DOE/NV/10872-T17

DOE/NV/10872--T17

DE92 014675

UNLV Information Science Research Institute

Quarterly Progress Report

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

T. A. Nartker
December 31, 1991

Table of Contents

APPENDICIES

UNLV Information Science Research Institute: Quarterly Progress Report

T. A. Nartker

December 31, 1991

I. Board of Advisors Activity

None.

II. Symposium Activity

Plans for our Symposium are proceeding well. Twenty-two papers have been accepted and an advanced program has been printed and is being mailed. A copy is attached as Appendix A.

III. Staff Activity

Recruiting

We have employed two new Software Engineer's. Mrs. Julie Borsack has accepted a position to manage text retrieval software systems and Mrs. Debbie Wallace will manage our GT1 database as well as new foreign language databases. Both will begin work in January.

Travel

T. Nartker attended the First International Conference on Document Analysis and Retrieval held in St. Malo, France in October.

Also in October, T. Nartker visited the Irish Science and Technology Agency (EOLAS), in Dublin, Ireland.

K. Taghva and T. Nartker attended the annual ACM SIGIR conference in Chicago in October. Kazem also attended the NSF workshop on Text-Retrieval which preceded the conference.

J. Kanai made a three week intensive tour of Japan to survey Japanese activity in both industrial and academic organizations in OCR. A summary report will be available in the next few weeks.

K. Taghva, J. Minor, T. Nartker, and S. Rice traveled to Washington, DC in December to participate in discussions at the DOE concerning testing of the JSPACE enhanced version of the BasisPlus text-retrieval system.

Papers accepted or presented

In October, R. Bradford and T. Nartker presented their paper titled "Error Correlation in Contemporary OCR Systems" at the International Conference on Document Analysis and Recognition in St. Malo, France.

Also in October, T. Nartker gave an invited presentation to the Information and Space Technologies staff at EOLAS (Ireland) on ISRI research.

Also in October, K. Taghva and T. Nartker presented an invited paper titled "Applications of Information Retrieval Technology: A Position Paper," at the NSF Workshop on Future Directions in Text Analysis, Retrieval, and Understanding, which was held in Chicago Illinois.

T. Nartker, J. Kanai, and S. Rice have submitted a paper titled "A Preliminary Report on OCR Problems in LSS Document Conversion" to the 2nd annual Nuclear Waste Management Conference to be held in Las Vegas in April of 1992. The paper has been accepted. A copy is attached as Appendix B.

IV. Document Analysis Program

OCR Devices

We have acquired copies of three software OCR systems during the subject quarter. These are:

Product	Company	Available Toolkit
Perceive	Ocron Corporation	Yes
Recognita Plus	Recognita Corporation	Yes
Type Reader	Expervision	Promised

Steve Rice is installing the first two of these at the the current time. Expervision promises to deliver a toolkit in January. We have finally made a contact within the Caere Corporation who has indicated a desire to provide us a copy of Omnipage PRO 2.0 which we can install and test. Mr. Serge Blanc at Caere has stated that, although a toolkit does not now exist, he hopes to provide a special version of their system tailored for our needs sometime this spring.

OCR Test system

Steve Rice has completed an enhanced version of our zoning utility. He continues to work with Allen Condit on the communications link portion (Ethernet TCP\IP) of the OCR server.

OCR Databases/GT1

With help from Lois Dickey, we have identified 38 tapes which contain LSS data not received last Fall. Andrew Bagdanov is making copies of these and returning the original reels to the DOE. Lois will work with Deborah Wallace to develop a salvage plan. Significant progress will require someone full-time (i.e. Debbie) to manage the effort.

OCR Databases/Foreign languages

R. Bradford, J. Kanai, F. Jenkins, & T. Nartker are conducting a literature survey of published work on OCR in Japanese, Cyrillic, Chinese, and Arabic. They are also investigating sources of material for Ground-Truth testing in each of these languages.

During the subject period, Dr. Huei Chuang at IDI provided a set of 5 pages of printed Chinese documents and corresponding text on diskette. A copy of these pages is attached as Appendix C. The number of characters is small, but they are our first non-English ground-truth database.

We expect to begin receiving other foreign language databases from ORD during the next quarter. Deborah Wallace will manage these databases.

Dr. Junichi Kanai visited Japan in December and made several valuable contacts there. Although the Japanese have developed databases for OCR research, they are of limited use for ORD test purposes. He suggests that data for ground-truth testing of Japanese OCR devices will need to be tailored to ORD requirements (see Dr. Kanai's Survey of Japanese OCR Technology).

OCR Experiments

None. Further experiments have been postponed until we have more devices to test and until more of our GT1 test data has been verified. The Caere Omnipage Board is currently not a competitive product. The current Caere software system (Omnipage PRO 2.0) is known to be a vastly superior product but, so far, it cannot be installed in our system. Similarly, the Toshiba Express Reader is not reliable enough to be a competitive OCR product. Until we install other competitive OCR (software) systems, we have only the Kurzweil and the Calera to test. Thus, we have chosen to postpone further device testing until more of the competing devices are available. We hope to begin our next round of device testing shortly after our Symposium in March.

OCR Technical reports/thesis

Mr. Tom Kurilla finished his M.S. Project paper, and passed his oral examination, in December. Tom submitted a literature survey on "Skew Detection and Page Orientation Methods in Optical Character Recognition." Tom's paper is attached as Appendix D.

Kevin Grover and Steve Rice have produced a report titled "GT0.5: Sample #1 Voting errors." Each character error produced by the majority voting algorithm on Sample #1 is shown. (all 3596 errors) For each word in error, the report shows the image of the word blown up by a factor of three, followed by the correct ASCII text, followed by the text output from the voting algorithm. On one hand, the report gives a very detailed view of the kinds of errors made by current devices. On the other, it demonstrates our ability to automatically synchronize text streams and to count recognition errors.

Interaction with OCR vendors

Representatives from several OCR companies visited us during the quarter. Mr. Denis Coleman, from Cognitive Technology Corporation visited in October. Mr. Dave Emmett and Ms. Mindy Bokser from Calera visited us in December. Dr. Sam Ang, from Expervision also visited in December. All three companies have expressed an interest in utilizing our experimental facilities in testing future versions of their product.

Interaction with OCR research organizations

Dr. Andreas Dengel from the German research center for artificial intelligence visited us in November. Dr. Dengle gave a seminar, titled "How to Build New Generation Reading Machines." In the future, there is some possibility of interaction with research projects at his center. Currently, he is focused on the automatic recognition and processing of business letters.

V. Text-Retrieval Program

As mentioned in previous reports, our text-retrieval program is scheduled to begin activities in 1992.

TR Software systems

Julie Borsack will begin installing the BasisPlus text retrieval system in January. She will be responsible for supporting text-retrieval software systems and associated experimental system development.

TR Test system

Installation of the SUN SPARCstation IPC, our text-retrieval file server (designated Champion), is complete.

TR Databases

No activity.

TR Experiments

Bryan Bullard, Shanti Kumar, and Xinnong Yang have continued their literature review. They are looking at Postgres and at LQ-text.

TR Technical reports/thesis

None

Interaction with TR vendors

In December, we met Dr. Earlene Busch of Information Access Systems and learned more about the JSPACE product during our visit to the DOE in Washington DC (see travel). Also, Bob Young from Centech visited us in November to keep us informed about their progress in adding JSPACE to BasisPlus. Our effort to help evaluate the potential DOE use of JSPACE continues.

VI. Institute Activity

Institute visitors

Date	Visitor	Agency
10/23/91	Dr. Alan Blandamer	Vitro Corporation
10/23/91	Mr. John Soloman	Input Solutions Inc.
10/23&24	Mr. Roger Bradford & Ms. Lois Dickey	SAIC
10/24/91	Mr. Denis Coleman	CTC Corporation
10/24/91	Mr. Robert Young	The Centech Group
11/06/91	Mr. Dan Graser	DOE
11/14/91	Dr. Maxim Bohlmann	Xerox Corporation
11/18/91	Dr. Andreas Dengel	German research center for artificial intelligence
11/18/91	(6 Russian Scientists)	Star Treaty, On-site team
12/03/91	Mr. Val Kerrigan & Dr. Wen-wu Shen	Federal Bureau of Investigation
12/04/91	Mr. Dave Emmett, VP Eng. & Ms. Mindy Bokser	Calera Corporation.
12/18/91	Dr. Sam Ang, President	Expervision

New agency contacts/ new research proposals

The following special contacts were made during the subject period:

Date	Agency	Contact
11/08/91	Cray Research	Dr. Bahram Nassersharif
11/14/91	Xerox Special Info. Services	Dr. Maxim Bohlmann
12/03/91	FBI	Mr. Val Kerrigan & Dr. Wen-wu Shen

Dr. J. Kanai & Dr. S. Latifi have submitted a proposal to Cray Research, Inc. for funds to conduct a research project in Arabic (and Farsi) OCR. They propose to attempt to segregate Arabic characters (and/or to recognize words) using neural-nets on the UNLV Cray Y/MP. A copy of this proposal is attached as Appendix E. We are optimistic about receiving funds from Cray for this project. Awards will be announced in early 1992.

Through our association with Larry Spitz, Dr. Maxim Bohlmann at Xerox Special Information Services has contacted us and indicated their interest in supporting our research. Xerox is willing to provide several different software systems, free of charge, which might be of significant value. We plan to visit Xerox in Pasadena and at Xerox PARC in January.

The FBI has a requirement to evaluate current OCR technologies which can accept Spanish language documents. During their visit, we discussed several types of projects in which we could be of significant help. Further discussions are expected early in 1992.

VII. Goals Achieved/Goals for Next Quarter

Goals from last quarter:

1) Continued recovery of prototype data has been postponed until Deborah Wallace can supervise this effort.

2) We have acquired our first Ground-Truth dataset for foreign language testing. (see Appendix C) We are investigating software available from Xerox to manage foreign language databases.

3) See report titled "GT0.5: Sample #1 Voting errors."

4) Perceive, Recognita Plus, and Type Reader have been received and are being installed.

5) OCR server architecture is in progress.

6) Julie and Deborah will start in January.

7) Early discussions with the FBI were promising. We have hopes that our experimental system can serve FBI needs.

Goals for next quarter:

1) Conduct our first "Symposium in Document Analysis & Information Retrieval."

2) Complete the orientation of Deborah Wallace and Julie Borsack. Begin definition of expanded data salvage and text-retrieval programs.

3) Acquire and install Omnipage Professional from Caere (& new product from CTC Corp. & LiOCR from Ligature) and expand device evaluation tests.

4) Complete version 2 of the vendor-independent interface.

5) Continue to investigate the availability of foreign language databases and software to support multiple-language databases simultaneously. We expect to begin receiving other foreign language databases for ground-truth testing from ORD during the next quarter.

6) J. Kanai & T. Nartker will again offer a graduate course on Pattern & Character Recognition during the Spring semester. Dr. G. Nagy, from RPI, will be spending a portion of the Spring visiting our Institute. Dr. Nagy will present a series of guest lectures on OCR. We are counting on his help in defining our long term research program in OCR.

7) Continue to pursue additional sources of support for institute research.

APPENDIX A.

Preliminary Technical Program:

"Symposium on Document Analysis and Information Retrieval"

Symposium on Document Analysis and Information Retrieval

March 16 - 18, 1992

Tropicana Hotel
Las Vegas, Nevada

Sponsored by the
Information Science Research Institute
and
The Howard R. Hughes College of Engineering
University of Nevada, Las Vegas

CONFERENCE SCHEDULE

Sunday, March 15, 1992

7:00pm - 11:00pm
Reception South Pacific 3

Monday, March 16, 1992

7:30am - 8:30am
Registration South Pacific 6

8:30am - 8:45am
Welcome South Pacific 6

Thomas A. Nartker, Director
Information Science Research Institute
Howard R. Hughes College of Engineering
University of Nevada, Las Vegas

Robert C. Maxson, President
University of Nevada, Las Vegas

William R. Wells, Dean
Howard R. Hughes College of Engir.eering
University of Nevada, Las Vegas

8:45am - 9:45am
Invited Session South Pacific 6

What Does a Machine Need to Know to Read a Document
George Nagy, Rensselaer Polytechnic Institute

9:45am - 10:00am
Refreshment Break South Pacific 6

10:00am - 12:00pm
Document Analysis Session I South Pacific 6

Skew Determination in CCITT Group 4 Compressed Document Images
A. Lawrence Spitz, Xerox Palo Alto Research Center

Visual Global Context: Word Image Matching in a Methodology for Degraded Text Recognition
Jonathan J. Hull, Siamak Khoubyari, and Tin Kam Ho, State University of New York at Buffalo

On Printed Music Score Symbol Recognition
Bharath R. Modayur, Robert M. Haralick and Linda G. Shapiro, University of Washington

10:00am - 12:00pm (CONTINUED)
Document Analysis Session I South Pacific 6

Document Structure Interpretation by Integrating Multiple Knowledge Sources
Suzanne Taylor, Mark Lipshutz and Carl Weir, Unisys Center for Advanced Information Technology

A Framework of Layout Recognition for Document Understanding
Toyohide Watanabe, Qin Luo, Kazue Sugino, Nagoya University, Japan; and Teruo Fukumura, Chukyo University, Japan

12:00pm - 1:30pm
Lunch (no host)

1:30pm - 2:45pm
Invited Session South Pacific 3

Retrieval From Large Text Databases
Bruce Croft, University of Massachusetts, Amherst

2:45pm - 3:00pm
Break

3:00pm - 5:00pm
Text Retrieval Session I South Pacific 3

A New Method for Full-Text Retrieval
Nassrin Tavakoli and Alan Ray, University of North Carolina at Charlotte

Using Conceptual Graphs for Information Retrieval: A Framework for Adequate Representation and Flexible Inferencing
Sung H. Myaeng, Syracuse University

Document Analysis Using Attribute Grammars
Karen A. Lemone, Worcester Polytechnic Institute

Adaptive Information Retrieval Systems in Vector Model
Jing-Jye Yang and Robert R. Korfhage, University of Pittsburgh

Filtering the Pravda With a Self-Organizing Neural Net
J. C. Scholtes, University of Amsterdam

6:30pm - 11:00pm
Dinner Great Hall
Tour of Facilities Beam
Engineering Bldg.
UNLV

Tuesday, March 17, 1992

8:30am - 9:45am
Invited Session South Pacific 6

Problems in the Recognition of Poorly Printed Text
Theo Pavlidis, State University of New York at Stony Brook

9:45am - 10:00am
Refreshment Break South Pacific 6

10:00am - 12:00pm
Document Analysis Session II South Pacific 6

Incorporation of a Markov Model of Language Syntax in a Text Recognition Algorithm
Jonathan J. Hull, State University of New York at Buffalo

Aspects of Knowledge Based Document Analysis
Thomas Bayer and Eberhard Mandler, Daimler Benz Research Center, Germany; Frank Hones and Andreas Dengel, German Research Center for Artificial Intelligence, Germany

Automatic Book Recognition
Luigi Stringa, Istituto per la Ricerca Scientifica e Tecnologica, Italy

Character Extraction From Half-Tone Background
Hong Yan, University of Sydney, Australia

A Structure Recognition Method for Japanese Newspapers
Qin Luo, Toyohide Watanabe and Noboru Sugie, Nagoya University, Japan

12:00pm - 1:30pm
Lunch (no host)

1:30pm - 2:45pm
Invited Session South Pacific 3

Building a User-Centered Database From the ACM Literature
Ed Fox, Virginia Polytechnic Institute and State University

2:45pm - 3:00pm
Break

3:00pm - 5:00pm
Text Retrieval Session II South Pacific 3

The Design and Development of a Database Model to Support Non-Linear Document Management
John A. Gawkowski, George Raudabaugh, Information Dimensions, Inc., Dublin, Ohio

3:00pm - 5:00pm (CONTINUED)
Text Retrieval Session II South Pacific 3

Design of Search Trees for Efficient Querying of Bit-String Oriented Information Systems
Ray Hashemi, University of Arkansas at Little Rock

Exploiting Linguistic Patterns for Information Retrieval
Sung H. Myaeng and Elizabeth D. Liddy, Syracuse University

An Investigation of a Conceptual Map of a Text Database for Retrieval
D. V. Rama, Bentley College; Padmini Srinivasan, University of Iowa

Using Classification in a Hypertext Information Retrieval System
C. Chrisment, C. Comparot, C. Julien, F. Sedes, C. Soule-Dupuy, Institute de Recherche in Informatique de Toulouse - Universite Paul Sabatier, France

Wednesday, March 18, 1992

8:30am - 9:00am
Host Presentation South Pacific 6

A Preliminary Report on UNLV/GT1: A Document/Image Library for Ground Truth Testing in Document Analysis and Character Recognition
Roger Bradford, SAIC; Barbara Cerny, DOE; Thomas Nartker, UNLV

9:00am -10:30am
Panel Session South Pacific 6

Databases for Research and Testing in Document Analysis and Information Retrieval

Moderator:	Roger Bradford
Panel Members:	Bruce Croft; Ed Fox; Jonathan Hull; George Nagy; Theo Pavlidis

10:30am - 10:45am
Refreshment Break South Pacific 6

10:45am - 11:30am
Document Analysis and Text Retrieval Session III South Pacific 6

Document Recognition Using Qualitative Reasoning
Hiroko Fujihara and Amit Mukerjee, Texas A & M University

Selection Criteria for the Text in Text-Based Systems
Alka Irani, National Centre for Software Technology, India

Invited Speakers

Bruce Croft is a Professor in the Department of Computer Science at the University of Massachusetts, Amherst, which he joined in 1979. He received the B.Sc. (Honours) degree in 1973, and an M.Sc. in 1974 from Monash University in Melbourne, Australia. His Ph.D. degree (in Computer Science) was from the University of Cambridge, England in 1979. His research interests are in formal models of retrieval for complex, text-based objects, text representation techniques, the design and implementation of text retrieval systems, computer-supported cooperative work, and user interfaces. He has published more than 60 articles on these subjects.

Dr. Croft was Chair of the ACM Special Interest Group on Information Retrieval from 1987 to 1991. He is an Associate Editor of the ACM Transactions on Information Systems and Information Processing and Management, and is a member of the editorial board of Knowledge Acquisition. He has served on numerous program committees and has been involved in the organization of many workshops and conferences.

Edward Fox received his B.S. from MIT and his M.S. and Ph.D. degrees from Cornell University. He serves as Associate Professor in the Department of Computer Science, and Associate Director for Research in the Computing Center, both at Virginia Tech (VPI&SU). His work with ACM includes: chair of SIGIR, member of the Publications Board, and Associate Editor for Transactions on Information Systems; previously he served as SIGIR vice chair, editor-in-chief of ACM Press Database and Electronic Products, and chair of the Electronic Publishing Committee. He began work in 1985 on CD-ROM, started in 1988 as project director for the Interactive Digital Video effort at VPI&SU (as a beta test site for DVI), was project director for the Virginia Disc series of CD-ROMs, and is involved in research and teaching regarding information storage and retrieval CD-ROM and optical discs, multimedia, hypertext and hypermedia, online public access catalogs, hashing, and computational linguistics. He now directs the Envision Project, a large National Science Foundation funded effort to build "A User-Centered Database from the Computer Science Literature," which will use ACM publications and deliver text and multimedia information.

George Nagy received the B.Eng. and M. Eng. degrees from McGill University, and the Ph.D. in Electrical Engineering from Cornell University in 1962 (on neural networks). For the next ten years he conducted research on various aspects of pattern recognition and OCR at the IBM T. J. Watson Research Center in Yorktown Heights. From 1972 to 1985 he was Professor of Computer Science at the University of Nebraska - Lincoln, and worked on remote sensing applications, geographic information systems, computational geometry, and human-computer interfaces. Since 1985, he has been Professor of Computer Engineering at Rensselaer Polytechnic Institute. He has held visiting appointments at the Stanford Research Institute, Cornell, the University of Montreal, the National Scientific Research Institute of Quebec, the University of Genoa and the Italian National Research Council in Naples and Genoa, AT&T Bell Laboratories, IBM Almaden, and McGill University. In addition to digitized document analysis and character recognition, his interests include solid modeling, finite-precision spatial computation, and computer vision.

Theo Pavlidis is a leading professor of Computer Science at the State University of New York at Stony Brook and a scientific advisor to Symbol Technologies. He received a Ph.D. in Electrical Engineering from the University of California at Berkeley in 1964. He has authored more than 150 technical papers and three books, including *Algorithms for Graphics and Image Processing,* (Computer Science Press, 1982) which has been translated into Chinese, German, Polish, and Russian. His general research interests are in the areas of Image Analysis, Pattern Recognition, and Computer Graphics. His current research is focused on optical character recognition and related problems of document processing, as well as, in problems arising from bar coding applications.

He was the editor-in-chief of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI) from 1982 to 1986, and has been a member of the editorial board of many other journals. He is the program chairman of ICDAR '93 (International Conference on Document Analysis and Recognition) and has served in the past as general chairman of the *Fifth International Conference on Pattern Recognition* (1980) and the 1988 *IEEE Conference on Robotics and Automation.* He is a Fellow of IEEE, and a member of ACM, and Sigma-Xi.

HOTEL REGISTRATION FORM

INFORMATION SCIENCE RESEARCH INSTITUTE
UNLV COLLEGE OF ENGINEERING

March 16-18, 1992
SUNDE

Symposium on Document Analysis and Information Retrieval

Reservations received after **February 16, 1992** will be accepted on a space available basis only.

Please reserve accommodations for:

NAME: ______________________________

COMPANY: ______________________________

ADDRESS: ______________________________

CITY: ______________ STATE: ______________ ZIP: __________

SHARING ROOM WITH: ______________________ NO OF PERSONS: __________

SIGNATURE: ______________________ PHONE NUMBER: __________

ARRIVAL DATE: ______________ TIME: __________ DEPARTURE: __________

The TROPICANA RESORT HOTEL AND CASINO can only confirm your reservation request when accompanied by one night's deposit (room rate, plus 8% Clark County Room Tax). This deposit may be made by check, money order or by American Express. If paying by check or money order, please include the arrival date on the face of the check. Refunds will be made when cancellations are received no less than two (2) days prior to your scheduled arrival date (be sure to keep your cancellation number).

AMERICAN EXPRESS CARD NUMBER: ______________________________

NAME AS IT APPEARS ON CARD: ______________________________

EXPIRATION DATE: ______________________________

$ 55.00 (plus tax)	Single/Double
$ 65.00 (plus tax)	Triple
$175.00 (plus tax)	One-Bedroom Suite
	(Clark County Room Tax is 8%)

Please return this reservation request to:

TROPICANA ROOM RESERVATIONS DEPARTMENT
P. O. Box 97777
Las Vegas, NV 89195-7777

NO SATURDAY ARRIVALS ARE AVAILABLE - PHONE: (800)634-4000

Symposium on Document Analysis and Information Retrieval

INFORMATION SCIENCE RESEARCH INSTITUTE
UNLV COLLEGE OF ENGINEERING
March 16-18, 1991

Conference Registration Form

Name: ______________________________

Title: ______________________________

Company: ______________________________ ______________________________

Address: ______________________________

City: ____________________ State: ____________ Zip: ____________

Telephone Number: (____) ______________________________

Registration Fees	Pre-Reg before 2/24/92	Regular after 2/24/92	Amount
Conference Registration (includes dinner Monday, 3/16/91)	$300.00	$375.00	$________
Dinner (Spouse/Companion)		$ 10.00	$________
Conference Proceedings (Extra)		$ 40.00	$________

SPECIAL REQUESTS: (Special meals are only available to pre-registered attendees)

Kosher Meal ____________ Vegetarian Meal ____________

Make checks/money orders payable to: UNLV Board of Regents

Mail completed conference registration form and check/money order to:

Mary C. Guirsch
Information Science Research Institute
University of Nevada, Las Vegas
4505 Maryland Parkway
Las Vegas, NV 89154

All checks/money orders should be in U. S. Dollars

APPENDIX B.

A Preliminary Report on OCR Problems in LSS Document Conversion

DRAFT: 11/25/91 To be presented at the Nuclear Waste Management Conference, Las Vegas, NV, April 1992.

A PRELIMINARY REPORT ON OCR PROBLEMS IN LSS DOCUMENT CONVERSION

T. A. Nartker, J. Kanai, and S. V. Rice
Information Science Research Institute
University of Nevada, Las Vegas

INTRODUCTION

To support licensing proceedings of the Nuclear Regulatory Commission, the Department of Energy is planning to construct a computer database of pertinent documents which can be accessed by all interested parties.[1,2,3] This computer system, called the Licensing Support System (LSS), is projected to cost $200,000,000 over a ten year life.[4]

One of the major costs associated with building this system is the cost of ASCII-text cleanup for documents processed by an "Optical Character Recognition" (OCR) device. The purpose of the current study has been to identify the predominant problems which contemporary OCR devices have in converting LSS document images into computer readable (ASCII) text. If these problems can be identified and corrected, it will be possible to save a large fraction of the projected cost.

DESCRIPTION

The present study is a first step in our attempt to understand the state of current OCR technology in a general, but quantitative way. We aspire to characterize the most difficult OCR problems by careful study of the errors made by several contemporary devices in recognizing text from a large sample of "typical" documents.

There are three different aspects to such an investigation. First, it is desirable to study the predominant errors made by all current systems and not just errors made by one or two vendor products. Second, it is desirable to select a large test data set which is "typical" of the eventual LSS input. Finally, it is desirable to have a quantitative measure of the frequency of different kinds of recognition problems.

In this study, we characterize "current OCR technology" not by any one OCR vice but instead by a combination of several devices. That is, the OCR errors we port are errors made by the voting scheme proposed in Bradford[6] and thus are, in me sense, errors made by the majority of a set of devices. The actual set of OCR vices used were the Calera RS9000, the Kurzweil 5200, the Toshiba ExpressReader, d the Caere OmniPage OCR board.

Second, UNLV has installed a version of the LSS prototype database (called GT1) r use in Ground-Truth testing of OCR technologies.[7] Because of the diversity of LSS ocument types (and the prevalence of non-original copies) these pages are a oderately difficult OCR input set. In this report, we study OCR errors made on a ndom sample of 240 pages from the GT1 database. These pages contained 278,786 SCII characters.

Finally, we enumerate a set of problem categories which can be identified by anual inspection of the input image. Thus, we quantify the state of "current OCR chnology" in terms of the percentage of total OCR errors caused by each different oblem category on the 240 page sample. In the tables below, we present the edominance of problems found with both ASCII and non-ASCII input text.

ESULTS

Table 1 shows the number of ASCII character errors (and the fraction of total SCII character errors) which were manually assigned to each problem category. e nine categories listed at the top of Table 1 are problems judged to have been used by flaws in the image. The categories at the bottom of Table 1 represent oblems not associated with flaws in the image but instead with inherently difficult aracter recognition situations.

Table 2 shows the number of occurrences of subscripts, superscripts, and non-SCII characters encountered in processing the 240 page sample. It is clear that, at ast for this sample, the effort needed to cleanup non-ASCII characters is small mpared to the effort needed to correct erroneous ASCII output.

Thus Table 1 shows the distribution of OCR problems found considering only e ASCII input characters. Most users of OCR devices are familiar with the problem tegories shown. The most surprising aspect of these distributions is the dominance errors caused by broken characters.

Problem Category	Number of Errors	Fraction of Total Errors
Broken characters	1872	52.1
Touching characters	734	20.4
Noise/ speckle	122	3.4
Skew (or curved baseline)	49	1.4
Broken & touching	186	5.2
Broken & noise	9	0.3
Broken & skew	33	0.9
Touching & noise	2	0.1
Touching & skew	10	0.3
Total	3017	83.9
Similar symbols (1,l O,0)	207	5.8
Wrong case	12	0.3
Stylized characters	46	1.3
Introduced spaces	79	2.2
Dropped spaces	39	1.1
Unknown cause	196	5.5
Total	579	16.1
Grand Total	3596	100

TABLE 1. Distribution of Estimated ASCII-OCR Problems

Problem Category	Number of Occurrences
Subscripts	157
Superscripts	127
Total sub & superscripts	284
Greek letters	69
Degree symbol	27
Bullet	20
Times symbol	17
Long hyphen/dash	13
Other (21 different problems)	64
Total Non-ASCII	210

TABLE 2. Number of Occurrences of Subscripts and Non-ASCII Characters

CONCLUSIONS

The results presented indicate that efforts to improve OCR technologies, at least from the point of view of the LSS, might provide substantial return if concentrated on a very few image problems. Based on this sample of pages, it is clear that a very large percentage of current OCR problems are associated with either broken or touching characters. All other problem categories combined contribute less than 25% of the ASCII conversion error measured.

Of the $200 million projected cost of the LSS, about one-half is associated with data acquisition.[4] Of this, approximately 60% is required for ASCII cleanup.[5] If the results shown in Table 1 are characteristic of the entire LSS database, the potential DOE savings from eliminating 100% of the broken character errors would be $31,260,000. Although eliminating 100% of these errors is probably not a realizable goal, it is clear that fractional improvements could result in significant savings for the DOE.

One of the most significant aspects of this study is the size of the test data set. Because of the effort involved in performing such tests, it is common to find papers comparing OCR technologies which are based on far fewer character conversions.[8] Even though the present study is based on a small sample of the GT1 database, the actual number of characters converted was 278,786. We plan to continue this study on the entire GT1 database to verify this result. At the same time, we plan to explore the use of image enhancement techniques on these data to determine overall effects and to confirm the potential for savings.

REFERENCES

[1] DOE Office of Civilian Radioactivc Waste Management, "Licensing Support System - Preliminary Needs Analysis," Feb. 1988.

[2] DOE Office of Civilian Radioactive Waste Management, "Licensing Support System - Preliminary Data Scope Analysis," Mar. 1988.

[3] DOE Office of Civilian Radioactive Waste Management, "Licensing Support System - Conceptual Design Analysis," May 1988.

[4] DOE Office of Civilian Radioactive Waste Management, "Licensing Support System - Benefit-Cost Analysis," July 1988.

[5] SAIC, "Capture Station Simulation & Lessons Learned: Final Report," DOE Contract #DE-AC01-87RW00084, 12, Nov. 1990.

[6] R. Bradford and T. Nartker, "Error Correlation in Contemporary OCR Systems," **Proc. First International Conference on Document Analysis and Information Retrieval**, St. Malo, France, Sept. 1991, pp 516-523.

[7] R. Bradford, B. Cerny, and T. Nartker, "A PRELIMINARY REPORT ON UNLV/GT1: A Document/Image Library for Ground Truth Testing in Document Analysis and Character Recognition," **Proc. First Symposium on Document Analysis and Information Retrieval**, Las Vegas, NV, March 1992.

[8] D. McClelland, "OCR - Teaching Your Mac to Read: Eight OCR packages recognized and ranked," MACWORLD, Nov. 1991, pp 169-175.

APPENDIX C.

Candidate Pages for Ground-Truth testing

with Chinese OCR devices.

瓜 科 Cucurbitaceae

絲 瓜

學名：Luffa cylindrica (L.)Roem.

別名：虞刺、洗鍋羅瓜、天絲瓜、天羅、蠻瓜、綿瓜、天羅瓜、菜瓜、水瓜、布瓜、
吊瓜、魚茲、天絡絲、倒陽菜、純陽瓜、天羅絮、天羅布瓜、紡線、縑瓜、絮瓜、蜀瓜、
瓜、米管瓜、竹篙瓜。

形態：本省栽培供蔬食之一年生攀援草本。幼時全株密被柔毛，老時近於無毛。莖柔
，具稜角，長可達 7～10 m；卷鬚 2～4 歧。葉互生，具長柄，三角形或近圓形，長 8～
cm，寬 15～25 cm，掌狀 3～7 裂，裂片三角形；邊緣具細齒；葉柄長 4～9 cm 。花單
，雌雄同株；雌花單生，具長梗，雄花為總狀花序，萼5深裂，裂片卵狀披針形；花冠5
裂，裂片廣倒卵形，黃色或淡黃白色；雄蕊5，花絲分離；子房下位，圓柱形，柱頭3，
大。瓠果長圓柱形，下垂，長 18～60 cm，具縱向淺槽或條紋。種子長卵形而扁，黑色，
緣有翅。(彩圖 289)

藥用部分：根、藤、葉、花、果肉、果皮、瓜蒂、種子、老瓜內的纖維、及莖中汁。

效用：1.按本草綱目卷二十八：瓜，甘、平、無毒，主治痘瘡不快，枯者燒存性，入
砂研末，蜜水調服，甚炒(震亨)。者食，除熱利腸，老者燒存性服，去風化痰，涼血解毒，
殺蟲，通經絡，行血脈，下乳汁，治大小便下血，痔漏崩中，黃積疝痛卵腫，血氣作痛，
疽瘡腫，齒䘌，痘疹胎毒(時珍)。

葉主治癬瘡，頻挼摻之，療癰疽丁腫卵癩(時珍)。

藤根主治齒䘌腦漏，殺蟲解毒(時珍)。

2.莖葉治瘡毒(佐佐木，1924)。

3.絲瓜皮治金瘡，疔瘡，坐板瘡(中藥大辭典，1982)。

4.絲瓜子利水，除熱。治肢面浮腫，石淋，腸風，痔瘻(中藥大辭典，1982)。緩下。

5.絲瓜花清熱解毒。治肺熱咳嗽，咽痛，鼻竇炎，疔瘡，痔瘡(中藥大辭典，1982)。

6.根甘、平、無毒。活血，通絡，消腫。治偏頭痛，腰痛，乳腺炎，喉風腫痛，腸風
血，痔漏(中藥大辭典，1982)。

7.絲瓜絡甘、平，通經活絡，清熱化痰。治胸脅疼痛，腹痛，腰痛，睪丸腫痛，肺熱
咳，婦女經閉，乳汁不通，癰腫，痔漏。炭：能止血，治便血、血崩(中藥大辭典，1982)

8.絲瓜藤舒筋，活血，健脾，殺蟲。治腰膝四肢麻木，月經不調，水腫，齒䘌，鼻淵，
牙宣(中藥大辭典，1982)。

9.絲瓜水(天羅水)，將絲瓜莖距地面約50公分處切斷，將切口插入瓶中，收集汁液，
鎮咳，解毒，清熱之功。

10.絲瓜葉治慢性氣管炎。

11.絲瓜蒂治咽喉腫痛。

石竹科 Caryophyllaceae

蠅子草

學名：Silene fortunei Vis.
　　　Silene kiruninsularis Masam.

別名：脫力草、古綿草、野蚊子草、水白參、蒼蠅花、小葉鯉魚膽、瞿麥沙參、白花
瓶、八月白、白花瞿麥、小仙桃草、蛇王草、白葫蘆、捕蟲古綿草、本瞿麥。

形態：自生於全省沿海至海拔 2000m 地區之多年生草本。莖直立，高 30～70 cm，
分枝。葉對生，線狀披針形或倒卵狀披針形，長 2～4 cm，寬 3～9 mm，先端銳尖，基
漸狹成細柄；全緣。聚繖花序頂生，花粉紅色或白色；苞片線形，對生，長 6～9 mm，
1 mm；花梗長 2～3.5 cm，被粘液狀腺毛；萼長管形，光滑，先端5齒，三角狀卵形；
瓣5，基部成爪，瓣片2裂，每裂片更細裂成窄條：雄蕊10；子房3室；花柱3。蒴果長
形，上部略膨大而下部狹小，呈棍棒狀，熟時先端6齒裂。種子腎形，具瘤狀突起。（彩
222）

藥用部分：全草。

效用：1.清熱利濕，補虛活血。治尿路感染，白帶，痢疾，病後體虛，扭、挫傷（中
大辭典，1982）。
　　　2.滋補強壯，解毒消腫。治體虛或病後虛弱少力，關節肌肉酸痛。

七葉一枝花

學名：Paris polyphylla Smith
　　　Paris formosana Hay.

別名：蚤休、蚩休、重樓、重台、草甘遂、螢休、紫河車、重樓金線、草河車、白甘遂、三層草、獨脚蓮、九道箍、鐵燈盞、七葉一盞燈、鶯蕎蟲、雙層樓、七子蓮、鐵燈台、多葉重樓、九層樓、枝花頭、螺絲七、海螺七、燈臺七、七層塔、八角盤、孩兒掏傘、金盤托荔枝、紅重樓、白河車、七葉遮花、金盤托珠、獨葉一枝花、金絲兩重樓、七葉蓮、青木、七厚蓮。

形態：全省山地自生之多年生直立草本。全體無毛，株高 30～120 cm 。根莖肥厚，結節明顯，黃褐色。單莖直立，青紫色或紫紅色，基部具膜質包莖葉鞘，葉 4～9片輪生莖頂，長橢圓形或橢圓狀披針形，長 15～30 cm，寬 8～10 cm ，先端漸尖，基部楔形，薄紙質，表面綠色，背面略帶紫色；全緣；基出脈3條。花單生莖頂，花梗長 3～6 cm；外列花被片 4～7片，綠色，葉狀，長橢圓狀披針形，長 6～9 cm ，寬約2cm；內列被片 4～7片，黃色或黃綠色，線形，長 2～6 cm ，寬約1 mm ，雄蕊 5～6 枚 ，與外片對生；花絲扁平；花藥線形，金黃色，縱裂；子房上位，具 4～6 稜；花柱短，先端 4～7 裂。蒴果球形，徑約 2～2.4 cm，熟時黃褐色，4～7 瓣裂。(彩圖 213)

藥用部分：根莖。

效用：1.按本草綱目卷十七：根苦，微寒，有毒。主治驚癇，搖頭弄舌，熱氣在腹中(本經)。癲疾，癰瘡陰蝕，下三蟲，去蛇毒（別錄）。生食一升，利水（唐本）。治胎風手足搐，能吐泄瘰歷（大明）。去瘧疾寒熱（時珍）。

2.為解毒藥，治毒蛇傷。根具抗癌作用(甘，1969)。

3.根莖清熱解毒，消腫散瘀。治毒蛇咬傷，瘡傷腫毒，小兒麻疹合併肺炎，流行性腮腺炎，高熱，痙攣，哮喘，癌症（中國草藥手冊）。

4.清熱解毒，平喘止咳，熄風定驚。治癰腫，疔瘡，瘰歷，喉痺，慢性氣管炎，小兒驚風抽搐，蛇蟲咬傷（中藥大辭典，1982)。

5.水煎，點酒水服。治婦人奶結，乳汁不通，或小兒吹乳。煲雞肉或豬肺服，治肺勞久咳及哮喘；醋磨塗患處，治耳內生瘡熱痛，脫肛；浸童便，洗淨晒乾研末，酒或開水送服，治新舊跌打內傷，止痛散瘀。研粉吞服，治喉痺；研末開水送服，另以鮮根搗爛，或加甜酒釀搗爛敷患處，治蛇咬傷。

6.治服痛，下疳，胃病，毒蛇咬傷（佐佐木，1924)。

吊竹草

學名：ZEBRINA PENDULA SCHNIZL

別名：吊竹梅、班葉鴨跖草、水竹草、吊竹菜、紫背金牛、白帶草、金瓢羹、血見愁、紅舌草、紅竹仔草、百毒散、鴨舌紅、紅鴨跖草、二打不死、花葉竹夾菜、紅苞鴨跖草、花葉鴨跖草、時線蓮、假金線蓮。

英名：HAPPY WANDERING JEW, INCH PLANT.

形態：歸化本省之多年生匍匐草本。莖柔弱，傾臥，結節落根，有毛。葉互生，卵形或長橢圓形，長3-7 cm, 寬1.5-3 cm, 先端短尖，基部鞘狀抱莖；表示紫綠色而雜以銀白縞紋，背面紫紅色。花玫瑰色，叢生於2片摺合狀之頂生苞片狀葉內；萼片3，合生成圓柱狀的管；花冠管狀3裂，裂片長約3mm；雄蕊6；子房3室。果為蒴果。（彩圖210）

藥用部份：全草。

成分：全株含CAFFEEYLIFERULOYL CYANIDIN 3,7,3'-TRIGLUCOSIDE (GARRIGUES, 1953; STIRTON AND HARBORNE, 1880; BROUILLARD, 1981) 長素(AUXIN)及 醋酸(INDOLEACETIC ACID)(AL-OMARY, 1968)。

效用：1.解毒、益陰、止血。清熱解毒。治咳血、白帶、慢性痢疾；外敷消毒癰。水煎服，治淋病

2.利水消腫，清熱解毒。治心臟性水腫、腳氣水腫、腎炎水腫、尿路感染及結石、扁桃體炎、咽喉炎、腸炎腹瀉、毒蛇咬傷、癰腫（中國草藥手冊）

3.本省山胞以葉烤熱後，貼患處，治腫瘍

番杏科　AIZOACEAE

粟米草

學名：MOLLUGO PENTAPHYLLA L.

MOLLUGO STRICTA L.

別名：朱子草、柚仔仁、出世老草、鐵鉤草、地麻黃、辣辣菜、地杉樹、鴨腳爪子草。

形態：自生於全省平地之一年生草本。莖纖細、有鈍稜，高10-30cm，分枝甚多。葉3-5枚輪生，披針形或線狀披針形，長1.5-3cm,寬3-7cm,先端鈍或稍銳形，基部銳形；全緣。夏秋開花黃褐色，細小，聚繖狀疏生；苞小型，膜質；小花梗絲狀，長1-4mm；萼片5，橢圓形，先端鈍或圓形，長約1.5mm。蒴果球形或長橢圓形，徑約2mm。種子多數，腎圓形略扁，徑約0.5mm，表面具微細之瘤狀突起。（彩圖219）

藥用部份：全草、葉、根。

成分：地上部分含粟米草　（MOLLUPENTIN 或8-C-α -L-ARABINOPYRANOSYLAPIGENIN) (CHOPIN ET AL., 1979) 及6-C-β -D'-XYLOPYRANOSY1 -8-C-α -L-ARABINOSYLAPIGENIN （融點220-30°（分解））[CHOPIN ET AL., 1982]。

效用：1.清熱解毒。治服痛泄瀉，皮膚熱疹，火眼（中藥大辭典，1982)

2.葉治毒蛇咬傷（林，1953)

3.根部與雞蛋酒水煎服，或單水浸，以其汁洗眼，治眼病（佐佐木，1924)

4.搗爛包寸口，治皮膚熱疹

5.印度用全草為緩和清瀉劑(BURKILL, 1955)

6.印尼以全草治熱帶性潰瘍性口內炎(BURKILL, 1935)

7.馬來用全草搗碎，敷眼痛(BURKILL, 1935)

APPENDIX D.

M.S. Project:

"Skew Detection and Page Orientation Methods in Optical Character Recognition."

Tom Kurilla

Detection and Page Orientation Methods in Optical Character Recognition

by

Thomas P. Kurilla

This paper is submitted in partial fulfillment of
the requirements for the degree of

Master of Science in Computer Science

Department of Computer Science
University of Nevada, Las Vegas
November 1991

Table of Contents

Table of Figures

Abstract

A survey of skew detection and page orientation methods in optical character recognition is presented. Eight papers are reviewed. Three types of skew detection are identified: Projection Profile, Fourier Transform, and Hough Transform. Many of the methods presented require a preprocessing stage that is vulnerable to skew. Only one method has been extensively tested. Accuracy achieved by this method is two minutes of arc (0.0333... degrees).

Combining two of the methods for skew detection seems the best direction for future research. One method, requiring no preprocessing and implemented in hardware, can be applied before digitization of the document page. The second method, implemented in software, can apply additional skew detection if necessary for each of the segments of the image.

No data on accuracy of page orientation methods are provided. A robust method of determining page orientation is not yet available.

1. Introduction

An automated optical character recognition (OCR) system is one that can digitize, analyze, and recognize document text reliably without human interaction. Skew detection is of vital importance to such systems because image skew angles exceeding ±2° greatly reduce accuracy of the recognition.

Automatic segmentation of digitized documents into columns, paragraphs, lines of text, and characters is a necessary part of document image analysis. Top-down methods that use some form of projection profile for segmentation are vulnerable to document skew of as little as one third of a degree. Bottom-up clustering techniques are relatively insensitive to skew angle, but are unreliable due to decisions made with low statistical confidence in small neighborhoods [Baird87].

Document skew also affects the pattern recognition methods used to recognize characters within a document; skewed and non-skewed characters exhibit different attributes. The detection of document image skew is an important obstacle to overcome. An accurate method for determining skew that is unaffected by a wide range of page layout would provide a secure starting point for reliable top-down segmentation, and enhance performance of bottom-up segmentation. Despite Baird's claim that image skew must be less than 1/3° [Baird87], most document images can be segmented correctly if the skew angle is less than 2°. A method must compute the skew angle within 2° of the actual skew angle to be useful.

Skew detection is a recent discipline. The earliest reference (specific to skew detection) that was found in this study was [Trincklin84]. Skew detection sometimes appears in reports dealing with layout analysis ([Hase85], [Casey90], [Srihari89]), making it difficult to know how many methods actually exist. The methods fall into three categories: Projection Profile, Fourier Transform, and Hough Transform. Each method is analyzed separately and presented by category. The merits and drawbacks of each method are discussed, and the methods are compared. Suggestions for further research are presented.

Determining page orientation is related to skew detection. A robust OCR system must determine whether lines of text on a page are vertical or horizontal (landscape or portrait); proper orientation is essential for character recognition. The author found no publications which dealt exclusively with page orientation. One method for determining page orientation is presented, and one method for determining whether text is upside-down is presented. Suggestions for further research are given.

2. Projection Profile Methods

A projection profile (of angle θ) is an accumulator array, perpendicular to θ, where each element contains the number of black pixels found on a line of angle θ through the image. Each succeeding element of the array corresponds to lines at angle θ through the image separated by one vertical pixel.

Projection Profile methods work as follows: Given an orientation angle θ, project black pixels onto an accumulator line perpendicular to the projection direction. The accumulator line is partitioned into m bins, and bin size is related to the height of characters to be recognized. The θ yielding the most rapid fluctuations in the accumulator line is the skew angle.

The accumulator array can be computed as follows:

```
For (x ranging over the columns of the digitized image)
    For (y ranging over the rows of the digitized image)
        If (pixel is black)
            For ($\theta$) {
                Calculate $\rho = y + x \tan(\theta)$
                Increment $A(\rho,\theta)$
            }
        }
```

This transform is usually applied to binary images; hence the need to check if a pixel is black.

2.1. The Method of H. S. Baird

Henry S. Baird [Baird87] presents an algorithm that estimates the skew angle of a document image by performing a projection-profile using character locations. Before the algorithm can be used, the document image must be "roughly" segmented into characters. Segmentation is performed by conducting a connectivity analysis on the single-pixel run-length-encoding of the image. A character is located at the mid-point of the bottom of its bounding box.

The algorithm is as follows: Given an orientation angle θ, project the locations of characters (abstracted to points) onto an accumulator line perpendicular to the projection direction. The accumulator line is partitioned into m bins, and bin size is set at 1/3 of the x-height of 6-point text. Let $c_i(\theta)$ denote the number of points projected into the i^{th} bin at angle θ.

A real-valued energy alignment function of θ is computed as

$$A(\theta) = \Sigma c_i^2(\theta), \qquad i = 1,\ldots,m$$

This function displays a global maximum at the correct skew angle. Locating the global maximum of $A(\theta)$ is straightforward and reliable, but expensive. An exhaustive search would require checking every 2 minutes of arc from -90° to +90° (5400 angles). Computing A(θ) for all 5400 angles required over 13 CPU minutes on a VAX 11/750 (approximately 0.6 MIPS). Baird derived a heuristic to locate the global maximum that requires about 40 sample angles for convergence.

Testing was performed on over 50 pages representing a variety of typographical layouts and layout styles. Samples were selected from books, journals, theses, and typewritten pages, and included multiple columns, sparse tables, mixed fonts (both fixed and proportional pitch), various text sizes, headers, trailers, footnotes, and scanner noise at margins. The machine used was a VAX 11/750, and the page images were monochrome, high contrast, 300 d.p.i. images of about 8.5 million pixels.

The author claims accuracy of 2 minutes of arc (0.0333... degrees), and processing time of 8.5 seconds on an 0.6 MIPS machine. Given the size of the image, and the speed of the CPU, this method is about 8 times faster, and 20 times more accurate than Postl's method (see 3.2, below). However, this performance does not factor in the 95 seconds necessary for preprocessing. The need for preprocessing is a weakness in this method. Since segmentation is somewhat vulnerable to skew, the accuracy of the method for severely skewed document images is probably not the 2 minutes of arc quoted. The method is also predisposed to "western" style typography. The bottom of a bounding box for "eastern" typography may not give the same accuracy; using the center of the located character, as in [Britt89], may produce slightly better results. The preprocessing time also causes this method to become about 5 times slower than Postl's method. However, for skew detection that must be done after segmentation (see 3.2.3 and 6) this method appears to be the method of choice.

2.2. The Method of R. H. Britt

This patent [Britt89] describes a system of hardware and firmware that performs OCR. The parts of the patent relevant to skew detection describe a projection-profile method similar to that of Baird (see 2.1) using the center, in the x- and y-direction, of the located characters instead of the mid-point of the base of the bounding box. Britt's projection, however, is *linear* rather than *superlinear* in form (the quantities projected in the accumulator bins are not squared), and therefore tends not to emphasize the larger values in the accumulator line.

The inventor makes no claims about accuracy, or processing time needed for his invention, but a few observations can be made. Only 30 milliradians of arc (≈1.7 degrees) are checked as possible skew angles. Deciding which of these angles is the skew angle is vague, but appears to be *linear* in the description. This approach makes choosing a "correct" angle more difficult, and more error-prone. As with [Baird87], the need for preprocessing is a significant limitation. As mentioned in 2.1, using the center of a character favors "eastern" typography more than a point on the base of the bounding box. Only checking ±1.7° of arc is a more serious limitation. Any skew determined by this method is already within the ±2° necessary for accurate character recognition.

Because of the required preprocessing, the limi<span>t</span>ed search arc, and the vague method of choosing the "correct" skew angle, Britt's invention, seems to be unusable in an automated system.

2.3. The Method of Casey and Wong

Richard G. Casey and Kwan Y. Wong [Casey90] present a general overview of document-analysis techniques, and include a skew detection method. The method presented is similar to Baird's [Baird87], but uses every pixel of the digitized page. Technical details are not given. Skew detection receives only one sentence in a section on component labeling. Skew detection can be done by "...iteratively examining small angle deviations from the normal direction to determine which angle gives the steepest variation of the projection profile."

Since no specific details are presented for the method, no direct comparison to the other methods can be made, and no analysis of attainable accuracy can be presented.

2.4. The Method of Nakano, et al.

Nakano, et al [Nakano90], present a refinement of a method presented in an earlier paper [Nakano86]. Nakano, et al., claim to use the Hough transform to find the skew angle of the document image. Close inspection of the code presented reveals that a simple projection profile is used. Nakano, et al, adapted the standard algorithm as follows:

a) Execute a preprocessing pass that performs connected component analysis.

b) Find the character bounding rectangles that meet a condition. Precise conditions are not given, but a general choice that excludes small noises or large image elements such as figures, tables, and photographs is suggested.

c) The lower-left corner of the rectangles are transformed using:

$$\rho = y + x \tan(\theta),$$

and $S(\rho,\theta)$ is incremented by the x-width of the rectangle. No bounds for θ were given.

d) Once the $S(\rho,\theta)$ has been calculated, the choice of the skew angle is made using a "sharpness" formula. Four different formulae were tested:

1) Sum of Absolute Differences

$$V_1(d) = \sum_{y=1}^{N-1} |S(y+1,d) - S(y,d)|$$

2) Maximum Voting

$$V_2(d) = \max_{y}\left[S(y+1,d) - S(y,d)\right]$$

3) Sum of Averaged Inclination

$$V_3(d) = \sum_{j=1}^{J}\sum_{y=y_1(j)}^{y_2(j)} \frac{\left|S(y+1,d) - S(y,d)\right|}{y_2(j) - y_1(j)}$$

4) Number of zeroes

$$V_4(d) = \sum_{y=1}^{N}\left[1 - u(S(y,d))\right]$$

Nakano, et al, claim 3 and 4 gave the best results, 2 was marginal, and 1 was unsatisfactory. The table of experimental results that these conclusions are based upon was not provided with the paper.

The authors claim an average processing time of 14.5 seconds on a 32-bit 68020 workstation (MIPS unknown). They do not elaborate on the efficiency of the method because they expect special-purpose hardware to be fabricated that will improve performance. They also claim accuracy of 0.1 degree for plain-text document images, and 0.2 degree for documents containing graphical elements. The accuracy certainly satisfies the 2° requirement. However, as with [Baird87] and [Britt89], the requirement for preprocessing is a weakness of this method. Also, the authors did not mention what portion of the image could be graphical in nature without degrading the accuracy beyond the 0.2 degrees quoted, nor did they state what angles of skew could be detected. As presented, this method seems less accurate than that of [Baird87]. Further testing on a wider range of images is required to demonstrate attainable accuracy.

3. Fourier Transform Methods

The two-dimensional Fourier transform extracts periodic properties of an image from the density and position of its pixels.

The discrete two-dimensional Fourier transform can represented as:

$$F(U,V) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) e^{-j2\pi\left(\frac{Ux}{M}+\frac{Vy}{N}\right)} \qquad (1)$$

where $f(x,y)$ is the density of the image at coordinate (x,y), U and V are the spatial frequencies along the x-axis and y-axis, respectively, M is the number of partitions of the image along the x-axis, and N is the number of partitions of the image along the y-axis. x and U take discrete values in m ($m = 0...M-1$). y and V take discrete values in n ($n = 0...N-1$).

The continuous two-dimensional Fourier transform can be represented as:

$$S(U,V) = \int_{x,y} s(x,y) e^{j2\pi\left(\frac{xU}{m}+\frac{yV}{n}\right)} dx\, dy \qquad (2)$$

where $f(x,y)$ is the density of the image at coordinate (x,y), U and V are the spatial frequencies along the x-axis and y-axis, respectively, m is along the x-axis, and n is along the y-axis.

Fourier transform methods have the advantage of ignorance of image composition; no preprocessing of images is necessary. The continuous transform can be implemented in real-time hardware; image skew can be eliminated before digitization.

3.1. The Method of Hase and Hoshino

Masahiko Hase and Yasushi Yokosuka [Hase85] present a top-down segmentation method relying on the Fourier transform for its decision-making. The digitized image exists in an (x,y) plane, where the density of the image at coordinate (x,y) is represented as $f(x,y)$. The discrete two-dimensional Fourier transform is used. Only square images are tested ($M=N$).

The Skew Angle is found as follows:

1) Digitize the document.

2) Convert to a binary image using a threshold.

3) Perform the discrete two-dimensional Fourier transform (1) on the binary image.

4) There is a peak at the origin in (U,V); the next highest peak is at coordinates (u,v).

5) The skew angle, θ, is defined as: θ = Arctan(u/v). see figure 1.

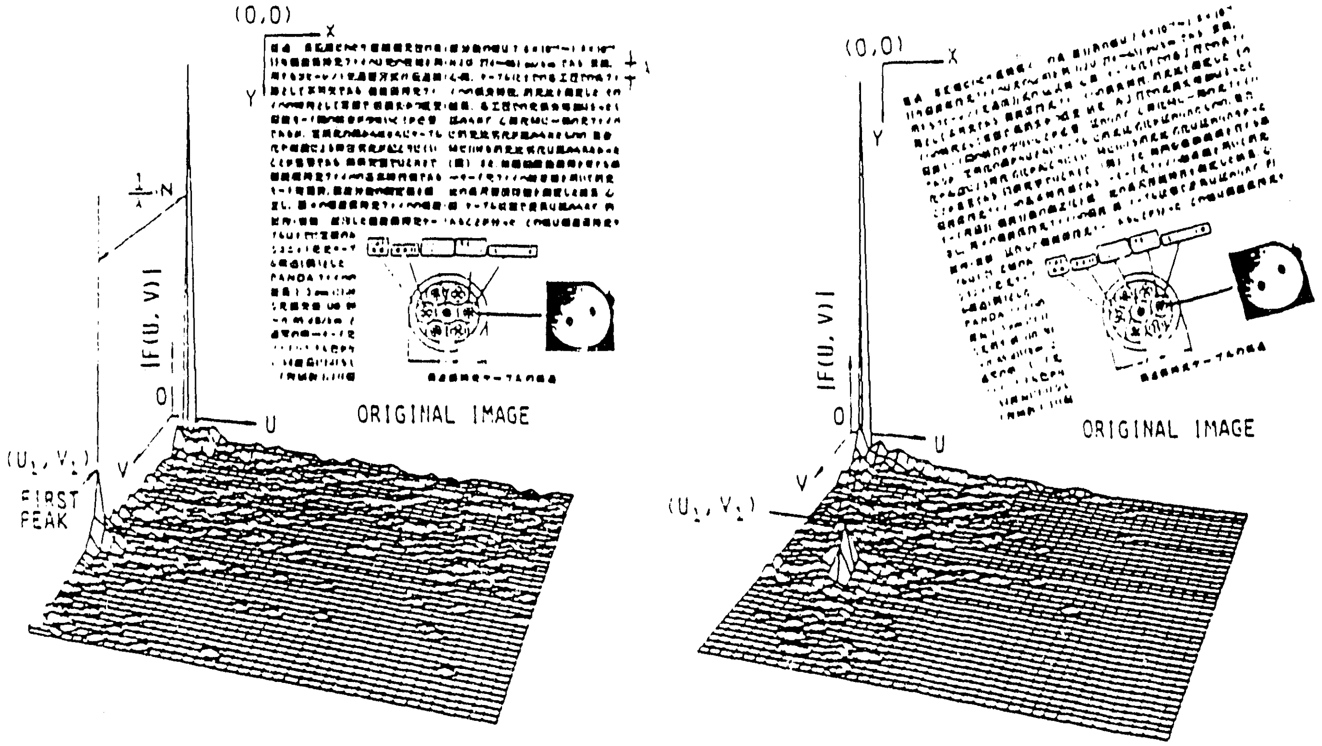


Figure 1: Discrete two-dimensional Fourier transformations.

Hase, et al, claim this method is accurate to within 2 degrees, even when the slope of the document image is large (greater than 10 degrees). They also claim that graphic regions covering over 50% of the document image do not affect the accuracy of the result; the second peak in (u,v) is shorter. These are impressive results. Unfortunately the method only applies to skew angles between 0° and 90°, and no limits are given for skew angles detected. It is not clear, for example, how much greater than 10 degrees of skew can be detected.

Characteristics of the Fourier transform suggest this method should be as accurate for skew angles of 90°±10°, which could allow detection of page orientation. Page orientation is not addressed.

Further work is necessary to determine the accuracy of the method over the range -90° to +90°, and whether page orientation can be accurately determined.

3.2. The Method of W. Postl

W. Postl [Postl86] specifically addresses skew angle in digitized documents. He describes experiments with Trincklin's method [Trincklin84], and two methods of his own design. The methods are reported to be equivalent. One is a discrete solution, and uses what Postl calls a "simulated skew scan" method, and the other is a continuous solution using a two-dimensional Fourier transform.

3.2.1. "Simulated Skew Scan" method

The document is represented as a reflectancy distribution D(x,y) positioned at the (yet unknown) angle α relative to the scanning field s(u,v). The document and scanner coordinates are denoted as (x,y) and (u,v) respectively. See figure 2.

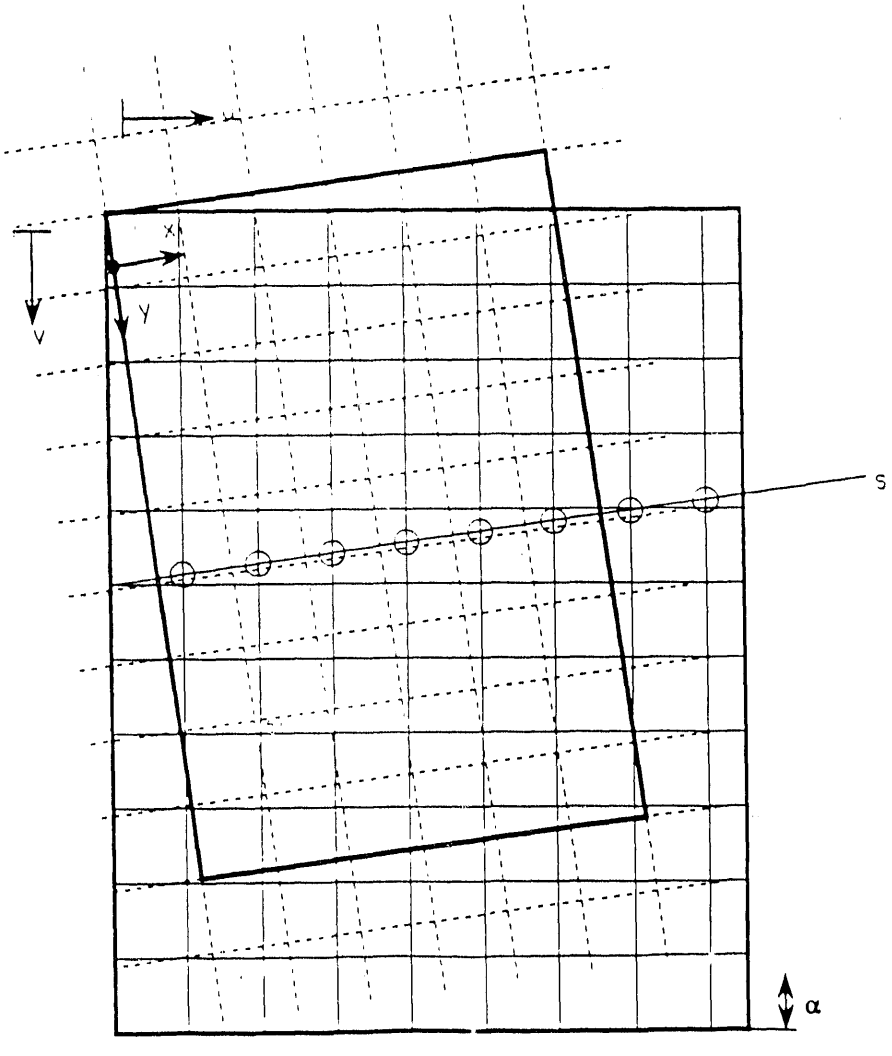


Figure 2. Document obliquely positioned in scan field
s = simulated scan line

The document and scan samples are related by:

$$D(x,y,\alpha) = s(u(x,y,\alpha),v(x,y,\alpha))$$

where

$$u(x,y,\alpha) = x\cos\alpha + y\sin\alpha \qquad (3)$$

$$v(x,y,\alpha) = y\cos\alpha - x\sin\alpha \qquad (4)$$

He defines an alignment premium, $P(\beta)$, for some search angle β that corresponds to search coordinates (ξ,η) as:

$$P(\beta) = \sum_{k=-\infty}^{\infty}[T((k+1)\times\Delta\eta,\beta) - T(k\times\Delta\eta,\beta)]^2 \qquad (5)$$

where $\Delta\eta$ and (below) $\Delta\xi$ are preset increments, and

$$T(\eta,\beta) = \sum_{k=-\infty}^{\infty} D(k\times\Delta\xi,\eta,\beta)\Delta\xi \qquad (6)$$

is the integrated trace of one simulated scan line, where $D(\ldots)$ is calculated using (3) and (4), assuming $s(u,v) = 0$ for any point outside the scan field. Non-integer coordinates returned by (3) and (4) are rounded to the nearest integer. Setting $\Delta\xi$ to correspond to a u-octet or a u-hextet will allow the infinite sum in (6) to be calculated with a table lookup.

Given the alignment premium function, $P(\beta)$, a sequence of skew scans is simulated at "search" angles α_i ($i = 0,1\ldots$) within a preset range $\alpha_{min} \leq \alpha_{max}$, starting with $\alpha_0 = 0$. Then for each search angle, calculate an alignment premium $P(\beta)$, thus generating premia P_i ($i = 0,1,\ldots$) where the next search angle

$$\alpha_{i+1} \leftarrow \bigcup_{j=0}^{i-1}\left(\alpha_j, P_j\right) \qquad (7)$$

is found following scme optimization strategy that will yield an ordered sequence of significant maxima of $P(\beta)$. The global maximum of $P(\beta)$ is defined as the skew scan angle. Other maxima identify other portions of the image that are skewed at different angles.

This is essentially looking for the maximum of a measure over a range of angles. The measure can be computed directly from the image as follows:

```
FOR every line inclined at angle $\theta$, $-45° \leq \theta \leq +45°$, every 3°
  BEGIN
    Compute integral density

    FOR each pair of neighboring scan lines
      Compute the difference of their densities

    Compute the sum of the square of these differences

  END

$\phi = \theta$ yielding the greatest sum
Use $\phi$ as the center of a second, identical search, ranging from $\phi$-3° to $\phi$+3°,
    every 1/3°.
```

The exact step of θ in this algorithm is not given in the paper. The 3° and 1/3° values quoted are estimated from figures present in the paper. Postl claims that this algorithm is "quite sub-optimal but robust" which implies that performance can be improved–perhaps using a heuristic similar to that of [Baird87]. The author claims that further improvement of this method is not necessary.

3.2.2. Continuous Fourier method

Let

$$S(U,V) = \int_{u,v} s(u,v)e^{j2\pi\left(\frac{uU}{m}+\frac{vV}{n}\right)}du\,dv \qquad (8)$$

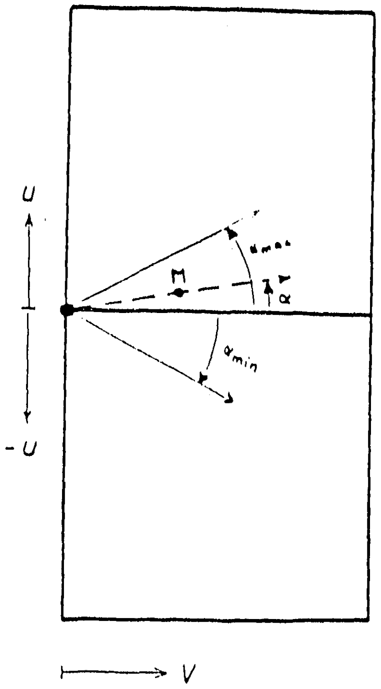
be the two-dimensional Fourier transform of the scan sample array, to be performed – i.e. approximated – before evaluation of the first premium. For the following steps, only one half-plane of the power spectrum coefficients

$$W(U,V) = S(U,V) \times S(U,V)^{*} \qquad (9)$$

needs to be stored ($V > 0$). The choice of m and n will affect the accuracy of processing in the same way that an additive overlay of the scan sample array were used with replicas shifted by m and n units of u and v, respectively.

The premium $P(\beta)$ is calculated by integrating $W(U,V)$ along a radius vector (search vector) inclined at angle β with respect to ordinate V. See figure 3.

This method is ideally suited for an optical hardware implementation, and as such can be used in skew detection *before* the document page is digitized.



Fourier transform space (right half plane) of scan field
M relative maximum of power spectrum of typical text document

Figure 3

3.2.3. Analysis

Testing of the discrete version of the method was performed on a SUN III workstation (68020 processor, ≈ 2-3 MIPS). The program required 5.1 seconds to detect the skew angle of a 2.4 million pixel object with an accuracy of <0.01 radians (0.57 degrees). Only every 16th line was evaluated during a simulated scan. The optimization strategy used was a coarse search from -45 to +45 degrees, followed by a fine search centered around the maximum found in the coarse search.

The accuracy of the method is acceptable. Characteristics of the method suggest that it should be accurate for skew angles of -90° to +90°, which could allow detection of page orientation. Unfortunately, accuracy figures are not presented for angles beyond ±45°, and page orientation is not addressed.

The most impressive feature of this method is that the contents of the document image do not need to be known. That is, no preprocessing of the image is necessary for this method to work; the original grey-scale image may be used. The continuous Fourier method is a valuable addition. It will allow an optical, real-time, hardware solution to be built that minimizes document image skew *before* digitization.

4. Hough Transform Methods

The Hough transform can be used to detect lines at any orientation. It consists of mapping points in Cartesian space (xy) to sinusoidal curves in $\rho\theta$ space via the transformation:

$$\rho = x \cos(\theta) + y \sin(\theta)$$

Each time a sinusoidal intersects another at a particular value of ρ and θ, the likelihood increases that a line corresponding to that $\rho\theta$ coordinate value is present in the original image. An accumulator array, $S(\rho,\theta)$, consisting of R rows and T columns is used to count the number of intersections at various ρ and θ values. Those cells in the accumulator array with the highest number of counts will correspond to lines in the original image. The transform is essentially:

```
For (x ranging over the columns of the digitized image)
    For (y ranging over the rows of the digitized image)
        If (pixel is black)
            For ($\theta$) {
                Calculate $\rho = x \cos(\theta) + y \sin(\theta)$
                increment $S(\rho,\theta)$
            }
        }
```

This transform is usually applied to binary images; hence the need to check if a pixel is black.

4.1. The Method of Hinds, et al.

Hinds, et al [Hinds90], present a method that uses the Hough transform to find the skew angle of an image. A special run-length transform is performed on the image prior to the Hough transform.

A "burst image" is created from the original binary image. The burst image is a grey scale image with each pixel's intensity representing the vertical run-length of a column of black pixels in the original binary image. More specifically, each individual vertical column (aligned with the scan axis) of L_i contiguous black pixels is replaced by L_{i-1} white pixels and one non-white pixel of value L_i positioned at the end of the original black run. The longer the column of black pixels in the original image, the darker (higher pixel value) the corresponding pixel (positioned at the bottom edge of a character or line) in the grey scale burst image. See figure 4.

Three important characteristics of the burst image are: 1) The number of black (non-white) pixels is significantly reduced (about a factor of 11), 2) It can be obtained readily from the compact form of an image which has been stored after application of the run-length encoding algorithm along the vertical axis, and 3) Greater emphasis is given to the bottoms of text lines.

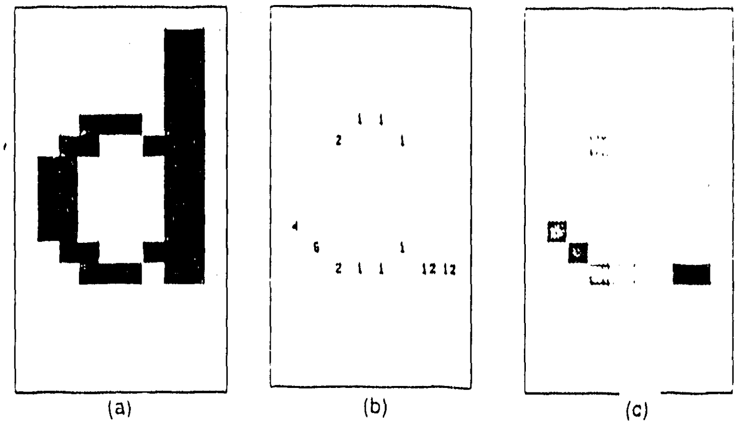


Figure 4. Creation of a burst image: (a) sample image; (b) burst image pixel values; (c) burst image as a grey scale image.

Figure 4

The accumulator array, S(R,T), is calculated similar to [Nakano90], except that the pixel values are used for the incrementing instead of the rectangle width. The skew angle is determined by searching S(R,T) for the cell with the largest value. The column that this cell belongs to is taken as the skew angle of the document. To increase the processing speed, only -15 to +15 degrees are searched.

Testing of the method was performed on a SUN 4-280 (10 MIPS). The authors claim a processing time of 67 seconds. There are no accuracy figures given; Hinds, et al, claim that the correct skew angle was determined.

It appears that the authors have drawn erroneous conclusions. They claim that their 67 seconds processing time is an improvement over 103.5 seconds needed in [Baird87]. However, Baird's method required 95 seconds on a 0.6 MIPS CPU, while their method needed 67 seconds on a 10 MIPS CPU. Actually, this method would take approximately 1100 seconds on the VAX 11/750 that Baird used! We conclude that this method does not represent an improvement over Baird's [Baird87]. The method of [Baird87] also searches a -45 to +45 degree arc, and this method only searches a -15 to +15 degree arc. The equivalent arc scan using this method would require over 3300 seconds! The results are also suspect because of the lack of any description of the accuracy achieved by the method. More careful work is necessary to properly evaluate this method.

4.2. The Method of Srihari and Govindaraju

Srihari, et al [Srihari89], describe the Hough transform, and how to use the accumulator array in detection of text skew angle.

The basis for all results derived in this paper is the Hough transform, which is exactly as described above, with θ ranging from 0° to 180°. Once the accumulator array has been computed, the projection-profile for an angle ϕ ($S(\rho,\phi)$, for all ρ) exhibits regular and rapid fluctuations when ϕ is close to the true skew angle of the text (see figure 5A). For ϕ not close to the true skew angle, the fluctuations are more gradual (see figure 5B).

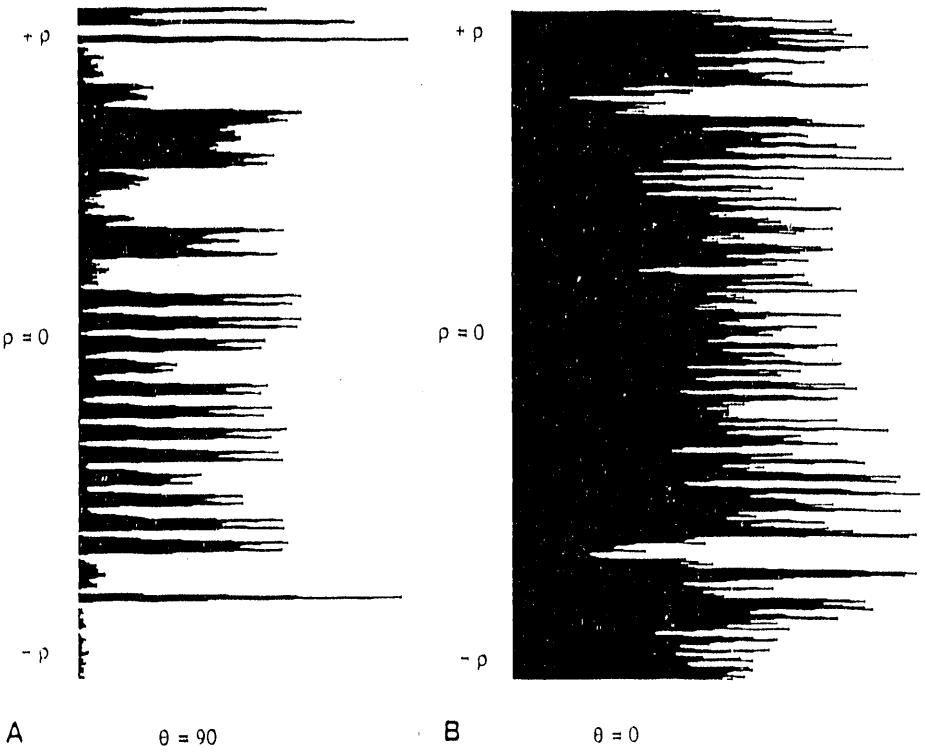


Figure 5. (A) Projection profile of the image in Figure 1A at Θ = 90°. The sharp valleys in the profile correspond to white spaces in between textual lines. (B) Projection profile of the image in Figure 1A at Θ = 0°. The valleys are not prominent in the profile.

Figure 5

This observation prompts the method of finding the sum of squares of the gradients R for the projection-profiles at all values of θ.

Let $g(r_j,\theta_i)$ be the rate of change in count value at (r_j,θ_i) Then

$$R(\theta_i) = \frac{1}{d(\theta_i)} \times \sum g(p_j,\theta_i)^2, \qquad j = 1, m(\theta_i)$$

Where $m(\theta_i)$ is defined as the number of "buckets" used for angle θ. The factor $1/d(\theta_i)$ helps scale the height and is defined as:

$d(\theta) = \cos(\theta), \qquad 0° < t \leq 45°$

$d(\theta) = \sin(\theta), \qquad 45° < t \leq 135°$

$d(\theta) = \cos(\theta), \qquad 135° < t \leq 180°$

Note: these values of d(θ) are the author's interpretation of ambiguous/erroneous values given in the paper.

Then, the skew angle, α = maximum $R(\theta)$.

This method is functionally similar to that used in [Nakano90], and corresponds closest to *formula 1* (sum of absolute differences) used in that method. There are three important differences: 1) The d(θ) term performs an anti-aliasing function, which provides a more accurate projection profile, 2) This method uses a square term instead of a linear term, making it more sensitive to the pattern being sought (i.e., rapid fluctuation), and 3) This method does not reduce the amount of data to be processed as in [Nakano90] which implies several orders-of-magnitude of additional processing.

Srihari, et al, provide no figures for speed or accuracy of this method. Analyzing the method can lead to some conclusions. As presented this method would require much more processing than [Nakano90] because each pixel in the digitized image is transformed for every θ; special hardware can reduce the processing. Even though this method most closely resembles *formula 1* in [Nakano90], which was found unacceptable, using anti-aliasing and a square term instead of a linear term may yield acceptable results. In any case, the accuracy of the skew detection will depend on how finely θ is divided between 0° and 180°.

Insufficient data are provided for this method. Further investigation is required to determine if this method is acceptable, and to compare this method to others.

5. Page Orientation

Detection of page orientation is related to detection of skew. Page orientation determines whether the document text is oriented "landscape" or "portrait" on the page. Though less important than skew detection, page orientation is a necessary part of any robust automated OCR system. No direct references to determining page orientation were found.

The only paper that addresses page orientation is [Hinds90]. Hinds uses run-length bursts (see 4.1) to determine page orientation. This is accomplished by computing vertical and horizontal run-length bursts. On the average, since stroke heights have longer run-lengths than stroke widths for a portrait-oriented document, a histogram of its horizontal run-length burst image will contain a larger number of counts corresponding to the run-lengths of character stroke widths than would a histogram of its vertical run-length burst image. Therefore, the orientation of a document image can be determined from a comparison of the counts of short run-lengths (which correspond to character width) in the histograms of vertical and horizontal burst images.

Although the experimental results seem poor, the tests include document images with large areas of non-textual data. For images that are predominantly text, the method seems accurate. As presented, this method will only work with document skew of no more than a few degrees, say ±15°. As the skew angle becomes greater, the distinction between the horizontal and vertical run-length bursts starts to blur, and becomes unreliable. Further testing is required to determine the accuracy of this method for a wide range of skew and document page formats.

Srihari, et al [Srihari89], do not provide a method for determining page orientation, but do provide a method of determining whether the text of a document is upside-down. Once the skew angle and approximate font-height are known, five projection lines are passed through the text. Characteristics of the roman alphabet suggest a pattern to the profile-projection of a line of text. If the pattern is inverted, then the text is upside-down. Certain degenerate patterns could give this method problems. However, if several sample line projections are taken, the correct pattern should be clear. Patterns for symbols other than the roman alphabet are not stated. This method seems promising, but further investigation is required to demonstrate attainable accuracy.

6. Conclusions

Of the methods presented, only [Britt89] seems unusable. The method is vague and checks very limited angles (±1.7° of arc).

Most of the remaining methods have some inherent weakness. [Casey90] provided no technical details. [Hase90] was not sufficiently tested, and only reports skew between 0° and 90°. [Hinds90] provided no data on the accuracy of the method, and depends on a preprocessing phase that is vulnerable to skew. Both [Baird87] and [Nakano90] require a preprocessing phase that is vulnerable to document skew. [Nakano90] appears to have the required accuracy, but was not tested as extensively as [Baird87].

Most promising of the the methods presented are those of [Baird87], [Srihari89], and [Postl86]. [Baird87] has the required accuracy, despite its vulnerability to skew during preprocessing, and is quite efficient. [Srihari89] is insensitive to skew, but further testing is required to demonstrate the attainable accuracy.

This author finds [Postl86] the most impressive method presented. It is independent of the contents of the document image. That is, no preprocessing of the image is necessary for this method to work. It is also reported to be more accurate than the method presented in [Hase90]. The continuous Fourier method is a valuable addition. It allows an optical, real-time, hardware solution to be built that could minimize document image skew *before* digitization. Then, since the image would have (nearly) zero skew, top-down segmentation can be performed with greater confidence and more accuracy. Due to the nature of typesetting (for older documents in particular) each segment may then need some form of skew detection. In these cases, since preprocessing has already been completed, Baird's method [Baird87] would yield the best results.

Bibliography

[Baird87] Baird, Henry S., "The Skew Angle of Printed Documents"; *Proc. of Society of Photographic Scientists and Engineers*, Vol. 40, pp. 21-24, 1987.

[Britt89] Britt, R.H., "Optical Character Reader with Skew Recognition"; U.S. Patent, 4,876,730, 1989.

[Casey90] Casey, Richard G., and Wong, Kwan Y., "Document-Analysis Systems and Techniques"; *Image Analysis Applications*, edited by Rangachar Kasturi and Mohan M. Trivedi; Marcel Dekker, Inc., New York and Basel, 1990, pp. 1-36.

[Hase85] Hase, Masahiko, and Hoshino, Yasushi, "Segmentation Method of Document Images by Two-Dimensional Fourier Transformation"; *Systems and Computers in Japan*, Vol. 16, No. 3, 1985, pp. 38-47.

[Hinds90] Hinds, S.C., Fisher, J.L., and D'Amato, P., "A Document Skew Detection Method using Run-length Encoding and the Hough Transform"; *10th Annual Pattern Recognition Conference*, IEEE, Atlantic City, NJ, 1990, 464-468.

[Nakano86] Nakano, Y., H. Fujisawa, and J. Higashino, "A Fast Algorithm for the Skew Normalization of Document Images"; *Trans. of Japanese Inst. of Electron. and Comm. Eng.*, Vol. J71-D, No. 10, pp. 1833-1834 (Nov. 1986, in Japanese).

[Nakano90] Nakano, Y., Shima, Y., Fujisawa, H., Higashino, J., and Fujinawa, M., "An Algorithm For The Skew Normalization of Document Image"; *10th Annual Pattern Recognition Conference*, IEEE, Atlantic City, NJ, 1990, 8-11.

[Postl86] Postl, W., "Detection of Linear Oblique Structures and Skew Scan in Digitized Document"; *8th International Conference on Pattern Recognition.* IEEE Computer Society Press, Paris, France, 1986, pp. 687-689.

[Srihari89] Srihari, S. and V. Govindaraju, "Analysis of Textual Images Using the Hough Transform"; *Machine Vision and Applic.*, Vol. 2, pp. 141-153, 1989.

[Trincklin84] Trincklin, J.P., "Conception d'un système d'analyse de documents etc."; Thèse, Université de Besançon, 1984. Ph.D. Thesis. This paper was referenced in [Postl86] and [Baird87]. Postl and Baird both described the skew recognition algorithm as being based on a piecewise least-squares fitting of vertical white run-lengths. However, it only works on bilevel patterns, and requires a clean left margin. This paper was unavailable.

APPENDIX E.

Proposal to CRAY Research:

Printed Arabic and Farsi Character Recognition using the Cray Y/MP

S. Latifi & J. Kanai

Printed Arabic and Farsi Recognition Using Cray Y/MP

Shahram Latifi, ECE Department
Junichi Kanai, CS Department

Howard R. Hughes College of Engineering
University of Nevada, Las Vegas

November 8, 1991

Abstract

The objective of this research is to develop algorithms for recognition of characters and words in printed cursive scripts. Among such scripts are printed Arabic, Bengali, and Farsi. We will focus on Arabic and Farsi because we are fluent in these languages and no commercial OCR device for these languages is currently available. Moreover, the demand for recognition of these languages has increased significantly due to recent events in the Middle East.

The variety of type faces and type sizes makes accurate recognition of detached printed characters, such as English texts, difficult. We face two additional problems in recognition of printed Arabic and Farsi texts. The first problem is extraction of individual characters from a word. Since characters in a word are connected, it is difficult to determine the beginning and ending of individual characters. Another problem is the position dependency of Arabic and Farsi character shapes. For instance, the Arabic alphabet consists of 29 characters, and each character takes a different shape depending on its position. Therefore, we must recognize not only character shapes but also their position in a word correctly.

To overcome these problems, we will investigate two approaches: character-based and word-based approaches. In the character-based approach, we will attempt to use a sophisticated segmentation algorithm to improve the accuracy of character extraction from a word. We shall use neural nets to classify characters. Artificial neural nets have been successfully used for a variety of pattern recognition tasks. We will build and train nets for Arabic and Farsi symbols using Neural Shell V3.0 on the Cray computer.

Our second approach treats a printed word as a pattern and attempts to classify the pattern without decomposing it into characters. This word-based approach eliminates the character extraction process. However, the number of possible classes will be large because a typical dictionary contains around

75,000 words (classes). We will take advantage of the power of the vector processors in the Cray Y/MP to build classifiers for Arabic and Farsi.

The results of this research would demonstrate the effectiveness of the Cray supercomputer in solving this class of problems.

1 Introduction

Advances in communication and computer technologies have allowed us to exchange and retrieve information quickly through the electronic media. Although some documents are generated by word processors and desktop publishing systems, the vast majority of published documents, especially in foreign countries, have not been integrated into an electronic environment. Optical character recognition (OCR) is the technique to convert printed symbols into computer readable form and is the key technology to convert information into a form suitable for electronic processing.

The objective of this research is to develop algorithms for recognition of characters and words in printed cursive scripts in which characters in a word are connected. Among such scripts are printed Arabic, Bengali, and Farsi. We will focus on Arabic and Farsi because we are fluent in these languages and no commercial OCR device for these languages is currently available. Moreover, the demand for recognition of Arabic and Farsi has increased significantly due to recent events in the Middle East.

Traditional OCR research focuses on texts that are made up predominantly of detached characters, such as printed English, Japanese, and Cyrillic. The Calera RS9000, one of the best OCR devices on the market today, recognizes approximately 98 percent of detached English characters. On the other hand, since it is difficult to isolate characters of a word in a cursive script, the best current recognition rate of printed Arabic is far from satisfactory, i.e., approximately 85 percent [Amin86]. This proposed research attempts to improve the recognition rate of scripts written in such languages.

This research will be cosponsored by the Information Science Research Institute (ISRI) at UNLV. ISRI was established in October 1990 when the Howard R. Hughes College of Engineering received funding from the U.S. Department of Energy for experimental research in electronic document analysis. ISRI is focused on improving OCR technology and testing new ideas in document analysis.

2 Methodology

The versatility of type faces and type sizes makes accurate recognition of printed English characters difficult. We face two additional problems in recognition of printed Arabic and Farsi texts. The first problem is extraction of individual characters from a word. Since characters in a word are connected, it is difficult to determine the beginning and ending of individual characters. Another problem is the position dependency of Arabic and Farsi character shapes. For instance, the Arabic alphabet consists of 29 characters, and each character takes a different shape depending on its position within a word: beginning shape, middle shape, end shape, and stand-alone shape. Therefore, we must recognize not only character shapes but also their positions in a word correctly.

Amin [Amin86] combined the character-based approach and a syntactic pattern recognition technique to achieve a recognition rate of 85 percent. His program first segments a word into subwords by decomposing the word image into a set of columns. Each subword is assigned to one of a predetermined set of tokens according to its shape; therefore, a word is converted into a string of tokens. A string is simplified by removing repeated tokens, and a substring of a simplified string corresponds to a character. Finally, each substring is parsed by his dictionary program and assigned to an Arabic character.

Amin's method has two weaknesses. He uses a rather simple segmentation algorithm which may cause OCR errors due to missegmentations. Another weakness is that his character classifier uses a small number of character features. Therefore, his method may fail to correctly classify characters whose overall shapes are similar but have minor differences in detail.

To overcome these weaknesses, we will investigate two approaches: character-based and word-based approaches. In the character-based approach, we will attempt to use a more sophisticated segmentation algorithm, such as [Tsji91], to improve the accuracy of character extraction from a word. We will use artificial neural nets to classify characters. Artificial neural nets have been used successfully for a variety of pattern recognition tasks [Nels91]. We will build and train nets, for Arabic and Farsi symbols using Neural Shell V3.0 on the Cray computer.

We will construct two types of artificial neural nets, namely Hopfield net and Hamming net, in this research. A Hopfield net has been used successfully to recognize detached characters [Hopf86]. The Hamming net is a neural net implementation of the optimum classifier for binary patterns corrupted by random noise [Lipp87]. Therefore, these nets are suitable for our OCR systems.

Our second method treats a printed word as a pattern and attempts to classify the pattern without decomposing it into characters. This word-based approach eliminates the character extraction process. However, the number of possible classes will be large because a typical dictionary contains around 75,000 words (classes). Designing a classifier for 75,000 classes is very computation intensive. We plan to exploit the computation power of the vector processors in the Cray computer to build classifiers for Arabic and Farsi.

3 Research Facilities

The Information Science Research Institute will provide the following image processing equipment for generating digitized pages:

- A Fujitsu M3096E+ Image Scanner will be used to digitize printed pages.
- A Seaport Imaging Deskewing board will be used to correct the alignment and rotation of page images.

- Workstations will be used to develop programs for extracting words and character from digitized pages.

The NSCEE will provide the following facilities for building and training OCR systems:

- The Cray Y/MP is used to handle the large number of computations.
- Neural Shell V3.0 routines will be used to construct and train neural nets.

The data will be transferred via the campus computer network from the ISRI to the NSCEE.

4 Timetable

We plan to conduct our research in three phases. The period of each phase as well as research the tasks are as follows:

Phase 1 (three months):

- Create training and test data.
- Develop programs for extracting words and characters from text-lines.

Phase 2 (six months):

- For the character-based approach, train neural nets.
- For the word-based approach, extract features from words and design a classifier.
- For the word-based approach, compare neural nets and feature-based classification techniques.
- Prepare the first semiannual report to Cray Research, Inc.

Phase 3 (three months):

- Evaluate the performance of the OCR systems.
- Prepare the second semiannual report to Cray Research, Inc.
- Document and publish the results.

5 Benefit to Cray

In the last few years, the accuracy of commercial optical character recognition (OCR) devices has improved substantially. For a typical collection of non-uniform English language documents, between 98 and 99% correct recognition (on a character basis) can be achieved by the best OCR devices currently available. For uniform documents, accuracies approaching 99.8% are not uncommon. Typical recognition accuracy of these devices for non-English language documents is significantly worse.

Also, in recent years, neural-net approaches to discrimination in high dimensional/highly non-linear problem domains have shown much promise [Nels91]. Especially in the area of OCR of foreign language documents, neural-nets have the potential of providing significant improvements in recognition accuracy.

One major impediment to the development of effective neural net recognizers is the effort required to select the neural architecture which can provide the most effective discrimination. To do this, several architectures must be explored. That is, various nets must be trained (for a large set of data) and compared for recognition efficiency. Proper training of even one network architecture is an extremely computation intensive task.

The authors believe that, using the Neural Shell 3.0 software on the Cray Y/MP supercomputer, it will be possible to develop and train a neural-net based OCR device which will provide a significant improvement in the best current reported accuracy for Arabic (and Farsi) documents.

The successful development of an improved neural-net Arabic OCR device using the power of the Cray Y/MP would benefit Cray in several ways. First, the results of this research would demonstrate the usefulness of the Cray supercomputer to develop better OCR algorithms. (The script languages proposed in this study are, of course, some of the most difficult languages to recognize.) Second, it would demonstrate the value of the power of the Y/MP in training neural-nets for many other kinds of image recognition problems. In essence, the Cray supercomputer is a powerful candidate to carry out the intensive computations inherent to this type of work.

6 Deliverables

We propose to produce the following items.

- Training and test data to conduct Arabic and Farsi OCR experiments.
- Neural nets for Arabic and Farsi text recognition running on the Cray.
- Two Master's theses.
- Archival publications in journals and conferences.
- Semiannual reports to Cray Research, Inc.

7 Budget

The Information Science Research Institute (ISRI) will match the requested budget. As mentioned in Section 3, it will also provide image processing equipment, workstations, access to its library, and office space for graduate students without charge. Moreover, ISRI will contribute the secretarial services needed for this project.

ISRI is currently funded by the United States Department of Energy to conduct research in the area of OCR of English language documents. It wishes to expand its research efforts to include OCR of foreign language documents.

Budget Table

Description	External	Cray	Total Budget
A. Personnel:			
A1. 1 month salary of Dr. Shahram Latifi	-	$5,800	$5,800
A2. 1 month salary of Dr. Junichi Kanai	$5,100	-	$5,100
A3. 1 year tuition and salary of two Grads.	$11,580	$11,580	$23,160
B. Fringe Benefits (%3)	$501	$521	$1,022
C. Total salaries and fringe benefits	$17,181	$17,901	$35,082
D. Travel (2 domestic conferences)	$1,000	$1,000	$2,000
E. Computer Resources:			
E1. Cray System Billing Unit	$2,420	$2,420	$4,840
E2. Permanent Storage	$50	$50	$100
E3. Sun 4/490 file server	$150	$150	$300
E4. Printing or other output	$25	$25	$50
F. Miscellaneous:			
F1. Materials and Supplies	$100	$100	$200
F2. Publication Charges	$ 200	$200	$400
H. Total costs	$21,126	$21,846	$42,972

Details of Budget:

A1. This item reflects 1 month of summer support for the Principal Investigator, Dr. Shahram Latifi.

A2. This item reflects 1 month of summer support for the Coprincipal Investigator, Dr. Junichi Kanai.

A3. This item covers the tuition and the stipend for two graduate students working on the project for the entire year (Tuition for nonresident:$4,380/semester, Stipend: $800 /month for 9 months).

B. This item reflects the fringe benefits charged by UNLV, and is calculated at 3% of A1+A2+A3.

C. Total salaries and fringe benefits (A1+A2+A3+B)

D. The proposed travel budget covers the cost of attending two domestic conferences ($1,000/Conference Trip).

E1. This item covers the cost of 20 System Billing Units at the rate $242/SBU.

E2. The cost of 50 megabytes Permanent storage for 180 days at the rate of $0.014/megabyte/day.

E3. The cost associated with 20 hours usage of the 4/490 CPU at the rate of $15 / CPU hour.

E4. The cost of approximately 500 pages of computer printout at the rate of $.10 /page.

F. This item covers miscellaneous costs such as: expenses associated with the printing of technical reports, ordering of conference proceedings, research monographs, postage, and telephone charges associated with the project.

8 References

[Amin86] Adnan Amin and Gerald Masini,"Machine Recognition of Multi Font Printed Arabic Texts," *Proceedings of 8th ICPR'* Paris, France, October 27-31, 1986, pp. 392-395.

[Hopf86]John J. Hopfield and David W. Tank, "Computing with Neural Circuits: A Model," *Science,* Vol. 233, August 8, 1986, pp. 625-632.

[Lipp87] Richard P. Lippmann, "An Introduction to Computing with Neural Nets," *IEEE ASSP Magazine,* April, 1987, pp. 4-22.

[Nels91] Marilyn McCord Nelson, and W.T. Illingworth, *A Practical Guide to Neural Nets,* Addison Wesley, 1991.

[Tsji91] Shuichi Tsujimotot and Haruo Asada, "Resolving Ambiguity in Segmenting Touching Characters," *Proceedings of ICDAR91*, Saint-Malo, France, September 30-October 2, 1991, pp. 701-709.

9 Biographies

Shahram Latifi received the MS in Electrical Engineering from the University of Tehran, Iran in 1980. He received the MS and the Ph.D. both in Electrical and Computer Engineering from Louisiana State University in 1986 and 1989 respectively. He is currently an Assistant Professor of the Electrical and Computer Engineering Department at UNLV. He has taught supercomputer architectures and related topics in the past few years. His current research interests include computer architecture, supercomputing, parallel processing, and pattern recognition. He is fluent in Farsi and Arabic. Dr. Latifi is a member of IEEE Computer.

Junichi Kanai received the BS in Electrical Engineering, M.Eng. and Ph.D. in Computer and Systems Engineering from Rensselaer Polytechnic Institute in 1983, 1985, and 1990 respectively. He is currently an Assistant Professor of the Computer

Science Department and a Research Scientist of the Information Science Research Institute at UNLV. He has been investigating document analysis and OCR techniques since 1985. His current research interests include supercomputing, document analysis, pattern recognition, and parallel processing. Dr. Kanai is a member of AAAI, ACM, and IEEE.

DATE
FILMED

07/8/92