced Mail Address Reader, a contextual processing machine for word recognition using omni-font. It reads the complete address and then checks it in a look-up table containing all acceptable street names and addresses. It selects the destination by contextual means if some alpha-numerics are unreadable.

The Postal Service is also experimenting with voice recognition techniques for sorting parcels using the first three significant digits of the zip code.

5. Social Security Administration

In 1964, the Social Security Administration began looking at OCR techniques for the processing of their quarterly #941 forms received from employers. The quarterly input was about 100 million line items, of which 50 million lines were OCR readable. The average page of #941 contains 22 lines; the maximum number is 44 lines per page. The standard reading width is 8 lines per inch. Social Security sent out a RFP specifically to process the #941 form. IBM responded and built a dedicated one-of-a-kind OCR reader (IBM #1975) on-site.

The IBM #1975 reads machine-printed #941 forms only, and requires timing tracks on the forms for picking up the lines. It reads three fields: Social Security number (numeric only); name of employee (alpha only); and the dollar amount (numeric only). An alpha or numeric character in the wrong field would generate an error. It reads at a 52% to 55% acceptance rate which is measured in the number of lines correctly read per page, not the numbers of characters read. The rejected lines are marked; these are later hand-

punched. It also rejects a complete page when the number of lines rejected reaches 50% of the total lines; this feature was incorporated so that time was not wasted reading bad pages. The substitution rate for the #1975 is about 4%. The recognition logic uses a "best guess" technique for characters. The reader does not read specific employer information which is also on the #941; this information has to be manually punched. The IBM #1975 has been permanently taken out of service as of 1977.

In 1972, the Social Security Administration released another RFP for a new OCR reader to supplement the IBM #1975. The basic requirements were for an acceptance rate (lines, not characters) of 65%; a scanning rate of 40,000 lines per hour; and the ability to also read the specific employer information on the #941 that the IBM #1975 could not read. Recognition Equipment, Inc., (REI) responded with their Input 80 OCR reader. This is an off-the-shelf reader, although it has been modified by REI to read about 400 machine-print fonts. It is not used to read handprint. The Input 80 also reads the employer information and the special box "check-marks" that the IBM #1975 could not read. If a character cannot be read, the machine places a red dot on that particular line, which is manually processed, and later merged with the original data. This reader can read between 35,000 and 38,000 lines per hour; it has an acceptance rate of 68% (lines), and a substitution rate of 4%. Social Security personnel manually process 1% of the data for verification and quality control. They currently have two operational REI Input 80's reading about 50 million lines per calendar quarter.

The Social Security Administration has one Scan Data #2250/1 OCR device for reading the 9-digit Social Security number in a constrained handprint format on Form #1002. This form requests information from an employer about a previous entry he made on the #914 form. The questions are typed in OCR-A font. The #1002 form is read and the data are stored in disk under control of a PDP-8 computer. All rejects are stored on a separate disk and are subsequently displayed on a CRT terminal for operator correction. The corrected data are then merged back on the disk with the accepted data. This is the only system the Social Security Administration has at this time which has a built-in edit capability and which will scan numeric handprint.

The Social Security Administration has now been charged with preparing the annual W-2 forms from employers. To support this new effort, they have released an RFP to obtain equipment to serve this task. Briefly, the basic requirements are to read forms

W-2 (three to a page), W-2P and W-3 with an acceptance rate of 96% for all numeric characters and 93% for all alphabetic characters. The substitution rate shall not exceed 3/4 of one percent for numeric characters, and 3 percent for alphabetic characters. This RFP is for machine-print fonts (400) only, and is not intended for handprint.

6. Bureau of the Census

The Census Bureau routinely sends various types of questionnaires to industry, farmers, and the general population to gather data. These forms are either filled out directly by the group concerned, or they are prepared by a field interviewer. In either case, the forms are prepared by filling in the appropriate box or circle.

The Census Bureau, working closely with the National Bureau of Standards, invented the hardware (mark sense readers) to accomplish this massive data gathering and processing effort. The special readers are called FOSDIC readers (Film Optical Scanning Device for Input to Computers). The Census Bureau has had a great deal of success using this data collection method.

They have tried using constrained handprint on some of their forms but feel that this method is not reliable enough for their use. This is because people do not follow the guidelines for making the characters. They require virtually 100% correct reading rates for their data, and the constrained handprint method does not meet this goal. It is easier for the public to fill in a box or circle than to duplicate a sample character set in a constrained manner. As a result, the Census Bureau is no longer involved in using constrained numeric handprint. They are continuing to use their FOSDIC readers.

III. OCR PRODUCTS

A. INTRODUCTION

The survey of commercial OCR systems that was conducted for this report concentrated on complete OCR systems and their present or directly modifiable capabilities. There was no attempt to survey and catalog all components that might be utilized in the assembly of a laboratory development OCR system for use in DMA applications. Such a catalog would not provide a representative picture of the state-of-the-art in the commercial OCR field since many components that might be included are not produced for specific OCR application in the first place.

Information for this part of the survey was gathered from two principal sources:

- Computer Decisions Magazine's OCR Manufacturers List
- Datapro Feature Report - "All About Optical Readers" (May 1977).

Additional background information was obtained from the survey papers of Harmon [38] and Ullmann [89].

A list of forty-two (42) OCR vendors was compiled from the Computer Decisions list and from lists and advertisements appearing in other computer equipment magazines and journals. This group of vendors was contacted for product information. A total of eighteen (18) replies were ultimately received from this survey, either directly or as a result of follow-up phone calls.

Subsequent to the above correspondence, the Datapro Feature Report of May 1977 was obtained. This report provides a comprehensive overview of the state-of-the-art in OCR systems as of May 1977 and as such provides the majority of the information necessary for the present survey. However, due to copyright restrictions, the summary table from the report cannot be included in this present report. Of the thirty-four (34) vendors covered in the Datapro Report, twenty (20) were included in the group of forty-two (42) originally circulated.

Analysis of the Datapro Report indicated fourteen (14) vendors whose products included handprint in the list of fonts recognized. Ten (10) of these vendors were among the list of respondents to the original vendor survey. Of the remaining four (4), IBM has been most reluctant to supply any information despite repeated phone calls and requests for assistance; a lack of clearance to release proprietary information has been quoted as the problem. The systems of the remaining three (3) vendors were determined, from the Datapro Report, to fall into the category of basic document and page readers as discussed below.

From the above outline of the OCR commercial product survey, it is felt that (with the exception of IBM) information on available OCR systems, with any handprint capability, is reasonably complete. One (1) additional system with handprint recognition capability was identified, and one (1) system that can read a wide variety of printed formats (books, letters etc.).

A list of vendors is included in Appendix C.

B. OCR SYSTEM TYPES

Three classifications of optical reader machines are normally used:

- Mark Readers
- Bar Code Readers
- Character Readers

In some usage the term OCR is applied to systems of all three types. However, OCR is more correctly used for character readers only, though in some instances, systems possess all three capabilities.

Mark and bar code readers are utilized in a restrictive and specialized range of applications where the input data is frequently in the form of a punched card (Mark reader) or retail industry product code (bar code reader). Such systems are clearly not appropriate for consideration for application to the DMA tasks. They are therefore not discussed further in this report other than to point out that the Datapro Report includes a comprehensive tabulation of available systems.

C. CHARACTER READERS

Optical character readers have been classified in the Datapro Report into five groups:

- Document Readers
- Page Readers
- Combination Document/Page Readers
- Self-Punch Readers
- Journal Tape Readers.

In each instance this classification is based on the size and form of the input data records. Most present-day OCR systems fall into the first three categories.

Since the present survey was aimed at systems possessing handprint recognition capability, character readers only handling machine printed or stylized fonts are not discussed further. Vendors of such readers are, however, listed in Appendix C.

D. HANDPRINT READERS

As indicated in Section III-A above, some fifteen (15) OCR systems were identified that included recognition of handprinted characters within their capabilities. These systems are listed in Table 3-1.

None of the systems listed can directly solve the immediate DMA problem of digitizing and recognizing smooth sheet data, nor would any of the systems, in their present form, be suitable for a laboratory development OCR system. However, the basic technology employed in several of the systems would allow solution of the smooth sheet problem, if appropriate modifications and developments took place.

A more detailed discussion of the systems listed in Table 3-1 follows, in particular those systems which were tried on a smooth sheet sample.

1. Smooth Sheet Sample Tests

As a result of information obtained on the various OCR systems with handprint reading capability and telephone conversations with vendors' representatives, four (4)

TABLE 3-1

Name of Company	Model	Scanner	Character Format	Comments
Cognitronics	System/70	Laser	Constrained	Sent smooth sheet sample
CompuScan	Alpha	Linear Array	Constrained	Witnessed system used for Newsprint type setting.
Control Data Corp.	959	Photocells	Constrained	Document and page reader
Cummins-Allison	Scanak 4216	Photocells	Constrained	Document reader
Hitachi-Zosen	XONDEX	Not Known	Constrained	Handprint on special coding form
IBM	Various	-	-	Information not available
Information International	GRAFIX I	CRT	Unconstrained	Sent smooth sheet sample
Input Business Machines	RIT 2000	Self Scan Array	Constrained	Document reader
Key Tronic	Data reader	Photocell	Constrained (numeric)	Document and page reader
Lundy	8400	Linear Array	Constrained	Page reader
National Computer Systems	OpScan 37	Photocells	Constrained (Numeric)	Document and page reader
Optical Business Machines	Laser OCR-ONE	Laser/ Photocell	Constrained	Document and page reader
Recognition Equipment	Input 80	Linear Array	Constrained	Sent smooth sheet sample
Scan-Data	2250	Flying Spot	Constrained	Document and page readers.
Scan-Optics	540	Image Dissection	Constrained (Numeric)	Sent smooth sheet sample

sample smooth sheets were sent out for trial reading and evaluation. The vendors who participated in these tests were:

- Cognitronics Corporation
- Information International, Incorporated
- Recognition Equipment, Incorporated
- Scan-Optics, Incorporated.

It was recognized that none of the systems possessed a data entry or scanning system suitable for handling the 37 inch x 42 inch smooth sheets. In addition, the orientation and positioning of the sounding data on the sheets were expected to be beyond the present capabilities of any of the systems.

The purpose of the tests was to determine if the recognition techniques used would be suitable for recognizing typical sounding characters and if the vendors felt that their equipment could be readily modified to handle this type of problem. It was also understood that the smooth sheet problem was only representative of numerous applications within DMA for a free-form OCR system.

Cognitronics made a brief analysis of the requirements and indicated that they have the technology to solve the problem. They estimated that the cost of a suitable system would lie in the region $100K - $125K. Their estimate was based on using a standard hardware configuration, developing specialized software and feeding the smooth sheet into the system in "slices."

Information International has developed the GRAFIX I Image Processing System with special application to the OCR area. A more detailed system functional description is included in Appendix B. This is a large-scale system based on a DEC-10 central control computer with a special purpose two-dimensional image processing computer, the Binary Image Processor (BIP). Since the GRAFIX I has been tested successfully on documents which were not created specifically for OCR applications (including unstylized mixed alphanumeric handprint), Information International feels that the system could be used to solve the smooth sheet problem with the development of new software for character detection and isolation.

Once again scanning and document handling pose a severe problem in that the GRAFIX I is configured for microfilm data entry. Thus "several dozen" photographs of a smooth sheet would be required. Datapro lists the typical price of a GRAFIX I system to be $2M. However the system would certainly have the computational power to address the DMA applications. Several papers have been published on the GRAFIX I and its applications, such as [32] and [33].

Recognition Equipment, one of the leaders in the development of OCR equipment, manufactures several document/page readers for general commercial applications. A functional description of the Input 80 System is contained in Appendix B. They also produce a hand-held OCR wand for use with a point-of-sale terminal. After study of the smooth sheet sample, they indicated that their present systems would not be suitable for handling that application. They recommended Information International as being the most likely company with a suitable product.

Scan Optics manufactures four (4) different document/page reader systems configured around an HP-21 minicomputer. Numeric handprint with a limited font of alpha and other handprinted characters can be read. Their conclusion, after study of the smooth sheet sample, was that their equipment would not be suitable for this application. The size of the document, together with orientation and position variances in the sounding characters, once again presented a fundamental obstacle.

2. Other Systems

As indicated in Section III-A, two (2) OCR systems with potentially interesting capabilities were identified, which were not covered in the Datapro Report, namely:

- Hitachi-Zosen XONDEX
- Kurzweil Computer Products Reading Machine

The XONDEX System was developed in Japan primarily for the purpose of transcribing computer programs from handwritten coding forms into punched cards thereby avoiding keypunch errors. Although this system claims 99.9% reading accuracy, and to have been designed for handwritten character recognition, the constraints on the character formation and positioning on the special coding form place it in the same class as other document/page readers as far as the DMA applications are concerned.

The Kurzweil Reading Machine was designed to convert printed narrative (books, letters, documents, journals etc.) into speech as a reading aid for the blind. Considerable success has been obtained with this system and a more conventional OCR system has since been marketed - the Kurzweil Data Entry Machine. With this system, a period of training is required to adapt the machine to the material being scanned. Once training is accomplished, production throughput is achieved. For good quality printed material, substitution errors of 1 in 20,000 (or better) are quoted. As yet handprint has not been included in this system's capabilities, although both Kurzweil machines are omnifont readers that can also accept type in a wide range of sizes. The system is turnkey in operation with proprietary software and costing $117K. The Kurzweil Reading Machine, which has fewer peripherals, sells for $50K.

E. CONCLUSIONS

It is clear from the smooth sheet evaluations, and the survey of commercial OCR systems in general, that no existing system is configured in a manner that could readily solve the smooth sheet problem or lend itself to a laboratory development system. The only exception to this is the GRAFIX I. However, this system is more properly termed an image processing system and, as such, is in a completely different class from the other systems, costing roughly four (4) times as much as the next most elaborate systems (Recognition Equipment Input 80C and Scan-Data 2250-1 425).

All of the existing systems which include handwritten characters within their recognition capability rely on the fact that such characters are placed in predetermined positions on the data entry form and conform to the American National Standards Institute (ANSI) format.

It is possible to separate the OCR process into three basic stages:

- Character Acquisition
- Character Isolation
- Character Recognition

In the commercial systems surveyed, the first two stages are accomplished through fixed format documents and constrained character formation. For the DMA applications, particularly smooth sheet digitization, character acquisition and isolation form a major part of the problem. It is in these specific areas, therefore, that the commercial OCR systems are not presently adequate or even easily modifiable.

In the character recognition stage, existing commercial techniques, such as feature analysis and matrix matching, are appropriate for DMA applications and give promise of low error rates, both reject and substitution. However, this part of the commercial OCR system normally consists of proprietary software installed on a general purpose minicomputer, and not available as a separate item.

The overall conclusions from the survey of commercial OCR systems are, therefore, the following:

- No existing OCR system suitable for use as laboratory development system.
- Recognition techniques developed with appropriate accuracy.
- Major part of DMA applications lie in character acquisition and isolation.
- Commercial OCR systems basically consist of a special purpose scanning and document handling device, a general purpose minicomputer (with or without peripherals) and proprietary software.
- Laboratory development system should be configured from specific components to meet DMA applications.

IV. SOFTWARE AND RESEARCH

A. GENERAL DISCUSSION

The field of optical character recognition has reached its present state of maturity during a period that has seen the development of very sophisticated and powerful computer graphics and text editing capabilities. It is natural, therefore, that there is some relationship. Commercially developed OCR equipment is used mainly in the text processing and graphic arts markets [14]. One example of both types of applications is supplied by the newspaper publishing field; some newspapers use OCR equipment for input of both news and advertising text.

The development of techniques in the OCR field with respect to the type of material to be read has not been uniform. Machine-produced print (e.g., typescript) can be read rapidly with a low substitution error rate, at least when certain constraints are observed. However, handprint reading ability is far less common in commercial OCR machines, and the error rates are higher. Reading cursive script is almost out of the question. (The reading of other alphabets will not be considered in this section.)

The reason for the disparity between the recognition of machine print and of handprint is easily described. Within a single machine-printed type font, samples of one character exhibit very little variability, while samples of different characters show large differences. So it is relatively easy to find a number of criteria for distinguishing between different characters. The situation is different in the case of handprint. Two samples of the same character, even by the same author trying for uniformity of style, show relatively large differences. So recognition schemes for handprint OCR must be able to allow for some latitude in the formation of each individual character, and at the same time be able to distinguish between different characters. Clearly this is not so easy to do.

The most highly developed OCR techniques deal with the simplest problem: recognizing members of a single, machine-printed font. The most common OCR hardware is designed to read one font (sometimes only one, otherwise one at a time from a small list of possibilities). There are even some special type fonts designed specifically for use with OCR equipment, such as OCR-A and OCR-B.

Considering the nature of the problems that have been (largely) solved, and those that remain, it is not surprising that most recent research in the field has dealt with reading more than a single type font (the so-called "multifont" or "omnifont" capability) and with recognizing handprint. (Cursive script reading is considered a special topic and will not be dealt with here.) Furthermore, in those cases where the nature of the work makes it meaningful to specify the type of text (such as the development of a complete OCR system), the majority of recent published work has dealt with handprint OCR.

Among those papers dealing with complete systems for recognizing machine print, the one by Al-Kibasi and Taylor [11] describes a device for omnifont applications (and handprint as well). Hård and Feuk [37] and Kooi and Lin [46] considered a single font, but they had a specific application in mind: a reading aid for the blind.

Albertsen, Munster and Ponsaing [8,9] and Cox, Blesser and Eden [22] described techniques applicable to multifont recognition. Several other articles on technique development and analysis restricted themselves to a single font, but again they had special purposes. The assessment of print quality was considered by Bohner [18, 19]. Hager [35] was concerned with speeding up a recognition algorithm without decreasing its accuracy. Patterson [61] and Troxel [87] wrote about specific techniques for portions of the problem.

Among descriptions of complete systems for handprint recognition, the paper by Schürmann [72] covered the character recognition component of a word recognition system. Other complete handprint OCR systems were reported on by Beun [15], Caskey and Coates [21], Fujimoto et al. [27], Hanaki, Temma and Yoshido [36], Huber [42], Krause, Schwerdtmann and Paul [49] and Yacyk [93].

A method for the assessment of handprint quality was given by Masuda [51]. In related work, but from a different point of view, the specification of a (machine) readable handprinting style was considered by Suen et al.[81] and Suen, Shinghal and Kwan [83]. Other papers on technique development and analysis for handprint recognition are by Ali and Pavlidis [10], Dasarathy and Kumar [23], Focht and Burger [26], Gaillat [30], Güdesen [34], Ichikawa and Yoshida [43], Kozlay [47], Kramer, Bergstrom and Ahlroth [48], Kwon and Lai [50], S. Mori et al.[52], T. Mori et al.[53], Ott [58], Parks et al. [60], Pavlidis and Ali [62], Powers [63], Sammon et al [69, 70), Shillman, Kuklinski and Blesser (74], Spanjersberg [76] and Suen and Shillman [82].

In the remaining parts of this section, a more detailed review of the recent literature on OCR software and research is provided. The intention is to convey an idea of the scope of work in the field, and to show the directions of current research interests.

B. CHARACTER RECOGNITION DEVICES AND TECHNIQUES

1. Introduction

There are four separate aspects to a practical character-recognition system, as pointed out by Harmon [38]. The first is presenting the text to the scanning device, which consists of some form of document handling (such as paper feeding). The second is scaning the text to convert the material into electrical signals that can be processed in order to interpret the content of the document. The third aspect consists of transforming the codes that directly represent the optical image into a form that can be operated on effectively by the decision logic. The final part is a decision-making process that applies certain criteria to the codes in order to effect a classification.

The first topic is only incidental to this discussion, and will not be covered here. Although a necessary adjunct, it is not fundamental to the process of recognition. Particular attention will be paid to the third and fourth aspects, the actual processing to extract information from the signal.

2. Scanners

Many types of devices which can be generally categorized as "scanners" have been used in character recognition systems and experiments. Hisdal et al. [40] used a commercial photometer, while Bohner et al. [18] built their own. Among all-electronic scanning systems, vidicon cameras were used by Huber [42], Beun [15], and Kooi and Lin [46].

A related device is the flying spot scanner, which was used by Fujita et al. [29] and Mori et al. [52]. Griffith [32, 33] described the use of a flying spot film scanner in the GRAFIX I OCR system built by Information International, Inc. A laser scanner was used by Fujimoto et al. [27]. Both types of device illuminate only a single small area of the document at one instant. A group of similar devices are those using photodiode and photocell arrays. They are solid state opto-electronic transducers that scan more than a single point at a time. Himmel [39], Schürmann [72] and Stringa [80] used photodiode matrices, while arrays of photocells were employed by Shurna et al. [75] and Caskey and

Coates [21]. Holt [41] described the Reticon RL-1872F, a page-width self-scanned photocell array. This device has 1872 photocells on 15 micron centers, with four parallel video output lines.

All of these devices have been used successfully. Huber [42] reported some difficulty in obtaining repeatable results, but the problem seemed to originate with the improvised mount rather than with the vidicon camera itself. Attention seems to be turning to the use of solid-state devices, particularly in commercial systems.

3. Encoding Signals

After the document to be read (or a portion of it) is scanned, the sequence of steps leading to the format operated on by the decision logic begins. In digital OCR systems each resolution element is normally represented by a single bit. There are a number of reasons for this, but the basic one is that a character is essentially a distribution of black on white (or one color on another, etc.). Consequently, it is generally agreed that having more gray levels does not increase the information content with respect to the basic problem of recognition. Degrees of blackness and whiteness, shading, and the like, while they may exist due to printing imperfections, paper quality and scanner errors, may be thought of as distortions. Because of such distortions, the threshold between digitized 0 and digitized 1 may need to be varied (perhaps as a function of position over a single document). Ullmann [89] discusses some techniques used for this purpose. The techniques reviewed by Ullmann are:

(1) Measurement of limb-width - The approximate limb-width (line thickness) of a character that has been binarized at an arbitrary threshold is measured. If it is found to be less than an ideal value, then the character can be binarized again at a threshold closer to white. If the limb-width is greater than the ideal value, then the character can be binarized again at a blacker threshold.

(2) Contrast determination - In these techniques it is usual first to process the video data so as to bring to a fixed predetermined level the analog signal that corresponds to background white. The binarizing threshold is then set at a level that depends on a global measurement of blackest black, or on the average of all gray-scale values blacker than a predetermined threshold.

(3) Fixed local contrast requirement - A pattern element is deemed black if it is more than a prescribed amount blacker than the average gray level of a set

of neighboring points (Laplacian binarizing technique), or if it is more than a prescribed amount blacker than at least one of a set of neighboring points.

Some of the research programs reported on considered single topics, such as classifying individual characters presented to some recognition software. But a practical system for reading text must first locate and isolate characters within the total field of view of the scanner. (Ordinarily individual characters are recognized, rather than groups. The main reason is that, for example, there are only 26 letters in the alphabet but thousands of words in the dictionary.) Some techniques for doing so are described in the literature. Albertsen, Münster and Ponsaing [8] developed techniques for detecting the locations of text lines and for figure (character) separation in an optical system. The method of character isolation used by Fujimoto et al. [27] used projection onto a horizontal line in which solid black projections separated by spaces give the horizontal extents of individual characters. Griffith [32, 33] described a commercial system in which the parameters specifying the text layout were programmed; however, they could be determined by a "layout analysis" step prior to the recognition pass. Huber's [42] system employed a routine that looked for closed-figures in one-half of a frame. Kooi and Lin [46] had several routines, among them a line isolation algorithm, and a character isolation algorithm that can handle variable-width characters, the separation of merged characters and the merging of split characters. Schürmann [72] used a hierarchy of different segmentation procedures. A line finder, a word separation routine and a character separation routine were listed in the description of the "data processing system" of Stringa [80].

The problem of character segmentation seems to be well in hand, at least when the characters are well separated. When characters are run together, the logic for segmentation becomes more difficult, possibly requiring the use of contextual information.

A number of preprocessing techniques are used in the OCR field, for what amounts to signal conditioning. One of the more commonly used is thinning (or "skeletonization" of) the lines making up the character until they are one sample wide, hopefully following the true contours of the character. Thinning was used by Beun [15], Fujimoto et al. [27], Gonzalez [31], Krause, Schwerdtmann and Paul [49], and Kwon and Lai [50]. Dasarathy and Kumar [23] discussed a technique for recognition only; however they assumed the characters were previously thinned.

Approximating the boundaries by well-behaved contours is a related technique. Pavlidis and Ali [62] (and [10] with the authors' names reversed) used polynomial approximations for this purpose. Sammon et al. [69, 70] applied smoothing to the boundary contours.

Normalization of character dimensions is another popular technique. This refers to adjusting one or both of the (horizontal and vertical) dimensions of all characters to standard values. A possible disadvantage of size normalization is that it might render an upper-case and a lower-case version of the same character indistinguishable. A way around this is to perform the vertical normalization using a scale factor determined from a character known to be upper-case. Size normalization was described by Focht and Burger [26], Huber [42], Krause, Schwerdtmann and Paul [49], and Sammon et al. [69, 70]. Schürmann [72] used this, along with normalizing varying stroke widths.

Several geometric transformations are applied. One of the simplest is centering, which normally means placing the center of gravity of the character pattern in the middle of the grid. Huber [42] and Schürmann [72] used centering. Focht and Burger [26] used "slant normalization", which combines shearing along the horizontal dimension with size normalization. The purpose of this is to compensate for different writing inclinations of different authors. Iijima, Genchi and Mori [44] described the use of geometric transformations, along with blurring to suppress noise, followed by a technique called "canonicalization" used to overcome the detrimental effects of blurring (more or less, the suppression of important detail).

The thickening of character lines to at least two samples wide was employed by Yoshida et al. [96], along with smoothing. Noise suppression techniques were used in their optical systems by Albertsen, Münster and Ponsaing [8], and Ozawa and Tanaka [59]. In his preprocessing, Beun [15] determined "special points" - end points and fork points - for later use. Approximation of contours or thinned lines by line segments with quantized directions was employed by Fujimoto et al. [27] and Sammon et al. [69, 70].

Güdesen [34] performed experiments to assess the effectiveness and computational effort of several preprocessing techniques. It is not clear to what extent his results depended on the particular choice of features used. The article by Shurna, Lashas and Gvildys [75] was on the implementation of several preprocessing techniques using optico-electric filters.

In practice, the choice of preprocessing techniques must necessarily be related to the recognition method. For example, template matching requires size-normalized (possibly

also slant-compensated) characters, while some other recognition techniques do not. A very simple recognition technique may require extensive preprocessing to obtain good accuracy; it is the total computational effort that must be considered in comparing techniques.

The decision-making process in general can be viewed as separating ensembles of points in a multidimensional space, each ensemble representing a class. Each member of a class is represented by a point, or vector, in this abstract space. An unknown sample is assigned to a class by determining in what part of the space the vector representing the sample lies. From this point of view it is clear that the single most important aspect of the design of a recognition system is the choice of a good set of features, the coordinates (axis labels) in this abstract space. A good set of features gives a transformed (in general) representation of the original measurement such that all members of a class have a common set of properties in feature space, distinguishable from other classes. No amount of ingenuity in the recognition logic can fully compensate for a poor choice of features.

The term "features" is commonly used to refer to two concepts: the distinguishing properties or characteristics of the patterns, and the numerical values of (or assigned to) these characteristics. The collection of features (numerical values) describing a pattern are often thought of as making up a feature vector. This gives rise to a geometric interpretation of pattern recognition, which was alluded to above.

"Feature selection" is a term that is used in two ways. In one sense it refers to an aspect of the design of a pattern recognition system, the choice of the features to be used. This (along with a distance measure) defines the space in which the decision-making mathematics will operate. The other (and more common) use of the term refers to a process of choosing a reduced subset of the totality of all features considered, to use in classification. This may be done, for instance, to reduce computational complexity and consequently increase the throughput rate.

Another similar-sounding term is "feature extraction." "Extraction" means drawing out, obtaining. Feature extraction is the process of obtaining the numerical values of features. It is the actual determination of the numbers, representing pattern characteristics, to be operated on by the decision logic.

The simplest features are the pattern points themselves. The successive binary values in a digitized raster scan, for example, may be thought of as making up the components of a feature vector. This approach was taken by Al-Kibasi and Taylor [11], Iijima, Genchi and Mori [44] (after extensive preprocessing), and Schürmann [72].

Typically, however, other quantities are chosen as features. The measurement space is often of high dimensionality, and it is frequently possible to transform to a feature space of lower dimension. This may lead to a major reduction in the amount of computation required for classification. There are other reasons for using as features something other than the original measurements, but the basic idea is to find a set of characteristics containing the essential information about the patterns in a convenient way.

One group of features frequently used in character recognition are topological features, loosely those that are related to the geometrical properties of patterns. Typically, they describe the strokes making up the characters. Focht and Burger [26] used correlations with sequences of masks for the strokes giving the basic outlines of reference patterns; this is a modified version of template matching. Gonzalez [31] investigated the use of topological features in great detail. Features of this type were also used by Nadler [54], Parks et al. [60] and Yoshida et al. [96].

Something roughly similar was used by Kozlay [47], who developed a hardware implementation of a combined feature selection - feature extraction system. In this system features are evaluated by looking at the sample character through a number of windows. This technique is sometimes known as peephole matching.

Characteristics of the external contour of the character, as seen from the sides, are also used as features. These features were used by Sammon et al. [69, 70] (in the latter paper, augmented by a few other shape characteristics) and by Spanjersberg [76].

Another category of features involves crossing counts: the numbers of times various vectors intersect the character. Kwon and Lai [50] used crossings as features, combined in various ways. Crossings were also employed as features by Spanjersberg [76] (who tested three classification systems).

Spanjerberg [76] also discussed a system using geometrical moments of character patterns as features. Geometrical moments were used by Tucker and Evans [88], as well. Wendling and Stamon [92] tested features involving certain characteristics of Hadamard and Haar transforms.

Two classes of feature extraction techniques go by the name "field". Fujimura and Tazaki [28] treated the sample points making up a character as particles interacting by means of a certain force function. In the equilibrium state of this pseudodynamical system all of the points collect in a few places, corresponding to end points, cross points and bending points of the original character. These are taken as features. A different field effect method was used by Mori et al. [52, 53] (the two papers have exactly the same authors, with their names listed in different orders). A field is associated with the pattern, and is used to extract concavities and enclosures by a specified operational process. An edge detection method was used as an adjunct.

The previous few paragraphs have reviewed papers using single types of features (or at least one at a time, as in the case of the paper by Spanjerberg). The use of several groups of features in one system is also quite common. Beun [15] employed the numbers of end points and fork points, with various other features added in an ad hoc manner. Caskey and Coates [21] made use of cavity, loop and spur information, perimeter data and other specialized measurements. The technique described by Dasarathy and Kumar [23] involved five separate groups of features. Fujimoto et al. [27] used as a basic feature set numerical codes specifying the direction of straight line segments approximating the contour, with 1 or 2 (different) local features also involved in each decision. Huber's [42] system employed information on both edges and character topology. Numerous features, relating to holes, areas, body lengths, loops, lines, concavities and angles were used by Pavlidis and Ali [62].

Most of the above features were specified by intuition or in some manner by the samples themselves (field effect, peephole matching). A different line of approach is discussed in a sequence of papers by Blesser et al. [16], Blesser, Kuklinski and Shillman [17], Shillman, Kuklinski and Blesser [74], and Suen and Shillman [82]. For recognizing unconstrained handprinted characters, they argue that approaches (such as template matching) that do not parallel human perceptual behavior should be rejected. They feel that feature detection is a better approach, and that the features should have some psychological significance. Rather than basing features on "archetypical" characters and treating ambiguous characters as "difficult cases," features should be based on properties that help separate the latter. Tests for studying and developing these concepts are described; some results are presented.

An approach for choosing a good set of features was described by Troxel [87]. This method is feature selection in both senses of the term. The procedure is largely automated. The article describes the use of the procedure for developing recognition systems for several machine-printed fonts in a relatively short time, achieving high classification accuracy.

This brings us to the topic of feature selection in the sense of choosing a subset of a larger collection of features, to use in the recognition logic. Ahlgren, Ryan and Swonger [7] developed a technique for selecting a particular "good" set of features from a much larger set of candidates, independent of the recognition method. Starting with a specified feature set, the problems associated with reducing this set to the minimum number of features were discussed by Gonzalez [31], who proposed an algorithm to perform this minimization. As referred to in the previous paragraph, Troxel [87] also considered this problem.

In the system described by Gaillat [30], part of the operation of the training phase of the classifier chooses a good feature subset to use. Hager [35] employed a least-mean-square procedure to select a group of polynomial terms to use in performing classification, effectively reducing the number of features. Sammon et al. [69, 70] used about half the total number of features in performing each recognition test.

In some recognition systems the logic is organized so that a different subset of features is employed in each test for class membership. In the majority of cases a sample can be classified without ever extracting all the features. This is true of the character classifiers presented by Beun [15], Caskey and Coates [21], Dasarathy and Kumar [23], Hanaki, Temma and Yoshido [36], Kozlay [47], Kwon and Lai [50], Pavlidis and Ali [62], and Sethi and Chatterjee [73]. Focht and Burger [26], whose method is to perform correlations of the unknown with masks representing individual strokes, used a different sequence of masks for each character definition.

In most of the material reviewed, there is not much detail on feature extraction. It is simply stated (or implied) that the values of the features are obtained by software routines. However, a few methods specifically developed for feature extraction (possibly extracting new types of features) were documented. There were field methods (Fujimura and Tazaki [28], S. Mori et al. [52], T. Mori et al. [53]), the use of principal component analysis (Ott [58]), and the reflection method (Yoshida et al. [96]). Feature extraction hardware was designed by Kozlay [47].

Quite a few different types of features have been used in OCR research, and a variety of them have given moderately good accuracy (80% - 95%) on relatively unconstrained handprint. Several authors have reported that their results were improved when the writers attempted to follow a few rules in forming their characters. (As was mentioned previously, special handprinting styles for OCR have been proposed.) In other words, the more nearly human-printed characters resemble machine-printed characters, the easier they are to recognize.

As has already been pointed out, nothing can compensate fully for a poor choice of features. Adding more features should increase the information presented to the recognition logic, but unless they are well-chosen they may increase the amount of computation required while adding very little to performance. (Furthermore, every increase in computation increases the probability of serious roundoff error.) A challenge facing the OCR field today is the reduction of the substitution error rate for unconstrained handprint from a few percent to a smaller value, preferably by changing most of these substitutions to correct identifications. This may be largely a problem in the choice of features.

Three trends can be discerned in this area. One of them is the use of feature selection. Starting with a pool of features, some ranking or evaluation technique can be used to determine which of them provide significant information to the tests used in recognition. Of course, the original pool must be sufficiently exhaustive in its information content. In a sense, feature extraction lets the data choose the features. Other data-directed techniques include the field-effect methods and those extracting sequences of stroke patterns, for example. The third approach is that based on human perceptual behavior.

Beyond the choice of a good feature set, some improvement in recognition accuracy may be obtained by increasing the sophistication of recognition methods. It is noted that human recognition of words is more accurate than recognition of isolated handprinted characters. The use of context, where possible, may increase the accuracy of recognition of individual characters.

4. Decision Logic

The final aspect of a character recognition system is classification, or identification. By applying a sequence of logic, a decision rule, to the feature vector (or other

representation of the unknown sample), the sample is assigned to a category or class, or it is determined that the sample's class membership is uncertain. (With some classifier formulations this sequence of operations may not appear to be taking place, but, conceptually at least, this is generally how character recognition systems work.)

OCR classifiers are generally supervised; supervision refers to the degree of knowledge about the data. A supervised classifier is one for which the categories (in this case, numerals, letters, punctuation marks, etc.) are known, as well as some information about how members of each look to the classifier. So at some point this information must be provided to the system. Sometimes such information is implicit in the design of the recognition logic, for example when a template is provided for each character in a machine that recognizes machine print in a fixed font. In other cases some or all of this information is provided to the classifier in a training phase prior to classification, possibly by providing the system samples of each class (labeled samples). Although the operational procedures are quite different in the two instances, in both the classifier may be thought of as a "student" that is "taught" how to recognize characters.

Classifiers may be parametric or nonparametric. If the forms of the multivariate probability distributions for each class are known (or assumed), the training phase is used to determine the values of the parameters (e.g. means, variances etc.) of those distributions. This is called "parametric learning." Once the probability distributions are determined in this way, they can be used to compute the probability that an unknown sample belongs to one class or another. Those classifiers for which the functional forms of the distributions are unknown (or not used) are called "nonparametric," even though some parameters may be involved. Most OCR classifiers are nonparametric.

There is considerable variation in the details of the operation of classifiers, though similarities appear when the methods are looked at in terms of the feature space representation. Some classifiers operate on the basis of discriminant functions $g_i(\underline{x})$, one for each class i. If $g_k(\underline{x}) > g_j(\underline{x})$ for all j not equal to k, where $\underline{x}$ is a feature vector, then the sample represented by $\underline{x}$ is assumed to belong to class k. (For parametric classifiers, the g_i can be probability densities.) By considering the geometric interpretation of this process in feature space, one is led to the concept of decision boundaries, surfaces in feature space separating the region occupied by one class from that of another (or all others). There may be a boundary for each pair of classes (particularly if the

functional forms of the boundary surfaces are simple), giving N (N-1)/2 boundaries, where N is the number of classes. Or successive dichotomy may be used, successively splitting the classes remaining into two groups. Then N-1 functions are needed. (More complicated functions may be required.) The dichotomy may be performed in any way. In particular, at each decision node, or some of them, one class may be separated from all the remaining ones.

In many cases, the distinction between discriminant functions and decision boundaries may be more formal than actual. If feature selection is performed, we may think of projecting the patterns onto the lower-dimensional space of the reduced feature subset. If the orientation of a decision surface is parallel to a feature axis, for instance, that feature plays no role in the decision and might as well be eliminated. The extension to more than one dimension follows. This is the geometrical interpretation of feature selection. (Analogies with two- and three-dimensional space are of limited usefulness; it is hard to visualize a higher-dimensional space.) It may be reasonable to use a different feature subset for each decision; some systems are designed this way.

Successive dichotomy leads to a tree structure for decision logic. Particularly when the features are binary or other discrete-valued variables, it may be convenient to make a decision simply based on the value of a feature (or a small number of them), rather than calculating a function. Logic trees are used frequently in OCR systems.

Sometimes a system of classification is used whereby one set of criteria is used to separate all the input samples into several groups, and another set of tests (possibly using different features) further breaks down the groups into individual classes. While this procedure has strong similarity with successive dichotomy, some OCR systems are explicitly designed as two-stage classifiers. For example, the first stage may determine to which of several initial subclasses the pattern belongs. With certain patterns it may be possible to make a decision at that stage, while with others it may be necessary to examine some additional features to resolve ambiguities.

An alternative to the feature extractor-categorizer approach to classification is the syntactic or linguistic approach. This type of approach makes use of a priori knowledge about the relationship between parts of a pattern. The pattern is considered to be a sentence in a language generated by a given grammar. Using the grammar, the sentence is analyzed to determine what class it belongs to. It is difficult to go further in

this explanation without going into the structure of formal language theory. A possible advantage of syntactic methods is that no labeled samples are required for "training." However, detecting the primitives (basic elements of the language) in the presence of noise is a serious problem.

Some hardware implementations of classification techniques were reported in the literature. Linear discriminants realized by arrays of resistors were described by Al-Kibasi and Taylor [11] and Huber [42]. Fujimoto et al. [27] implemented "nonlinear elastic matching," which appears to be a type of prototype classification. Prototype classification measures the degree of distortion of a sample from an idealized example of each class as a distance, and assigns the sample to that class to whose prototype it is closest. Schürmann [72] described a system using quadratic discriminants, built for the Federal German Post Office.

Logic trees were a popular technique for handprint recognition. Their use was reported by Beun [15], Caskey and Coates [21], Dasarathy and Kumar [23], Kozlay [47], Kwon and Lai [50], Pavlidis and Ali (in one case, with a second stage using linear decision boundaries) [62], Sethi and Chatterjee [73]. Hanaki, Temma and Yoshido [36] developed an interactive tree building system, for easy design and modification of trees, and used it to develop systems for recognizing three alphabets.

Gonzalez [31] developed a formulation in which a large class of nonlinear discriminants could be treated as linear discriminants in a transformed space. His classifier is a two-stage system involving nonlinear discriminants in the first stage and special tests to resolve some ambiguities in the second. Nonlinear polynomial discriminant classifiers were described by Hager [35] and Ott [58]. Sammon et al. [69, 70] used two-stage logic in which the first stage makes a partial classification, and the second finishes the job using a different feature set with linear decision boundaries.

Spanjersberg [76] worked with three systems. Two of them employ linear decision boundaries. The third classifies a sample into the class for which the probability of membership is greatest, based on empirical statistics derived during the training phase. Tucker and Evans [88] described a classifier using a (parametric) normal probability decision rule, which leads to quadratic discriminants.

Some authors classified a sample into the class for which the normalized (by the class variance) distances from the empirical class distribution centroids is minimum.

Since features are extracted in a variety of ways and represent different things, it is essential to use care in properly weighting or normalizing the features when computing distances. Kooi and Lin [46] and Patterson [61] used this approach.

Ali and Pavlidis [10] used a two-level method of syntactic classification. Syntactic methods were also used by Powers [63] and Rajasekaran and Deekshatulu [64].

A technique for the design of classifiers when the measurement variables are discrete was given by Stoffel [79], who used a character recognition example. Swain and Hauska [84] presented the basic concepts of decision tree classifiers, two methods for designing decision trees and a discussion of the advantages and disadvantages of the two methods.

C. SPECIAL TOPICS IN OCR

The research efforts reviewed so far have concerned the development of techniques and devices for the automated recognition of machine-produced characters and handprint. When one considers the possibility of cursive script recognition it is immediately obvious that the problems are of a different order entirely. Not only does the idea of font have less meaning (if any) than for handprint, but the problem of segmenting individual characters written with a continuous line must be dealt with.

Nevertheless, some work has been done in this field. It is of interest at least because it represents a frontier in the OCR field. The approaches to the problem may be the best source of new ideas in character recognition. Work on the recognition of cursive script was reported by Ehrich [25], Hisdal et al. [40], Nagel and Rosenfeld [55], and Sayre [71].

Substantial work has been performed in development of OCR techniques for foreign alphabets, particularly Chinese. Since a number of the problems are different (varying stroke width may be significant, for example) this work is reported separately. Research in foreign alphabet recognition was reported by Fujimoto et al. [27], Fujita, Nakanishi and Miyata [29], Hanaki, Temma and Yoshido [36], Nakano et al. [56], Sakai et al. [68], Sethi and Chatterjee [73], Stallings [78], Tanaka and Ozawa [86] (the method is that described in Ozawa and Tanaka [59]), Wang [90], Wang and Shiau [91], Yamamoto, Nakajima and Nakata [94], Yoshida and Eden [95], and Yoshida et al. [96].

Doster [24] described a postprocessing system making use of context for word recognition (a "word" may be a logical grouping of numerals). The system is part of an

automatic postal address reading machine developed for the Federal German Postal Service. Candidate words are compared with entries in a dictionary containing 16,000 different postal place names with ZIP code and postal district numbers. The individual character recognition portion of the system was described by Schürmann [72]. Other work relating to ZIP code readers was reported by Albertsen, Münster and Ponsaing [8, 9] and Focht and Burger [26]. Harmon [38] points out that fully automatic mail processors have been in use in Japan since before 1970, reading handprinted ZIP codes.

Most of the research described so far has involved digital techniques. A parallel analog system was described by Tanaka [85]. All-optical techniques, coherent and incoherent, were used by Albertsen, Münster and Ponsaing [8, 9], Armitage and Lohmann [13], Brown and Lohmann [20], and Hård and Feuk [37]. Not very much interest has been shown in optical techniques. Optical processors, while elegant and well-suited to certain tasks, are often regarded as inflexible, require a skilled operator, and may involve expensive (and difficult to obtain) optical components.

The character recognition system of Hård and Feuk [37] was intended as a component in a system to enable the blind to read printed text. Kooi and Lin [46] developed a system for reading printed text and converting it to speech, for this purpose. There is a commercial system, the Kurzweil Reading Machine (described in another part of this report), that accomplishes this task in real time.

D. RELATED TOPICS

It was mentioned earlier in this section that techniques have been developed for assessing handprint quality, and machine-readable handprinting styles have been proposed. Also, Bohner [18] described a technique and Bohner, Sties and Bers [19] designed a measurement device, for the evaluation of the quality of (machine) printed characters. Apsey [12] conducted tests on the ability (and cooperation) of people to follow certain rules in the formation of handprinted characters. His conclusion was that severe constraints are impractical.

Spanjersberg [77] reported on several experiments in the use of OCR for input of handwritten numerical data into the "Postal Giro Service," a public institution that carries out payment orders received from its account holders. Based on these experiments a reading machine was developed and installed as a device for the automatic verification of payment orders. Over 200,000 account holders now use the special cards designed for this

system. The machine operates at a rate of 10,000 cards per hour, with an individual character accuracy rate of 94%. The fraction of payment orders read correctly is 60%. The substitution error rate is 1 in 10,000 payment orders. This is lower than the rate achieved by a manual procedure in which the cards are handled by two different human operators.

Character recognition by man and by machine were compared by Niemann [57]. He found out that it depends highly on the recognition task whether humans or the (near-optimal) machine performed better. His conclusion was that, in many cases, significant improvements of the error rate of existing recognition systems are possible. Niemann proposed that as a challenge for further research.

V. SUMMARY

A broad overview of recent developments and current capabilities in OCR technology has been presented in this report. This summary is intended to provide an assessment of the applicability of the techniques and systems reviewed to the DMA OCR applications.

As indicated throughout the report, it is understood, by the authors, that the DMA OCR applications can broadly be characterized by the unconstrained nature of the data to be read, both in character format and character position. For applications exemplified by the oceanographic smooth sheet problem under study by NORDA, the historical nature of the data allows no relief in either character format or position. Where present or future data sources are concerned, standards may be imposed on character formation, for example the ANSI standard referred to in Section III-E. However, character position and/or orientation will continue to present a problem for the majority of map-related applications.

With only two exceptions, the commercial OCR systems reviewed were designed explicitly for use in applications where the input data are presented in a tightly controlled format on standard forms. The two exceptions are the GRAFIX I image processing system and the Kurzweil Data Entry Machine. In each of these cases, greater freedom is allowed in data entry through the capability to read printed pages of a narrative style. However, even here, the format of regularly spaced lines of characters provides the necessary constraint to achieve effective scanning.

Once character acquisition and isolation have been achieved, the recognition techniques employed on many of the commercially available systems are effective in providing high reliability recognition. Many of the systems utilize feature analysis methods which have and are being researched for the NORDA project. Thus, in this third stage of the overall recognition process, current technology supports the techniques being developed for DMA applications. However, it is concluded that for the first two stages - character acquisition and isolation - a scanning and data handling system tailored to the specific needs of DMA applications is required. This is not readily available off-the-shelf.

The survey of government applications of OCR reveals that the great majority of current requirements are for OCR systems that will address problems in the business or bureaucratic areas rather than in scientific types of problems such as those posed by DMA. It is also clear that the major vendors of commercial OCR systems are able to meet that need, for example, Recognition Equipment Inc., IBM and Scan-Data Corporation.

Government-sponsored research and development into handwritten character recognition has had mixed success. Several of the applications studied have had constrained data characteristics that fall within the capabilities of present-day commercial systems. However, two projects, both dealing with varied handprint material, concluded that specialized systems are required for such problems.

In general, the results of the government OCR applications survey support the conclusion that the characteristics of the DMA problems dictate appropriate system development both in hardware and software as opposed to direct adaptation of earlier work on existing systems.

The survey of recent research activities in OCR indicates that, at the present stage of development, acceptable recognition accuracy can be obtained on unconstrained font handprinted characters. In order to meet the requirement of reading text unconstrained both in format and font, overall systems design and planning coupled with technique development are necessary. Most of the individual elements are available, but apparently nobody has addressed the complete problem.

Almost any method of scanning that possesses sufficient resolution should be usable. Character location and isolation is an area of work requiring further development. Little information was found on the location of randomly placed characters, as opposed to lines of text or placing characters in prespecified locations. Character segmentation (isolation) may need special attention since the spacing is arbitrary, including cases where characters (possibly from different groups) may run together.

The variety of recognition methods applied to the problem of recognizing handprint evidences a realization that the techniques used successfully for fixed-font machine print recognition are inadequate. Though more work needs to be done in this area, some of

the papers reviewed in Appendix A give an idea of the success that has been attained by newer methods. Attention is also drawn to the discussion in Section IV-B-3 of modern trends in choosing features. It is generally agreed that relatively simple decisions applied to well-chosen features should give better results than complex and powerful statistical analysis techniques working on features that may not possess sufficient information content. Feature selection techniques, of course, may be applied in conjunction with any methods of choosing and extracting features. The contention that more work needs to be done in recognition technique development is supported by Niemann's [57] optimistic suggestion that "significant improvements of the error rates of existing recognition systems are possible" in order for machines to equal the performance of the human visual system.

This report has presented a panoramic view - admittedly incomplete, but omitting no important areas - of the current status of the field of optical character recognition. Significant accomplishments have been achieved, and fruitful new work is still being performed. The gap between human and machine recognition performance is still wide, but challenging problems like those posed by the DMA requirements help point the way toward closing the gap.

APPENDIX A.

REVIEWS OF SELECTED PAPERS

1. R. C. Ahlgren, H. F. Ryan & C. W. Swonger, "A Character Recognition Application of an Iterative Procedure for Feature Selection," IEEE Transactions on Computers C-20, 1067-1075 (1971).

The authors developed a technique for feature selection, independent of the recognition method. Some experimental results are given for a character recognition application, using a standard data base of machine-printed characters. They did not develop a complete OCR system.

The authors were with the Computer Research Dept., Cornell Aeronautical Laboratory, Inc., Buffalo, N.Y. The work described in the paper was sponsored by the Bureau of Research and Engineering, U.S. Post Office Dept., Washington, D.C. (Contract RE101-68).

The method developed chooses a particular "good" set of features from a much larger set of candidate features. No particular recognition technique is required. The feature selection technique can be used with any classification logic. The concept of a Bayes' classifier* was used in assessment of feature sets. A linear discriminant classifier was used in tests. The method applies in principle to any feature set. The features used in the tests were pattern points (pixels) themselves in size-normalized 24 x 24 characters.

There are two feature selection procedures - a search method and an evaluation method. Both make use of a performance index, the Shannon information content measure.** The search method generates a sequence of candidate sets of features, of the desired size, from the large pool. These sets are the best obtainable under the circumstances using a suboptimal selection method. (The only known method of finding the optimum subset is an exhaustive search through all subsets of the given size, which is not computationally feasi-

---

*A Bayes' classifier uses a statistical method of deciding on class assignment, in which a likelihood ratio (ratio of probability densities) is compared with a threshold that involves the costs of misassignments.

**The Shannon information content measure is a statistical quantity that measures the ability of the features to separate the pattern classes.

ble.) Each set is then evaluated by computing its performance index. An estimate of the performance of a Bayesian classifier and the actual performance of a linear classifier were also used.

The training procedure for the recognition algorithm used in the test was not specified. No hardware requirements were specified either, although it was stated that training was done on a special-purpose computer.

The feature selection procedure was applied to an eight-class problem. 19,000 samples of eight alphanumeric characters (B, E, 3, L, I, J, 2, Z) were chosen from a data base of 100,000 typed samples. The 8-class discrimination problem was formulated as a tree of 11 two-class problems. Eleven sets of 25 pattern points (features) were first chosen by the search procedure. A typical running time for this for one class pair (two-class problem) using 5000 training patterns was quoted as 15 min on an IBM 360/65, using 535K bytes. Next, 11 sets of ten points each were obtained using the first sets as the pool. Evaluations were made of the two-class problems. The linear classifier gave the following results:

Performance Rates in Percent

	110 Features			275 Features		
Category	Lower Bound	Estimate	Upper Bound	Lower Bound	Estimate	Upper Bound
Correct	93.45	93.94	94.07	95.06	95.56	95.83
Rejected	4.27	4.53	4.97	3.35	3.57	4.04
Errors	1.37	1.53	1.84	0.75	0.87	1.08

The author speculates that the results are not better because of the inadequacy of the linear classifier, rather than the performance of the feature extraction process.

The algorithms should be available, since the work was done under U.S. Government contract.

2. Belur V. Dasarathy and K. P. Bharath Kumar, "CHITRA: Cognitive Handprinted Input Trained Recursively Analyzing System for Recognition of Alpha-Numeric Characters," to be published, International Journal of Computer and Information Sciences 7 No. 2 (1978).

The authors developed an algorithm for the recognition of handprinted characters in the form of a decision tree with two major segments, one for numerals and the other for upper-case alphabetic characters. The concept is capable of wider application. Scanning and pre-processing were not included.

The work described in the paper was conducted at the School of Automation, Indian Institute of Science, Bangalore, India. Both authors are now in the United States. B. V. Dasarathy is with M&S Computing, Inc., Huntsville, Alabama. K. P. B. Kumar is with the Department of Electrical Engineering, University of Hawaii, Honolulu, Hawaii.

The algorithm was designed for the recognition of handprinted characters in general, without restricting to a particular application. A design goal was that it should be usable in practice, by being simple enough to be implemented on a modest computer and fast enough to be useful. The algorithm employs a decision tree for which only one or a few features are extracted for the test at each node.

The result of one test determines which branch is followed, and hence which test is to be performed next; or it may result in a class assignment at that point. Each input sample is taken on a path through the tree for which only a fraction (sometimes only one) of the total number of tests must be performed, and for which it is seldom necessary to extract all the features.

There are five categories of features, designed to be relatively insensitive to minor changes in the formation of a character, while responding to the qualitative differences between different characters. The five types of features are:

(1) Features relating to the density of the character.

(2) Abastract features (based on vector crossings).

(3) Features relating to the external contour.

(4) Features based on topology.

(5) Morphological features.

The characters are assumed to be thinned prior to processing; some of the feature definitions rely on this. Characters of a fixed size were used in the tests. However,

none of the features are size-dependent, so size normalization is not required. Breaks in the lines making up the characters may lead to unpredictable results.

Each type of features describes certain characteristics of the characters. The idea is to combine the sets, so as to give a complete description, and at the same time, only use the features that are actually necessary at any one point. As an example, consider the use of one of the features, the presence or absence of a sharp protrusion on the right side of the character. This feature separates the numeral 4 from 1, 7 and 9, but it is of no help in discriminating between 0, 6 and 8. Feature selection was accomplished in the design of the decision tree; the paper presents a complete tree. The tree designed by the authors was designed by a manual, trial-and-error procedure.

The design of the tree constituted training. Samples of characters produced by 50 individuals were used. The number of sets written by each individual was not given. However, it is assumed that the number was not very large, since the data had to be digitized manually.

The authors used a software implementation of the processing, including feature extraction and the decision logic. The computer used was not specified. In an appendix, they considered the feasibility of hardware implementation of the feature extractors, and presented block diagrams for two sets of features. The method should not require a large-scale computer system.

The algorithm was tested against the data base used in the design of the tree. An accuracy of 100 percent was obtained with numeric characters, and 99.3 percent with the alphabetic set (the errors were not broken down into substitutions and rejections). The total throughput, including card reader input and line printer output (apparently on-line) was stated as 3-4 characters per second. (As was pointed out above, the computer used was not identified.) It can be concluded that the feature extraction-recognition logic itself was much faster than this, and could probably be made faster still by careful programming.

The article presents an already-designed logic tree for the recognition of handprinted alphanumeric characters. But, more significant that this, it formulates an approach to

the problem. That approach is the combination of feature selection and sequential testing in a tree structure, using a large feature set composed of several categories of features. The tests of the algorithm were not adequate, since the design data set was used as the test data set. Also, the data used did not contain any scanner noise, artifacts induced by preprocessing, etc. In general, though, the approach appears to have merit.

3. Y. Fujimoto et al, "Recognition of Handprinted Characters by Nonlinear Elastic Matching", Proceedings of the Third International Joint Conference on Pattern Recognition, IEEE Computer Society, November 8-11, 1976.

This paper describes a complete system to recognize handprinted characters. It was primarily designed for reading computer coding forms. The system also includes a context checker to detect grammatical errors in FORTRAN programs.

The authors are with the Central Research Laboratory, Hitachi Ltd., Kokubunji, Tokyo, Japan.

The system was designed to recognize handprinted members of the FORTRAN character set (upper case letters, numerals, arithmetic operators and other special symbols) or other sets of about 50 characters. Basically, the method is a type of template matching that is relatively insensitive to small local expansion and contraction of line segments. The concept is based on the analysis of shape variations.

Individual characters are isolated by the use of pattern projection onto a horizontal line. Next, the character lines are thinned to a one-bit-thick skeleton. Then the thinned lines are approximated by straight line segments of quantized direction, which are represented by the Freeman code. The Freeman code uses the numerals 1-8 to represent short straight line segments at increments of 45° in direction.

The next step in the processing of individual samples is "topological classification", determination of the number of branches, number of loops and number of components (sets of the branches connected). A group of candidate standard character patterns is picked from the same topological class as the input pattern by "branch correspondence". The standard patterns, or templates, include a variety of ways of forming each character. The input pattern is compared with each of the candidate patterns using a similarity measure that is invariant to small changes in length of component lines, and that can measure the

differences in line segment directions as a distance. This step is called "nonlinear matching". The character is recognized as the one with the smallest or next-smallest similarity measure, or not recognized, depending on the relative sizes of the measures. (Since similarity is calculated as a distance, smaller values denote greater similarity.) If both possibilities give small and not sufficiently distinct similarity measures, the system attempts to make a decision using local features.

The system is "trained" by the definition of the standard patterns. The details of this were not specified. Templates are distinguished by morphological variations. Changes in proportion (i.e., line lengths) are accommodated by the similarity measure.

The system consists of an H-8959 OCR system (laser scanner, paper-handling mechanism, control and recognition unit, memory unit) with additional nonlinear matching hardware and an extended memory unit, keyboard display, magnetic tape unit, card punch and typewriter. The H-8959 is apparently manufactured by Hitachi Ltd.

The authors reported a 0.06 percent error (substitution) rate and a 0.20 percent rejection rate with a sampling of 26,400 characters by six trained writers. No figures were given on recognition speed.

The system is now in use at the Hitachi central laboratory, as a computer input device going directly from coding sheets, bypassing the keypunch step. Dr. Fujimoto gave a negative response to a request for more information about one aspect of the system, citing failure of the company in obtaining a patent for the process in question. It would appear, therefore, that the algorithms themselves may not be available. The complete system itself presumably is.

4. William A. Huber, "Handprint Reader", Research and Development Technical Report ECOM-4087, U.S. Army Electronics Command, March 1973.

A complete system for the machine recognition of relatively unconstrained (in font) handprint was developed and tested.

The author is with the U.S. Army Electronics Command, Fort Monmouth, N.J. 07703. The work described in the report was sponsored by the U.S. Army Electronics Command, AMSEL-NL-A-1, Fort Monmouth, N.J. 07703.

The prototype system developed was adapted to read capital letters, numerals and 10 special symbols written in blocks on computer coding sheets. The author proposed some military applications for handprint readers using the concepts developed in this work. Mathematically, the handprint reader appears to be a feature extractor-classifier system.

The feature set used is not specified clearly. The features describe certain information about edges and character topology. The latter category refers to spurs, concavities and enclosures.

The data stream from the scanner (a vidicon camera was used in the prototype system) is sampled and digitized, giving a 120 x 120 array. The character is then centered and size-normalized in a 24 x 24 bit array. The details of determining that a character in fact exists within the field were not specified.

After scanning and preprocessing, the features are extracted and expressed in a 100-bit word, which is effectively a feature vector. The feature vector is then processed by "statistically weighted networks". These are resistor networks with 100 input terminals, one per bit. The bit values (logic levels) and the resistances determine the currents flowing in each branch. The 100 branches are summed to give the network output. There are 46 networks, one for each class. The network producing the greatest output specifies the class to which the sample is assigned.

The machine is (normally) trained for each author's handwriting. The networks consist of variable resistors whose values are automatically adjusted under computer control (algorithm not specified). The procedure operated by adjusting all network resistor values iteratively until correct classifications of the entire training set, or as much of it as possible, are obtained.

A vidicon camera system was used for scanning. The remainder of the system was "an adaptive learning machine operating under computer control and utilizing software preprocessing". No details were given, except for a brief description of the weighted resistor network. The author cites references that describe the hardware.

Five authors each handprinted 40 "alphabets" consisting of the 46 characters, under minimum restrictions. Thirty of the 40 alphabets produced by each were used for training; the remaining 10 were used as test data sets. The overall results:

Set	Classification Accuracy
Training	95.6 ± 0.67%
Test	91.4 ± 0.69%

No rejections were allowed.

As a result of this test, certain handprinting characteristics were determined whose correction, it was felt, should improve recognition. Five more authors were asked to provide another group of handprinted characters of the same size as the first, with additional constraints. The data were used in the same way. The results were:

Set	Classification Accuracy
Training	98.9 ± 0.57%
Test	98.0 ± 0.75%

The author considered a rejection strategy employing a threshold that must be exceeded in order to cause assignment to a class. He found that (for the second group of data) a threshold could be specified that would reduce the substitution rate to zero, at the expense of rejecting 2 percent of the characters that were correctly read initially.

The recognition logic could be implemented in software, if desired. The operation of the resistor networks appears to be mathematically equivalent to forming the scalar product between the 100-bit feature vector word and each of 46 weighting vectors, and choosing the largest result. *This describes classification using linear discriminant* functions.

The author pointed out that the performance of the vidicon scanning system was unsatisfactory. The problem was in the camera mount and associated positional control, rather than in the electronic system itself. However, the author recommends using a solid state scanner with appropriate paper handling mechanism.

This work was performed by the U.S. Army, so is available to the Government.

5. John W. Sammon, Jr., et al, "Handprinted Character Recognition Techniques," RADC-TR-70-206, Rome Air Development Center, October 1970.

This report describes research into logic for machine recognition of handprinted alphanumeric characters. The authors did not develop a complete OCR system.

The authors were with Computer Symbolic Inc., 310 E. Chestnut Street, Rome, NY 13440. The work described in the paper was sponsored by Rome Air Development Center (EMBDP), Griffiss Air Force Base, NY 13440 (Contract F30602-69-C-0374).

The report describes the design of logic for the automatic machine recognition of relatively unconstrained (in font) handprinted alphanumeric characters. The recognition technique developed employed a two-stage nonparametric feature extraction and classification logic, basically employing linear decision boundaries. Some feature selection is used, in that only half the total number of features are used for any one boundary.

For the first stage of recognition, the feature set consists of the number of positive and negative convexities as seen from each of the four sides of the enclosing box. (Only two of the four character contours are used for a single test.) For the second stage, the features used are five measurements made on each of the convexities. Each contour is forced to have 1, 3, or 5 convexities.

Each character is converted from its input format (not clearly specified) to a 24 x 24 raster. Next the character is stretched in the vertical direction to a height of 24 units. Some smoothing is performed; then the external contours of the character are approximated by directed line segments with quantized directions (0°, 180°, 225°, 270°, 315°). This gives a "string" representation for the contours, which is used in determining the convexities.

After preprocessing and string generation the number of "bumps" along each contour is determined. This is followed by second-stage feature extraction. The second-stage recognition logic then takes place. It consists basically of $N(N-1)/2$ tests, where N is the number of possible labels for the unknown sample (N=10 for numerals, 36 for alphanumeric data). The tests are to discriminate between one possible assignment and another; there is one test for each pair of possibilities. For each pair, it is necessary to decide which two contours are to be used in making the decision. (In the numeric-only logic the left and right contours are always used. This leads to some simplification.)

For each I/J (decision between any two possible assignments I and J) there is a separate test for each "sort group" (a, b). Here a is the number of convexities along the

first contour used and b is the number along the second. There are 9 sort groups, so the complete logic in this part of the second stage of classification consists of 9N (N-1)/2 tests. This amounts to 5670 tests that must be implemented for the alphanumeric logic, of which 1/9 or 630 are performed in classifying each sample. The possible implementations of each test I/J are:

Decide I.

Decide J.

No vote.

Apply mathematical formula (linear discrimant) to decide between I and J.

The actual implementation of each test was determined from the training data. (A fifth decision rule, piecewise linear discriminants, was never used for alphanumeric logic because of lack of development time. It was used in 2 percent of the numeric pairwise tests.)

After all tests are performed, the total number of "votes" received by each possibility is used to decide on a class assignment for the sample. Different rejection strategies are possible; however, no rejections were allowed in the experiments.

The rationale for the features used is not stated in any detail. Each test makes use of about half of the features extracted for the sample. "At an early point in this research it was decided that at most two of the four contours . . . would be actually needed to discriminate any pair of characters . . . Even though one contour is often sufficient, our algorithm utilized two prespecified contours for each pair of characters." is all the authors have to say on this point. The method by which the pairs of contours to use for each test were determined is not described. All of the second-stage features for the two contours employed are used.

OLPARS, the On-Line Pattern Analysis and Recognition System implemented at RADC, was used to design the logic used in the tests. The decision algorithm then runs on some other computer system.

The authors estimate that a PDP-8 or any minicomputer with 12 or more bits/word would be capable of implementing the logic using integer arithmetic. They feel that 8K words of storage should be adequate for the numeric-only logic.

Using 1568 of a set of 1640 numeric characters (the others were judged not recognizable by humans) for training and a separate set of 500 characters for testing, the cited results were:

	Correct	Incorrect	Rejected
Training Set	99.2%	0.8%	0%
Test Set	98.6%	1.4%	0%

When all 2068 of these characters were used for training,

Training Set	99.3%	0.7%	0%

With the full alphanumeric logic, using 4883 characters both for training and testing, the results were

Training Set	99.2%	0.8%	0%

All of these characters were handprinted. Times were not given.

For some tests the first-stage features alone are enough to make the decision (decide I, decide J, no vote). In all other cases about half (numeric only) or all (alphanumeric logic) the total number of features must be calculated because they are almost certain to be needed in the large number of tests performed. There was apparently no attempt at determining optimum feature subsets of the given size for the individual discriminants. Also, there was no indication that the choice of features to use was based on any previous research, by the authors or by anyone else. The choice of features is normally the most difficult and the most important part of a pattern recognition problem.

In a nonparametric N-class problem there are basically two approaches. In one, each class is separated from each of the others (the approach used here). There are $N!/(N-2)!2! = N(N-1)/2$ tests to perform. Some of them may be trivial, e.g., "decide I". (The authors state that 65% of the tests were of this type.) However, in the implementation of this paper each possible pair must be "handled".

The other approach is to separate one class from all the others. This is a natural sequential method in which it is possible for recognition to come after only a few (even one) tests, instead of requiring all the tests to be performed. The maximum number of tests is (N-1). It is clear that if feature selection is also employed, feature extraction may be sequential; it may seldom be necessary to extract all the features in classifying a sample.

The authors estimate that the throughput rate for numeric (only) logic on a PDP-8 with extended arithmetic hardware could be about 25 characters per second after the features have been computed. Time for feature extraction is assumed negligible. These estimates are not supported by any tests. No estimates were made of the time for data acquisition, character isolation, etc.

Nearly all of the experimental results in tests of classifier performance were obtained by classifying the same data set as was used for training. The sole exception was a small data set of 500 characters. Nearly 100 percent accuracy is expected in classifying the set used to train the algorithm. Some classification algorithms will always give 100 percent accuracy when applied to the design set.

This work was performed under U.S. Government contract, so it should be available.

6. John W. Sammon, et al, "Handprinted Character Recognition", RADC-TR-72-329, Rome Air Development Center, January 1973.

This report documents a continuation of research into logic for machine recognition of handprinted alphanumeric characters. There was no specific application and a complete system was not developed.

The authors were with Pattern Analysis and Recognition Corporation, 128 East Dominick Street, Rome, NY 13440. The work described in the report was sponsored by Rome Air Development Center (ISCP), Griffiss Air Force Base, NY 13441 (Contract F30602-71-C-0331).

This document reports on a continuation of the work documented in the report whose review immediately precedes this one. Therefore, much of the general description is the same. Some exceptions:

1. The feature set was modified.
2. The recognition logic was organized more efficiently.
3. Rejections were allowed. Four rejection strategies were tested.
4. A larger data set was used for training and an independent set was used for testing.
5. Recognition accuracy and error rates were poorer than in the earlier reports.
6. No time, memory or other hardware requirements were given.

The second-stage feature set of the previous report was changed to a maximum of 21 for each contour; 4 of the 25 measurements taken from contours with five convexities were eliminated because of certain redundancies. However, eight new features, not measured from the contours, were added. They deal with various other shape characteristics. No rationale was given for adding the new features.

The operation of the algorithm is the same as described in the earlier report, except for the extraction of a slightly different set of features and a more efficient strategy for sequencing through the tests. For the latter, rather than performing all the tests in arbitrary order and summing the "votes" to determine the "winner", the order depends on the results of previous tests. For example, suppose that with K classes the number of votes required for assignment is K - 1, the maximum possible. Suppose that the first test performed is A versus B, and A receives the vote. Then B cannot possibly be the winner and it is futile to perform any more tests involving B. The extension of this idea should be clear. The pairwise tests should otherwise be ordered in accordance with class probabilities; there was no discussion of how to estimate these probabilities. The expected number of tests for various values of K was given. For example, if all classes are equally likely and K = 36, the expected number of tests is 50.5, substantially less than the number when all are performed, K (K - 1)/2 = 630.

Training was performed using a data set that was edited to relabel mislabeled characters, remove noise and delete totally illegible characters. The edited data set contained 33,128 alphanumeric characters. The test data set was a separate, unedited group of 6127 samples.

Four strategies for rejecting samples (i.e., deciding that they could not be recognized) were tested:

Strategy A. Require at least 35 votes to assign the sample to that class, reject ties (the maximum possible number of votes with 36 alphanumeric classes is 35).

Strategy B. Require at least 34 votes; reject ties.

Strategy C. Require at least 33 votes; reject ties.

Strategy D. Simply reject a character in case of a tie.

The experimental results were:

Strategy	Substitution Rate	Rejection Rate
A	9.14%	16.97%
B	11.45%	10.18%
C	12.87%	6.92%
D	18.67%	2.01%

The authors point out that most of the substitution errors come from "confusion pairs", e.g., V and U, S and 5, K and X. Discounting the confusion pairs reduces the substitution rate for strategy A to 1%. The authors assert that a substitution rate of 1% with a rejection rate of 16.97% on unconstrained alphanumerics compares favorably with human performance.

The assertion that a rejection rate of 16.97% compares favorably with human performance does not agree with results reported by other authors (e.g., [57]). Also, other authors have been able to obtain substitution rates lower than 9.14% without having to ignore "confusion pairs". Since the recognition logic used here is mathematically sound, suspicion must be directed at the feature set employed. Concerning the choice of features, this report refers only to the earlier report. Both reports are distinguished by a lack of references to previous work in character recognition.

The features used are introduced without any quantitative discussion of their efficiency in discrimination, such as figures of merit. Although from 18 to 50 features are used in each second-stage test in the present work, the recognition rates obtained are only around 78 percent. A subset of the total feature set is used in each test, but there is no evidence presented to show that these subsets are in any way optimum. Since all of the features are computed before the recognition tests are performed, restricting the features used in each test to only those along two of the four contours - all of those features, and no others except the eight "special" features - is somewhat artificial. It is possible that a different group of features - possibly even a smaller one - might be more effective in a particular case.

This work was performed under U.S. Government contract, so it should be available.

8. Jürgen Schürmann, "Multifont Word Recognition System With Application to Postal Address Reading", Proceedings of the Third International Joint Conference on Pattern Recognition, IEEE Computer Society, November 8-11, 1976.

The paper describes techniques used in a complete system for postal address reading (multifont machine-printed characters: upper- and lower-case alphabetic, numerals).

The author is with AEG-Telefunken, Research Institute, D-7900 Ulm, West Germany. The work described in the paper was sponsored by the Federal German Postal Service (West Germany).

The complete system was designed for word recognition, as opposed to simple character recognition. The text is relatively unconstrained, as are font, print quality and character size.

The topics of character recognition and word recognition in this system will be taken up separately in the following.

A. SINGLE-CHARACTER RECOGNITION

The system uses supervised nonparametric classification employing a quadratic discriminant in each of three channels (capital letters, small letters, numerals). Features are the binary values of raster points in a 16 x 16 normalized character array.

The preprocessing operations performed are:

(1) Underlining suppression.

(2) Line-skew correction by shearing.

(3) Segmentation (separation into individual characters).

(4) Centering to the center of gravity.

(5) Normalizing of varying stroke widths and sizes.

The resulting standardized black and white raster pictures are tagged with their origin, line and character position, and are buffered before further processing.

The samples are processed by all three channels. The single character recognition subsystem output consists of the first three choices in each channel, along with a reliability measure for each channel. Depending on the actual values that the three reliability measurements have, a variable number of different choices is to be processed further.

For training the system, a labeled training set is extracted from live mail. Labeling is done primarily by using a preliminary recognition system of the same kind. Only uncertain patterns are labeled manually. Training consists of determining the discriminant polynomial coefficients using a least mean square error approach.

The single character recognition subsystem is implemented using a "specially prepared" microprocessor system, not otherwise specified. The operating speed is given as 1000 characters per second. The complete subsystem is housed in one 19-inch-wide chassis with 27 printed circuit boards.

B. CONTEXTUAL POSTPROCESSING

The basic steps performed in this subsystem are:

1. Formation of words from individual characters.
2. Selection of correct channel output from each 3-channel single-character recognition.
3. Word recognition.

A clustering procedure is applied to the set of gap width measurements collected from the line being tested, in order to group the characters into words (a multidigit number is a "word"). Words may be numeric or alphabetic; if the latter, all letters may be capitals, only the first may be capitalized or all the letters may be small. Using these considerations, decision theoretic logic is applied to the reliability measures for the characters in the word, in order to select the correct channel for each character. Using sets made up of alternative choices for individual characters, then the problem is to choose the right combination. This is done by comparison of the alternatives with a dictionary containing all legitimate words (including misspellings) - e.g. city names, street names.

The following sections refer to the complete system.

The system developed in this work is designated as the AEG-Telefunken AL 880 Address Reader. It was designed to handle 60,000 pieces of mail per hour.

Tests were performed in 1975 on live mail. (The quantity was not stated. Also, the recognition speed on individual characters was not stated.) The results cited were: Segmentation error rate = 1.3%. Recognition error rate (characters) = 1.4% (apparently no rejections were allowed). For word recognition, with a dictionary of 16,000 entries, word recognition rate = 98%, word error rate = 1%, word rejection rate = 1%. Using postcode (apparently the German equivalent of zip code) as an error check converts the error rate almost completely to a rejection rate of the same value.

The prototype machine was scheduled to be installed in Wiesbaden, Germany, in the middle of 1977. No information on its performance could be learned, either through the U. S. Postal Service or the Consulate General of Germany. It is not known whether this work is easily available to the U. S. Government.

APPENDIX B.

REVIEWS OF SELECTED OCR SYSTEMS

1. Information International GRAFIX I System

The GRAFIX I is an OCR system which accepts microfilm input (16 or 35mm) and is capable of multifont plus alphanumeric handprint recognition. This system was originally developed to convert documents prepared for human use into computer format so that the information they contained could be retrieved, updated, and republished in a more efficient manner.

The first GRAFIX I system was placed in operation for the U.S. Naval Air Systems Command. That system, installed in Jacksonville, Florida, is in daily use converting, updating, and republishing technical manuals used at Naval Air Rework Facilities.

The second GRAFIX I system was sold to the Department of Health and Social Security in the United Kingdom. It reads a combination of computer lineprint and unconstrained alphanumeric handprint. GRAFIX I enables the accommodation of substantial increases in the quantities and kinds of data which are required to be gathered in the course of administering their Social Security System.

The basic system includes a central control computer (DEC-10), a specialized image processing computer, a microfilm scanner, and satellite processors which perform such functions as controlling display subsystems. A description of each of these system components and the handprint character recognition capability follows.

MICROFILM SCANNER

GRAFIX I reads microfilm images rather than paper forms. This is done for several reasons:

(1) Transmitted light can be measured for contrast more accurately than reflected light.

(2) There are no problems during scanning due to bad form feeds, skew, torn pages, or paper jams.

(3) Original documents can vary in size. Proper film image size for different application results from use of proper lenses in the microfilm camera.

(4) In a classified environment, original pages may be filmed at their normal storage location.

(5) After filming turnaround documents, the originals can be destroyed, avoiding paper archiving problems.

The input scanner used in GRAFIX I is a Programmable Film Reader manufactured by Information International. The scanner accepts 16mm or 35mm roll microfilm and transports film images to a scanning aperture where a flying-spot CRT beam is directed to page areas which contain information to be recognized. Scanning converts visual images on microfilm to gray-scale picture elements at speeds as high as 1 million points per second. The gray scale data is subsequently converted to binary images (thresholded) and passed to recognition.

The scanner is driven by a Triple-I 15 computer which is connected to the DEC-10 memory. Major components of the scanner subsystem are:

Optical/Mechanical Unit

The Optical/Mechanical Unit includes the Programmable Light Source (PLS), a film transport, light collection and measuring devices, and optics. A special circuit protects the PLS against excessive beam current and power failure; a deflection computer dynamically corrects for the size of the PLS spot and for pincushion distortion. PLS spot spacing is dynamically variable over a 65K square raster. Spot size and intensify time are dynamically adjustable under program control.

Video Processor Unit

The video processor's density output format features a sliding, variable-width, variable-resolution "window," which allows the selection of a density format that can be programmably optimized. Density measurement steps are adjustable under program control over the range from 0.0 to 2.56 in 0.005 increments.

Density data is automatically packed and written into memory. One to nine bits of dnesity data per scanned point can be provided under program control. Digital outputs from both the film path and reference path photomultipliers are available in both linear and logarithmic form.

Scan Generator

The scanner's monitor CRT has a digital raster (deflection) generator separate from that which controls the deflection of the precision scanning CRT. The two raster generators can be operated in synchronism. When operated in synchronism, the size, location, orientation, and shape of the two rasters can be independently controlled, allowing enlarged playback of the area being scanned.

Correlator

The scanner includes a high-speed automatic correlator which is used to assist film reading programs in locating and measuring (primarily) lines and edges. The correlator contains two memories, one of which is loaded by the program to contain a sequence of numbers which define the density profile of the expected trace. The other memory is used to store a continuously updated history of the 32 most recent density measurements. As each new point is scanned, the expected trace is convolved with the density history, and coordinate data for the point of best correlation is determined.

Operator Console

The Scanner Operator Console consists of the Monitor CRT, Operator Display Console, and an ASR-33 teletype.

The Monitor CRT may be used to display alphanumeric data, graphics, and raster displays, with intensity modulation derived from digital data stored in memory or analog video directly from the film being scanned. The Monitor CRT can be programmed to display a full-screen raster while the precision CRT (PLS) simultaneously scans a smaller area located anywhere on the film frame.

Triple-I 15 Control Computer

The Triple-I Series 15 computer is an integrated circuit, 18-bit binary word, general-purpose computer designed and built by Information International.

Memory Access Logic

This device allows block transfers between the Triple-I 15 18-bit memory and the main system's 36-bit memory.

CONTROL COMPUTER

The Central Processor in GRAFIX I is a Digital Equipment Corp. DEC-10. It runs under control of a timeshared operating system developed by DEC and modified by Information International to accommodate the special-purpose peripheral devices that have been added to the system. The Central Processor features protection and relocation registers, multiprogram protection, dynamic core allocation and reentrant programs. Maximum memory size is 256K, 36-bit words (1024K bytes). The main memory has 550-nanosecond access time, 4-way interleaving, and a cycle time of 950 nanoseconds.

On-line random access storage is provided by a disc controller and up to eight disc drives, each with a capacity of 10 million 36-bit words.

Magnetic tape I/0 is performed by a tape controller and up to eight tape drives. The tape controller is connected directly to memory through a data channel. Maximum data transfer rate at 1600 bpi is 240,000 bytes per second.

Two data channels connect the disc and tape controllers directly to memory. These channels allow data transfers between auxiliary memory and core memory to occur simultaneously with central processor computation.

A 96-character line printer is used to produce hardcopy output.

On-line CRT terminals are used for display of system status information by the Job Control Program, and entry of commands by the operator.

IMAGE PROCESSING

The Binary Image Processor (BIP) is a special-purpose, stored-program, serial computer for the manipulation of two-dimensional arrays of numbers, especially binary numbers. The BIP's major functions include measurement of basic topological properties of binary arrays such as area, line width, Euler number, and character height and width; cross-correlation of two arrays; and creation of new arrays as a function of one or two other

arrays. Array transformation capability includes gray-scale to binary image conversion, a powerful set of 3 x 3 neighborhood operations on binary arrays, and a complete set of Boolean operations on two source arrays and the result of neighborhood computations.

The BIP serves as a special-purpose slave processor, performing inner-loop tasks at very high speed while the central processor carries out system control and decision-making.

The BIP performs image transformations, producing one output image as a function of one or two input images. A second major function is the measurement of images, including measurement of absolute local properties of one image and of one image with respect to another (e.g., cross-correlation).

The BIP performs these computations up to several thousand times as fast as typical medium-large computers. Its greater speed is due to several principal factors:

(1) Its organization closely corresponds to image geometry; two-dimensional array structures and neighborhoods are built in.

(2) It is a serial machine, which makes counting of features in arrays very simple.

(3) While most computers spend as much time accessing and decoding instructions as accessing and managing data, the BIP has a very high ratio of data operations to command operations.

(4) The image processing section consists of a long pipeline containing many array points at different stages of processing; this allows high throughput.

(5) The BIP consists of integrated circuits in a special packaging arrangement, permitting a large amount of pipeline logic to run synchronously at over 35MHz.

(6) The BIP has local, high-speed, 8K by 36-bit semiconductor memory for private use.

The BIP executes a program after being given initialization data by the DEC-10 CPU. A typical command causes the BIP to fetch several hundred bits of detailed control data, or parameters; following this, the image process itself is started based on the parameters that were loaded.

During the image process, one or two arrays of data, denoted the Unknown (U) and Mask (M), are usually accessed from sequential locations in main system memory. An output array, denoted the Result (R), is stored either in local, high-speed BIP memory or in main system memory. At the same time the three images U, M and R are being measured in various ways. When the image process is complete, the measurements are returned to main system memory.

Major BIP Functions

During OCR, the BIP performs the image manipulations required for character recognition. Major functions of the BIP, in the sequence which they are performed on a single image, are:

Thresholding

The scanner outputs gray-scale data that must be converted to binary character images prior to recognition. This conversion step is called thresholding. In this process the density value output for each scanned point is evaluated as greater or less than a cut-level (threshold) value. All gray-scale values equal or less than the cut-level are represented as ones in the binary result image. Gray-scale values greater than the cut-level are represented as zeroes or the complement. The cut level is not necessarily fixed, but can be any sequence of bytes of the appropriate size.

The cut-level used in thresholding should yield a binary image that contains continuous solid strokes corresponding to the strokes in the unknown character. However, there are many instances where uneven inking of the original impression creates a binary image that contains "holes," or noise dots, or non-smooth contours. This condition can be determined during image measurement and corrected by enhancement techniques, using the BIP's neighborhood processing capabilities.

Normalization

In some applications, especially handprint, images of nonstandard dimensions can be expanded or shrunk to conform to norms using the BIP in a one-step process.

Array Correlation (Masking)

One of the basic techniques in recognition is the correlation of the unknown with a stored set of binary masks. The position of the unknown within its character envelope is

known approximately but not exactly; since cross-correlation is highly position-sensitive, the BIP simultaneously correlates the unknown with the mask at nine relative offsets. It has a feature that enables it to search rapidly for the best positional match of the two images over the nine offsets. Correlation values used in recognition decisions are based on the best-fit correlation value.

SATELLITE SUBSYSTEMS

Reject Conversion Subsystem

Reject Conversion is the process of visually identifying degraded character images that the OCR program could not identify. Reject Conversion terminals provided for this purpose are controlled by a Triple-I 15 which is connected on-line to the DEC-10. The Triple-I 15 displays unknown characters and context, and sends corrected files back to the main system.

The main elements of this subsystem are:

Triple-I 15

This is a general-purpose minicomputer connected to the DEC-10 by an on-line interface. The interface allows block transfers between the two memories.

Display Controller (DPC)

This is a special-purpose computer whose instructions cause displays to appear on the Reject Conversion CRTs. The display refresh does not tie up main system memory or the Triple-I 15 memory.

Interactive Terminals

Each terminal consists of a CRT and keyboard. The keyboard includes a full character set plus special symbols and control keys. The character set is stored in a special display memory and is easily changed. Up to 172 different character shapes may be stored in a single memory load. Characters have a smooth, uniform appearance. Each terminal has individual controls for display positioning, intensity and focus.

Data Tablet Subsystem

The Data Table subsystem is used in applications that must read variable format pages. The Data Tablet process creates page descriptor files used by the OCR program.

The main elements of this subsystem consist of a triple-I 15 and graphic CRT terminals. Each interactive terminal is supplemented by a digitizer tablet, function menu, and electronic stylus.

The digitizer tablet contains registration pins that align with holes punched in each page at the page preparation step. The electronic stylus causes the digitizer tablet hardware to generate current-point coordinate data which is read by the data tablet program.

Handprint Character Recognition

Handprint characters are recognized by two separate programs. The top-level recognition program is called the Filter. It is very fast because it makes heavy use of the BIP, and can correctly identify the unknown in a majority of cases. The Filter makes recognition decisions based on character edge characteristics.

In mixed alphanumeric handprint recognition, ambiguous recognition situations occur. The letters "8" and "B" look similar to the eye, and to the Filter. So do other alphanumeric pairs such as "S" and "5". Ambiguities are important when the Filter is unable to differentiate between two or more possible identifications for the unknown. In these cases, the second recognition program, called the Verifier, is used. This program consists of one subprogram for each of the symbols in the handprint character set. The Filter calls appropriate Verifiers to distinguish between possible matches for the unknown. The Verifier then applies specific tests which analyze subtle differences between character shapes. In the case of an S-5 ambiguity, the Verifier concentrates, among other things, on the upper left corner of the character and decides between the two possibilities.

The Handprint recognition system has been designed to read the character shapes that people write. The only constraints imposed on the writer are that straight lines be straight, curved ones curved, and characters placed within the areas provided for them. The power of the Handprint recognition system has been proven by its ability to read mixed alphanumeric handprint written by a large number of clerks without artificial character shaping rules, extensive training, or reduced clerical performance.

2. Recognition Equipment Input 80 System

The Input 80 is an OCR system that accepts page size documents as input and is capable of reading multifont and handprinted information. The system consists of a Page/

Document Transport, System Controller (Processor, magnetic tape and peripheral interface), Recognition Unit and Operator Communication device. A brief description of each major system component including the Input Sensor follows.

PAGE/DOCUMENT TRANSPORT

Functions of the Input 80 transport are to feed, align and optically read source documents; add sequence numbers and create a permanent record on 16mm microfilm if desired; direct documents to any one of three output stackers, and record data on magnetic tape.

Batches of input documents to be processed are loaded into the input hopper on the transport and automatically conveyed through the system at constant speed. Document speeds are operator selectable.

The top document on the input stack is selected by the feeder mechanism, aligned, and tested to ensure that only a single document has been fed. The document is transported past a solid-state page-width optical scanner where the data to be read is converted into video signals. These electronic images are processed by the Video Processor Unit (part of the System Controller) and sent to the Recognition Unit for character identification.

When Input 80 options for line marking/sequence numbering or microfilming are selected, these functions occur after the data is read. Line marking/page sequence numbering provides a capability to mark lines containing unrecognized characters as well as to perform program controlled page-sequence numbering. Normally, the sequence numbering printed on the document is also written on the tape record, permanently cross-referencing the two items.

Two microfilm options are available to provide permanent records of documents on 16mm film with no reduction in throughput: one performs microfilming of the front side of the document; a second allows concurrent front and back microfilming.

As the document nears the end of the transport belt drive assembly, it is directed to one of three transport output stackers under program control.

SYSTEM CONTROLLER

The System Controller for the Input 80 includes a Programmed Controller, Video Processor Unit, peripheral controllers and magnetic tape transport(s).

The control element of the System Controller is a general purpose, stored program, digital computer. Key features of the computer are: 24-bit word length, fully buffered I/O channels and single address capability. Hardware multiply, divide and square root capability are also provided. A repertoire of approximately 600 instructions is featured; up to 32,768 words of memory may be addressed directly by all instructions. Basic memory is 16,384 words expandable to 32,768 words in 8,192 word increments.

OPERATOR CONSOLE CRT TERMINAL

The Operator Console CRT Terminal provides overall control of the Input 80 system. It consists of a CRT display and keyboard housed in a table-top cabinet.

The CRT portion of the Operator Console provides visual display of alphanumeric characters utilizing a five-by-seven dot matrix character pattern. Twenty-four lines of up to 80 characters per line can be displayed. Data to be displayed may originate from the keyboard or may be received from the programmed Controller. A standard 64-unit ASCII character set is employed.

In-Line Reentry capability can be added to the Operator Console. This optional capability provides a display of unrecognized characters for the operator to view, recognize and manually reenter using single keystrokes.

RECOGNITION UNIT

Input 80 utilizes two proven recognition units. The Template Recognition Unit is for reading machine-printed alphanumeric characters and the optional Feature Recognition Unit is for recognizing numeric handprinting.

The Template Unit offers singlefont, multiple-font or multifont recognition capability for a wide variety of machine-printed fonts and is capable of accurate recognition of degraded characters. The Feature Unit offers numeric handprint recognition and, with

options, will accomplish several types of mark-sensor recognition. Together, they are capable of recognizing machine and handprinted data intermixed on the same line.

Each Recognition Unit accepts video signal patterns stored within the Video Processor Unit. These patterns are compared with the Recognition Unit's vocabulary. If the video pattern matches a pattern within the vocabulary, an output code is generated corresponding to the character pattern matched. If a character is unrecognizable, a reject character code is generated. Both character and reject codes are transmitted to the System Controller for processing and storage.

Reentry of Unrecognized Characters

Two options are offered for key entry of unrecognized characters. In-line Reentry (ILR) provides low-cost single-terminal display of unrecognized characters for operator entry. Total Data Entry can be utilized for multiterminal display and key entry of unrecognized characters.

INPUT SENSOR

The "eye" of Input 80 is the Integrated Retina, a high-resolution optical sensor with two 1872 photodiode arrays incorporated on slices of silicon approximately one inch long. The sensors convert the optical character images to electrical signals. Large scale integration (LSI) techniques eliminate the need for thousands of semiconductors and connections.

The Integrated Retina approximates the human eye in its reading resolution. A character is divided into rectangular "cells" .007" high by .0035" wide. Each cell is classified - not as simply black or white but as one of 16 different shades of gray. This gray scale value is used to compare a cell to its surrounding cells. Data from the Integrated Retina is transmitted to the Recognition Unit, where each cell is compared with surrounding cells and identified as either black or white. This process enables the character images to be cleaned up, with weak strokes filled in, smudges ignored, and contrast sharpened, adding significantly to the reading performance.

The Input 80 System with its Integrated Retina also works much like the human eye in compensating for size differences in the material it is reading. The character images are electronically "normalized" to a common size enabling Input 80 to switch almost instantaneously between different sizes of machine-printed and even handprinted characters, even when they appear on the same line.

APPENDIX C.

OCR SYSTEM VENDORS

Company Name	Product	Handprint
Ball Computer Products, Inc.	Mark reader	
Bell and Howell Company	Mark reader	
Bourns Management Systems	Mark reader	
Burroughs Corporation	Document reader	
Chatsworth Data Corporation	Mark reader	
Cognitronics Corporation	Document/page reader	X
CompuScan, Inc.	Document/page reader	X
Computer Entry Systems Corp.	Document reader	
Context Corp.	Page reader	
Control Data Corp.	Document/page reader	X
Cummins-Allison Corp.	Document reader	X
Datatype Corp.	Bar code reader	
Dest Data Corp.	Page reader	
Documentation, Inc.	Mark reader	
ECRM, Inc.	Page reader	
Entrex, Inc.	Document reader	
Hendrix Electronics	Document/page reader	
Hewlett-Packard Company	Mark reader	
Hitachi-Zosen	Page reader	X
Honeywell Information Systems, Inc.	Document reader	
IBM Corp.	Document/page reader	X
Information International, Inc.	Document/page (film) reader	X
Input Business Machines, Inc.	Document reader	X
Key Tronic Corporation	Document/page reader	X
Kimball Systems	Mark and bar code reader	
Kurzweil Computer Products	Document/page reader	
Lundy Electronics and Systems, Inc.	Document/page reader	X

APPENDIX C.

OCR SYSTEM VENDORS (Continued)

Company Name	Product	Handprint
National Computer Systems, Inc.	Document/page reader	X
Optical Business Machines, Inc.	Document/page reader	X
Peripheral Dynamics, Inc.	Mark reader	
Recognition Equipment, Inc.	Document/page reader	X
Rockwell International	Page reader	
Scan-Data Corporation	Document/page reader	X
Scan-Optics, Inc.	Document/page reader	X
Univac Div., Sperry Rand Corp.	Document reader	
Westinghouse Learning Corp.	Mark reader	

APPENDIX D.

NTIS ABSTRACTS

1. CHARACTER RECOGNITION SYSTEM USING A SPATIAL FILTER.
Tanaka, Kokichi; Tamura, Shinichi; Mike, Shigehiko; Ozawa, Kazumasa.
Osaka University, Toyonaka, Japan
February 1976

Observations of the Fourier transform images of character patterns made up of many straight lines suggest that measuring the intensity of directional components is quite useful for character recognition. The use of a combination of spatial and band-pass filters for such character recognition is reported. The experiments were made with fifteen classes of typewritten characters, twenty-four classes of typewritten characters and ten classes of handwritten characters. It is concluded that the system proposed is simple in construction and has a rather satisfactory recognition rate, being less susceptible to translations, variations in size, and slight rotations of the input patterns.

2. OPTIMIERUNG MUSTERKENNENDER SCHICHTSTRUKTUREN DURCH MERKMALSSYNTHESE
(Optimization of Pattern Recognition Layer-Lattice Structures by Feature Synthesis)
Giebel, Hayo
Technical University, Munich, Germany
December 1975

A procedure for nonlinear transformation of patterns is proposed which simplifies classification. The main goal is not a selection but a synthesis of features, using two steps: elementary features are evaluated by AND gates and then combined into groups by OR gates. These steps can be optimized with the aid of certain functions, derived by considering sufficient conditions for linear classification. Repetitive application approaches these conditions successively. Thus a hierarchical system

is obtained, using a linear classifier as the last stage. An example of handwritten alphanumeric characters illustrates the procedure and yields quantitative results.

3. APPLICATION OF CCD's TO DOCUMENT SCANNING

Simms, T.

Canadian Post Office, Ottawa, Ontario

December 1975

Recent advances in solid state technology provide means of improving the designs of document scanning subsystems for use in facsimile communications, optical character recognition and high-speed electronic mail transmission equipment. A CCD optical imaging array which provides an alternative to the traditional mechanical complexity of such equipment is described. It is shown that the application of charge-coupled technology promises to impact the design of communications equipment in general in the areas of high capacity digital storage and signal processing.

4. STRUCTUAL CHARACTER RECOGNITION BY FORMING PROJECTIONS.

Breuer, P.; Vajta, M. Jr.

Research Institute for Telecommunications, Budapest, Hungary

1975

A procedure for recognizing handwritten numerals is suggested. The procedure starts with placing the numerals onto a raster field. On this basis, as usual, a matrix with zero and one entries is associated with the character. Data reduction (feature extraction) is obtained in two steps: first, integer valued projections are defined, and then identical runs of appropriate size in these projections are used to form a feature of three components. The small size of the feature space permits making a decision using a dictionary. The features are easy to calculate. A formal description of the feature extraction is given by using a goal-oriented, context-free grammar.

5. TOSHIBA'S OCR REVIEW AND FORECAST.

Nakayama, Naoto; Takebe, Hisao

Tokyo Shibaura Electric Co., Yanagicho Works, Japan

December 1975

Described is a series of optical character reader (OCR) systems which can read alphanumerics, with no regard to whether they are machine printed or handwritten, and can provide multifont reading ability. Two OCR technologies involving feature extraction and multiple similarity are employed in the system.

6. EXTRACTION OF CONCAVE AND CONVEX STRUCTURES BY THE CLOSURE RATE FIELD

Mori, Terunori; Mori, Shunji; Shimizu, Shinichi

1975

Described is a character recognition system based on a topological line segment method using the closure rate field. Feature extraction is performed both along the character line and normal to the character line. The line segments are concave, convex, and enclosed segments. The method has been successfully applied to about 5,000 alphanumeric and 48 special character categories written by about 100 untrained subjects in as close conformity as possible to the standardized characters. The results of recognition show that the correct reading rate, reject rate and error rate are 99.1%, 0.8% and 0.1% respectively in one-half of the set of characters (training data), with 97.5%, 2%, 1% and 0.4% respectively in the other half (test data). (In Japanese with English abstract).

7. RECOGNIZE HAND-PRINTED CHARACTERS WITH A SIMPLE ALGORITHM.

Whetstone, Albert; Domyan, Stephen

Summagraphics Corp. Fairfield, Connecticut

February 1, 1975

Described is a character recognition scheme which permits a logic circuit to recognize hand-printed characters. The circuit need only record the regions where a hand-printed character begins and ends. These first and last regions, properly encoded, then address a programmed read-only memory look-up table to provide the ASCII code for the hand-printed number. The characters are drawn on a paper form pre-printed with boxes, in which the characters are hand-printed. Each box is divided into nine areas. All the standard numbers plus four other symbols can be identified.

8. ALGORITHM FOR A LOW COST HAND PRINT READER.

Holt, Arthur W.

Arthur Holt, Inc., Annapolis, Maryland

February 1974

An algorithm based on the use of a central constraint line for handwriting input for optical character recognition equipment provides insensitivity to shape distortion and low cost implementation. The algorithm, called Snow White, converts a shape measurement to a simple topological measurement. The algorithm is immune to pieces of dirt on the paper or places where the pencil point was inadvertently placed. It is easy to teach.

9. LEFT-SIDE DETECTION SEGMENTATION

Baumgartner, R. J.; Buettner, J. A.; Miller, G. D.

July 1974

In optical character recognition machines which read uncontrolled input, the task of isolating individual characters becomes extremely difficult. This segmentation scheme provides a solution by detecting when the left side of the character is present, by measurements designed to detect features characteristic of individual classes.

10. NEW PAGE OPTICAL CHARACTER READER, OCR-V.

Yoshizawa, Masaaki; Asami, Hiroaki

Ome Works, Japan

June 1974

A general-use page optical character reader has been developed that can process sheet sizes of from 105x148 mm to 364x364 mm at a 250 sheet per minute rate. Reading speed is 1600 characters per second. Handwritten characters mixed with stylized fonts can be accommodated. Scanning is by photodiode array. The recognition method employs feature extraction and partial pattern matching.

11. LARGE-SCALE OPTICAL CHARACTER RECOGNITION SYSTEM SIMULATIONS.

Himmel, David P.; Peasner, David

Recognition Equipment Inc., Dallas, Texas

January 1974

A simulation is described which provides a vehicle for the synthesis and subsequent performance analysis of optical character recognition (OCR) algorithms aimed at solving some of the most difficult problems encountered in OCR. The simulation treats the problems of linking complex interactive algorithms together and processing large real-world data files under economic constraints. The simulation is written in Fortran V and is installed on University Computing Company's Dallas 1108 facility. It is comprised of a main program, four major subroutines, 29 supporting routines, and the system library routines. It presently requires 104,500 words of memory (418,000 bytes). Results of the computer simulation on real-world handprint character data are presented.

12. EXPERIMENTAL PROCEDURE FOR HANDWRITTEN CHARACTER RECOGNITION.
Dutta, Asoke Kumar
Indian State Institute, Calcutta, India
May 1974

A description is given of a recognition scheme which is independent of the dynamics of writing and is suitable both for on-line and off-line systems. New recognition criteria, namely the distribution of intensity of marking along and perpendicular to the direction of writing, are used. Methods for correcting the inclination of script, determination of zonal limits and segmentation of the continuous script into a set of curve elements were developed. The method of correlation is used for classification of the curve elements. A simple grammar for the reconstruction of letters from classified segments is presented. A significant recognition was found.

13. EXPERIMENTAL STUDY OF INFORMATION MEASURE AND INTER-INTRA CLASS DISTANCE RATIOS ON FEATURE SELECTION AND ORDERINGS.
Michael, Mark; Lin, Wen-Chun
Case Western Reserve University, Cleveland, Ohio
March 1973

The algorithms are first presented and then they are applied and compared to recognize handprinted alphanumeric characters. Both Highleyman's data and raw data obtained in the Signal Processing Laboratory at Case Western Reserve University,

Cleveland, Ohio, were used for the study. It is believed that the criteria can be used for other applications and can especially be used where the statistical independence among features is not assumed.

14. TRADEOFFS IN MONOLITHIC IMAGE SENSORS: MOS VS CCD.
Melen, Roger
Stanford Electronics Laboratory, California
May 1973

The article compares the capabilities of the charge-coupled device to the capabilities of the older MOS photodiode image sensor. Both types of monolithic image sensors offer fundamental improvements over earlier imaging methods, especially for optical character recognition, facsimile systems, and video communications, where high-voltage devices often requiring high light levels are being used. Optical character recognition and facsimile displays require only small arrays, while the CCD display appears to be the only one of the two suitable for television applications, both at high and unusually low light levels.

15. RECOGNITION OF HANDWRITTEN WORDS USING A LINE-FOLLOWER
Wilson, J. D.
Royal Military College of Canada, Kingston, Ontario
October 1970

A hybrid technique for machine recognition of cursive handwriting is described. The method imposes little constraint on the writer and requires no special fields or marks. A line-following procedure is used to control a flying - spot scanner. The route taken along the pattern-lines is dependent upon accumulated knowledge of the pattern. A description vector, generated sequentially during the line-following phase, is subjected to base-alignment and stroke length normalization. Contextual aid is drawn from constraints inherent within diagrams and trigrams. The method was tested by simulation and proved successful with reasonably well-formed writing.

16. RECOGNITION OF HANDPRINTED NUMERALS BY TWO-STAGE FEATURE EXTRACTION
Chuang, P. C.

Recognition Equipment, Inc., Dallas, Texas
April 1970

An optical character recognition system for handprinted numerals of noisy and low-resolution measurement is proposed. The system consists of the two-stage feature extraction process. In the first stage a set of primary features insensitive to the quality and format of a black-white bit pattern is extracted. In the second stage, a set of properties capable of discriminating the character classes is derived from primary features. The system is simple and reliable in that only three kinds of primary features are needed to be detected. The recognition is based on the decision tree which tests the logic statements of secondary features.

17. FEATURE DETECTION METHOD FOR OPTICAL CHARACTER RECOGNITION
Hosking, K. H.
1969

A method of optical character recognition is presented that may be suitable for the recognition of hand-printed alphanumeric characters. It overcomes some of the problems of distortion, misregistration, orientation and sensitivity to variations in stroke thickness. The technique has been simulated on Myriad and results with machine printed upper-case alphanumeric character are encouraging.

18. SYNTAX DIRECTED ON-LINE RECOGNITION OF CURSIVE WRITING.
Kim, Yung Taek; Evans, David
July 1968

A syntax organization for recognition of handwritten connected words is studied in the work. Each writing is cut out into strokes at the middle point of every down curve of the writing, and the strokes are named using their directional characteristics and relative size among the strokes. A syntax is organized using the hierarchy of the stroke characteristics and self-iteration for the error corrections. The strokes are classified by the hierarchical characteristics. The lowest level of hierarchy collects those strokes which cannot be combined into characters by their solid stroke characteristics and organizes a two dimensional family relation for relative combination of the strokes into characters. The local classifying routines are called for

those stroke relations which require the evaluation of the relative characteristics between the strokes for the optimal decision.

19. RECOGNITION OF HANDPRINTED SYMBOLS FOR COMPUTER-AIDED MAPPING.
Nolan, B. E.
December 1971

The report describes a study of the recognition of handprinted symbols for a computer-aided mapping system, which was performed under a contract with USAETL, Fort Belvoir, Virginia. Software is used to recognize a constrained set of handprinted symbols which have been digitized on a drum scanner, and to identify their X-Y coordinates for creating point symbols or descriptive information on map overlay negatives. Pattern classification is accomplished using discrete Fourier transformations. Recognition rates for the character sets averaged out at about 96%.

APPENDIX E

CONFERENCE PROCEEDINGS, JOURNAL ARTICLES, TECHNICAL REPORTS

1. Proceedings of the First International Joint Conference on Pattern Recognition (Washington, D.C., Oct. 30-Nov. 1, 1973), IEEE Catalog Number 73 CHO 821-9C, 1973.

2. Proceedings, Second International Joint Conference on Pattern Recognition (Copenhagen, Denmark, Aug. 13-15, 1974), IEEE Catalog Number 74 CHO 885-4C, 1974.

3. Conference Record, 1976 Joint Workshop on Pattern Recognition and Artificial Intelligence (Hyannis, Mass., June 1-3, 1976), IEEE Catalog Number 76CH1169-2C, 1976.

4. Proceedings, Cybernetics and Society International Conference (Washington, D.C., Nov. 1-3, 1976), IEEE Catalog Number 76 CH1137-9 SMC, 1976.

5. Proceedings, the Third International Joint Conference on Pattern Recognition (Coronado, Calif., Nov. 8-11, 1976), IEEE Catalog Number 76CH1140-3C, 1976.

6. Proceedings, IEEE International Conference on Cybernetics and Society (Washington, D.C., Sept. 19-21, 1977), IEEE Catalog Number 77CH1259-1 SMC, 1977.

7. R. C. Ahlgren, H. F. Ryan, and C. W. Swonger, "A Character Recognition Application of an Iterative Procedure for Feature Selection," IEEE Transactions on Computers C-20, 1067-1075 (1971).

8. O. Albertsen, E. Münster and P. Ponsaing, "A Low-Cost Zipcode Reader Using Optical Line Detection," in [2], 401-408.

9. O. Albertsen, E. Münster and P. Ponsaing, "Optimal Information Economy in the Multifont OCR Input System," in [5], 127-134.

10. Farhat Ali and Theodosios Pavlidis, "Syntactic Recognition of Handwritten Numerals," IEEE Transactions on Systems, Man, and Cybernetics SMC-7, 537-541 (1977).

11. K. T. Al-Kibasi and W. K. Taylor, "The UCLM 3 Programmable Pattern Recognition Machine," in [2], 241-246.

12. Robert S. Apsey, "Human Factors of Constrained Handprint for OCR," in [4], 466-470.

13. D. Armitage and A. W. Lohmann, "Character Recognition by Incoherent Spatial Filtering," Applied Optics 4, 461-467 (1965).

14. Gordon B. Baty, "OCR-Where It's Going," Journal of Systems Management, 10-15 (Feb. 1976).

15. M. Beun,"A Flexible Method for Automatic Reading of Handwritten Numerals," Philips Technical Review 33, 89-100 (Part 1) and 130-137 (Part 2) (1973).

16. B. Blesser et al, "A Theoretical Approach for Character Recognition Based on Phenomenological Attributes," in [1], 33-40.

17. Barry A. Blesser, Theodore T. Kuklinski and Robert J. Shillman, "Empirical Tests for Feature Selection Based on a Psychological Theory of Character Recognition," Pattern Recognition 8, 77-85 (1976).

18. M. Bohner, "A Quantitative Method for Font Assessment," in [2], 529-533.

19. M. Bohner, M. Sties and K. H. Bers, "An Automatic Measurement Device for the Evaluation of the Print Quality of Printed Characters," Pattern Recognition 9, 11-19 (1977).

20. B. R. Brown and A. W. Lohmann, "Complex Spatial Filtering with Binary Masks," Applied Optics 5, 967-969 (1966).

21. David L. Caskey and C. L. Coates, "Machine Recognition of Handprinted Characters," in [1], 41-49.

22. C. Cox, B. Blesser and M. Eden, "The Application of Type Font Analysis to Automatic Character Recognition," in [2], 226-232.

23. Belur V. Dasarathy and K. P. Bharath Kumar, "CHITRA: Cognitive Handprinted Input Trained Recursively Analyzing System for Recognition of Alpha-Numeric Characters," to be published, International Journal of Computer and Information Sciences 7 (1978).

24. Wolfgang Doster, "Contextual Postprocessing System for Cooperation With a Multiple Choice Character Recognition System," in [5], 653-657.

25. Robert W. Ehrich, "A Contextual Post-Processor for Cursive Script Recognition-Summary," in [1], 169-171.

26. Louis R. Focht and Alfred Burger, "A Numeric Script Recognition Processor for Postal ZIP Code Application," in [4], 489-492.

27. Y. Fujimoto et al, "Recognition of Handprinted Characters by Nonlinear Elastic Matching," in [5] 113-118.

28. Sadao Fujimura and Eiichi Tazaki, "Feature Extraction From Characters by Pattern Deformation due to Field," in [6], 577-580.

29. T. Fujita, M. Nakanishi and K. Miyata, "The Recognition of Chinese Characters (Kanji), Using Time Variation of Peripheral Belt Pattern," in [5], 119-121.

30. Gérard Gaillat, "An On-Line Character Recognizer with Learning Capabilities," in [2], 305-306.

31. Rafeal C. Gonzalez, "Pattern Recognition via Topological Feature Extraction," Ph.D. Dissertation, University of Florida, 1970.

32. Arnold K. Griffith, "From Gutenberg to GRAFIX I-New Directions in OCR," The Journal of Micrographics 9, 81-89 (1975).

33. Arnold K. Griffith, "The GRAFIX I System and Its Application to Optical Character Recognition," in [5], 650-652.

34. A. Güdesen, "Quantitative Analysis of Preprocessing Techniques for the Recognition of Handprinted Characters," Pattern Recognition 8, 219-227 (1976).

35. Klaus Hager, "Experience With Sequential Polynomial Pattern Classifiers," in [2], 210-214.

36. Shin-Ichi Hanaki, Tsutomu Temma and Hiroshi Yoshido, "An On-Line Character Recognition Aimed at a Substitution for a Billing Machine Keyboard," Pattern Recognition 8, 63-71 (1976).

37. S. Hård and T. Feuk, "Character Recognition by Complex Filtering in Reading Machines," Pattern Recognition 5, 75-83 (1973).

38. Leon D. Harmon, "Automatic Recognition of Print and Script," Proceedings of the IEEE 60, 1165-1176 (1972).

39. David P. Himmel, "Some Real-World Experiences with Handprinted Optical Character Recognition," in [4], 479-484.

40. Ellen Hisdal et al, "Structural Recognition of Handwriting, in [5], 140-143.

41. Arthur W. Holt, "The Impact of New Hardware on OCR Designs," Pattern Recognition 8, 99-105 (1976).

42. William A. Huber, "Handprint Reader," Research and Development Technical Report ECOM-4087, U.S. Army Electronics Command, March 1973.

43. Tadao Ichikawa and Jun Yoshida, "On-Line Recognition of Handprinted Characters With Associative Read-Out of Patterns in a Memory," in [2], 206-207.

44. Taizo Iijima, Hiroshi Genchi and Kenichi Mori, "A Theory of Character Recognition by Pattern Matching Method," in [1], 50-56.

45. A. G. Kegel, J. K. Giles and A. H. Ruder, "Observations on Selected Applications of Optical Character Readers for Constrained, Numeric Handprint," in [4], 471-475.

46. Robert Kooi and Wen C. Lin, "An On-Line Minicomputer-Based System for Reading Printed Text Aloud," in [4], 509-513.

47. Douglas Kozlay, "Feature Extraction in an Optical Character Recognition Machine," IEEE Transactions on Computers C-20, 1063-1066 (1971).

48. Henry P. Kramer, John Bergstrom and Jon Ahlroth, "The Recognition of Handprinted Symbols and its Relation to Formal Languages," in [1], 57-58.

49. P. Krause, W. Schwerdtmann and D. Paul, "Two Modifications of a Recognition System with Pattern Series Expansion and Bayes Classifier," in [2], 215-219.

50. S. K. Kwon and D. C. Lai, "Recognition Experiments with Handprinted Numerals," in [3], 74-83.

51. Isao Masuda, "Assessment of Handprinted Character Quality and its Application to Machine Recognition," in [2], 202-203.

52. S. Mori et al, "Recognition of Handprinted Characters," in [2], 233-237.

53. Terunori Mori et al, "Recognition of Handwritten Alphanumeric and Special Characters," Systems Computers Controls 6, 37-46 (1975).

54. Morton Nadler, "Structural Codes for Omnifont and Handwritten Characters," in [5], 135-139.

55. Roger N. Nagel and Azriel Rosenfeld, "Steps Toward Handwritten Signature Verification," in [1], 59-66.

56. Y. Nakano et al, "Improvement of Chinese Character Recognition Using Projection Profiles," in [1], 172-178.

57. Heinrich Niemann, "A Comparison of Classification Results in Character Recognition by Man and by Machine," in [5], 144-150.

58. Rainer Ott, "On Feature Selection by Means of Principle Axis Transfrom and Nonlinear Classification," in [2], 220-222.

59. Kazumasa Ozawa and Kokichi Tanaka, "Character Recognition Using Correlation Technique on Spatial Complex-Frequency Plane," in [2], 409-410.

60. J. R. Parks et al, "An Articulate Recognition Procedure Applied to Handprinted Numerals," in [2], 416-420.

61. John D. Patterson, "The Linear Mean Square Estimation Technique of Recognition," in [1], 67-76.

62. Theodosios Pavlidis and Farhat Ali, "Computer Recognition of Handwritten Numerals by Polygonal Approximations," IEEE Transactions on Systems, Man, and Cybernetics SMC-5, 610-614 (1975).

63. V. Michael Powers, "Pen Direction Sequences in Character Recognition," Patter Recognition 5, 291-302 (1973).

64. S.N.S. Rajasekaran and B. L. Deekshatulu, "A Sequential Template Matching Procedure for the Recognition of Complex Line Patterns," in [2], 109-110.

65. Azriel Rosenfeld, "Picture Processing: 1974," Computer Graphics and Image Processing 4, 133-135 (1975).

66. Azriel Rosenfeld, "Picture Processing: 1975," Computer Graphics and Image Processing 5, 215-237 (1976).

67. Azriel Rosenfeld, "Picture Processing: 1976," Computer Graphics and Image Processing 6, 157-183 (1977).

68. Kunio Sakai et al, "An Optical Chinese Character Reader," in [5], 122-126.

69. John W. Sammon, Jr., et al, "Handprinted Character Recognition Techniques," RADC-TR-70-206, Rome Air Development Center, October 1970.

70. John W. Sammon et al, "Handprinted Character Recognition," RADC-TR-72-329, Rome Air Development Center, January 1973.

71. Kenneth M. Sayre, "Machine Recognition of Handwritten Words: A Progress Report," Pattern Recognition 5, 213-228 (1973).

72. Jürgen Schürmann, "Multifont Word Recognition System With Application to Postal Address Reading," in [5], 658-662.

73. I. K. Sethi and B. Chatterjee, "Machine Recognition of Constrained Handprinted Devanagari," Pattern Recognition 9, 69-75 (1977).

74. R. J. Shillman, T. T. Kuklinski and B. A. Blesser, "Experimental Methodologies for Character Recognition Based on Phenomenological Attributes," in [2], 195-201.

75. R. Shurna, A. Lashas and J. Gvildys, "Preprocessing in Optical Character Recognition Systems," in [2], 394-395.

76. A. A. Spanjersberg, "Combinations of Different Systems for the Recognition of Handwritten Digits," in [2], 208-209.

77. A. A. Spanjersberg, "Experiments With Automatic Input of Handwritten Numerical Data Into a Large Administrative System," in [4], 476-478.

78. William Stallings, "Approaches to Chinese Character Recognition," Pattern Recognition 8, 87-98 (1976).

79. James C. Stoffel, "A Classifier Design Technique for Discrete Variable Pattern Recognition Problems," IEEE Transactions on Computers C-23, 428-441 (1974).

80. L. Stringa, "MARS: A Hand and Machine-Written Address Reader Implemented on a Multimini Based Associative Processor," in [2], 443-447.

81. C. Y. Suen et al, "Reliable Recognition of Handprint Data," in [3], 98-102.

82. Ching Y. Suen and Robert J. Shillman, "Low Error Rate Optical Character Recognition of Unconstrained Handprinted Letters Based on a Model of Human Perception," IEEE Transactions on Systems, Man, and Cybernetics SMC-7, 491-495 (1977).

83. C. Y. Suen, R. Shinghal and C. C. Kwan, "Dispersion Factor: A Quantitative Measurement of the Quality of Handprinted Characters," in [6], 681-685.

84. Philip H. Swain and Hans Hauska, "The Decision Tree Classifier: Design and Potential," IEEE Transactions on Geoscience Electronics GE-15, 142-147 (1977).

85. Kokichi Tanaka, "Some Studies on Parallel Processing for Character Recognition," in [1], 179.

86. Kokichi Tanaka and Kazumasa Ozawa, "Imprinted Character Recognition Using Template Matching on Spatial Complex Frequency Plane-Especially for Katakana Letters and Numerals," in [5], 663-668.

87. D. E. Troxel, "Feature Selection for Low Error Rate OCR," Pattern Recognition 8, 73-76 (1976).

88. N. D. Tucker and F. C. Evans, "A Two-Step Strategy for Character Recognition Using Geometrical Moments," in [2], 223-225.

89. J. R. Ullmann, "Picture Analysis in Character Recognition," in A. Rosenfeld, Ed., Digital Picture Analysis, Springer-Verlag, Berlin, 1976.

90. Paul P. Wang, "The Topological Analysis, Classification, and Encoding of Chinese Characters for Digital Computer Interface-Part II," in [1], 180-186.

91. Paul P. Wang and Robert C. Shiau, "Machine Recognition of Printed Chinese Characters via Transformation Algorithms," Pattern Recognition 5, 303-321 (1973).

92. S. Wendling and G. Stamon, "Hadamard and Haar Transforms and Their Power Spectra in Character Recognition," in [3], 103-112.

93. Joseph Yacyk, "Alphabetic Handprint Reading," in [4], 485-486.

94. S. Yamamoto, A. Nakajima and K. Nakata, "Chinese Character Recognition by Hierarchical Pattern Matching," in [1], 187-196.

95. Makoto Yoshida and Murray Eden, "Handwritten Chinese Character Recognition by an Analysis-by-Synthesis Method," in [1], 197-204.

96. M. Yoshida et al, "The Recognition of Handprinted Characters Including 'Katakana,' 'Alphabets' and 'Numerals,'" in [5], 645-649.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NORDA Technical Note 217	2. GOVT ACCESSION NO. A131341	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) An Overview of Optical Character Recognition (OCR) Technology and Techniques		5. TYPE OF REPORT & PERIOD COVERED Final 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) L.K. Gronmeyer, B.W. Ruffin, M.A. Lybanon, P.L. Neeley, and S. E. Pierce Jr.		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Computer Sciences Corporation NSTL Station, Mississippi 39529		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS PE 63701B
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Ocean Research and Development Activity Ocean Science and Technology Laboratory NSTL Station, Mississippi 39529		12. REPORT DATE June 1978 13. NUMBER OF PAGES 87
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release, distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Optical Character Recognition OCR Mapping, Charting, and Geodsey

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report presents the results of a survey of 1978 Optical Character Recognition (OCR) technology conducted by NORDA Code 302, the Mapping, Charting, and Geodsey Development Group. Three principal areas of OCR technology development were reviewed:

- Government application of OCR.
- Commercial OCR products.
- Software and basic research.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

END

FILMED

9-83

DTIC