LOAN DOCUMENT

PHOTOGRAPH THIS SHEET

AD-A254 458

DTIC ACCESSION NUMBER

LEVEL

1

INVENTORY

AFOSR-TR-92-0719

DOCUMENT IDENTIFICATION

DISTRIBUTION STATEMENT

ACCESSION FOR
NTIS GRA&I
DTIC TRAC
UNANNOUNCED
JUSTIFICATION

BY
DISTRIBUTION/
AVAILABILITY CODES

DISTRIBUTION	AVAILABILITY AND/OR SPECIAL	
A-1		

DISTRIBUTION STAMP

DTIC ELECTE
S AUG 3 1992
C

DATE ACCESSIONED

DTIC QUALITY INSPECTED 5

DATE RETURNED

92 7 31 002

DATE RECEIVED IN DTIC

370 743 861 PG
92-20723

REGISTERED OR CERTIFIED NUMBER

PHOTOGRAPH THIS SHEET AND RETURN TO DTIC-FDAC

DTIC FORM 70A JUN 90

DOCUMENT PROCESSING SHEET

PREVIOUS EDITIONS MAY BE USED UNTIL STOCK IS EXHAUSTED.

LOAN DOCUMENT

HANDLE WITH CARE

UNITED STATES AIR FORCE

1989 RESEARCH INITIATION PROGRAM

Conducted by

UNIVERSAL ENERGY SYSTEMS, INC.

under

USAF Contract Number F49620-88-C-0053

RESEARCH REPORTS

VOLUME IV OF IV

Submitted to

Air Force Office of Scientific Research

Bolling Air Force Base

Washington, DC

By

Universal Energy Systems, Inc.

June 1992

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	25 Jun 92	ANNUAL 1 Jan 90 to 31 Dec 90

4. TITLE AND SUBTITLE
US Air Force 1989 Research Initiation Program Conducted by Universal Energy Systems, Inc, VOL # 2

5. FUNDING NUMBERS
F49620-88-C-0053
2305/D5

6. AUTHOR(S)
Rod Darrah

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Universal Energy Systems, Inc
Dayton OH

8. PERFORMING ORGANIZATION REPORT NUMBER
C-1157

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NI
Bolling AFB DC

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
(U)

12b. DISTRIBUTION CODE
(U)

13. ABSTRACT (Maximum 200 words)

This program is for follow-on research efforts for the participants in the Summer Faculty Research Program. Funding is provided to establish RIP awards to about half the number of participants in the SFRP. Participants in the 1989 SFRP competed for funding under the 1989 RIP. Evaluation of the proposals were made by the contractor. Evaluation criteria consisted of: 1. Technical excellence of the proposal 2. Continuation of the SFRP effort 3. Cost sharing by the university. The list of proposals selected for award was forwarded to AFOSR for approval of funding and for research efforts to be completed by 31 December 1990. The following summarizes the events for the evaluation of proposals and award of funding under the RIP. A. RIP proposals were submitted to the contractor by 1 November 1990. The proposals were limited to $20,000 plus cost sharing by the universities. The universities were encouraged to cost share, since this is an effort to establish a long term effort between the Air Force and the university. B. Proposals were evaluated on the criteria listed above and the final award approval was given by AFOSR after consultation with the Air Force Laboratories. C. Subcontracts were negotiated with the Universities. There were a total of 122 RIP awards made under the 1989 program.

14. SUBJECT TERMS	15. NUMBER OF PAGES
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
(U)	(U)	(U)	(U)

UNITED STATES AIR FORCE

1989 RESEARCH INITIATION PROGRAM

Conducted by

UNIVERSAL ENERGY SYSTEMS, INC.

under

USAF Contract Number F49620-88-C-0053

RESEARCH REPORTS

VOLUME IV OF IV

Submitted to

Air Force Office of Scientific Research

Bolling Air Force Base

Washington, DC

By

Universal Energy Systems, Inc.

June 1992

TABLE OF CONTENTS

INTRODUCTION

Research Initiation Program - 1989

AFOSR has provided funding for follow-on research efforts for the participants in the Summer Faculty Research Program. Initially, this program was conducted by AFOSR and popularly known as the Mini-Grant Program. Since 1983 the program has been conducted by the Summer Faculty Research Program (SFRP) contractor and is now called the Research Initiation Program (RIP). Funding is provided to establish RIP awards to about half the number of participants in the SFRP.

Participants in the 1989 SFRP competed for funding under the 1989 RIP. Participants submitted cost and technical proposals to the contractor by 1 November 1989, following their participation in the 1989 SFRP.

Evaluation of these proposals were made by the contractor. Evaluation criteria consisted of:

1. Technical excellence of the proposal
2. Continuation of the SFRP effort
3. Cost sharing by the university

The list of proposals selected for award was forwarded to AFOSR for approval of funding. Those approved by AFOSR were funded for research efforts to be completed by 31 December 1990.

The following summarizes the events for the evaluation of proposals and award of funding under the RIP.

A. RIP proposals were submitted to the contractor by 1 November 1989. The proposals were limited to $20,000 plus cost sharing by the universities. The universities were encouraged to cost share, since this is an effort to establish a long term effort between the Air Force and the university.

B. Proposals were evaluated on the criteria listed above and the final award approval was given by AFOSR after consultation with the Air Force Laboratories.

C. Subcontracts were negotiated with the universities. The period of performance of the subcontract was between October 1989 and December 1990.

Copies of the final reports are presented in Volumes I through IV of the 1989 Research Initiation Program Report. There were a total of 122 RIP awards made under the 1989 program.

STATISTICS

PROGRAM STATISTICS

Total SFRP Participants	168
Total RIP Proposals submitted by SFRP	132
Total RIP Proposals submitted by GSRP	2
Total RIP Proposals submitted	134
Total RIP's funded to SFRP	94
Total RIP's funded to GSRP	2
Total RIP's funded	96
Total RIP Proposals submitted by HBCU's	9
Total RIP Proposals funded to HBCU's	5

LABORATORY PARTICIPATION

Laboratory	Participants	Submitted	Funded
AAMRL	12	10	6
WRDC/APL	10	8	6
ATL	9	9 (1 GSRP)	9 (1 GSRP)
AEDC	10	8	8
WRDC/AL	7	5	4
ESMC	0	0	0
ESD	3	2	1
ESC	11	8	7
WRDC/FDL	9	7	5
FJSRL	7	5	4
AFGL	12	10	6
HRL	12	10 (1 GSRP)	8 (1 GSRP)
WRDC/ML	9	7	5
OEHL	4	1	1
AL	12	10	6
RADC	15	11	8
SAM	17	16	9
WL	8	7	3
WHMC	1	0	0
Total	168	134	96

LIST OF PARTICIPATING UNIVERSITIES

Alabama, University of	-	1	New York, State University of	-	2
Alfred University	-	1	North Carolina State University	-	1
Arkansas-Pine Bluff, Univ. of	-	1	Northern Arizona University	-	1
Auburn University	-	1	Northern Illinois University	-	1
Bethel College	-	1	Northwestern University	-	1
Boston College	-	1	Notre Dame, University of	-	1
Brescia College	-	1	Ohio State University	-	2
California Polytechnic	-	1	Oklahoma, University of	-	3
California State University	-	2	Old Dominion University	-	1
Cincinnati, University of	-	2	Pennsylvania State University	-	1
Denver, University of	-	1	Pittsburgh, University of	-	1
Eastern Kentucky University	-	1	Rhode Island, University of	-	1
Florida Atlantic University	-	1	San Diego State University	-	1
Florida Institute	-	1	San Jose State University	-	1
Florida, University of	-	4	Savannah State College	-	1
Hamilton College	-	1	Scranton, University of	-	1
Harvard University	-	1	Southern Oregon State College	-	2
Illinois Institute of Technology	-	1	Southwest Texas State University	-	1
Illinois-Rockford, University of	-	1	Tennessee State University	-	1
Illinois State University	-	1	Tennessee Technological Univ.	-	1
Indiana-Purdue, University of	-	1	Texas A&M University	-	6
Kansas State University	-	2	Texas Southern University	-	1
Lawrence Technological University	-	1	Texas-San Antonio, University of	-	3
Long Island University	-	1	Transylvania University	-	1
Lowell, University of	-	1	Trinity University	-	1
Massachusetts, University of	-	2	US Naval Academy	-	1
Michigan, University of	-	1	Utah State University	-	1
Minnesota-Duluth, University of	-	2	Utica College	-	1
Mississippi State University	-	2	Vanderbilt University	-	1
Missouri-Rolla, University of	-	1	Washington State University	-	1
Murray State University	-	1	West Virginia University	-	2
Nebraska-Lincoln, University of	-	2	Wisconsin-Platteville, Univ. of	-	1
New Hampshire, University of	-	1	Worchester Polytechnic Institute	-	1
New York Institute of Technology	-	1	Wright State University	-	6
			Total	-	94

PARTICIPANTS LABORATORY ASSIGNMENT

AERO PROPULSION AND POWER DIRECTORATE
(Wright-Patterson Air Force Base)

Dr. Jerry Clark
Wright State University
Specialty: Physics

Dr. Frank Gerner
University of Cincinnati
Specialty: Mechanical Engineering

Dr. Thomas Lalk
Texas A&M University
Specialty: Mechanical Engineering

Dr. Baruch Lieber
State University of New York
Specialty: Aerospace Engineering

Dr. William Schulz
Eastern Kentucky University
Specialty: Analytical Chemistry
760-7MG-079 and 210-10MG-095

Dr. Richard Tankin
Northwestern University
Specialty: Mechanical Engineering

ARMAMENT DIRECTORATE
(Eglin Air Force Base)

Dr. Peter Armendarez
Brescia College
Specialty: Physical Chemistry

Dr. Joseph Brown
Mississippi State University
Specialty: Mechanical Engineering

Dr. Roger Bunting
Illinois State University
Specialty: Inorganic Chemistry

Dr. Satish Chandra
Kansas State University
Specialty: Electrical Engineering

Dr. David Cicci
Auburn University
Specialty: Aerospace Engineering

Mr. William Newbold (GSRP)
University of Florida
Specialty: Aerospace Engineering

Dr. Boghos Sivazlian
University of Florida
Specialty: Operations Research

Dr. Steven Trogdon
University of Minnesota-Duluth
Specialty: Mechanics

Mr. Asad Yousuf
Savannah State College
Specialty: Electrical Engineering

ARMSTRONG LABORATORY
(Brooks Air Force Base)

Dr. Robert Blystone
Trinity University
Specialty: Zoology

Dr. Carolyn Caudle-Alexander
Tennessee State University
Specialty: Microbiology

Dr. James Chambers
University of Texas - San Antonio
Specialty: Biochemistry

Dr. Mark Cornwall
Northern Arizona University
Specialty: Human Performance

Dr. Vito DelVecchio
University of Scranton
Specialty: Biochemical Engineering

Dr. Gwendolyn Howze
Texas Southern University
Specialty: Molecular Biology

Dr. Harold Longbotham
University of Texas-San Antonio
Specialty: Electrical Engineering

Dr. Ralph Peters (1987)
Wichita State University
Specialty: Zoology

Dr. Raymond Quock
Univ. of Illinois at Rockford
Specialty: Pharmacology

Dr. Ram Tripathi
University of Texas-San Antonio
Specialty: Statistics

ARNOLD ENGINEERING DEVELOPMENT CENTER
(Arnold Air Force Base)

Dr. Brian Beecken
Bethel College
Specialty: Physics

Dr. Stephen Cobb
Murray State University
Specialty: Physics

Dr. John Francis
University of Oklahoma
Specialty: Mechanical Engineering

Dr. Orlando Hankins
University of North Carolina State
Specialty: Nuclear Engineering

Dr. Lang-Wah Lee
University of Wisconsin-Platteville
Specialty: Mechanical Engineering

Dr. Chun Fu Su
Mississippi State University
Specialty: Physics

Dr. Richard Tipping
University of Alabama
Specialty: Physics

Dr. D. Wilkes
Vanderbilt University
Specialty: Electrical Engineering

AVIONICS DIRECTORATE
(Wright-Patterson Air Force Base)

Dr. David Choate
Transylvania University
Specialty: Mathematics

Dr. R. H. Cofer
Florida Institute
Specialty: Electrical Engineering

Dr. Dar-Biau Liu
California State University
Specialty: Applied Mathematics

Dr. Robert Shock
Wright State University
Specialty: Mathematics

CREW SYSTEMS DIRECTORATE
(Wright-Patterson Air Force Base)

Dr. Thomas Lockwood
Wright State University
Specialty: Toxicology

Dr. Ethel Matin
Long Island University
Specialty: Experimental Psychology

Dr. Randy Pollack
Wright State University
Specialty: Anthropology

Dr. Donald Robertson (1987)
Indiana University of Pennsylvania
Specialty: Psychology

Dr. Michael Stanisic
University of Notre Dame
Specialty: Robotics

Dr. Chi-Ming Tang
State University of New York
Specialty: Mathematics

Dr. Ebo Tei
University of Arkansas-Pine Bluff
Specialty: Psychology

ENGINEERING AND SERVICES CENTER
(Tyndall Air Force Base)

Dr. William Bannister
University of Lowell
Specialty: Organic Chemistry

Dr. Emerson Besch
University of Florida
Specialty: Animal Physiology

Dr. Avery Demond
University of Massachusetts
Specialty: Civil Engineering

Dr. Kirk Hatfield
University of Florida
Specialty: Civil Engineering

Dr. Kim Hayes
University of Michigan
Specialty: Environmental Engineering

Dr. Deborah Ross
University of Indiana-Purdue
Specialty: Microbiology

Dr. Dennis Truax (1987)
Mississippi State University
Specialty: Civil Engineering

Dr. George Veyera
University of Rhode Island
Specialty: Civil Engineering

ELECTRONIC SYSTEMS DIVISION
(Hanscom Air Force Base)

Dr. Stephen Kolitz (1986)
University of Massachusetts
Specialty: Operations Research

Dr. Sundaram Natarajan
Tennessee Technical University
Specialty: Electrical Engineering

FLIGHT DYNAMICS DIRECTORATE
(Wright-Patterson Air Force Base)

Dr. Kenneth Cornelius
Wright State University
Specialty: Fluid Mechanics

Dr. Arnold Polak
University of Cincinnati
Specialty: Aerospace Engineering

Dr. Nisar Shaikh
University of Nebraska-Lincoln
Specialty: Applied Mathematics

Dr. William Wolfe
Ohio State University
Specialty: Engineering

Dr. Lawrence Zavodney
Ohio State University
Specialty: Mechanical Engineering

FRANK J. SEILER RESEARCH LABORATORY
(United States Air Force Academy)

Dr. Robert Granger
US Naval Academy
Specialty: Mechanical Engineering

Dr. Timothy Troutt
Washington State University
Specialty: Mechanical Engineering

Dr. Clay Sharts
San Diego State University
Specialty: Chemistry

Dr. Hung Vu
California State University
Specialty: Applied Mechanics

GEOPHYSICS DIRECTORATE
(Hanscom Air Force Base)

Dr. Phanindramohan Das
Texas A&M University
Specialty: Geophysical Science

Dr. Thomas Miller
University of Oklahoma
Specialty: Physics

Dr. Alan Kafka
Boston College
Specialty: Geophysics

Dr. Henry Nebel
Alfred University
Specialty: Physics

Dr. Charles Lishawa
Utica College
Specialty: Physical Chemistry

Dr. Craig Rasmussen
Utah State University
Specialty: Physics

HUMAN RESOURCES DIRECTORATE
(Brooks, Williams and Wright-Patterson Air Force Base)

Dr. Kevin Bennett
Wright State University
Specialty: Applied Psychology

Mr. John Williamson (GSRP)
Texas A&M University
Specialty: Psychology

Dr. Deborah Mitta
Texas A&M University
Specialty: Industrial Engineering

Dr. Michael Wolfe
West Virginia University
Specialty: Management Science

Dr. William Smith
University of Pittsburgh
Specialty: Linguistics

Dr. Yehoshua Zeevi
Harvard University
Specialty: Electrical Engineering

Dr. Stanley Stephenson
Southwest Texas State University
Specialty: Psychology

Dr. Robert Zerwekh
Northern Illinois University
Specialty: Philosophy

MATERIALS DIRECTORATE
(Wright-Patterson Air Force Base)

Dr. Donald Chung
San Jose State University
Specialty: Material Science

Dr. Kenneth Currie
Kansas State University
Specialty: Industrial Engineering

Dr. Michael Resch
University of Nebraska-Lincoln
Specialty: Materials Science

Dr. James Sherwood
University of New Hampshire
Specialty: Aerospace Mechanics
210-9MG-088 and 210-10MG-098

Dr. Michael Sydor
University of Minnesota-Duluth
Specialty: Physics

OCCUPATIONAL AND ENVIRONMENTAL HEALTH DIRECTORATE
(Brooks Air Force Base)

Dr. Stewart Maurer
New York Institute of Technology
Specialty: Electrical Engineering

ROCKET PROPULSION DIRECTORATE
(Edwards Air Force Base)

Dr. Lynn Kirms
Southern Oregon State College
Specialty: Organic Chemistry

Dr. Mark Kirms
Southern Oregon State College
Specialty: Organic Chemistry

Dr. Faysal Kolkailah
California Polytechnic
Specialty: Mechanical Engineering

Dr. Vittal Rao
University of Missouri-Rolla
Specialty: Control Systems

Dr. Larry Swanson
University of Denver
Specialty: Mechanical Engineering

Dr. Roger Thompson
Pennsylvania State University
Specialty: Engineering Mechanics

ROME LABORATORIES
(Griffiss Air Force Base)

Dr. Charles Alajajian
West Virginia University
Specialty: Electrical Engineering

Dr. Ian Grosse
University of Massachusetts
Specialty: Mechanical Engineering

Dr. Henry Helmken
Florida Atlantic University
Specialty: Physics

Dr. Michael Klein
Worcester Poly Institute
Specialty: Physics

Dr. William Kuriger
University of Oklahoma
Specialty: Electrical Engineering

Dr. Khaja Subhani
Lawrence Tech. University
Specialty: Electrical Engineering

Dr. David Sumberg (1987)
Rochester Institute of Tech.
Specialty: Physics

Dr. Donald Ucci
Illinois Institute of Technology
Specialty: Electrical Engineering

Dr. Kenneth Walter (1988)
Prairie View A&M University
Specialty: Chemical Engineering

Dr. James Wolper
Hamilton College
Specialty: Mathematics

WEAPONS DIRECTORATE
(Kirtland Air Force Base)

Dr. Harry Hogan
Texas A&M University
Specialty: Mechanical Engineering

Dr. Arkady Kheyfets (1988)
North Carolina State University
Specialty: Mathematical Physics

Dr. Duc Nguyen
Old Dominion University
Specialty: Civil Engineering

Dr. Duane Sanders
Texas A&M University
Specialty: Civil Engineering

RESEARCH REPORTS

MINI-GRANT RESEARCH REPORTS

Technical Report Number Volume I	Title and Mini-Grant Number	Professor
Rome Laboratories		
1	Optimal Design of Finite Wordlength FIR Digital Filters for an Analog Transversal Filter with Tap Weight Circuitry Defects Using Adaptive Modeling 210-10MG-123	Dr. Charles Alajajian
2	Automatic Adaptive Remeshing for Finite Element Reliability Assessment of Electronic Devices 210-10MG-129	Dr. Ian Grosse
3	Ionospherically-Induced Phase Distortion Across Wide-Aperture HF Phased Arrays 210-10MG-047	Dr. Henry Helmken
4	A Study of Interacting Tunneling Units with Possible Application to High Temperature Superconductors 210-10MG-057	Dr. Michael Klein
5	Reduced Bandwidth Binary Phase-Only Filters 210-10MG-052	Dr. William Kuriger
6	Computer Modeling of GaAs/AlGaAs MQW Devices for Optical Properties 210-10MG-107	Dr. Khaja Subhani
7	Fiber Optic Distribution System for Phased Array Antennas 760-7MG-113	Dr. David Sumberg (1987)
8	Continuation Study of a Communications Receiver for Spread Spectrum Signals 210-10MG-067	Dr. Donald Ucci
9	Development of a System to Deposit Thin Films of Titanium Carbide Using Atomic Layer Epitaxy 219-9MG-113	Dr. Kenneth Walter (1988)

10	Neural Networks for Invariant Pattern Recognition 210-10MG-061	Dr. James Wolper

Arnold Engineering Development Center

11	The Performance of IR Detectors Illuminated by Monochromatic Radiation 210-10MG-029	Dr. Brian Beecken
12	Sodium Fluorescence Studies for Application to RDV of Hypersonic Flows 210-10MG-076	Dr. Stephen Cobb
13	Report Not Publishable At This Time 210-10MG-086	Dr. John Francis
14	NOT PUBLISHABLE AT THIS TIME 210-10MG-134	Dr. Orlando Hankins
15	An Experimental Approach for the Design of a Mixer for an Arc Heater 210-10MG-027	Dr. Lang-Wah Lee
16	No Report Submitted (1986) 760-6MG-099	Dr. Arthur Mason
17	Laser-Induced Fluorescence of Nitric Oxide 210-10MG-054	Dr. Chun Fu Su
18	Spectroscopic Monitoring of Exhaust Gases 210-10MG-099	Dr. Richard Tipping
19	Transient Analysis of Parallel Distributed Structurally Adaptive Signal Processing Systems 210-10MG-084	Dr. D. Wilkes

Electronic Systems Division

20	Reliability in Satellite Communication Networks 760-6MG-094	Dr. Stephen Kolitz (1986)
21	Comparison of Testability Analysis Tools for USAF 210-10MG-065	Dr. Sundaram Natarajan

Engineering and Services Center

22	Anomalous Effects of Water in Fire Fighting: Facilitation of JP Fires by Azeotropic Distillation Effects 210-10MG-115	Dr. William Bannister
23	Effect of Simulated Jet Aircraft Noise on Domestic Goats 210-10MG-119	Dr. Emerson Besch
24	Migration of Organic Liquid Contaminants Using Measured and Estimated Transport Properties 210-10MG-025	Dr. Avery Demond
25	Laboratory Investigations of Subsurface Contaminant Sorption Systems 210-10MG-064	Dr. Kirk Hatfield
26	Effects of Surfactants on Partitioning of Hazardous Organic Components of JP-4 Onto Low Organic Carbon Soils 210-10MG-125	Dr. Kim Hayes
27	Biodegradation of Hydrocarbon Components of Jet Fuel JP-4 210-10MG-018	Dr. Deborah Ross
28	760-7MG-079; See 210-10MG-095 Report # 71 (Aero Propulsion and Power Directorate)	Dr. William Schulz
29	Pretreatment of Wastewaters Generated by Firefighter Training Facilities 760-7MG-105	Dr. Dennis Truax (1987)
30	Stress Transmission and Microstructure in Compacted Moist Sand 210-10MG-019	Dr. George Veyera

Frank J. Seiler Research Laboratory

31	No Report Submitted (1985) 760-0MG-008	Dr. Hermann Donnert

52	Improved Modeling of the Response of Pressurized Composite Cylinders to Laser Damage 210-10MG-008	Dr. Harry Hogan
53	Relativistic Effects in Global Positioning 210-9MG-114	Dr. Arkady Kheyfets (1988)
54	No Report Submitted (1987) 760-7MG-047	Dr. Barry McConnell
55	Parallel and Vector Processing for Nonlinear Finite Element Analysis 210-10MG-051	Dr. Duc Nguyen
56	Resonant Scattering of Elastic Waves by a Random Distribution of Spherical Inclusions in a Granular Medium 210-10MG-085	Dr. Duane Sanders

Volume III

Wright Laboratory

Armament Directorate

Aero Propulsion and Power Directorate

79	Numerical Study of Surface Roughness Effect on Hypersonic Flow Separation 210-10MG-056	Dr. Arnold Polak
80	Ultrasonic Stress Measurements and Craze Studies for Transparent Plastic Enclosures of Fighter Aircraft 210-10MG-126	Dr. Nisar Shaikh
81	210-9MG-088, See 210-10MG-098 Report # 87 Materials Directorate	Dr. James Sherwood
82	Experimental Determination of Damage Initiation Resulting from Low Velocity Impact of Composites 210-10MG-094	Dr. William Wolfe
83	The Response of Nonlinear Systems to Random Excitation 210-10MG-093	Dr. Lawrence Zavodney

Materials Directorate

84	The In-Situ Deposition of High Tc Superconducting Thin Film by Laser Ablation 210-10MG-116	Dr. Donald Chung
85	Self-Improving Process Control for Molecular Beam Epitaxy of Ternary Alloy Materials on GaAs and InPh Substrates 210-10MG-030	Dr. Kenneth Currie
86	Detection of Fatigue Crack Initiation Using Surface Acoustic Waves 210-10MG-120	Dr. Michael Resch
87	Investigation of the Thermomechanical Response of a Titanium Aluminide Metal Matrix Composite Using a Viscoplastic Constitutive Theory 210-10MG-098	Dr. James Sherwood
88	No Report Submitted (1985) 760-0MG-067	Dr. Robert Swanson

99	Adenosine Modulation of Neurotransmitter Release from Hippocampal Mossy Fiber Synaptosomes 760-7MG-091	Dr. Ralph Peters (1987)
100	Behavioral and Neurochemical Effects of Radiofrequency Electromagnetic Radiation 210-10MG-035	Dr. Raymond Quock
101	An Investigation of Dioxin Half-Life Estimation in Humans Based on Two or More Measurements Per Subject 210-10MG-068	Dr. Ram Tripathi

Crew Systems Directorate

102	No Report Submitted (1985) 760-0MG-049	Dr. John Flach
103	Degradation of the Renal Peritubular Basement Membrane in Relation to Toxic Nephropathy from Compounds of Military Interest 210-10MG-101	Dr. Thomas Lockwood
104	Parametric Studies of the Breakdown of Total Information Processing Time into During-Display and Post-Display Components 210-10MG-024	Dr. Ethel Matin
105	A Blackboard Architecture for Landmark Identification on 3-Dimensional Surface Images of Human Subjects 210-10MG-077	Dr. Randy Pollack
106	Effect of System Reliability on Probabilistic Inference 760-7MG-094	Dr. Donald Robertson (1987)
107	Stable Grasping with the Utah/MIT Dexterous Robot Hand 210-10MG-034	Dr. Michael Stanisic
108	Articulated Total Body (ATB) "View" Program 210-10MG-053	Dr. Chi-Ming Tang

109	Explorations into the Visual Perceptual Factors Operating in High-Speed Low-Altitude Turns 210-10MG-105	Dr. Ebo Tei
110	No Report Submitted (1985) 760-0MG-071	Dr. Yin-min Wei

Human Resources Directorate

111	Computer-Based Training for Complex, Dynamic Tasks 210-10MG-015	Dr. Kevin Bennett
112	Report Not Publishable (1987) 760-7MG-100	Dr. Ronna Dillon
113	No Report Submitted (1986) 760-6MG-134	Dr. Stephen Loy
114	Advancing User Interface Capabilities in an Integrated Information Environment: A Fisheye Browser 210-10MG-110	Dr. Deborah Mitta
115	An Intelligent Teacher's Associate for Network Theory Based on the Heuristic of Polya 760-6MG-032	Dr. Philip Olivier
116	An Assessment of the Effects of CONFER: A Text-Based Intelligent Tutoring System Designed to Enact Tutorial Conversation and to Increase a Student's Sense of Intertextuality 210-10MG-003	Dr. William Smith
117	The Effect of Student-Instructor Interaction on Achievement in Computer-Based Training 210-10MG-006	Dr. Stanley Stephenson
118	No Report Submitted (1985) 760-0MG-030	Dr. Christian Wagner
119	An Evaluation of Stereoscopic and Other Depth Cues in Computer Display 210-10MG-112	Mr. John Williamson
120	New Architectures for WISIWYSWIWSWYS 210-10MG-028	Dr. Michael Wolfe

1989-1990 USAF-UES RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the Universal Energy Systems, Inc.

FINAL REPORT

Confirmation of the Possible Role of Lipopolysaccharide in Expressing an Abelson Murine Leukemia Virus in RAW 264.7 Macrophage Cells

Prepared by:	Robert Vernon Blystone
Academic Rank:	Professor of Biology
Department and	Biology
University:	Trinity University
Research Location:	USAFSAM/RZP (now AL/OEDR) Brooks AFB San Antonio, TX 78235 and Dept. of Biology Trinity University San Antonio, Texas
USAF Researcher:	Dr. Johnathan Kiel
Date:	July 1, 1991
Contract No:	F49620-88-C-0053/SB5881-0378

Confirmation of the Possible Role of Lipopolysaccharide in Expressing an Abelson Murine Leukemia Virus in RAW 264.7 Macrophage Cells

by

Robert V. Blystone

ABSTRACT

The original thrust of this research was modified in order to develop techniques that would allow better visualization of RAW 264.7 macrophage cells while in their culture environment. This was deemed necessary because the measurement of RAW cell response to such low quantities of applied lipopolysaccharide demanded very sensitive techniques to measure cell growth in culture. Techniques for monitoring live cell growth with an optical flatbed scanner were developed. Such techniques were correlated with morphological investigation of the same cultured cells with light and electron microscopy. Low cost methods for rendering gray scale growth data in three dimensions were developed. Techniques are now in place to monitor and visualize viral production by RAW cells.

I. INTRODUCTION:

Research initiated in Dr. Kiel's lab during the summer of 1989 resulted in the discovery of viral expression by RAW 264.7 mouse macrophage cells. The expression of this probable Abelson leukemia virus was speculated to be linked to the presence of added lipopolysaccharide (LPS) to the RAW cell culture media. The intent of the minigrant was to confirm the relationship between LPS exposure and viral expression. Subsequent observations made before and during the course of the research period suggested changes in the research questions asked and answered. Given these changes in research direction, the role of LPS induction of viral expression in RAW macrophages was neither confirmed or denied.

With the concurrence of Dr. Kiel, the focus of the study was shifted to better understand the growth habits of cultured RAW cells. Experimentation revealed that RAW growth could be affected by cell number at seed, by age of cells, and by manipulation of cells during culture. During the research period, additional image analysis equipment was acquired. This addition allowed for more ambitious techniques to be developed for the image processing of RAW cells. A major new goal was the three dimensional reconstruction of a serially sectioned RAW cell that had been treated with LPS. Therefore, the goals of the research period were widened and shifted from those stated in the original proposal.

A second USAF-UES Summer Faculty Research fellowship was awarded for Summer 1990. This document reflects, in part, material previously reported

for the 1990 summer fellowship period and additional research information collected during the Fall and Winter of 1990.

II. OBJECTIVES OF THE RESEARCH EFFORT:

The revised objectives for the project were to better understand the dynamics of growth in RAW cells in culture and to interface the techniques of scientific visualization with the growth and morphology data of *in situ* cells. These objectives were designed to complement the evolving growth factor work of Dr. Kiel. Additionally, Dr. Kiel was impressed by initial results in visualizing macrophage foci in culture. He encouraged the development of visualization techniques to this end.

III. APPROACH TO THE PROJECT:

The RAW 264.7 mouse macrophage cells (referred to as RAW cells) were cultured in RPMI 1640 culture medium fortified with 10% fetal calf serum and antibiotics added. The cells were typically incubated at 37°C in a 5% CO_2 atmosphere. When numbers of cells were desired, or for holding cells for a future experiment, culture was performed in T-75 plastic flasks. For experimental purposes, cells were grown in LabTek 8-chamber culture slides. The small surface area and volume of these 1 cm-on-side multiple growth chambers allows for constrained use of supplies, manageable scan areas for microscopy, and many repetitions of an experiment. Typically, the RAW cells were passaged every 4 to 10 days with 7 day intervals most common.

Early Spring 1990 work suggested viral release varied under differing culture conditions. Continuing through the summer, it was obvious that RAW cells had unique growth characteristics exclusive of LPS and viral parameters. These unique qualities included the following:

a. irregular growth over the surface of the slide (high and low density areas);

b. cell derived from confluent monolayers (termed old) behaved differently in growth than cells from log phase cultures (termed young); and

c. cells seemed to grow in preferential patterns depending on the well position among the eight wells of the slide.

An immediate question based on these initial observations revolved about the extent that LabTek slide culture wells influenced the growth of RAW cells.

Morphology at the microscopic level is made difficult by the expanse produced by magnification. With a surface area of less than one square centimeter and by being on glass, the morphology of cultured cells can readily be sampled by several techniques of microscopy on the same constrained area. By being able to recognize and examine the same cells at magnifications ranging from 1 to 100,000 times, a more complete morphological picture can evolve. That is a major reason why the LabTek culture well slides were used as the principle growth environment. Experience gained with these constrained growth environments quickly revealed that larger environments such as the T-75 flasks could also show preferential growth patterns. This digression into the relationship of growth chamber to cell growth would later prove interesting.

IV. THE OPTICAL METHODOLOGY FOR MORPHOLOGY:

The experimentally treated cells were optically examined at several levels.

A. Gross culture morphology-

The glass slide was placed on an HP ScanJet optical scanner. The culture (living or fixed) was digitized providing a 300 dot per inch TIFF file. This image can be analyzed by several means.

B. Low power individual cells-

The same glass slide could be placed in a light microscope and examined. Only fixed cultures can be viewed because the plastic chamber devise must be released from the slide for viewing in a standard compound microscope. The microscope has an attached digital camera (Javelin RGB chip) which can send an RGB signal to a Data Translation DT2255 frame grabber board. The image is, therefore, digitized and can be analyzed by several means.

C. High power individual cells-

The same glass slide can be prepared and placed in a scanning electron microscope (SEM). Here micrographs can be made at moderate magnification ranges (200X to 2,000X). The resultant photomicrographs can be digitized on a HP ScanJet optical scanner. These images can be analyzed by several means.

D. High power views of individual cell components-

The same glass slide can be prepared for transmission electron microscopy (TEM). The cells are released from the glass, sectioned, and examined at magnifications ranging from 4,000X to 40,000X. Resultant transmission electron micrographs are digitized on a HP

ScanJet optical scanner. The now digital images can be analyzed by several means.

This approach worked well except for the last step (D) involving TEM. The cells degraded too much from prior handling. The goal was to follow a cell culture in situ by optical means. The approach was successful up to the last step.

Cell were fixed and stained by the following method.

1. Primary fixation in 3% glutaraldehyde in 0.1 M sodium phosphate buffer.
2. Washed in 0.1 M sodium phosphate buffer.
3. Stained in 1% OsO_4 in 0.05 M sodium phosphate buffer.
4. Dehydrate in a graded ethanol series.

5A. For SEM, the cells were dried in amyl acetate and coated with gold.

5B. For TEM, the SEM cells were embedded in plastic and released from the glass culture slide. The resultant blocks were sectioned with glass knives and post stained with lead citrate.

Microscopes used were the following: LM, Zeiss BK-2; SEM, ISI-40; and TEM, Hitachi HS-8.

The digital images were processed in a Mac 2ci operating under either Mac OS or AUX-2. The software packages used in analysis and visualization were

NIH Image (various versions), Enhance 1.0, SpyGlass (in various forms), Renderman, CricketGraph 1.3, StatView II, Excel 2.2 and Word 4.0.

V. COMPUTER ANALYSIS AND VISUALIZATION:

Optical scanning of whole LabTek slide ultimately revealed a wealth of information. Each chamber is gray scale averaged. That is, each surface produces an average optical density ranging from between 0 and 255 on the gray scale. (See Figure 3.) One chamber of each 8-chamber slide serves as a background control by having no cells in it, only culture medium. The background density is subtracted from the other 7 chambers. The gray scale value corresponds to the density of the cells. In the case of living cells, the values move to the white range because the living cells become white with increasing density. With the dead, stained cells, the gray scale values move to the black because the cells are stained with osmium.

A variety of analytical techniques are possible with the gray scale data. The gray scale image data can be converted to a 3-D representation which provides an immediate visual confirmation of the distribution of growth over the surface of the chamber. (See summer 1990 fellowship report for a figure.) The image may also be reduced to a predetermined gray scale threshold value and converted to a binary image. The binary image provides data on cell coverage; that is, what proportion of the surface is covered with cells. The gray scale information and image can also be pseudocolored which can provided additional visual confirmation of growth distribution within the well.

Cell images captured with the light microscope can be treated in a similar fashion, however, they represent smaller sampling areas due to the magnification of the microscope. These smaller areas have greater resolution and can be compared with the optically scanned areas for confirmation of data. The correlation was good. Cell diameters and cell morphology could be quickly measured for large numbers of cells. Foci are easily resolved at this level of magnification.

SEM provides a detailed look at the cells' morphology. Successive SEM photomicrographs were assembled into movie sequences showing overflights of the cell culture surface. Although dramatic appearing, more work needs to be done to pull information from this visualization technique.

Not much time was spent processing TEM photomicrographs. Since the viral aspects of the project were delayed in terms of investigation, computer visualization of TEM images was not a priority in the time available.

Figure 1 and 2 does provide an example of what computer imaging can accomplish with the growth data. The first view shows a single well of an RAW culture that represented five days growth. The actual area represented in the photo is one centimeter on the side. The data in the picture was rendered into a three dimensional gray scale and the resultant view is seen in figure 2. A significant amount of time was spent learning how to produce such views of the data. Project time ran out and work is progressing in applying these techniques to experiments.

VI. OTHER ACCOMPLISHMENTS:

Two research presentations were made based in part from information collected during the project period. A paper was given at the XII International Congress for Electron Microscopy. The printed abstract is attached. A second paper was given at the 31st annual meeting of the American Society for Cell Biology. The printed abstract for this paper is attached. Two formal papers detailing the work outlining the optical scanning and rendering techniques are in preparation. One paper concerning the growth of the cells monitored by these techniques is in an early stage of preparation.

Two undergraduate students were hired with resources provided by this grant during the spring and summer of 1990. Both students (T. Romo and C. Collumb) described their research experiences at an undergraduate research conference at the University of Chicago in September of 1990. T. Romo presented an additional research paper to an undergraduate research conference at Cal Tech during March of 1991. Three students were hired with resources provided by the grant during the Fall of 1990 (T. Romo, M. Wentland, and H. Papaconstantinou). All of these students were able to interface with Dr. Kiel at Brooks Air Force Base.

VI. CONCLUSIONS AND RECOMMENDATIONS:

The work presented here primarily describes the establishment of new techniques to monitor cell growth using computer image analysis. Also a cell culture facility is being established at Trinity University based on

experienced gained during this research period. Additional equipment is being acquired to allow for light microscopy of live RAW cells. Many of the objectives of a second RIP grant will be met in the summer of 1991. With these objectives met, the original protocol envisioned in this RIP grant can be addressed: confirmation of LPS in viral induction of RAW cells.

ACKNOWLEDGMENTS

My deepest appreciation is extended to Dr. Johnathan Kiel. He has opened his lab to me and my undergraduate students. Our discussions concerning the progress of my research is has been most helpful.

Dr. Jill Parker at Brooks has provided patient guidance in cell culture technique.

Sgt. Alls has helped make many preparations and solutions for us.

Dr. R. Burton, chief science officer for USAF/SAM has provided a friendly and interactive environment on the base.

To all of these kind people, thanks.

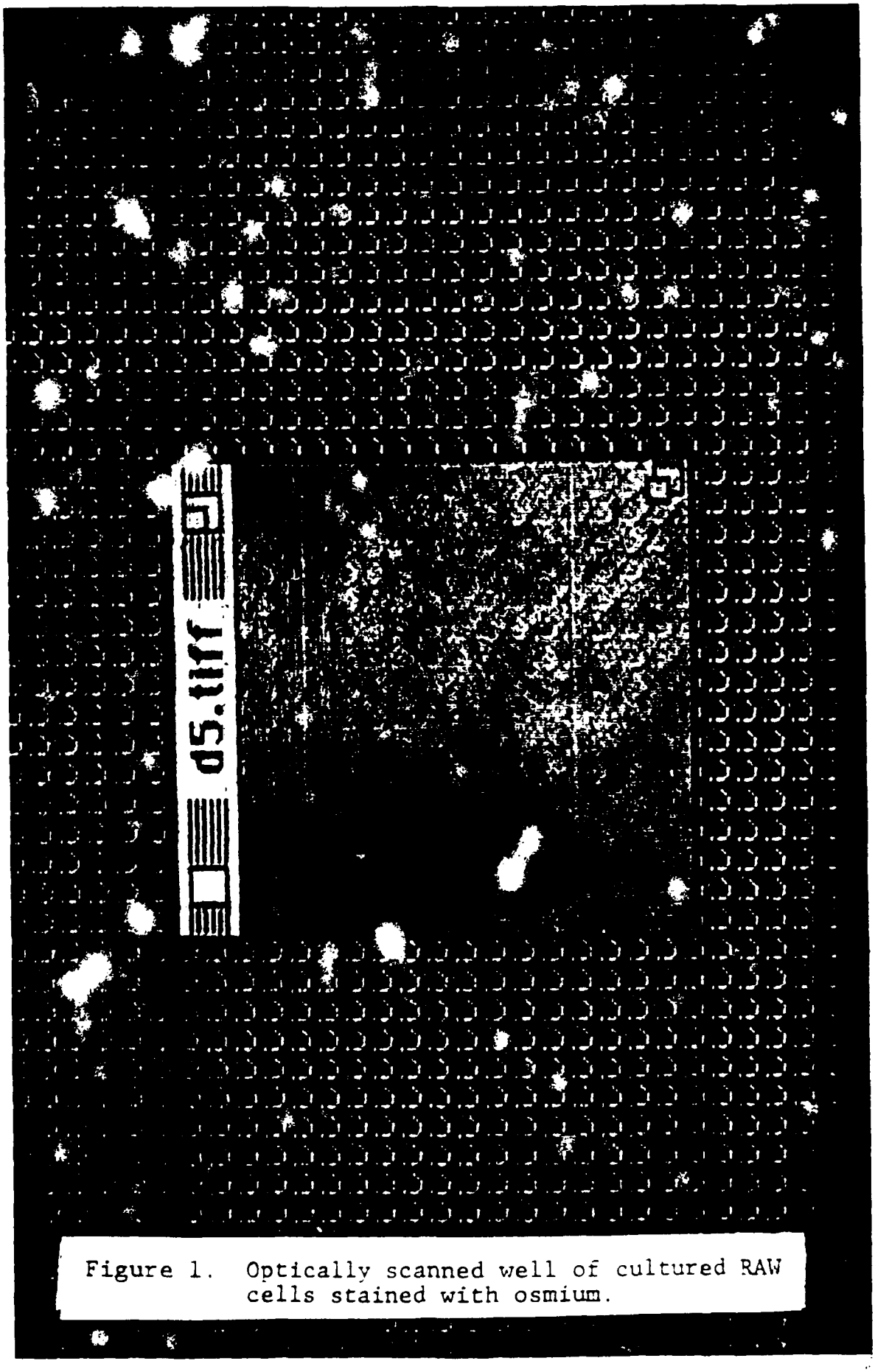


Figure 1. Optically scanned well of cultured RAW cells stained with osmium.

Figure 2. Computer visualized version of Figure 1.

Effect of Microwave Radiation on Cultured Cells

Final Report

Sponsored by the
Air Force Office of Scientific Research

Conducted by
Universal Energy Systems, Inc.

Contract Number: F49620-88-C-0053/SB5881-0378

Prepared by: Carolyn Alexander-Caudle, Ph.D.

Department/University: Department of Biological Sciences
Tennessee State University

USAF Collaborators: Dr. Johnathan L. Kiel
Dr. Jill Parker

I. ABSTRACT

The potential hazardous nature of radiofrequency radiation (RFR) has been a great concern for several years. The human population has been increasingly exposed in modern times to nonionizing radiation, specifically microwave. Current research suggest that biochemical and pathological changes may occur after cells are exposed to nonionizing radiation. The harmful effects of RFF at low exposure levels are unclear and reports in the literature are contradictory. This study is designed to determine the effects of non-ionizing radiation on mammalian cells. The focus is on Tumor Necrosis Factor (TNF), a cytokine which is rarely found in measurable amounts in body fluids of healthy individuals.

II. ACKNOWLEDGEMENTS

I gratefully acknowledge the Air Force Systems Command and the Air Force Office of Scientific Research for the sponsorship of this project; and Universal Energy Systems for administration of the grant.

Special gratitudes are extended to Drs. J. Kiel and J. Parker for collaborating with me, providing direction and support, and test samples. Colleagues at Tennessee State University: Dr. A. Adibi, for the use of $C0_2$ incubator and hood, Dr. E. L. Myles for daily interactions and use of equipment and supplies, R. Nesby for technical assistance, and T. Gray for administrative assistance and support.

III. INTRODUCTION

Carswell et al. (1975) identified a cytotoxin (CTX) in the serum of mice infected with bacille calmette Guerin (BCG) and subsequently treated with endotoxin. They named this CTX tumor necrosis factor (TNF) because of its ability to induce selective tumor necrosis. It is now apparent that CTX can be produced by a wide variety of cell types, but activated macrophases and mononuclear phagocytes are the major TNF-producing cells (Carswell et al., 1975).

Studies, in vivo and in vitro, show that the endotoxin LPS (lipopolysaccharide) appear to be one of the most potent inducers of TNF (Meager, 1989). Other substances such as phorbol esters, interlukin-2 (IL-2) and sendai viruses have been reported as inducers in vitro. The production of TNF may be effected by other cell types such as natural killer (NK) cells, T lymphocytes, lymphokins and macrophage colony-stimulating factor (m-CSF) (Meager, 1989)).

Tumor necrosis factor, which has been previously described as a cytolytic agent capable of killing transformed cells with a high degree of selectivity (Beulter et al. 1985), may play a critical role in many physiologic and pathophysiologic responses (Parat, O, 1989).

Detectable concentrations of TNF may be found in body fluids during episodes of acute microbial infections and during the course of chronic invasive diseases (Meager, 1989)). TNF has been reported to be found in serum, cerebrospinal fluid, synovial fluid, bronchial lavage fluid, vesical

fluid, lymph and urine. TNF is also one of the major signals in regulating the production of interleukin-l (Feldmann, 1990).

TNF consist of two proteins designated TNF-α (also called cachetin (Paul and O'Hara, 1987) and TNF-β (also called LT (lymphotoxin) (Parker, 1982). TNF-α is produced primarily by mononuclear phagocytes and T cells in human, and in mouse both the $T_{H}1$ and $T_{H}2$ CD_4^{+T} cell subsets and CD_8^{+T} cells can synthesize the protein.

Two forms of TNF-α have been reported, a membrane form thus far observed only in human cells has a molecular weight of 26K Da and a soluble form with a molecular weight of 17K Da. Human TNF-α has not been found to be glycosylated but mouse TNF-a possesses one potential n-glycosylation site, but no significant level of glycosylation has been reported.

IV. OBJECTIVES

This study was designed to investigate the effects of microwave radiation on the production of tumor necrosis factor (TNF) in cultured cells; in the presence of the metabolic stimulant LPS (lipopolysaccharide) alone or in combination with the metabolic inhibitor 3-amino-L-tyrosine 3(AT).

V. RESULTS

Supernatant samples sent to me from Dr. Kiel were tested for TNF activity using a biological assay. The activity of TNF was monitored by a lytic assay using L929 (mouse connective tissue cell line) cells as described by Aggarwal et al., (1985). Trypsinized L929 cells were counted and resuspended in RPMI-1640 media supplemented with 3% Fetal Calf serum to a cell density of 2 X 10^5 cells/ml. The cells (0.1 ml volume) was grown in 96 well sterile tissue culture treated Falcon plastic microtiter plates for 20-24 hours to form a monolayer at 37°C in 5%$C0_2$. Supernatant samples were serially diluted in RPMI in the presence of 1 mg/ml of actinomycin D and added to the monolayers. At the end of an additional 16-24 hours of incubation, the media was aspirated off the L929 cell monolayer and the cell stained with 50µl of 1% solution of crystal violet for 3-5 minutes at room temperature. The crystal violet was first dissolved in 95% ethanol, diluted with deionized water to 50% and filtered through a 0.2 µm filter to remove undissolved particles. After 3-5 minutes, the excess crystal violet was removed by vigorously rinsing the microtiter plates with tap water. The plates were inverted to remove excess water and allowed to dry. After the plates were dried, 100µl of 95% ethanol containing 40mM HCl was added to each well to re-dissolve the crystal violet that stained the L929 cells. The crystal violet was evenly re-suspended in the HCl-Ethanol solution by gently agitating the microtiter plates. The end point of the microtiter plates was determined by a microplate reader (Model 700). The number of cells in the well was directly proportional to the amount of absorbance. The cells

exposed to culture media alone were at 0% lysis and those exposed to 3mM quanidine hydrochloride was 100% lysis. One unit of TNF is defined as the reciprocal of the dilution required for 50% cell lysis.

L929 cells initially obtained from Dr. Kiel presented a big problem for me so L929 cells were obtained from Dr. Gale A. Granger (University of California, Irvine.)

L929 cells should be passed twice weekly by adding 0.25 X 10^6 cells in 12 ml of RPMI 1640 with 3% fetal calf serum in T-25cm^2 flask or 0.75 X 10^6 cells in 30 ml of media in T-75 cm^2 flask. When confluent, media was decanted, rinsed with 10 ml PBS and 1-2 ml of 0.1% trypsin was added and incubated at 37°C for 2-5 minutes. The action of trypsin was stopped by the addition of 10ml of RPMI with 3% serum.

Many supernatant samples appeared to be positive from the biological assay but results are in the process of being repeated for reliability and reproducibility. Several factors were found to be critical for the assay. Similar factors have also been reported by Meager et al. 1989. The inoculation cell density is very important. Sensitive assays are obtained if cells are taken from the period of logarithmic growth and seeded in 96 well microtiter plates at a low density (1-4 X 10^4 cells/well). The concentration of serum may also affect the outcome of assays. Variation in results can occur when serum containing media is used for the dilution of samples especially if the TNF activity is low. Results in Figure I from one-dimensional SDS polyacrylamide gel electrophoresis as described by Laemmli (1970) suggest that TNF may be present in serum.

Sensitivity of the biological assay may be increased by including one or more metabolic inhibitors such as actinomycin D (AmD), mitomycin C or emetin in the culture medium. The most common inhibitor has been reported (Meager,1989) to be AmD and the sensitivity of L929 cell increased ten fold and the assay length shortened when 1 μg/ml of AmD was used.

The widespread availability of mouse L929 cells has insured its popularity for the biological assay but differences in sublines of the L929 have been found and this will also affect sensitivity of the assay. Cells kept in continuous passage have been found to show a decline in susceptibility. Other mouse cell lines such as L-M (American Type Culture Collection, CCL 12); EMT-6 and the WEHI 164 clone 13 have also been found to be suitable for this assay (Meager, et al. 1989).

Cytotoxic assays are not most ideal for measuring biological activity of TNF because the outcome of such assays can be affected by numerous endogenous and environmental factors. These factors can distort results to a degree that the wrong conclusions are drawn and data may not reflect the relative concentrations of TNF in test samples.

Supernatant samples were also tested using an immunoassay. ELISA test kits for mouse tumor necrosis factor mTNF-α) were obtained from Genzyme Catalog #1509-00. The mouse TNF-α ELISA is a solid phase immunoassay employing the multiple antibody sandwich technique. A hamster monoclonal antibody specific for murine TNF was allowed to attach overnight to the 96-well microtiter plates. The following day the monoclonal antibody was removed by aspiration and plates were washed three times with wash buffer provided in the kit. TNF present in the

standard samples or supernatants samples from Dr. Kiel was added to plates for a total of 2 hours. Plates were washed to remove test samples and a goat polyclonal anti-mouse TNF antibody was added for 90 minutes. A third antibody, donkey anti-goat Ig peroxidase conjugated, was added for one hour. The peroxidase enzyme acted with peroxidase substrate and OPD (chromagen) to produce a yellow color if TNF was present. The reaction was stopped by the addition of 2N H_2S0_4 and the absorbance was read at 490 nm using an ELISA reader. The measured absorbance was proportional to the concentration of TNF that was present in the original sample. A reference curve is obtained by plotting different concentrations of TNF (3.2 - 0.05 ng/ml). The TNF concentration in the test samples was determined by comparing the absorbance with that obtained from the known amount of TNF in the standards. The results from this study are listed in Table I.

The detection limit of this ELISA Assay is 100 pg TNF/ml. This was established by observing that the mean absorbance obtained with 100 pg TNF/ml was greater than two standard deviations higher than the mean baseline absorbance obtained from replicative negative control wells. Two standard deviations was chosen as the background cut off of statistical interpretations of the ELISA results (Parkinson et al. 1988).

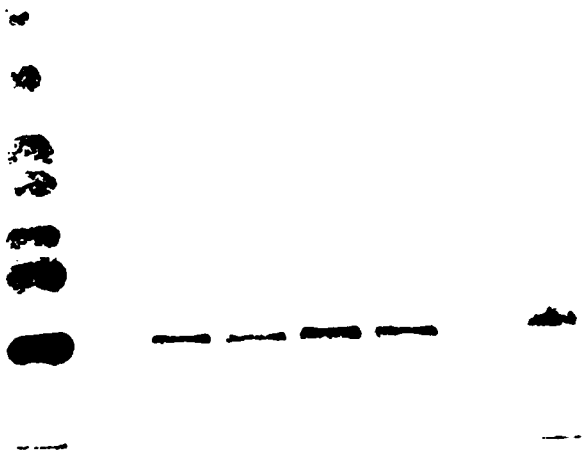


Figure I. SDS PAGE

Lane I	7.5μg of molecular weight standards
Lane II	TNF
Lane III	3T3 Supt. 0.5 LPS
Lane IV	3T3 Supt. 0.1 LPS
Lane V	36 Sham Supt
Lane VI	36 Exp
Lane VII	5μg molecular weight standards
Lane VIII	Serum

SAMPLES ASSAYED FOR TNF BY
MOUSE TUMOR NECROSIS FACTOR - ALPHA ELISA

	ng/ml TNF
4(A)4) 1/25/90 supt	.54
RPMI 3% Serum	.027
0.01 X 100 (6) supt RAW	>3.2
3A(6)5% 2/1/90 3T3 Supt	n.d.a.
Supt 3T3 11/15/89 10% 0.01 24h LPS	.210
0.1 X 1000 1/16/90 7 day	1.36
3T3 supt 1 9/15/89 0.0% LPS 24h	.211
NIH 3T3 supt 11/15/89 0.01% LPS Supt	.347
NIH 3T3 supt Cont (24 hr) no LPS Supt. 11/15/89	.344
Supt 3T3 11/15/89 0.05 10% LPS 24h	.191
0.05 1/16/90 4 day	.936
Supt 3T3 11/15/89 0.05 LPS (LPS 24h)	.430
4B6 5% 3T3 supt 1.2.90	n.d.a.
5A 3T3 supt 2/5/90	n.d.a.
Supt 6 day from 3T3 0.05 50% 1/22/90	n.d.a.
Supt NIH 11/15/89 (0.01 LPS 24h)	1.3
3C(6) 5% 3T3 supt 2/1/90	n.d.a.
0.1 1A 1/18/90 6 day	1.229
Supt 3T3 11/15/89 0.05 LPS (LPS 24h)	.544
0.01 X 1000 2E 1/18/90 6 day	1.14
Control 6 day from 3T3 10% 1/22/90	.0584
6 supt NIH 3T3 11/15/89 0.01 (LPS 24h)	1.14
0.1 1/16/90 4 day supt	.661
3T3 148 (0.05) 301	n.d.a.
Supt 6 day from 3T3 0.01 X 1000 10% 1/22/90	n.d.a.
0.05 X 1000 2B 6 day	1.145
Supt 6 day from 3T3 0.05 X 1000 10% 1/22/90	n.d.a.
Expt 59 Shan 6/26/89 Supt	.967
Expt 59 RFR 6/26/89 supt	.809
Expt 59 inch 6/26/89 supt	.474
Expt 44 incub 5/17/89 supt	.934
Expt 44 RFR 5/17/89 supt	.848
Expt 44 Sham 5/17/89 supt	1.103
5B (3T3 supt 2/5/90)	
7B RAW Filtered supt	.183U
3T3 0.1 2/8/24	.108
.1 LPS	.714
3A6 supt 3T3 10%	.0648
6B filtered supt 24190	.106
Expt 71 sham 8/2/90 supt	.694
Supt 3T3 11/15/89 0.05 1% LPS (24hr)	.587
3C (6) 10% 3T3 Supt 2/1/90	.111

SAMPLES ASSAYED FOR TNF BY MOUSE TUMOR NECROSIS FACTOR - ALPHA ELISA

	ng/ml TNF
10% Cont (6) 3T3 Supt 2/1/90	n.d.a.
Expt 59 RFR 2/19/89	.356
6B 3T3 supt 2/5/90	.038
3B (6) 5% 3T3 supt 2/1/90	n.d.a.
3T3 supt 11/15/89 0.1 LPS 24 H	.453
3T3 supt 4C (6) 10% 2/1/90	.093
0.01 1/16/90 1 day supt	1.03
4B (4) 1/25/90 supt	.363
Cont 1/16/90 4 day	.6349
Supt 6 day from 3T3 .05 10% 1/22/90	n.d.a.
0.05 X 1000 1/16/90 4 day	.897
Supt cont NIH 3T3 11/15/89 no LPS	.102
0.1 X 1000 2A 1/18/90 6 day	1.22
0.01 X 1000 1/16/90 7 day	1.093
.01 6 day 1C 1/18/90	1.876
Supt 6 day from 3T3 0.015% 1/22/90	.596
3T3 Supt 10/15/89 0.1 (LPS 24h)	1.03
3A (4) 1/25/90 supt	.162
NIH 3T3 supt 11/15/89 0.05 10% LPS (24h)	.391
1/25/90 cont	.255
0.05 X 1000 (6) supt 4B(6) RAW 1/29/90	1.217
4C (A) 1/25/90 supt	.160
3C (7) 1/25/90 supt	.208
7A 3T3 supt 2/5/90	n.d.a.
Supt from 3T3 0.1 5% 1/22/90	n.d.a.
0.05 1B 6 Day 1/18/90	1.27
4B (6) 10% 3T3 Supt 2/1/90	n.d.a.
5% cont (6) 2/1/90	n.d.a.
Supt 6 day from 3T3 0.01 10% 1/22/90	n.d.a.
3T3 supt cont	.046
3T3 supt 3A (4) 1/27/90	n.d.a.
Supt 6 day from 3T3 0.1 10% 1/22/90	.0866
0.01 (6) 3C supt RAW 1/29/90	2.23
Control 1D 6 day 1/18/90	1.44
3T3 supt 3C (4) 1/29/90	n.d.a.
3T3 supt 3B (4) 1/29/90	.419
4(A) (6) 10% 2/1/90 3T3 supt	.056
Cont (6) 1/29/90	n.d.a.
6CA1 Filtered RAW supt 2/1/90	.310
5A Filtered supt 2/1/90	n.d.a.

SAMPLES ASSAYED FOR TNF BY MOUSE TUMOR NECROSIS FACTOR - ALPHA ELISA

	ng/ml TNF
3T3 cont supt 1/12/90	n.d.a.
3B(6) supt 2/1/90 10%	.0848
Supt 3T3 11/15/89 0.05 LPS (LPS 24h)	1.86
0.05 (6) 1/29/90 RAW	.391
3T3 Supt New RAW's 1/29/90	n.d.a.
RAW 5(B) Filtered supt 1/2/90	.0686
7A Filtered RAW supt 2/1/909	.0636
3T3 supt 4A (4) 1/29/90	n.d.a.
3T3 supt 4B (4) 1/29/90	n.d.a.
3T3 0.05 2/8/2A	n.d.a.
New RAW supt to 3T3 2/12/90 filtered	n.d.a.
0.1(6) 3A6 1/29/90 supt RAW	.747
6A 3T3 supt 2/5/90	n.d.a.
7B 3T3 supt unfiltered 2/5/90	n.d.a.
7B 3T3 supt 2/5/90	n.d.a.
Supt from untreated 3T3 supt	n.d.a.
4A 5% 2/1/90 3T3 supt	n.d.a.
4C 5% 3T3 supt 2/1/90	n.d.a.
New RAW supt to 3T3 2/12/90	n.d.a.
0.1 X 1000 (6) 4A 6 1/27/90 supt RAW	.5006
RAW new + P4 supt	.253
5% .05 LPS 3T3 supt	1.356
3T3 12/8 30' .01	n.d.a.
10% .05 LPS 3T3 supt	.0725
3T3 1/8 .1 301	n.d.a.
.05 3T3 11/14/89	n.d.a.
.05 LPS	.986
Supt from 3T3 6 day control	.076
Cont 3T3 supt	n.d.a.
.01 3T3 11/14/89	.599
.1 3T3 1/14/89	.486
5% .1 LPS 3T3	.663
Cont 11/14/89	.769
3T3 .01 12/8	n.d.a.
1% .05 LPS 3T3	n.d.a.
.01 LPS	.566
5% .01 3T3	.594
Expt 81 INCB 10/25/87 supt	.731
Expt 81 Sham 10/25/90 supt	.467
Expt 81 RFR 10/25/87 supt	.560
Expt 74 Sham 8/23/89 supt	2.01

SAMPLES ASSAYED FOR TNF BY MOUSE TUMOR NECROSIS FACTOR - ALPHA ELISA

	ng/ml TNF
Expt 74 RFR 8/23/89 supt	2.02
Expt 76 Sham 10/2/89 supt	2.09
Expt 73 RRF 8/23/89 supt	.426
Expt 73 Sham 8/22/89 supt	.830
Expt 87 Sham 11/1/89 supt	.971
Expt 87 RFR 11/1/89 supt	.391
Expt Sham 10/31/89	.405
Expt 86 RFR 10/31/89 Supt	.354

VI. RECOMMENDATIONS

The lack of specificity and reproducibility in the biological assay allows me to make the following suggestions:

(1) All test sample for TNF should be screened first by an immunoblot kit for Genzyme. Positive samples should be further quantitated with the ELISA Kits. The shorter periods for the assay and precision should compensate for the amount of the Kits.

(2) Culture media containing 1-10% serum may induce alterations in absorbance when compared to dilution buffer in the ELISA Assay Kit (Genzyme). The biological assays should be repeated using serum free media. Kraner and Carver (1986) replaced a serum-containing media with a serum free preparation to increase reducibility using L-M cells in media 199 supplemented with 0.5% Bacto-peptone and 24mM Hepes Buffer. Attempts will be made to use this cell line in our lab.

VII. REFERENCES

Aggarwal, B., W. Kohn, P. Hass, B. Moffat, S. Spenser, W. Hanzel, T. Gringman, G. Nedwin, D. Goeddel and R. Hawkins, (1985). Human Tumor Necrosis Factor. J. of Biol. Chem 260: 2345

Beutler, B. J., N. Mahoney, P. Pekala, A. Cerami. (1985). Purification of Cachetin: A Lypoprotein lipase suppressing hormone secreted by endotoxin induced suppressing RAW 284.7 Cells. Exp. Med. 161: 984

Carswell, E.A., L.J. Old, R. L. Kassel, S. Green, N. Fiore and B. Williamson. (1975). An endotoxin induced serum factor that causes necrosis of tumors. Proc. Natl. Acad. Sci. USA 72: 3666

Flick, D.A. and G. E. Gifford. (1984). Comparison of in vitro cell cytotoxic assay for tumor necrosis factor. J. Immunol. Med. 68: 167

Parker, D.C. (1982). Separable helper factor support B cell proliferation and maturation to Ig Secretions. J. Immunol. 129: 469

Parkinson (1988). The interpretation of ELISA results by means of the standard deviation ration. J. Immunol. Methods 15: 105

Paul, W. E. and J. O'Hara. (1987). B-cell stimulatory factor/interleukin 4. Ann. Rev. Immunol 5: 429

Porat, O. (1989). Tbe effect of tumor necrosis factor alpha on the activity of lipoprotein lipase in adipose tissue. Lymphokine Res. 8:459

Laemmli, U.K. (1970). Cleavage of Structural protein during assembly of phage T4 head. Nature Vol 227:880

Meager, A., H. Leung and J. Woolley. (1989). Assay for tumor necrosis factor and related cytokines. Journal of Immunol 116: 1

Sheehan, R., N. Ruddle, R.D. Schreibes (1989). Generation and characterization of Hamster monoclonal antibodies that neutralize murine tumor necrosis factors. J. Immunol. 142: 3884.

Tracey, K. J., A. Cerami (1990). Metabolic response to cachectin/TNF. A Brief Review. Ann N.Y. Acad Sci 587:325

Yamazaki, S., E. Onishi, E. Anami, K. Natori, M. Kohase, H. Sakamoto, M. Tanouchi, H. Hayashi. (1986). Proposal of standardized methods and reference of assaying recombinant human tumor necrosis factor. J. Med. Sci Biol 39: 105

USAF-UES RESEARCH INITIATION GRANT

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

IN VIVO PROCESSING OF TETRAISOPROPYL PYROPHOSPHORAMINE

Prepared by:	**James P. Chambers**
Academic Rank:	**Professor**
Department:	**Division of Life Sciences**
University:	**The University of Texas at San Antonio**
USAF Researcher:	**Dr. James P. Chambers**
Date:	**June 24, 1991**
Contract No:	**F49620-88-C-0053/SB5881-0378**

I. INTRODUCTION:

Cholinesterases (ChEs) are present in the blood of most animals, as well as in liver, intestinal mucosa and other tissues. They exist in several molecular forms and other than the role of acetylcholinesterase (AChE) in neural transmission, have no physiological role but are known to hydrolyse a range of choline esters and a number of drugs [1].

In rat serum, two different enzymes hydrolyze choline esters. Acetylcholinesterase (AChE, EC 3.1.1.7) and a second enzyme butyrylesterase (BChE, EC 3.1.1.8) which hydrolyzes butyrylthiocholine in addition to many other esters.

In addition to the ChEs, rat serum contains high levels of carboxylesterase (CaE, EC 3.1.1.1) activity [2]. The functional role of CaEs is obscure, except possibly in the metabolism of certain xenobiotic esters [3]. The ChEs and CaEs have been shown to exhibit widely different and overlapping inhibitor and substrate specificity [4]. The use of various substrates in response to inhibitors following separation of mixtures of esterase activity may be useful in compartmentalization and quantitation of inhibitor effects upon specific esterase pools. Previously, we observed that titration of rat serum with CBDP [2-(o-Cresyl)-4H-1:3:2-benzodioxaphosphorin-2-oxide], BNPP [bis-p-nitrophenyl-phosphate], and Iso-OMPA [tetraisopropyl pyrophosphoramide] indicated CBDP to be a potent inhibitor of naphthylacetate esterase activity [4]. In contrast, Iso-OMPA and BNPP had no effect upon the hydrolysis of naphthylacetate, yet Iso-OMPA was observed to inhibit hydrolysis of naphthylacetate following injection suggesting that a possible metabolite of Iso-OMPA was required for inhibition of CaE. Initially, we had proposed to compare in vivo and ex vivo effects of Iso-OMPA. However, subsequent experiments carried out in this laboratory indicated the major esterase activities of rat serum could potentially be

separated from one another. Therefore, we chose to address the issue of possible metabolism of inhibitors to inhibitory metabolites by first separating and characterizing these respective activities. Our primary concern was eliminating the possibility that the inhibitor may be removed via noncatalytic processes thus lowering the "effective" concentration of inhibitor rather than generation of a unique inhibitory metabolite.

II. OBJECTIVES:

Because binding to and/or inhibition of esterase activity can reduce free concentrations of specific substrates or inhibitors, separation of component activites may afford a clearer view of inhibition patterns. The goal of the proposed work was to determine if prior separation of component activities affected inhibition of serum esterase activities by Iso-OMPA and CBDP using different substrates.

III. MATERIALS AND METHODS:

All electrophoretic reagents were of highest quality and purchased from Bio-Rad Laboratories, Richmond, California. CBDP [2-(o-cresyl)-4H-1:3:2:bensodioxaphosphorin-2-oxide] was obtained from the US Army Medical Research Institute of Chemical Defense through the courtesy of Dr. Donald Maxwell, Aberdeen Proving Ground, Maryland. Iso-OMPA (tetraisopropylpyrophosphoramine) and BW286C51 [1,5-bis(4-allyldimethylammoniumphenyl)pentan-3-one dibromide)] were purchased from Sigma Chemical Company, St. Louis, Missouri. Electrophoresis was carried out using a Mini-Protean II Electrophoresis Gel system powered by a Model 250/2.5 Bio-Rad power supply.

Preparation of Serum: Animals were decapitated and trunk blood collected. Blood was allowed to clot on ice for 30 mins and serum carefully removed by aspiration. Where indicated, serum was diluted using 0.01 M Tris buffer, pH 7.4.

Addition of Inhibitors: Indicated amounts of CBDP and Iso-OMPA were added to serum and incubated for 30 mins prior to electrophoresis. Compound BW286C51 (100 uM), a reversible inhibitor of AChE, was added to gel immersion reaction mixture in the absence of substrate following electrophoresis (15 mins at room temperature). Gels were treated in identical fashion as control samples with no inhibitor added.

Preparation of Gels: Gels (7.5 %, v/v, separating overlayed with 4 %, v/v, stacking) were prepared as described by Korenovsky and coworkers [5].

Electrophoresis: Prior to application of sample, gels were subjected to electrophoresis at 200 volts for 1 hour. Serum (10 ul of 1:10 diluent for determination of CaE activity) or 5 ul of undiluted serum (for determination of ChE activity) was loaded into respective wells and pre-electrophoresed at 10 volts for 1 hr. Under these conditions, densitometric linearity is maintained with regard to protein and hydrolysis of substrate. Depending upon the enzyme of interest, electrophoresis is continued at 1) 200 volts for 1 hr at 25 ^{o}C or 2) 50 volts for 22 hours at 5 ^{o}C.

Gel Staining:

CaE Activity: Naphthyl acetate has previously been shown to be a very useful substrate for detection of CaE activity [6]. The procedure of Baldwin and Hochachka [7] as modified by Korenovsky and coworkers [5] was used for detection of naphthylacetate hydrolysis.

C. E Activity: The procedure of Hodgson and Chubb [8] as modified by Korenovsky and coworkers [5] was used to determine hydrolysis of ACh and BCh substrates.

Densitometric Scanning: Scanning of developed gels was accomplished using a Hoefer Scientific Instruments model GS-300 Transmittance/Reflectance Scanning Densitometer

interfaced with a Jenco 286-12 computer and GS-300 Data Analysis System. Respective peak areas were integrated using a compensating Polar Planimeter.

IV. RESULTS:

As shown in Figure 1, three esterase activities of rat serum were electrophoretically separated from one another. Naphthylacetate was hydrolyzed by 1) a very rapidly migrating species that electrophoresed as a symmetrical peak of esterase activity after 1 hr electrophoresis (Fig. 1A) and 2) an esterase activity electrophoresing essentially as a symmetrical peak with a leading shoulder that migrated slightly off the origin after 22 hours (Fig. 1D). Electrophoresis of serum for 22 hours and subsequent development of gels in the presence of butyryl- and acetylthiocholine revealed the presence of two esterase components, one slow and one fast (Fig. 1 E and F) in addition to the rapidly migrating species following 1 hour electrophoresis (Fig. 1 A).

To further characterize these activities, the effects of known CaE, AChE, or BChE inhibitors (CBDP, BW284C51 and Iso-OMPA, respectively) were determined. As shown in Figure 2 (left column), esterase activity resolved after 1 hour electrophoresis that hydrolyzes naphthylacetate is very sensitive to CBDP (approximately 50 % inhibition in the presence of 0.10 uM) but less sensitive to Iso-OMPA (Fig. 2, Right Column). Hydrolysis of both acetyl- and butyrylthiocholine by the 22-hour slow electrophoretic species was significantly reduced in the presence of 8 uM CBDP (Figure 3). In contrast, the 22-hour fast component which was detected only in the presence of acetylthiocholine (Fig. 3, Left Column), appears to be slightly sensitive (approximately 25 %) to CBDP at concentrations very inhibitory (> 85 %) to the slow migrating activity (Figure 3).

The effects of Iso-OMPA, a known inhibitor of BChE is shown in Figure 4.

Hydrolysis of butyrylthiocholine is significantly reduced (> 90 %) by 20 uM Iso-OMPA (Fig. 4, Right Column). When assayed with acetylthiocholine, Iso-OMPA (20 uM) inhibited (>95 %, Fig. 4, Left Column) the slow species with little effect upon the fast component (Fig. 4, Right Column). Shown in Figure 5, Left Column is the effect of the reversible AChE inhibitor BW286C51 upon hydrolysis of acetylthiocholine indicating complete inhibition of the 22-hour fast component with little or no effect upon the slow migrating activity. Shown in Figure 5, Right Column is the effect of BW286C51 upon hydrolysis of naphthylacetate by the rapidly migrating species following 1 hour electrophoresis. As indicated in the figure, significant inhibition (45 %) was observed.

V. DISCUSSION:

Our data indicate that rat serum esterase activity is comprised of three major electrophoretic species. A rapidly migrating esterase activity was resolved following 1 hr electrophoresis and readily hydrolyzed naphthylacetate but did not utilize either acetyl- or butyrylthiocholine as substrate. In contrast, electrophoresis of long duration (22 hrs) separated two additional activities, one fast and one slow. The slower migrating species (22-hour slow), hydrolyzed both acetyl- and butyrylthiocholine, while the faster component (22-hour fast) utilized only acetylthiocholine as substrate. Furthermore, when naphthylacetate is used as substrate, the 22-hour fast species was not detected. Iso-OMPA, a known inhibitor of BChE [9] inhibited only the 22-hour slow component. Furthermore, this activity was detected in the presence of acetyl- and butyrylthiocholine substrates in contrast to the 22-hour fast species which was detected only when gels were developed in the presence of acetylthiocholine. Immersion of gels electrophoresed for 22-hours in a reaction mixture containing the irreversible AChE inhibitor compound BW286C51 prior to

development in the presence of acetylthiocholine revealed inhibition of only the fast component (Figure 5). In contrast, gels developed in the presence of butyrylthiocholine following incubation with compound BW286C51, indicated the slow component to be unaffected (> 90 % control level, data not shown).

Surprisingly, CBDP, a potent inhibitor of CaE activity [4], significantly inhibited the 22-hour slow activity at concentrations approximately 20 times that needed for comparable inhibition of CaE activity. It would appear that CBDP inhibition of the 22-hour slow component is not representative of CaE activity since the very fast migrating species (separated following 1 hr electrophoresis) will not hydrolyze either acetyl- or butyrylthiocholine. Also noteworthy is the significant inhibitory effect (45 %) of compound BW286C51 on the rapidly migrating species that hydrolyzed neither acetyl- or butyrylthiocholine.

The ChE data presented here are in agreement with those of Korenovsky and coworkers [5]. Using naphthylacetate as a substrate, these investigators observed significant rat plasma esterase activity that did not hydrolyze acetylthiocholine. Assuming the most rapidly migrating species (resolved following electrophoresis for 1 hr) in our hands to be representative of this activity, we would conclude that this band is not a ChE isoenzyme due to its inability to hydrolyze either acetyl- or butyrylthiocholine (Fig. 1B and 1C).

The apparent disparity between in vivo and ex vivo effects of Iso-OMPA may arise from partitioning in plasma of components that specifically bind and remove soluble constitutents that when in the presence of the inhibitor will bind the inhibitor thus lowering the effective concentration. The most likely candidates would be either clotting and/or cellular constituents of plasma. The fact that following electrophoresis, the

inhibitory effects of both Iso-OMPA and CBDP are manifested at lower concentrations would support this premise. Furthermore, serum that has been electrophoresed and then subjected to Iso-OMPA has no more effect than serum initially exposed to inhibitor followed by electrophoresis. This observation does not support the catalytic conversion of Iso-OMPA to inhibitory metabolites. Simple dilution can not account for these effects because BChE assays are carried out in the presence of more protein per assay yet is inhibited by substantially less Iso-OMPA (20 uM) when compared to inhibition of CaE by Iso-OMPA (>100 uM). Because of assay linearity requirements, serum must be diluted tenfold when assaying CaE.

VI. CONCLUSION AND RECOMMENDATIONS:

Summarized in Table I are substrate reactivity and inhibitor sensitivity of the three electrophoretically distinguishable species described in this report. From data summarized in Table I, we conclude that the very rapidly migrating species resolved following 1 hour electrophoresis is representative of CaE activity and the 22-hour fast and slow components are representative of AChE and BChE activities, respectively.

Previously, we demonstrated Iso-OMPA had little inhibitory effect upon naphthyl acetate hydrolysis at 250 uM [4]. However, we report here that 100 uM Iso-OMPA reduces CaE activity greater than 80 % (Figure 2, Right Column). BChE which also hydrolyzes naphthylacetate is also inhibited at very low Iso-OMPA concentration (20 uM) (Fig. 4, Right Column). However, at this inhibitor concentration, CaE activity is insensitive to Iso-OMPA (data not shown) and requires significantly more inhibitor (100 uM) to reduce CaE activity (Fig. 2, Right Column). Like inhibition of BChE by Iso-OMPA, very low concentration of CBDP (0.2 uM) was required for inhibition of CaE activity (Figure 2,

Left Column). Thus, binding of inhibitors to other serum components, may effectively reduce the free concentration of Iso-OMPA and CBDP otherwise available to inhibit these esterases.

We feel that gel electrophoresis provides a much more sensitive method for assessing the complex interactions between esterases, inhibitors and substrates. Only after separation of these enzymes and subsequent testing can one begin to understand specificities of esterases for both substrates and inhibitors. This information should aid in design and interpretation of blood and tissue biochemical analyses.

VII. FIGURE AND TABLE LEGENDS:

Figure 1: Electrophoretic separation of esterase activities of rat serum. Electrophoresis was carried out for the indicated times and gels developed as previously described under "Materials and Methods".

Figure 2: Effects of increasing amounts of CBDP and Iso-OMPA upon hydrolysis of naphthylacetate. Serum was incubated in the presence of indicated amounts of inhibitors, electrophoresed for 1 hr and developed in the presence of naphthylacetate substrate as previously described under "Materials and Methods".

Figure 3: Effects of increasing amounts of CBDP upon hydrolysis of Acetyl- and Butyryl-thiocholine following electrophoretic separation. Serum incubated in the presence of indicated amounts of CBDP was loaded onto gels, electrophoresed for 22 hrs and developed in the presence of substrates as previously described under "Materials and Methods".

Figure 4: Effects of increasing amounts of Iso-OMPA upon hydrolysis of Acetyl- and Butyrylthiocholine f of indicated amounts of Iso-OMPA was loaded onto gels, electrophoresed for 22 hrs and developed in the presence of substrates as previously described under "Materials and Methods".

Figure 5: Effects of increasing amounts of compound BW286C51 upon hydrolysis of Acetyl-, Butyrylthiocholine and Naphthylacetate following electrophoretic separation. Serum was loaded upon gels, electrophoresed for either 1 and 22 hrs and incubated for 30 mins in the presence of 100 uM inhibitor. Gels were developed in the presence of substrates as previously described under "Materials and Methods".

Table I: Results and Conclusion Summary. Data were derived from gels developed in the presence of substrates and inhibitors as previously described under "Materials and Methods". AtC, Acetythiocholine; BtC, Butyrylthiocholine; NA, Naphthylacetate; CBDP, [2-(o-Cresyl)-4H-1:3:2-benzodioxaphosphorin-2-oxide]; Iso-OMPA, tetraisopropyl pyrophosphoramide; BW, Compound BW286C51 [1,5-bis(4-allyldimethylammoniumphenyl)pentan-3-one dibromide]; CaE, Carboxylesterase; BChE, Butyrylcholinesterase; AChE, Acetylcholinesterase; -, no effect; +, slight hydrolysis/inhibition; + +, moderate hydrolysis/inhibition; + + + +, very significant hydrolysis/inhibition.

VIII. REFERENCES:

1. Taylor, Palmer (1991) Minireview: The Cholinesterases J. Biol. Chem. 266: 4025-4028.

2. Sterri, S.H. and Fonnum, R. (1984) Detoxification of organophosphorous compounds, In: M. Brzin, E.A. Barnard and D. Sket (Eds.), Cholinesterase: Fundamental and Applied Aspects, Walter de Gruyter, Berlin, pp. 389-400.

3. Bryson, P.K. and Brown, T.M. (1985) Reactivation of carboxylesterase hydrolase following inhibition by 4-nitrophenylorganophosphinates. Biochem. Pharmacol. 34: 1789-1794.

4. Chambers, J.P., Hartgraves, S.L., Murphy, M.R., Wayner, M. J., Kumar, N. and Valdes, James J. (1991) Effects of three reputed carboxylesterase inhibitors upon rat serum esterase activity. Neuroscience and Biobehav. Rev. 15:85-88.

5. Korenovsky, A., Laev, H., Mukherjee, S. and Mahadik, S.P. (1990) Quantitative Analyses of Plasma Cholinesterase Isozymes in Haloperidol Treated Rats. Biol. Psychiatry 27:871-883.

6. Ecobichon, D.J. (1970) Characterization of the esterases of canine serum. Can. J. Biochem. 48: 1359-1367.

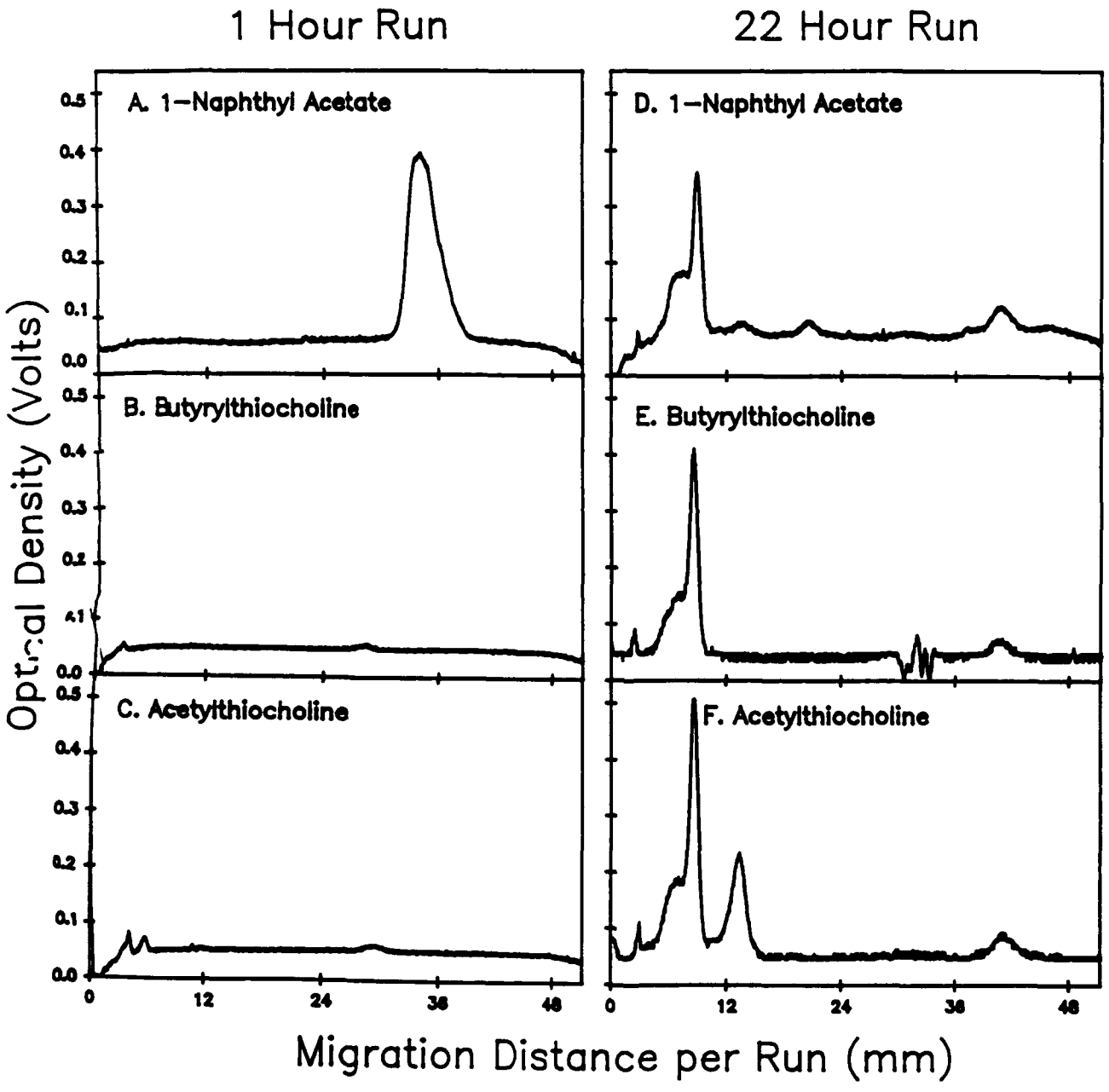
7. Baldwin, J. and Hochachka, P.W. (1970) Functional significance of isoenzymes in thermal acclimatization: AChE from adult brain. Biochem. J. 116: 883-887.

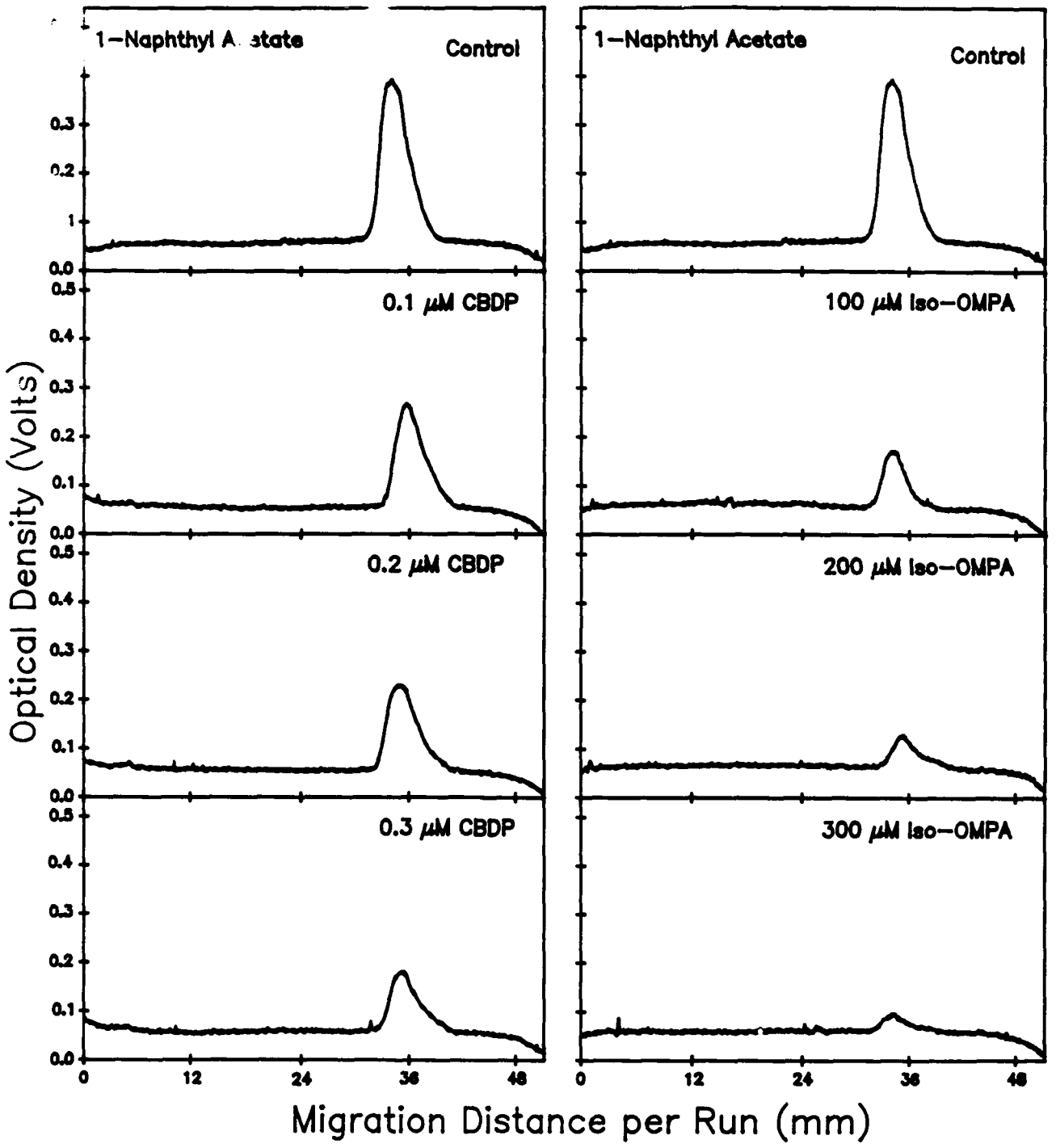
8. Hodgson, A.J. and Chubb, I.W. (1983) Isolation of the secretory form of soluble acetylcholinesterase by using affinity chromatography on edrophonium-sephadex. J. Neurochem. 41: 654-662.

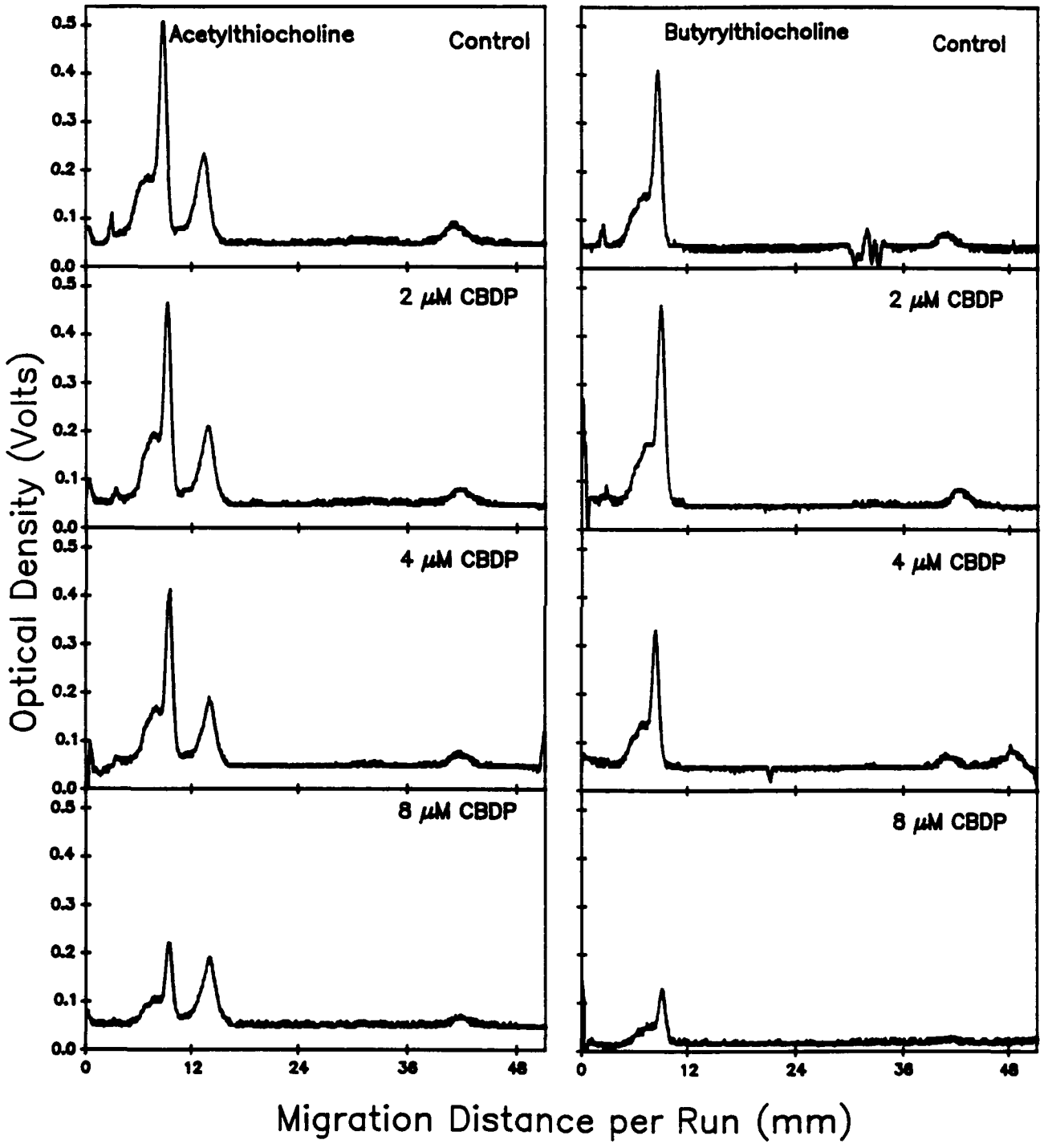
9. Berman, H.A., Decker, M.M. and Jo, S. (1987) Reciprocal regulation of acetylcholinesterase and butyrylcholinesterase in mammalian skeletal muscle. Dev. Biol. 120, 154-161.

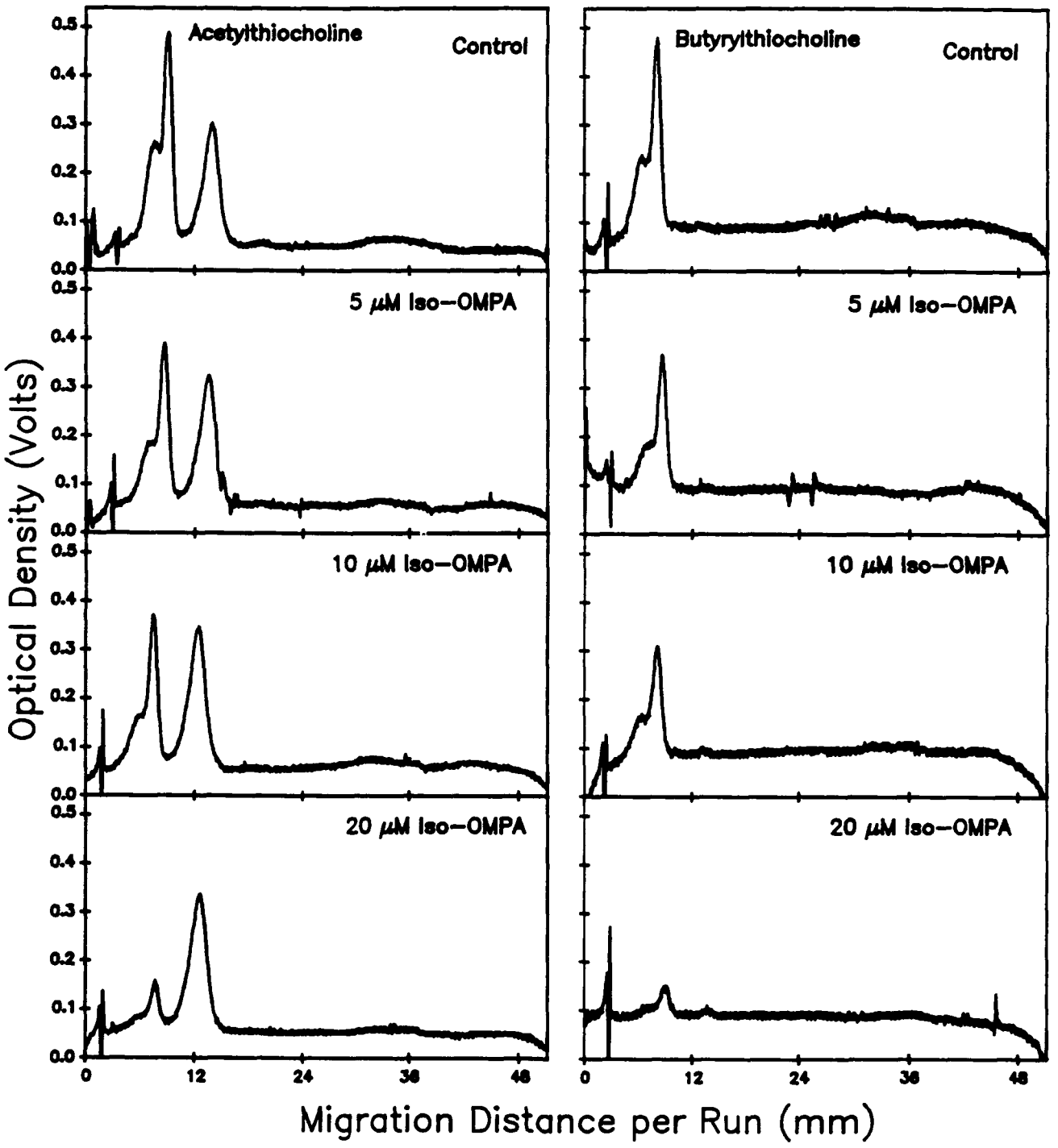
Table 1: Results and Conclusion Summary

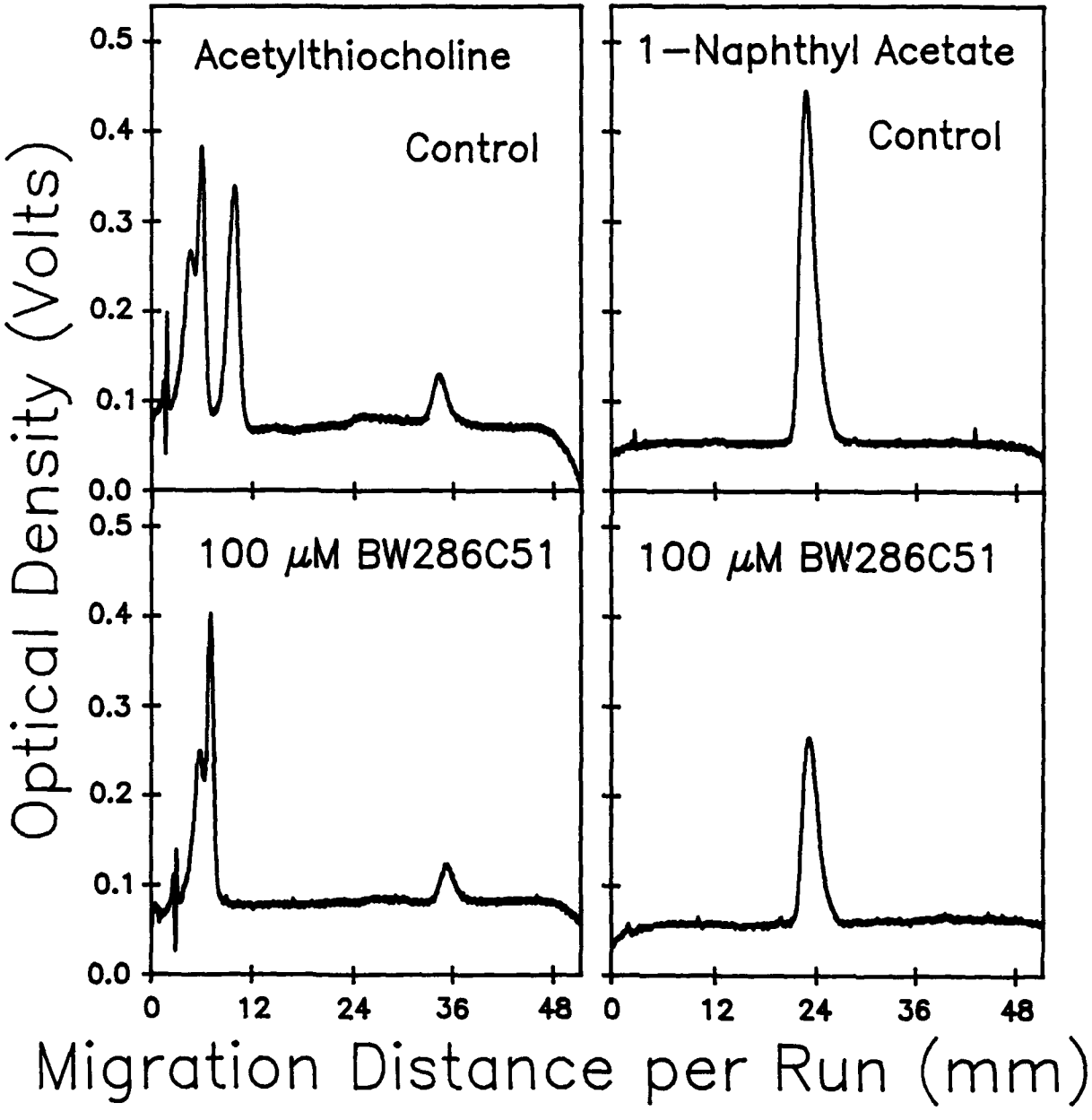
	Substrate Reactivity			Inhibitor Sensitivity			Conclusion
Electro-phoretic Species	AtC	BtC	NA	CBDP	Iso-O	BW	
1 hour	-	-	++++	++++	++++	++	CaE
22 hour fast	++++	-	-	+	-	++++	AChE
22 hour slow	++++	++++	++++	++++	++++	-	BChE


1 Hour Run
22 Hour Run
A. 1–Naphthyl Acetate
B. Butyrylthiocholine
C. Acetylthiocholine
D. 1–Naphthyl Acetate
E. Butyrylthiocholine
F. Acetylthiocholine
Optical Density (Volts)
0.0
0.1
0.2
0.3
0.4
0.5
0
12
24
36
48
Migration Distance per Run (mm)


1–Naphthyl A. state
Control
1–Naphthyl Acetate
Control
0.1 μM CBDP
100 μM Iso–OMPA
0.2 μM CBDP
200 μM Iso–OMPA
0.3 μM CBDP
300 μM Iso–OMPA
Optical Density (Volts)
Migration Distance per Run (mm)
0
12
24
36
48


Acetylthiocholine
Control
Butyrylthiocholine
Control
2 μM CBDP
2 μM CBDP
4 μM CBDP
4 μM CBDP
8 μM CBDP
8 μM CBDP
Optical Density (Volts)
0.0
0.1
0.2
0.3
0.4
0.5
0
12
24
36
48
Migration Distance per Run (mm)


Acetylthiocholine
Control
Butyrylthiocholine
Control
5 μM Iso-OMPA
5 μM Iso-OMPA
10 μM Iso-OMPA
10 μM Iso-OMPA
20 μM Iso-OMPA
20 μM Iso-OMPA
Optical Density (Volts)
0.0
0.1
0.2
0.3
0.4
0.5
0
12
24
36
48
Migration Distance per Run (mm)


Acetylthiocholine
Control
1—Naphthyl Acetate
Control
100 μM BW286C51
100 μM BW286C51
Optical Density (Volts)
0.0
0.1
0.2
0.3
0.4
0.5
0
12
24
36
48
Migration Distance per Run (mm)

1990 USAF RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

EMG ANALYSIS OF MUSCULAR FATIGUE AND RECOVERY FOLLOWING ALTERNATING ISOMETRIC CONTRACTIONS AT DIFFERENT LEVELS OF FORCE

Prepared by:	Mark W. Cornwall, P.T., Ph.D.
Academic Rank:	Assistant Professor
Department:	Department of Physical Therapy
University:	Northern Arizona University Flagstaff, Arizona 86011

Date:	December 1, 1990
Contract No:	F49620-88-C-0053/SB5881-0378

EMG ANALYSIS OF MUSCULAR FATIGUE AND RECOVERY FOLLOWING ALTERNATING ISOMETRIC CONTRACTIONS AT DIFFERENT LEVELS OF FORCE

by

Mark W. Cornwall, PT, PhD

ABSTRACT

The purpose of this study was to document the amount and rate of fatigue during alternating levels of isometric contraction similar to that found during the SACM. In addition, the time needed to recover from such an exercise was examined. Twenty males between the age of 22 and 35 years performed an isometric contraction at alternating levels of tension (20 and 50% maximum voluntary contraction) until exhaustion. The time at each contraction level was 10 seconds. Subjects were then randomly assigned to one of six recovery intervals (10, 20, 40, 60, 120 and 240 minutes) followed by a repeat of the previously described exhaustive exercise. The results of the study showed that endurance time (ET) was significantly ($p<.05$) recovered (90.96%) within 60 minutes after stopping the activity. Analysis of the myoelectric signal showed that amplitude (RMS) increased while frequency content of the signal (MPF) decreased during the fatiguing activity. Although MPF returned to its pre-test value after each of the recovery intervals, the normalized RMS value did not.

Acknowledgements

I wish to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsoring this research effort.

I would also like to express my appreciation to Dr. Larry Krock for his willingness and enthusiasm to serve as a sounding board for ideas related to this project. The project would not have had its present clarity of design without him.

INTRODUCTION

Pilots of high performance aircraft are frequently required to perform in a high-gravitational stress ($+G_z$) environment. It is imperative in these situations that the pilot maintain complete mental and physical control of the aircraft. Acceleration stress tolerance training has been an active area of research interest for a number of years.(2,8,10,17) A primary aim of this research has been to minimize the physical demands required in high-G situations so that the cognitive tasks required by the pilot can be focused upon. Among the current methods employed to assist pilots to withstand the high $+G_z$ forces is the Anti-G Straining Maneuver (AGSM). The AGSM requires the pilot to perform near-maximal activation of numerous muscle groups in response to the acceleration stress.(17) Although the AGSM has been found to significantly contribute to the tolerance of sustained $+G_z$ stress, it is considered to be extremely fatiguing.(8,10,17) It is obvious that as the muscles involved in the AGSM become fatigued, their effectiveness is diminished. Failure to accomplish an effective AGSM, therefore, could lead to impaired performance and even loss of consciousness (LOC).(9) Unfortunately, the level of fatigue and its rate of occurrence and recovery during $+G_z$ exposure have not been well studied. Studies which have included an assessment of muscle fatigue during acceleration stress have been general in nature and at times

subjective.(10,17) A further question that has not been adequately addressed in the literature is related to a pilots ability to perform multiple, successive AGSMs.

LITERATURE REVIEW

Electromyographic Changes with Fatigue

Power spectral analysis of the myoelectric signal has been advocated in the literature as an index of local muscle fatigue.(33) During fatigue, the power in the higher frequency components decreases with a concomitant increase in the power at the lower frequencies.(30,33,39) Three major theories concerning the cause of this spectral shift exist: 1) slowing of the muscle action potentials(34), 2) increased motor unit synchronization(18) and 3) changes in neuromuscular control(50). Of the three theories, only the first has experimental proof.

Using the center frequency of the power spectral density function, numerous investigators have documented both the magnitude and rate of muscle fatigue during exercise.(1,21,26,37,41-43,53) The decrease in the central frequency of the electromyographic (EMG) signal has been shown to be linear throughout the duration of a fatiguing isometric contraction.(40,43,60) At the point when the muscle can no longer generate the required tension, the frequency shift then plateaus.(40,60) Based on the above research, spectral analysis of the EMG signal is considered

to be both useful and reliable as an indicator of localized muscle fatigue.(43,61)

In contrast to frequency, the amplitude of the EMG signal has been shown to continually increase during submaximal fatiguing contractions.(21,26,34,38,61) In contrast, Arendt-Nielsen and associates reported that EMG amplitude increased during sustained isometric contractions, but then fell when the muscle could no longer maintain the required tension.(1) The increase in EMG amplitude has been attributed to either recruitment of additional motor units or higher discharge frequencies of the recruited units as they fatigue.(15)

The literature concerning the recovery of the various EMG signal parameters indicates that both center frequency and amplitude follow a recovery curve that is fairly rapid.(32) Recovery times of 5 minutes or less have been reported in the literature.(30,32,41)

Muscle Fatigue and Recovery

Localized muscle fatigue can be defined as the loss of the force generating capacity of a muscle.(4) Based on this definition, investigators have found that the rate at which a muscle fatigues is dependent upon the level of tension required, and the type of contraction being performed. A number of investigators have shown that the length of time before loss of force is dependent upon the level of tension during the contraction. Muscles contracting at higher

tensions fatigue quicker compared to those at lower tensions.(20,42,51) In addition, muscles fatigue quicker under isometric versus isotonic contractions.(1) A similar relationship between type of contraction and level of tension has been found with the recovery from fatiguing exercise. Although recovery curves are similar for isometric and isotonic exercise, the rate of recovery is more rapid for isotonic contractions.(23,51) In addition, the rate of recovery of a muscle's endurance capacity is inversely related to its tension.(20)

Recovery of a muscle's strength following fatiguing exercise has been shown to be relatively quick. A majority of studies indicate that recovery is complete within 10 to 20 minutes.(16,20,30,37,41,57,63) The ability of the muscle to recover its endurance capacity, however, is much slower. This endurance capacity of a muscle has often been referred to as a muscle's endurance time (ET). Funderburk and associates found that the recovery of ET was 85-90% complete 40 minutes following a fatiguing exercise.(20) A study by Stull and Kearney reported recovery from a 50% maximal isometric contraction to be 90% complete after 45 minutes.(52) Similar findings have been reported by Yates, Kearney, Noland and Felts. They found that recovery of ET was just under 90% complete, 42.5 minutes following an exhaustive exercise.(63) On the other hand, in a recent article by Kroon, recovery times from a force level of 40%

MVC were reported to be greater than 24 hours.(29) Unlike the previous studies, Kroon and associates performed repeated assessments of the subject's fatigue in the same day. This could account for the discrepancy between their results and those of previous investigators. The prolonged recovery time for endurance capacity of a muscle compared to its strength has been hypothesized by Yates and associates to be primarily the result of continued re-establishment of the muscle's acid- base and temperature homeostasis.(63)

PURPOSE AND SIGNIFICANCE

Research in the area of $+G_z$ tolerance has typically employed what is called the Simulated Aerial Combat Maneuver (SACM). This is a protocol conducted on a human centrifuge which attempts to simulate the levels and durations of $+G_z$ exposure that would be encountered during aerial combat. It typically consists of alternating levels of accelerative stress of 10 second durations at either 4.5 and 7 or 5 and 9 $+G_z$, depending upon the cockpit configuration and/or the support devices used. The nature of the SACM requires the subject to alternate the level of their AGSM in response to the changing $+G_z$ stress. Although a number of investigations have been conducted in the area of muscle fatigue and recovery, they have all dealt with either an isotonic exercise at a constant workload or an isometric contraction at a constant level of tension. No published

articles could be found dealing with alternating levels of tension in the same exercise. Such alternating tension levels would be similar to those found during an SACM.

The purpose of this study was to document the amount and rate of muscle fatigue during alternating levels of isometric contraction similar to that found during a SACM. In addition, the time needed to recover from such an exercise was investigated. This information will allow researchers to better understand the musculoskeletal demands required of high $+G_z$ acceleration stress and can assist in the development of exercise programs to further improve a pilot's G-tolerance.

METHODS AND PROCEDURES

Subjects

Twenty male volunteers between the age of 22 and 35 years participated in this study. Subjects were recruited from the Northern Arizona University, Flagstaff, Arizona community and conformed to the USAF height/weight ratios. Subjects with a history of a musculoskeletal or neuromuscular condition that would prevent them from performing sustained isometric contractions were excluded from the study. All subjects signed an informed consent prior to participation in the study. A summary of the demongraphic characteristics of the subjects used in this study is found in table 1.

Instrumentation

The isometric force exerted by the quadriceps femoris muscle group was recorded using a Cybex II isokinetic dynamometer.[a] The subjects were positioned in a chair with their hips in 80 degrees of flexion and their right knee in 60 degrees of flexion. Stabilizing straps were used to help maintain this position during all experimental testing. EMG activity from the vastus lateralis muscle was recorded using bipolar Ag-AgCl surface electrodes[b] having a recording area of 4 mm and an inter-electrode distance of 25 mm. The electrodes were arranged parallel to the muscle fibers and positioned so that the recording electrodes were proximal and distal to the muscle's motor point. Day-to-day consistency in the placement of the electrodes was assured by first electrically determining the motor point(31) at the start of the experiment and then marking that point with a permanent marker.(22) The EMG signal was first band-pass filtered to 5 and 500 Hz and then digitized at 1000 samples per second using an analog-to-digital converter[c]. All digitizing and analysis was performed using and a microcomputer.

Experimental Protocol

---

[a] Cybex Inc., Ronkonkonma, NY 11779

[b] IN-VIVO Metric, Healdsburg, CA 95448

[c] Scientific Solutions, Inc., Solon, OH 44139

The data for this study was collected in the Biomechanics/Motor Control Laboratory at Northern Arizona University. All subjects were tested on six separate occasions. Each test session consisted of 1) assessment of the subject's maximal voluntary strength (MVC), 2) successive isometric contractions of the quadriceps femoris muscle until fatigue and 3) repetition of the isometric contraction following a pre-determined recovery period.

MVC was determined as the mean force generated during three successive isometric contractions of the right quadriceps femoris muscle. Each contraction lasted three seconds and a rest interval of at least 30 seconds was provided between contractions. At the conclusion of the final rest period, subjects were asked to contract their right quadriceps femoris muscle at alternating tension levels of 20 and 50% of their previously determined MVC. The amount of time at each tension level was 10 seconds. The contractile tensions and durations at each level was selected in order to simulate the type of muscular contraction thought to be used by individuals during a SACM.(10) Subjects continued to alternate between the two tension levels until exhaustion. For this study, exhaustion was defined as the point where the subject either refused to continue with the contractions or they could not achieve and/or maintain the desired tension level. The target and subject-produced tension levels were displayed on an

oscilloscope[d] to ensure that the desired muscle tension was produced and maintained.

Following the completion of the fatiguing exercise, subjects were given one of the following six recovery intervals: 10, 20, 40, 60, 120 or 240 minutes. Each subject received all of the recovery intervals, but at least 24 hours elapsed between subsequent tests. The order in which subjects received the recovery intervals was randomly determined. At the conclusion of the pre-determined recovery interval, the subject again performed the previously outlined fatiguing exercise.

Data Analysis

From the digitized samples of the EMG signal, the root mean square (RMS) and the median power frequency (MPF) of the power spectral density function were calculated at 1 second intervals. The RMS values were normalized to the amount of activity recorded during the three maximal isometric contractions. The RMS and MPF values were then averaged over each of the 10 second contraction levels. The resulting means at each contraction level were then fit to a third degree polynomial equation. This best fit equation was used to obtain interpolated amplitude and frequency values at 10 percent intervals during the entire fatiguing contraction. These interpolated values were used to

[d] Model TEK 2221, Tektronix, Inc., Beaverton, OR 97077

determine if differences existed between the EMG activity during the two different contraction levels, over the course of the fatiguing contraction or between the six different recovery intervals. These tests for differences were conducted using a repeated measures analysis of variance (ANOVA) design. In addition, the effect of the six recovery intervals on ET was tested using a repeated measures ANOVA. Finally, an ANOVA test was performed using both RMS and MPF activity to assess when, if ever, they recovered following the initial exhaustive exercise. An alpha level of 0.05 will be used for all tests of statistical significance. Between-day reliability for MVC, ET, RMS and MPF was assessed using an intraclass correlation coefficient (ICC).(47) Between-day reliability for MVC and ET was found to be excellent. The ICC for MVC was .961 while the ICC for ET was .922. Between-day reliability for RMS and MPF measures was found to be moderatly good, being .823 for RMS and .777 for MPF.

RESULTS

Endurance Time

The mean pre-test ET for each of the six experimental conditions are shown in figure 1. Although a progressive increase is evident over the six conditions, there was no statistically significant difference ($p>.05$) between any of the means. The result of the ANOVA (Table 1) showed that

there was a statistically significant difference ($p<.05$) between the pre-test and post-test means. The mean pre-test ET was 174.02 (sd=48.57) seconds while the mean post-test ET was 156.8 (sd=52.17) seconds. In addition, a statistically significant difference ($p<.05$) was found between the six recovery intervals. Finally, the ANOVA indicated that there was a significant ($p<.05$) interaction between the six recovery intervals and the pre-test and post-test values. The nature of this interaction is illustrated in figure 2. A post hoc analysis using Tukey's HSD test showed that mean post-test ET was significantly different ($p<.05$) from the pre-test at the 10, 20 and 40 minute recovery intervals. The mean pre-test and post-test ET were not significantly different ($p>.05$) at the 60, 120 and 240 minute recovery intervals. At the 60 minute recovery interval, mean ET for the post-test was 90.96% of the pre-test value. Although not statistically significant, non of the mean post-test ET fully recovered to its pre-test value (Table 3).

Electromyographic Changes

EMG Amplitude. The results of the ANOVA for changes in EMG amplitude showed that the mean normalized RMS values increased significantly ($p<.05$) during the course of the fatiguing activity. Mean RMS values increased from 15.04 to 31.45 percent during the time of the fatiguing exercise. This increase represents an average change of 109.1 percent. The ANOVA also demonstrated that there was a significant

difference ($p<.05$) between the RMS values at the two contraction intensities. At 50% MVC, the mean amplitude of the vastus lateralis muscle was 41.09 percent of the maximum RMS value. Under the 20% MVC condition, the mean amplitude of the EMG signal was only 7.56 percent of maximum. Finally, there was a significant interaction ($p<.05$) between the level of contractile intensity and the time of contraction. This significant interaction indicates that EMG amplitude as measured by the RMS value of the myoelectric signal during a fatiguing exercise is dependent upon the level of tension exerted by the muscle. The pattern of the RMS response to a fatiguing exercise was adequately described using a third degree polynomial equation. Although the pattern of response at each contraction intensity is exponential, the slope of the regression line is more steep when the muscle is contracting at 50% of the subject's MVC (Figure 3). The mean increase observed at 50% MVC was 234.6% while at 20% MVC it was only 124.8%.

Finally, a two-way repeated measures ANOVA test was performed to assess whether EMG amplitude returned to its pre-test value following any of the six experimental recovery periods. Figure 4 shows the pre-test and post-test RMS values for each of the experimental conditions. The result of that test showed that for each of the six recovery intervals, pre-test and post-test RMS values at the start of

the fatiguing activity were statistically significant ($p<.05$).

EMG Frequency. The results of the ANOVA for changes in EMG frequency showed that the mean MPF values decreased significantly ($p<.05$) during the course of the fatiguing activity. MPF decreased from 60.23 to 52.63 Hz during the course of the fatiguing activity. This change represents an average decline of 12.6 percent. As with EMG amplitude, there was also a statistically significant difference ($p<.05$) found between the muscle's response at the two contractile intensities. The mean MPF during the 50% MVC contractions was 56.38 Hz compared to 53.79 Hz at 20% MVC. Again, as with EMG amplitude, there was a significant interaction ($p<.05$) found between the level of contractile intensity and the time during the fatiguing activity. As can be seen in figure 5, the pattern of decline in the muscle's MPF during contractions at 20% MVC is less steep and more linear compared to during 50% MVC. Under the 50% MVC contraction level, MPF decreased 13.0%. The overall decrease during tension at 20% MVC was 12.2%. It is interesting to note, however, that at the end of the fatiguing activity, both values are nearly identical.

Again, a two-way repeated measures ANOVA test was performed to assess whether EMG frequency returned to its pre-test value following each of the six experimental recovery periods. The pre-test and post-test MPF values for

each of the experimental conditions are shown in figure 6. The result of the ANOVA test showed that across the six recovery intervals, pre-test and post-test MPF values at the start of the fatiguing activity were not statistically significant ($p>.05$). There was, however, a significant ($p<.05$) interaction found between the recovery intervals and the two tests. A post hoc analysis using Tukey's HSD test revealed that pre-test and post-test values were significantly different ($p<.05$) for only the 20 minute recovery interval.

DISCUSSION

Endurance Time

The mean pre-test ET of 174.02 (sd=48.57) seconds obtained in the present study is similar to G-tolerance times reported in the literature for SACMs of 4.5 and 7.0 G_z. In 1982, Epperson, Burton and Bernauer investigated 24 individuals divided into 3 separate exercise training groups (runners, weight lifters and control).(17) Prior to the start of each group's training program, their G-tolerance was assessed using the SACM at 4.5 and 7.0 G_z. The observed means for each group were found to be 180 (sd=31), 232 (sd=33) and 195 (sd=34) seconds respectively. Two years later, using the same SACM protocol and 11 subjects, Tesch and Balldin reported a G-tolerance time of 242 (sd=116) seconds.(54) Finally, Burton, Whinnery and Foster, again, using the same SACM protocol, reported a mean G-tolerance time of 112 (sd=18.1) seconds.(10) Although the mean endurance time reported in the present study is lower than the majority of the previously reported G-tolerance times, it certainly falls within their reported variance. The small discrepancy that does exist may be accounted for by differences in subject selection. Clarkson, Kroll and McBride demonstrated that power trained individuals fatigue up to five times faster than endurance trained individuals.(12) Furthermore, acceleration studies have shown that strength training significantly improves G-

tolerance times compared to endurance training.(17,55) One explanation of these findings it is that muscle fiber type significantly contributes to a person's ET.(12,19,27,46,58) Given these findings, it is possible that if prior training and/or muscle fiber type had been more closely controlled by this, and other studies, the ET values would be even more similar.

The finding of similarity of endurance time lends support, first to the estimates by Burton and his colleagues regarding the level of isometric muscle contraction required during the execution of the SACM at 4.5 and 7.0 G_z.(10) This finding also indicates that at least under some controlled conditions, research regarding muscular activity during acceleration stress can be performed at 1 G_z with reasonably good success. Such research would be advantageous because it would be less expensive and would allow tighter research controls.

Although the physical activity required in the present study is unique, comparison of the rate of ET recovery with previous literature does show dramatic similarities. The majority of investigators have reported that following an exhaustive isometric exercise, the recovery of ET is exponential with an initially steep rate of recovery.(20,41,51,63) The observed initially rapid recovery of ET has been theorized in the literature to be the result of the restoration of blood flow to the

muscle(51) and the rapid removal of lactic acid.(24) Others feel that the rate of recovery is dependent upon the capillary density of the contracting muscle.(57) In addition to this initial rapid recovery, there seems to be a consensus in the literature that recovery of ET is approximately 80 percent complete after 10-12 minutes(41) and then plateaus to 85-90 percent after 40 minutes.(20,63) As can be seen from table 3 and figure 2 an almost identical finding was observed in the present study. Only one study has looked at the amount of recovery beyond one hour. That one study showed that ET was not fully recovered, even 25 hours after the exercise.(29)

Electromyographic Changes

EMG Amplitude. The finding of the present study indicating that there was a significant difference in the RMS value of the myoelectric signal at 20 and 50% MVC was not surprising. It has been known for some time that the amplitude of the EMG signal is related (linearly or non-linearly) to the amount of force produced by a muscle.(3,6,45,48,64) Thus, as the level of contraction increases, so does the amplitude of the EMG signal. This is certainly the case in the present study and is illustrated in figure 3.

The alteration of the RMS values in the present study during the course of the fatiguing exercise (figure 3) has been reported in the literature by numerous other

investigators.(1,11,21,23,26,32,33,38,41,42) The exact pattern of the increase, however, is in less agreement. Although the majority of researchers have shown that the increase is exponential or logrithmic in shape,(1,11,21,23,26,33,38,41) others have reported a linear increase.(32,42) The increase in EMG amplitude is generally considered to be the result of increased motor unit recruitment and a decrease in the motor unit recruitment threshold. These changes as the muscle attempts to maintain a certain level of tension.(35,38) The different response patterns seen in the present study between 20 and 50% MVC (Figure 3) has also been observed by other researchers.(11) The more dramatic change in RMS values at 50% MVC is probably related to the presence of eschemia in the muscle. It has been shown that blood flow is occluded at contraction intensities greater than 40% MVC and contributes to the rapid build-up of lactic acid.(56,65)

The results concerning recovery of EMG amplitude indicates that even after 4 hours, it has not returned to its pre-test values. These findings are supported by Kroon and Naeije who reported that EMG amplitude had not returned to its pre-exercise level after 25 hours(29) and by Kramer and associates who found that amplitude was not recovered for several hours following a fatiguing activity.(28) Other investigators, however, have shown that EMG amplitude recovers within a matter of minutes.(37,41) Several factors

may account for the apparent discrepancy between my study and the published literature. It may be the uniqueness of the fatiguing contraction performed in the present study compared to those previously conducted or possibly the increased measurement sensitivity of present study because of the greater number of subjects tested.

EMG Frequency. The differences in MPF at the two levels of muscular contraction in the present study has been observed previously by Stulen in 1980.(49) The higher frequency seen at 50% MVC is most likely related to the recruitment and firing patterns of the muscle's motor units. Several investigators have demonstrated that MPF of a muscle is related to the muscle's motor unit firing rate,(30,57,59) but they also feel that it should not neccessarily be attributed to motor unit synchronization.(5,13,25,59) As was pointed out previously, during the course of a fatiguing exercise, motor unit recruitment and firing rate increase to maintain the desired level of tension. This increased firing rate not only increases the observed amplitude of the EMG signal, but its frequency content as well.

The observed decrease in the frequency content of the EMG signal in the present study during the course of the fatiguing exercise has also been well documented in the literature(1,13,14,21,23,26,33,38,41-44,62) and it is felt to be the direct result of slowed muscle fiber conduction velocity.(1,13,30,50,57,65) There is no agreement, however,

whether the decrease is linear or non-linear. The amount of decrease has been reported to be as much as 25 to 50% which is much greater than the 12.6% reported in the present study. This discrepancy is most likely the result of the alternating contraction levels used in my study. The alternating contractions at 20% MVC allowed blood flow to be restored to the muscle for short periods of time, thus dissipating lactic acid and dampening the overall effect observed in the rate of MPF decline. This argument is somewhat supported by an examination of figure 5. As can be seen, the rate of decline increased during the later portion of the exercise and indicates the possibility that such an accumulative effect is present. As lactic acid increases, the conduction velocity of the muscle fibers will decrease, thus decreasing the frequency composition of the myoelectric signal.(56,65)

Unlike the previous discussion of EMG amplitude, the literature is in agreement with regard to the rate of recovery of MPF following a physically exhaustive activity. Previous investigations have shown that MPF recovers very quickly and is fully recovered in approximately 5 minutes or less.(30,36,37,42-44) The results of the present investigation lends even further credence to these studies. In fact, my study indicated that MPF had not only recovered, but was even higher at the conclusion of the recovery period (figure 6). The cause of the rapid return of MPF to their

pre-exercise values has been shown to be closely associated with the return of localized blood flow and the removal of lactic acid.(7,24,37,65)

CONCLUSIONS

It appears from the above discussion, that exhaustive exercise at alternating contraction levels of 20 and 50% MVC is similar to previous research that has focused upon isometric fatiguing exercises at a constant force level. The following conclusions can be drawn from this study regarding exhaustive isometric exercise at alternating tension level of 20 and 50% MVC.

1. The ET with such an activity closely resembles the G-tolerance times reported in the literature during an SACM at 4.5 and 7.0 G_z. The mean ET found in this study was 174.02 (sd=48.57) seconds.
2. Following an exhaustive exercise at alternating contraction levels of 20 and 50% MVC, recovery of a person's ET is 90.96% complete at 60 minutes and 98.96% complete at 120 minutes. Statistically, ET was recovered at 60 minutes.
3. During the course of an exhaustive exercise at alternating contraction levels of 20 and 50% MVC, EMG amplitude (RMS) increases exponentially. The mean increase was 234.6% at 50% MVC and 124.8% at 20% MVC.
4. During the course of an exhaustive exercise at

alternating contraction levels of 20 and 50% MVC, EMG frequency decreases. The mean decrease at 50% MVC was 13.0% while at 20% MVC it was 12.2%.

5. Recovery of a muscle's MPF is rapid, being less than 10 minutes. EMG amplitude, on the other hand, was still increased above its pre-exercise level, 4 hours after stopping the exhaustive exercise.

RECOMMENDATIONS

This study should be repeated on other muscles and different levels of tension to further substantiate its ability to simulate muscle contractile properties under high acceleration stress. In addition, this study should be replicated using a human centrifuge at 4.5 and 7.0 G_z. If the findings of such a study were similar to the ones in this study, additional research regarding muscle function under acceleration stress could be performed at +1 G_z instead of in a centrifuge. This would enable more research to be conducted with tighter controls, more safety to human subject and lower expense. Other studies that should be done in the future include investigating whether a muscle endurance or muscle strength training program would significantly increase a person's ET to exercise at alternating contraction levels.

REFERENCES

1. Arendt-Nielsen L, Mills KR. Muscle fibre conduction velocity, mean power frequency, mean EMG voltage and force during submaximal fatiguing contractions of human quadriceps. Eur J Appl Physiol 1988; 58:20-25.
2. Balldin UI, Myhre K, Tesch PA, Wilhelmsen U, Anderson HT. Isometric abdominal muscle training and G tolerance. Aviat Space Environ Med 1985; 56:120-124.
3. Bigland B, Lippold OCJ. The relationship between force, velocity and integrated electrical activity in human muscles. J Physiol 1954; 123:214-224.
4. Bigland-Ritchie B, Woods JJ. Changes in muscle contractile properties and neural control during human muscular fatigue. Muscle and Nerve 1984; 47:691-699.
5. Blinowska A, Verroust J, Cannet G. An analysis of synchronization and double discharge effects on low frequency electromyographic power spectrum. Electromyogr Clin Neurophysiol 1980; 20:465-480.
6. Bouisset S. EMG and muscle force in normal motor activities. In: Desmedt JE, ed. New Developments in Electromyography and Clinical Neurophysiology. Basel, Switzerland: Karger, 1973; 457-583.
7. Broman H. An investigation on the influence of a sustained contraction on the succession of action potentials from a single motor unit. Electromyogr Clin Neurophysiol 1977; 17:341.

8. Burns JW, Balldin UI. Assisted positive pressure breathing for augmentation of acceleration tolerance time. Aviat Space Environ Med 1988; 59:225-233.

9. Burton RR, Whinnery JE. Operational G-induced loss of consciousness: something old; something new. Aviat Space Environ Med 1985; 56:812-817.

10. Burton RR, Whinnery JE, Forster EM. Anaerobic energetics of the simulated aerial combat maneuver (SACM). Aviat Space Environ Med 1987; 58:761-767.

11. Clamann HP, Broecker MS. Relation between force and fatigability of red and pale skeletal muscles in man. Am J Phys Med 1979; 58:70-85.

12. Clarkson PM, Kroll W, McBride TC. Plantar flexion fatigue and muscle fiber type in power and endurance athletes. Med Sci Sports Exer 1980; 12:262-267.

13. DeLuca CJ. Myoelectrical manifestations of localized muscular fatigue in humans. CRC Crit Rev Biomed Engr 1800; 11:251-279.

14. DeLuca CJ, Sabbahi MS, Roy SH. Median frequency of the myoelectric signal: effects of hand dominance. Eur J Appl Physiol 1986; 55:457-464.

15. Edwards RG, Lippold OCJ. The relation between force and integrated electrical activity in fatigued muscles. J Physiol (Lond) 1956; 132:677-681.

16. Edwards RHT, Hill DK, Jones DA, Merton PA. Fatigue of long duration in human skeletal muscle after exercise.

J Physiol 1977; 272:769-778.

17. Epperson WL, Burton RR, Bernauer EM. The influence of differential physical conditioning regimens on simulated aerial combat meneuvering tolerance. Aviat Space Environ Med 1982; 53:1091-1097.

18. Fex J, Krakau ET. Some experiences with Walton's frequency analysis of the electromyogram. J Neurol Neurosurg Psychiat 1957; 20:178-184.

19. Fugl-Meyer AR, Gerdle B, Eriksson BE, Jonsson B. Isokinetic plantar flexion endurance: reliability and validity of output/excitation measurements. Scand J Rehabil Med 1985; 17:47-52.

20. Funderburk CF, Hipskind SG, Welton RC, Lind AR. Development of and recovery from fatigue induced by static effort at various tensions. J Appl Physiol 1974; 37:392-396.

21. Gamet D, Maton B. The fatigability of two agonistic muscles in human isometric voluntary submaximal contraction: an EMG study. I. assessment of muscular fatigue by means of surface EMG. Eur J Appl Physiol 1989; 58:361-368.

22. Graham GP. Reliability of electromyographic measurements after surface electrode removal and replacement. Percept Mot Skills 1979; 49:215-218.

23. Hagberg M. Muscular endurance and surface electromyogram in isometric and dynamic exercise. J

Appl Physiol 1981; 51:1-7.

24. Harris RC, Hultman E, Sahlin E. Glycolytic intermediates in human muscle offer isometric contraction. Eur J Physiol 1981; 389:277-282.

25. Jones NB, Lago PJA. Spectral analysis and the interference EMG. IEEE Proc 1982; 129:673.

26. Jorgensen K, Fallentin N, Krogh-Lund C, Jensen B. Electromyography and fatigue during prolonged, low-level static contractions. Eur J Appl Physiol 1988; 57:316-321.

27. Karlsson J, Sjuodin B, Jacobs I, Kaiser P. Relevance of muscle fibre type to fatigue in short intense and prolonged exercise in man. Ciba Found Symp 1981; 82:59-74.

28. Kramer H, Lun A, Mucke R, Kuchler G. Changes in mechanical and bioelectrical muscular activity and in heart rate due to sustained voluntary isometric contractions and time required for recovery. Part I: contractions at constant level of bioelectrical activity. Electromyogr Clin Neurophysiol 1979; 19:381-386.

29. Kroon GW, Naeije M. Recovery following exhaustive dynamic exercise in the human biceps muscle. Eur J Appl Physiol 1988; 58:228-232.

30. Kuorinka I. Restitution of EMG spectrum after muscle fatigue. Eur J Appl Physiol 1988; 57:311-315.

31. Licht S. Electrodiagnosis and Electromyography. Baltimore: Waverly Press, Inc., 1971.

32. Lind AR, Petrofsky JS. Amplitude of the surface electromyogram during fatiguing isometric contractions. Muscle and Nerve 1979; 2:257-264.

33. Lindstrom L, Kadefors R, Petersen I. An electromyographic index for localized muscle fatigue. J Appl Physiol 1977; 43:750-754.

34. Lindstrom L, Magnusson R, Petersen I. muscular fatigue and action potential conduction velocity changes studied with frequency analysis of EMG signals. EMG 1970; 4:341-356.

35. Maton B, Gamet D. The fatigability of two agonistic muscles in human isometric voluntary submaximal contraction: an EMG study. II. motor unit firing rate and recruitment. Eur J Appl Physiol 1989; 58:369-374.

36. Merletti R, Sabbahi MA, DeLuca CJ. Median frequency of the myoelectric signal: effects of ischemia and cooling. Eur J Appl Physiol 1983; 52:258.

37. Mills KR. Power spectral analysis of electromyogram and compound muscle action potential during muscle fatigue and recovery. J Physiol 1982; 326:401-409.

38. Moritani T, Muro M, Nagata A. Intramuscular and surface electromyogram changes during muscle fatigue. J Appl Physiol 1986; 60:1179-1185.

39. Moxham J, Edwards RHT, Aubier M, DeTroyer A, Farkas G,

Macklem PT, Roussos C. Changes in EMG power spectrum (high-to-low ratio) with force fatigue in humans. J Appl Physiol 1982; 53:1094-1099.

40. Petrofsky JS. Computer analysis of the surface EMG during isometric exercise. Comp Biol Med 1980; 10:83-95.

41. Petrofsky JS. Quantification through the surface EMG of muscle fatigue and recovery during successive isometric contractions. Aviat Space Environ Med 1981; 52:545-550.

42. Petrofsky JS, Glaser RM, Phillips CA, Lind AR, Williams C. Evaluation of the amplitude and frequency components of the surface EMG as an index of muscle fatigue. Ergonomics 1982; 25:213-223.

43. Petrofsky JS, Lind AR. Frequency analysis of the surface electromyogram during sustained isometric contractions. Eur J Appl Physiol 1980; 43:173-182.

44. Petrofsky JS, Lind AR. The influence of temperature on the amplitude and frequency components of the EMG during brief and sustained isometric contractions. Eur J Appl Physiol 1980; 44:189-200.

45. Philipson L, Larsson PG. The electromyographic signal as a measure of muscular force: a comparison of detection and quantification techniques. Electromyogr Clin Neurophysiol 1988; 28:141-150.

46. Prince FP, Hikida RS, Hagerman FC. Human muscle fiber types in power lifters, distance runners and untrained

subjects. Eur J Physiol 1976; 363:19-26.

47. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86:420-428.

48. Siegler S, Hillstrom HJ, Freedman W, et al. Effects of myoelectric signal processing on the relationship between muscle force and processed EMG. Am J Phys Med 1985; 64:130-149.

49. Stulen FB. A technique to monitor localized muscular fatigue using frequency domain analysis for the myoelectric signal. Cambridge, MA: PhD Thesis, Massachusetts Institute of Technology, 1980.

50. Stulen FB, DeLuca CJ. Frequency parameters of the myoelectric signal as a measure of muscle conduction velocity. IEEE Trans Biomed Eng 1981; 28:515-523.

51. Stull GA, Clarke DH. Patterns of recovery following isometric and isotonic strength decrement. Med Sci Sports 1971; 3:135-139.

52. Stull GA, Kearney JT. Recovery of muscular endurance following rhythmic or sustained activity. J Mot Behav 1974; 6:59-66.

53. Tarkka IM. Power spectrum of electromyography in arm and leg muscles during isometric contractions and fatigue. J Sports Med Phys Fitness 1984; 24:189-194.

54. Tesch PA, Balldin UI. Muscle fiber type composition and G-tolerance. Aviat Space Environ Med 1984;

55:1000-1003.

55. Tesch PA, Hjort H, Balldin UI. Effects of strength training on G tolerance. Aviat Space Environ Med 1983; 54:691-695.

56. Tesch PA, Karlsson J. Lactate in fast and slow twitch skeletal muscle fibers of man during isometric contraction. Acta Physiol Scand 1978; 103:413.

57. Tesch PA, Wright JE. Recovery from short term intense exercise: its relation to capillary supply and blood lactate concentration. Eur J Appl Physiol 1983; 52:98-103.

58. Thorstensson A, Karlsson J. Fatiguability and fibre composition of human skeletal muscle. Acta Physiol Scand 1976; 98:318-322.

59. Verroust J, Blinowska A, Cannet G. Functioning of the ensemble of motor units of the muscle determined from global EMG signal. Electromyogr Clin Neurophysiol 1981; 21:11-24.

60. Viitasalo JT, Komi PV. Signal characteristics of EMG during fatigue. Eur J Appl Physiol 1977; 37:111-121.

61. Viitasalo JT, Komi PV. Interrelationships of EMG signal characteristics at different levels of muscle tension and during fatigue. EMG Clin Neurophysiol 1978; 18:167-178.

62. Wolf E, Blank A, Shochina M, Gonen B. Effect of exercise of the lower limbs on the non-exercised biceps

brachii muscle. Am J Phys Med 1984; 63:113-121.

63. Yates JW, Kearney JT, Noland MP, Felts WM. Recovery of dynamic muscular endurance. Eur J Appl Physiol 1987; 56:662-667.

64. Zuniga EN, Simons DG. Nonlinear relationship between averaged electromyogram potentials and muscle tension in normal subjects. Arch Phys Med Rehabil 1969; 50:613-620.

65. Zwarts MJ, Arendt-Nielsen L. The influence of force and circulation on average muscle fibre conduction velocity during local muscle fatigue. Eur J Appl Physiol 1988; 58:278-283.

Table 1. Demographic Information Concerning the Twenty Subjects Used in the Study.

	Age (yrs)	Height (in)	Weight (lb)	Mean MVC (lb)
Mean	26.70	69.65	175.15	271.07
SD	3.54	2.43	18.55	43.81

Table 2. Summary of the Repeated Measures Analysis of Variance Test for Endurance Time.

SOURCE OF VARIATION	SS	df	MS	F
SUBJECTS	387878.59	19	204119.66	14.33*
TEST	17805.48	1	17805.48	12.50*
ERROR 1	27064.92	19	1424.47	
RECOVERY	32361.28	5	6472.26	5.20*
ERROR 2	118203.91	95	1244.25	
TEST X RECOVERY	6074.42	5	1214.88	3.49*
RESIDUAL	33050.72	95	347.90	

* $p < .05$

Table 3. Percentage of Endurance Time Recovered During each of the Six Experimental Recovery Intervals.

Interval (min)	Pre-Test (sec)	Post-Test (sec)	Recovery (%)
10	165.45	132.48	80.07*
20	166.45	143.55	86.24*
40	172.83	151.13	87.44*
60	174.76	158.97	90.96
120	181.18	177.93	98.96
240	183.49	176.74	96.32

* $p < .05$

Figure 1. Pre-test Endurance Times for Each of the Six Experimental Recovery Intervals.

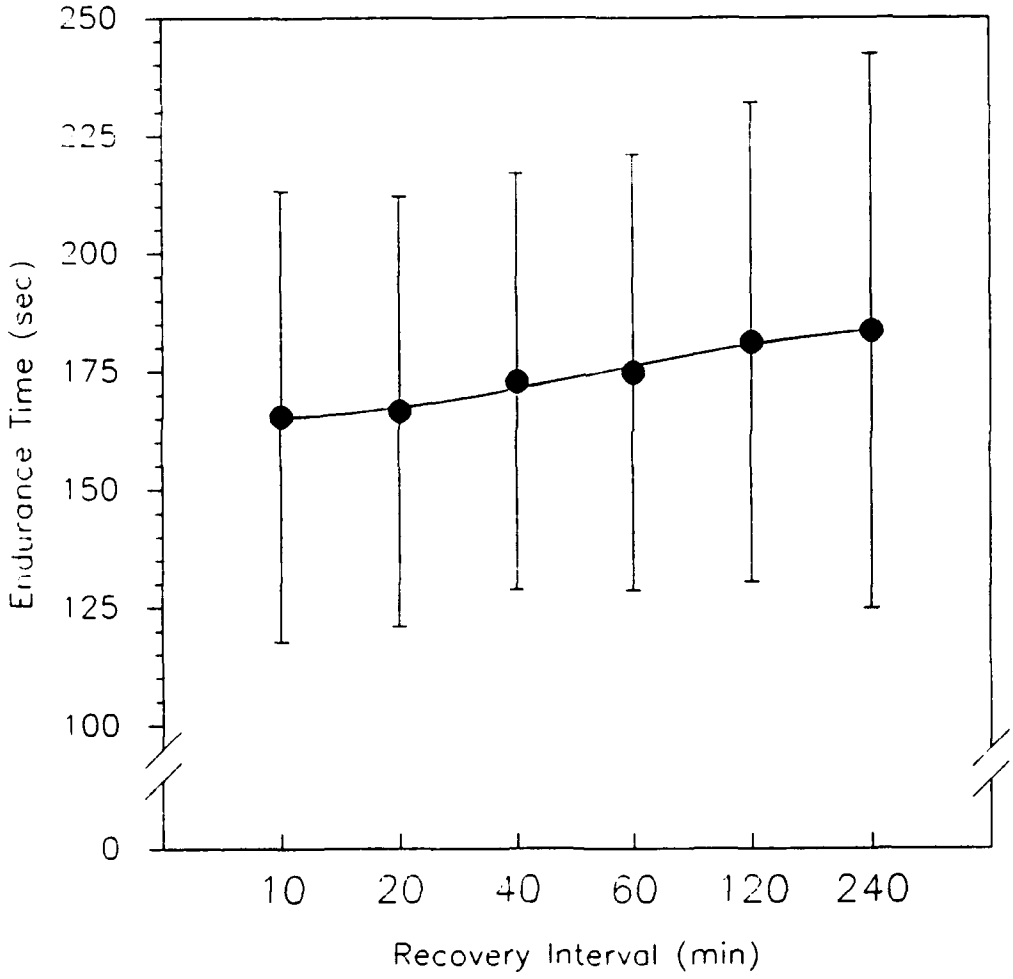


Figure 2. Pre-test and Post-test Endurance Times for the Six Experimental Recovery Intervals During the Fatiguing Activity.

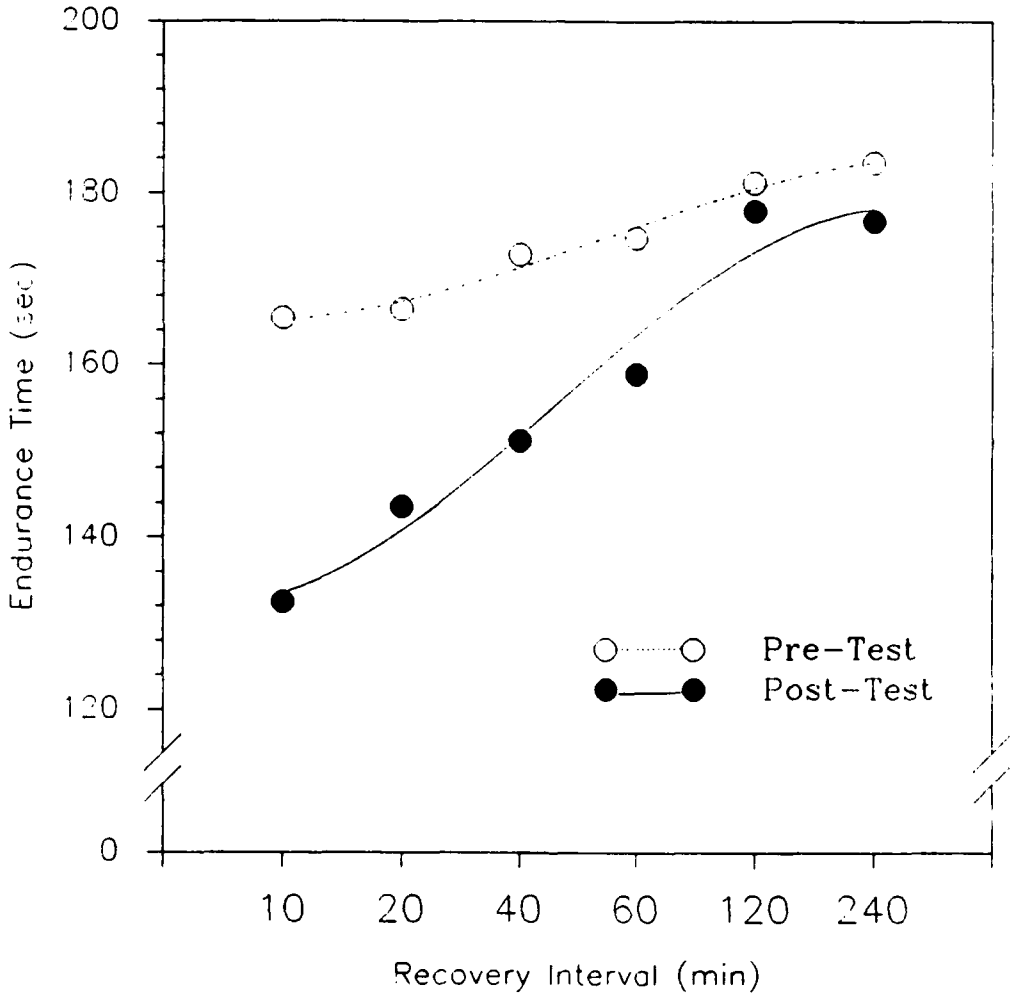


Figure 3. The Effect of Contraction Level on Percent of Maximum RMS During the Fatiguing Isometric Contraction.

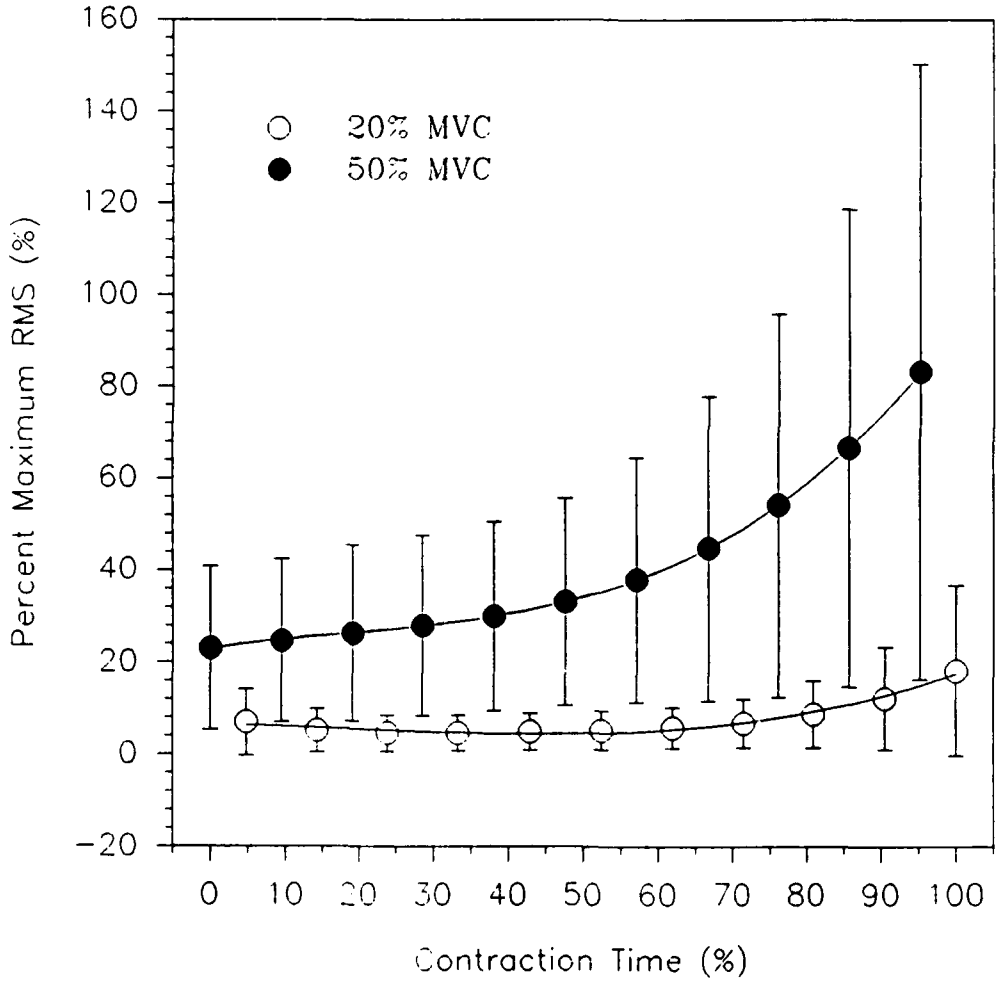


Figure 4. Percent Maximum RMS at the Start of the Pre-Test and Post-Test for Each of the Six Experimental Recovery Intervals.

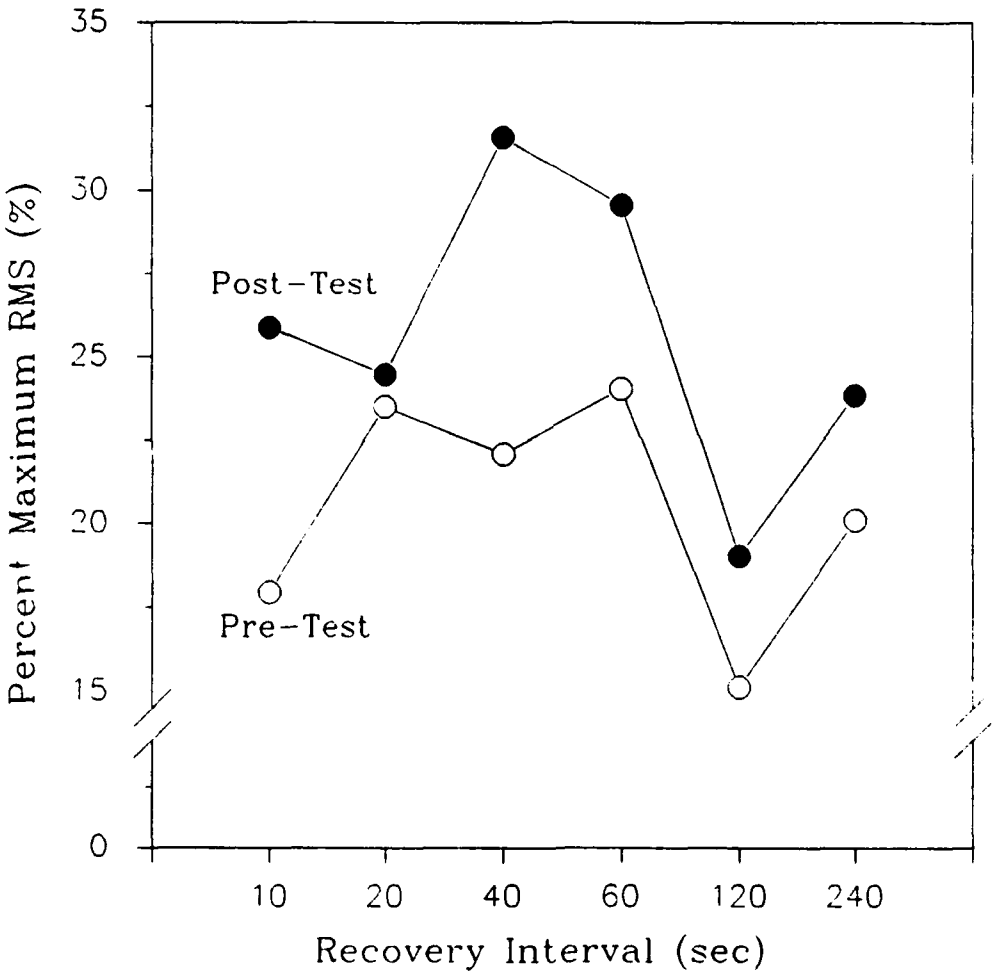


Figure 5. The Effect of Contraction Level on MPF During the Fatiguing Isometric Contraction.

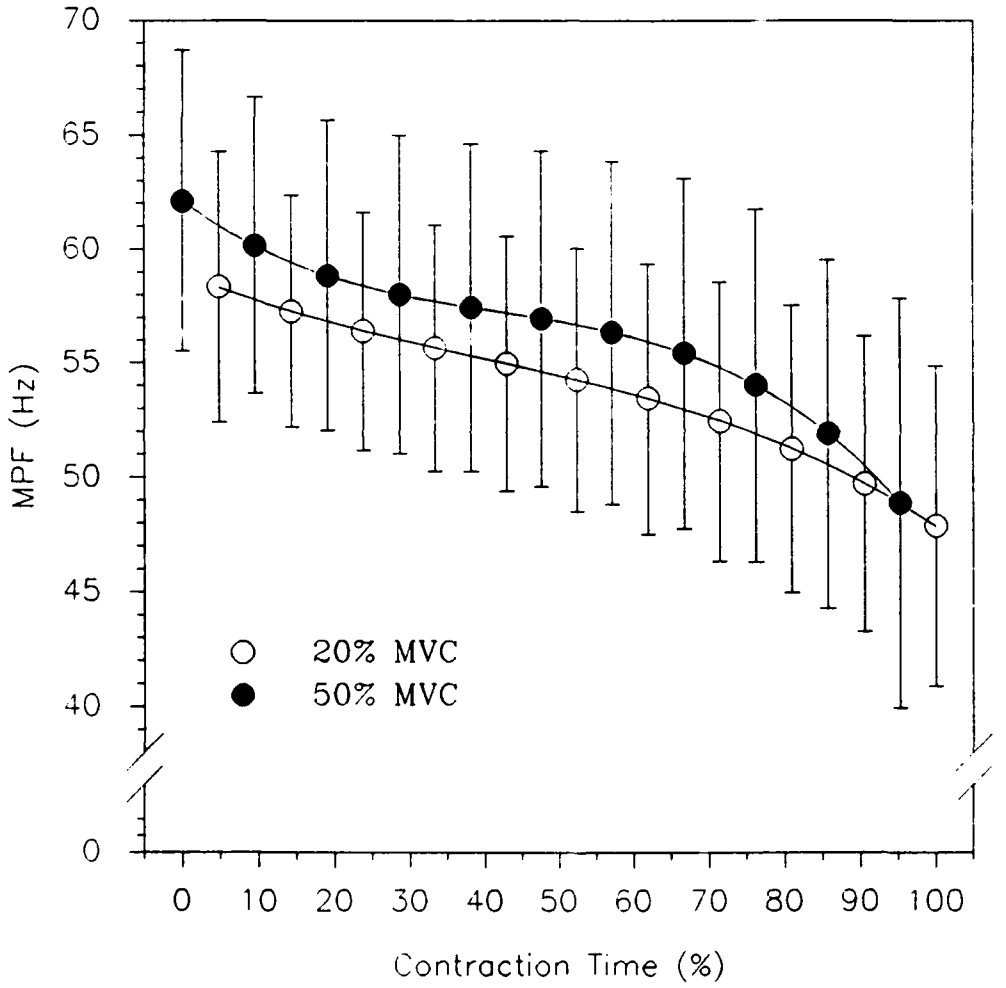
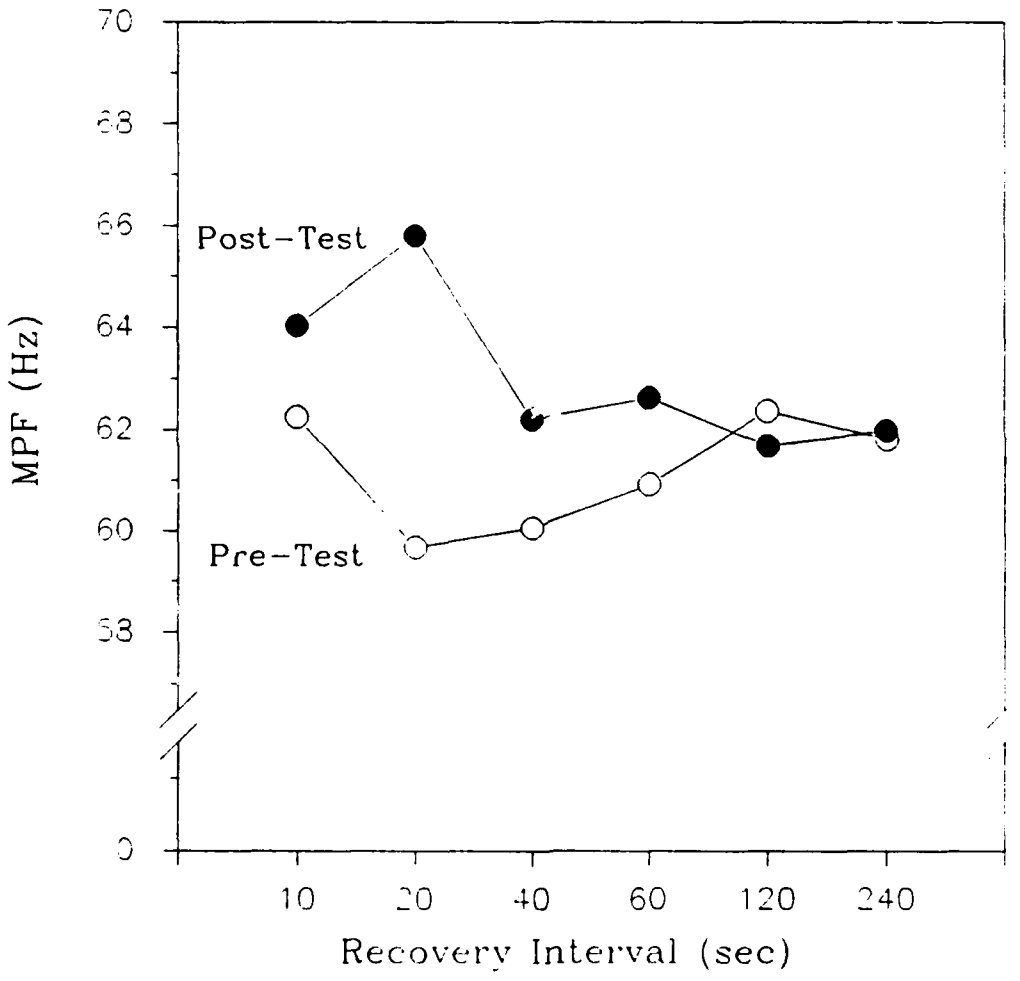


Figure 6. MPF at the Start of the Pre-Test and Post-Test for Each of the Six Experimental Recovery Intervals.



1990 AFOSR Research Initiation Program

Conducted by:

Universal Energy Systems, Inc.

FINAL REPORT

PCR Analysis of Specific Target Sequence of *Mycoplasma hominis* and *Ureaplasma urealyticum*

210-10MG-013

Principal Investigator:	Prof. Vito DelVecchio
University:	University of Scranton
USAF Researcher:	School of Aerospace Medicine Brooks AFB, TX

FINAL REPORT (1990 MINI GRANT PROGRAM)

PCR Analysis of Specific Target Sequence of *Mycoplasma hominis* and *Ureaplasma urealyticum*.

Sponsored by AFOSR and conducted by UES, Inc.
UES Grant 546

Target sequence of *U. urealyticum* and *M. hominis* which were previously shown to be species specific by dot-blot and *in situ* hybridization analysis, were subjected to PCR analysis. The sequences consisted of a 1350 bp portion of *M. hominis* genomic DNA and a 740 bp sequence from *U. urealyticum*. A 177 bp span from the *U. urealyticum* DNA was sequenced and targeted by PCR as well as 137 bp of the *M. hominis* DNA portion. Using the sequenced portions, PCR primers were chosen with the hope that each would have absolute specifically. Four sets of *Ureaplasma* and 3 sets of *Mycoplasma* primers were tested for cross reactivity with each other as well as with *M. salivarium*, *M. genetalium*, and *M. hyorhinis*. After varying the temperature, time periods, number of cycles, magnesium concentration, and adding specificity enhancers, it was determined that the 177 bp and 137 bp segments were too yield limiting for PCR systems which are specific. Each set of primers was sensitive enough to detect target sequences in DNA isolated from clinical samples.

To expand our capabilities for PCR analysis the 740 bp segment of *U. urealyticum* was sub-cloned into pUC 119 so that more efficient sequencing can be accomplished. Nested deletions of the 740 bp segment will be made using the Erasabase system. Religation and cloning of the various deleted recombinant molecules will permit dideoxy sequencing of the entire segment. This is presently in progress. Once accomplished, a series of PCR systems can be run simultaneously to increase the specificity of the procedure.

RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

Studies On Melanocytes and Melanins

Prepared by:	Gwendolyn B. Howze, Ph.D. Principal Investigator
Academic Rank:	Associate Professor
Department and University:	Department of Biology Texas Southern University
Research Location:	Texas Southern University 3201 Wheeler Ave. Houston, Texas 77004 (713) 527-7095
USAF Research	Johnathan L. Kiel, DVM, PhD USAFSAM/RZP Brooks AFB San Antonio, Texas 782325
Date:	January 1, 1990 - May 31, 1991 (Extended)
Contract No.	F49620-88-C-0053/SB5881-0378

ABSTRACT

Studies on Melanocytes and Melanin

DOPA-melanin was synthesized by auto-oxidation of L-3,4-dihydroxyphenylalanine and used in several types of experiments which are described below.

The DOPA-melanin (DOPAM) binds to hemoproteins to form complexes which have absorption maxima at 416.9nm, 420nm, and 423nm for complexes with cytochrome C, hemoglobin and myoglobin respectively. The hemoproteins tested were cytochrome C, hemoglobin, myoglobin, horseradish peroxidase, and catalase, all of which form complexes with DOPAM. In the native state they have have absorption maxima in the narrow range of 402 to 408. When complexed with DOPAM, the absorption maxima shift to longer wavelengths (see above). In addition to binding to hemoproteins, DOPAM chemically reduces cytochrome C and suppresses enzyme activity of horseradish peroxidase. DOPAM cytochrome C complexes were precipitated by ethanol and the dried intrinsicly conductive ppt. was visualized in the scanning electron microscope (SEM).

Electrophoretic studies indicate that the DOPAM hemoprotein complexes are electronegative at pH 8.6 irrespective of the net charge on the unbound hemoprotein.

The DOPAM is a conductive polymer which upon evaporation yields intrinsicly conductive films and powders, which can be visualized by (SEM) without the usual requirement of a coating with gold or similar conductive metal.

Melanocytes from tissue culture were made conductive enough to permit SEM visualization by application of a DOPAM coating. The resolution obtained was comparable to that characteristic of cells coated with gold by the conventional preparation procedure. Observations on melanocyte growth patterns are also described.

DOPAM and DOPAM-hemoprotein complexes are conductive polymers which might be of some interest to basic science as wells as materials science. They may also have immediate technical applications in the fields of scanning electron microscopy and scanning tunneling microscopy. The fact that melanin forms complexes with hemoproteins and also serum albumin has medical implications.

1.0 OBJECTIVES AND USAF RELEVANCY

There are three related objectives of this research on the biopolymer melanin. Firstly, the objective has beens to expand and solidify the results obtained during the 1989 United States Air Force Summer Faculty Research Program (SFRP). Thus, the experiments on interactions between melanin and hemoproteins are a direct outgrowth of studies begun during the 1989 SFRP. The second objective was to assess the feasibility of an approach to study of melanin biosyntheis using 3-aminotyrosine as a regulator of polymer elongation combined with infrared absorption spectroscopy, proton nuclear magnetic resonance, and gel electrophoresis. The third objective was to study melanocytes grown in tissue culture: 1) by scanning electron microscopy, 2) by extraction of melanin and quantification of extracted melanin.

It is expected that the data obtained will have relevance to health and safety issues and guidance for Air Force personnel who work with emitters of radiofrequency radiations. The original 1989 SFRP research was a project which studied effects of microwaves, and was conducted in the Mechanisms Function Laboratory, in the Radiation Physics Branch of the Radiation Sciences Division, USAF SAM, Brooks AFB.

2.0 INTRODUCTION

2.1 Rationale for the Study

This project is an outgrowth of research done in the laboratory of Dr. J.L. Kiel at Brooks AFB, during the United States Air Force 1989 Summer Faculty Research Program (SFRP).

During the 1989 SFRP, preliminary studies on the effects of microwave radiations on melanins and melanocytes were begun. Some of the data obtained suggested that there were detectable interactions between microwaves and synthetic melanin in a noncellular model system. The present "follow-on" experimentation investigates some of the nonradiation aspects of the SFRP research. In particular, the studies of the interactions between melanin and hemoproteins is an outgrowth of the rediscovery that melanin reduces cytochrome C in vitro. The original discovery was made In 1983 by Slawinska, Slawinski and Ciesla (14) The studies on the interactions between melanin and cytochrome C have been expanded to include other hemoproteins. The hemoproteins studied in this project include the following: cytochrome C, hemoglobin, myoglobin, catalase, lactoperoxidase, and horse radish peroxidase. The effects of melanin on the: 1) absorption spectrum, 2) difference spectrum, 3) chemical activity, and 3) migration in argarose gel electrophoresis have been studied.These experiments were deemed important because the data was expected to complement and solidify the results obtained during the 1989 SFRP. These experiments were also of interest as a possible indication of potential human pathology. Inside normal cells the two types of biopolymers are compartmentalized and probably do not have an opportunity to interact. It is conceivable however that melanin breakdown products and precusors (were they to enter the systemic circulation) might pose a potential hazard to a variety of hemoproteins. Thus the data will also have relevance to certain pathological conditions wherein melanin is released into the systemic circulation.

Preliminary studies on melanin biosynthesis versus chemical synthesis were continued. The effects of 3-aminotyrosine on biosynthesis by melanocytes and also in vitro chemical synthesis were begun. Melanocytes grown in tissue culture and melanosomes were studied by scanning electron microscopy. Observations made during the 1989 SFRP indicated that pigment biosynthesis might be a useful phenomenon to study in trying to characterize the interactions between melanin and microwave radiations. While the current project does not employ radiation, it is expected that it will yield baseline information which will be available for future studies of radiofrequency effects on melanocytes.

The following types of experiments were proposed: 1) experiments studying the effects of soluble DOPA melanin on several hemoproteins; 2) studies of melanocytes by SEM including the feasibility of detecting changes in melanosomes by scanning electron microscopy; 3) studies on the feasibility of polymer

size regulation by 3-amino-L-tyrosine; 4) studies on the feasibility of determination of relative size of melanin polymers; 5) comparison of the NMR and IR spectra of three types of melanin polymers; 6) extraction and solubilization of melanin from cultivated nonirradiated cells and quantification of the extracted melanin.

2.2 Background Information

Melanin is of particular interest because it absorbs energy in a wide range of the electromagnetic energy spectrum, e.g., ultra violet, visible, infrared, ultra sound (4,5,7). It also responds to applied electrical fields. At specific combinations of hydration, temperature and applied electric field, synthetic and natural melanins exhibit conductivity changes compatible with biological semiconduction (6,7). Ultrasound absorption by melanin in the frequency range of 0.1 to 10 MHz has been described. Melanin samples absorbed ultrasound in an efficient resonant transfer mechanism.It was speculated that the melanosome might function in two different modes, as energy storage device and as an energy transducer(7).

Due to its versatility as an absorber of electromagnetic radiation it was speculated that melanin and melanin containing cells should be good systems for attempting to detect nonthermal interactions between microwaves and living systems. Melanocytes synthesize melanin in membrane bound cytoplasmic organelles called melanosomes. The melanin, which is a major component of the melanosome, is complexed with membranes which are composed of protein and lipid. In this project scanning electron microscopy is used to probe the structure of the melanocyte including the melanosome

The question asked during the 1989 SFRP was whether melanin absorbs microwaves, and whether there are consequent detectable nonthermal effects due to this interaction. Using a model system, the tentative answer seems to be affirmative. The model system used takes advantage of the observation that soluble synthetic DOPA melanin interacts with oxidized cytochrome C. Since the absorption spectrum of cytochrome C (cytC) changes to the profile expected for the reduced form of the molecule, it is assumed that melanin reduces cytC. The reaction rate is concentration dependent and can be followed by measuring absorption in the range 546 to 550 nm. In the model experiment, the effects of microwave irradiation on polymerization was studied by assaying for an effect on the rate of cytC reduction, i.e., increased absorption at 550nm.

Samples from the polymerization mixture were either irradiated or sham irradiated, at 37 degrees in both instances. Solutions of cytC were mixed with either irradiated or sham irradiated samples and absorption at 550nm was measured. The irradiated samples produced the larger absorption reading. Since the measured

temperature was the same in both treatments, it is assumed that the effect was due to the microwave irradiation, rather than thermal heating. It is believed that the irradiation speeded up the polymerization reaction, which resulted in an increased concentration of DOPA melanin in the irradiated sample, thus causing more cytC molecules to be reduced. Two types of evidence support the hypothesis of increased rate of polymerization: 1) the irradiated specimen had a higher absorption throughout the spectrum from 200 to 800nm, indicating higher melanin concentration; 2) the pH was lower in the irradiated specimen. In the aereated auto-oxidation polymerization, the pH drops as the polymerization proceeds. Therfore a lower pH would be expected in the specimen with the higher rate of polymerization.

Melanin has both oxidative and reductive properties (11,13,14). Melanin oxidizes reduced NAD (13), and it reduces cytC The reduction of cytC by melanin was rediscovered and exploited during the 1989 SFRP. Slawinska, Slawinski and Ciesla originally discovered this effect in 1983 (14). It has been reported that melanin oxidizes hydroxylamine to nitroxide and also slowly reduces nitroxides to their hydroxylamines (15).

The redox feature of melanin was exploited during the 1989 SFRP to detect microwave interactions with melanin. The redox property might, however, be of importance in a pathological aspect. There are pathological conditions wherein melanin is circulated in the blood and excreted in the urine. Under such conditions, cytochrome C and possibly other hemoproteins, such as catalase and peroxidase, might be at risk. It is also possible that small amounts of soluble melanin, produced as a result of melanocyte turnover, might under certain conditions enter the systemic circulation with ill effects that have not yet been attributed to melanin. For example, melanin not only binds to the hemoproteins described in the results section, it also binds to bovine serum albumin. The presence of melanin in the systemic circulation could be a factor in auto-immune disease by rendering normal body proteins antigenic.

Although microwaves are not used in the current project ,it is expected that the acquired data will be useful for planning future studies on the interactions occuring between microwaves / or other radiofrequency radiations and melanocytes wherein pigmentation is the most significant parameter.

3.0 RESEARCH PLAN

3.1 Specific Aims

The objectives were addressed by the following approaches.

1) Interactions between melanin and hemoproteins
The specific aim is to determine if melanin causes chemical changes in hemoproteins other than cytochrome C, and to study those changes by spectroscopy, gel electrophoresis, and assay of a characteristic chemical function of the hemoproteins.The hemoproteins which are being studied are: cytochrome C, hemoglobin, myoglobin, lactoperoxidase, horse radish peroxidase, and catalase.

2) polymer synthesis and elongation
A specific aim is to determine if 3-aminotyrosine can be used to regulate the size of the melanin polymer, in vitro, i.e., cell culture and chemical synthesis.
A third specific aim is to study several types of melanin polymers (synthetic-DOPA, synthetic-tyrosine, natural, and 3-aminotyrosine truncated) by proton magnetic resonance, infrared spectroscopy and gel electrophoresis.

3) Studies on melanocytes in tissue culture
The 1989 SFRP studies suggested that microwave irradiation can influence melanin polymerization. Therefore a reliable method of quantification of melanin biosynthesis is needed. A fourth specific aim is to test several published methods of isolation and quantification of melanin content in nonirradiated cells.
In most cells melanin is complexed with lipoprotein in a cytoplasmic organelle called the melanosome. Since this is the active form, in situ, most reactions must therefore be solid state interactions occurring in a very compact particle. The fifth aim is to study melanocytes/melanosomes by high resolution scanning electron microscopy (SEM). The main significance of the SEM studies is the attempt to exploit the electronic features of melanin to promote visualization in the SEM.

3.2 Experimental Design: summary of methods

3.2.1 Interactions Between Melanin and Hemoproteins

The effect of melanin on hemoproteins
Several hemoproteins were studied: cytochrome C, myoglobin, hemoglobin, catalase, lactoperoxidase, and horse radish peroxidase. Each protein was mixed with soluble melanin. The mixtures were studied by absorption and difference spectrophotometry and by chemical assays.
Experimental prototype #1: absorption spectrophotometry protocol.
1. Use a concentration of hemoprotein which will allow all peaks to be visualized at an absorbance which will be on scale for the system in use.
2. Obtain an absorption spectrum, 200 nm to 900nm.
3. Treat hemoprotein with melanin. In the prototype experiments the combination of 30 micro molar (uM) cytochrome C and 0.2mg/ml melanin were used. For myoglobin and hemoglobin, use 15 uM and 7.5 uM respectively.
4. Obtain an absorption spectrum of hemoprotein melanin mixture.

HEMOPROTEIN	abbr.	MOLECULAR WEIGHT*
Cytochrome C	cytC	13,370 (bovine)
Hemoglobin (oxidized)	Hb	64,500 (human)
Myoglobin	Mb	16,900 (horse)
Horse Radish Peroxidase	HRP	40,000
Lactoperoxidase	LPO	77,500
Catalase	Cat	250,000

* information provided by suppliers, (Sigma:cytC, Hb, Mb) (Worthington:HRP, LPO, Cat)

Experimental prototype #2: difference spectrophotometry protocol.
1. Use a concentration of hemoprotein which will allow all peaks to be visualized at an absorbance which will be on scale for the system in use. For titration experiments 7.5 uM cytochrome C, 7.5 uM and 1.875 uM myoglobin and hemoglobin respectively are maximal on a 4.0 absorbance scale. Treatment with as little as 0.0025 mg/ml melanin produces a change in the difference spectrum.
2. Obtain the baseline, using a solution of the hemoprotein in both the reference and sample cells.
3. Add melanin to the sample cuvette and an equal volume of solvent to the reference cuvette, mix and record the results.
Experimental prototype #3: assay of melanin effect on chemical activity of horse radish repoxidase (HRP).
1. Prepare the reaction mixture using 2.9 ml per cuvette.
2. Prepare a blank containing 2.9 ml of reaction mixture and 0.1 ml of water (the solvent for the enzyme solution).
3. Add 0.1 ml of the diluted enzyme mixture to the sample tube containing the reaction mixture, mix and read at 510 nm at 30 second intervals for 5 minutes (16).
In order to assay for a melanin effect, the enzyme is treated with melanin prior to adding to the reaction mixture.

Horse radish peroxidase (HRP) is a hemoprotein which catalyzes the oxidation by hydrogen peroxide of a number of substrates. the above described assay of enzyme activity, 4-aminoantipyrine is the hydrogen donor. A rose pink color developes as the hydrogen peroxide decomposes during the oxidation reduction. One unit of activity results in decomposition of one micromole of hydrogen peroxide per minute at 25 degrees centigrade at pH 7.0.

Experimental prototype #4:electrophoretic migration of complexes of melanin and hemoprotein.
1. Prepare stock solutions of hemoprotein at 4 mg/ml
2. Prepare melanin stock solutions of 4mg/ml
3. Prepare the complex by mixing equal quantities of of the two stock solutions.
4. Add 15 micro liters of the mixture to the gel well.
5. Conditions for electrophoresis: 1.2 % agarose gel, 185 V, one hour. The Tris-glycine buffer should be pH 8.6.

3.2.2 Polymer Synthesis and Elongation

The effect of 3-amino-L-tyrosine on melanin synthesis
Experiment #1: melanocytes, cultivated as described below in part 3.2.3 of this section, were grown in 24 well plates and tested with a range of 3-Amino-L-tyrosine (3-AT) concentrations.
Experiment #2: cell cultures were be grown to maximal density. After there was visual evidence of pigment production, the cultures were segregated into two groups: untreated controls and 3-AT treated. After the controls had matured to maximum pigmentation (the time varied with cell inoculum). The melanin was extracted from the cultures by methods described below in part 3.2.3 of this section.It was thought that agarose gel electrophoresis of the two pigment fractions should give an indication of the effect of 3-AT. The parameters of interest are rate of migration, number of bands and band size. In preliminary experiments it was found that the pigment yield from these experiments was to low for reliable and consistant detection by the relatively insensitive method available in this laboratory. These experiments were discontinued. A spectrofluorometric method of greater sensitivity has been described in the literature. The The effect of 3-AT on chemical synthesis of DOPA melanin was studied using the type of protocol described below.
Experiment #3: DOPA melanin was synthesized by the method of Sealy (10), as in past experiments. One gram of dihydrophenylalanine (DOPA) is added to 200 ml of distilled water. The pH is adjusted to 8.8 and the mixture is stirred with aeration for three days. When the solution is adjusted to pH 6 with HCl, melanin precipitates. The precipitate is collected and dried to produce powdered melanin. This will be referred to as type I melanin. In some cases the complete synthetic product was evaporated to dryness to produce type II powder. Type III melanin powders were obtained by using Amicon concentrators and evaporation.

Experiment #4: three reaction beakers were prepared as in experiment #3 above and incorporated the following modifications. 1) The control specimen was prepared exactly as above. 2) In the T0 sample, one gram of 3-AT was added simultaneously with the DOPA. 3) In the T15 spemimen, the 3-AT was added 15 minutes after stirring had begun. The control specimen was already dark after 15 minutes and black by 30 minutes. The T15 was also dark after 15 minutes, but took longer to turn black. The T0 sample was much retarded. It was a light orange at 30 minutes, subsequently turned orange brown, and turned black after an hour. The three samples were visibly identical after three days of stirring. The pH was checked and adjusted to 8.8. By one hour after the start the pH had stabilized. Samples T15 and T0 required more NaOH to stabilize the pH at 8.8. The type III powdered melanins from the three samples were visually indistinguishable. The type II melanins were different in that it has not been possible to produce a powder from T0 and T15. T0 and T15 evaporate into a tacky tar-like material rather than a powder.

Proton magnetic resonance and infrared spectroscopy of melanins
It was expected that at least three types of melanins would be studied; DOPA(chemically synthesized) melanin, as described above, tyrosine (chemically synthesized) melanin (Sigma M8631), natural melanin from Sepia (Sigma 2649) and, if the time allows, 3-AT treated melanin. Preliminary NMR studies have been done. The Varian EM-360 60 MHz NMR spectrometer was used. Tetramethylsilane (TMS) absorbance peak was used as reference point. Chloroform was selected as solvent for the first run. Only the Sigma-tyrosine melanin was soluble in the chloroform. Therefore we were not able to obtain a spectrum for the DOPA-melanin. All of the melanins tested are soluble in dimethylsulfoxide (DMSO). Therefore, in the future DMSO will be used as solvent for the NMR spectroscopy. The DMSO spectrum is known, it absorbs in a region which should not confuse the analysis of the spectra. Preliminary infrared spectroscopy has also been done. Insolubility in chloroforn did not pose a problem in this case, since the solid chemical could be fused to a KBr pellet for solid state observation. The results from the infrared and NMR preliminary studies are presented in the results section. Dr. John Sapp, professor of chemistry, provided important advice for the IR and NMR experiments.

3.2.3 Studies on Melanocytes in Tissue Culture

The Cloudman S91 melanoma, clone M-3, cell strain (A.T.C.C.) has been used. Ham's F10 medium supplemented with 2.5 % fetal calf serum and 15 % horse serum has been the growth medium. The medium RPMI is actually better at promoting growth, but not as good for pigment production. The Sera were not heat inactivated. The cells produce (visually detectable) pigment as the culture becomes dense and crowded. The population doubled in ca. three to four days. Stock cultures were grown in plastic T-flask.

Extraction and estimation of melanin content in melanocytes
The method for extracting the pigment is the method of Friedman and Gilchrest (1). In summary: Detach cell from wall of flask; count cells; pellet cells, dissolve cells in small volume 0.2N NaOH ; solubilize melanin in DMSO.; read O.D. at 475. A standard curve based on soluble synthetic tyrosine melanin (Sigma)is used. The concentration range of the standard curve is 0 to 20.0 micrograms per ml. and the solvent is DMSO.

SEM studies of melanocytes and melanosome
The culture conditions for scanning electron microscopical (SEM) studies varied. When using the conventional visualization regimen, see below, the cells were grown in well plates on glass slips or in LAB TEK chamber slides for tissue culture. The cells remainded attached to the substratum during the fixation. When using nontraditional visualization, cells attached to glass were less likely to give good results. Although these methods are still evolving, the most successful approach is described below

1)The conventional procedure follows.
The basic fixation utilizes 2% glutaraldehyde in 0.1 M pH 7.4 phosphate buffer for 2 hours, followed by a thorough rinse in the solvent buffer and post fixation in 1% OsO for three to four hours. The solvent for the osmium is 0.1 M cacadylate, which is also used to wash away the excess osmium after the osmification. The fixed and rinsed cells are dehydrated in a graded series (25%,50%, 75% 95%, 100%(3 rinses) of ethanol. The dehydrated cell are desiccated by critical point drying. The desiccated cells are attached to a SEM stubb. The stubb with cells is stored in a vacuum desiccator until ready for conductive coating by the traditional method of deposition of a 20nm gold coat. Some of the fixed cells were also cryofractured, but during subsequent steps the cells were lost and that part of the project was discontinued. An Hitachi S-570 model, high resolution SEM was used. An accelerating voltage of 20kv was used uniformly. The magnification was varied depending on the subject of interest. The specific magnifications will be supplied upon request. Polaroid type 55 P/N film was used. The resultant micrographs obtained by the traditional method variant described above is being used as a baseline for comparing the products resulting from the three methods which are described below.

2) Three methods are given which do not use gold coating to make the specimen conductive (a prerequisite for visualization in SEM). The significance of the three prototype methods described below is the attempt to exploit the electronic attributes of melanin tc make the cell and melanosome conductive without requiring metallic coating.In The first method the redox feature of melanin is used to reduce silver nitrate (17). It was expected that the silver precipitated at the reaction site should make the melanosome (and hopefully melanocyte) conductive, thus the usual

metallic coating would not be necessary. The potential for using this type of method to detect tyrosinase activity and the site of other redox products and enzyme activities via SEM should be emphasized. The second method uses tannic acid as a mordant and osmiun to enhance conductivity (18).The third and most successful method utilizes a concentrated melanin solution to coat the cell, followed by air drying.

Method #1: silver nitrate (17,18)

1. Fix cells in a mixture of 4% glutaraldehyde, 12% DMSO, 0.68% KH PO and 0.087% K HPO and 0.1% BSA for at least 45 min.
2. Transfer to 50% glutaraldehyde for 3 hours.
3. Rinse well in distilled water (DH2O).
4. Treat for 18 to 24 hours in the dark with methenamine/ silver nitrate solution.
5. Repeat step 3.
6. Wash two min with 5% sodium thiosulfate
7. Repeat step 3.
8. Dehydrate in ethanol series, as in the traditional method described above.
9. Remove ethanol by critical point drying.
10. Mount dessicated specimen onto SEM stubb.
11. Store stubb-specimen in vacuum dessicator at least 24 hours.
12. Observe in SEM.

Method #2: osmium/tannic acid (18)

1. Fix for 30 min in 1% glutaraldehyde in pH 7.4, 0.078 M cacadylate buffer.
2. Fix for 30 min to 1 hour in a mixture of 3% glutaraldehyde in 8% tannic acid- 0.1 M cacadylate solvent.
3. Rinse 5 min in 0.1 M cacadylate buffer.
4. Post-fix for 30 min in 0.5% osmium tetroxide in 0.1 M cacadylate.
5. Rinse six times in distilled water (DH20), at least 15 min.
6. Rinse for one hour in 5% tannic acid.
7. Wash quickly in DH2O.
9. Incubate one hour in 0.5% aq. OsO .
10. Rinse six times in DH2O, over a 15 min period..
11. Dehydrate as in method one above.
12. Critical point drying should be done as in method one above.
13. Mount specimen on SEM stubb.
14. Store specimen as in method above.

Method #3: coating with soluble melanin, an evolving procedure

1. Harvest mitotic cell from rapidly growing cultures by the mitotic shake maneuver.
2. Gently sediment using a manual centrifuge.
3. Coat aluminum foil with 1% lysine hydrobromide (Sigma).
4. Spread thick suspension of mitotic cells on the coated foil, shiny side coated.
5. Allow a short time for binding.
6. Fix in 2%, pH7.2 phosphate buffered, glutaraldehyde for one hour.

7. Rinse prep. in buffer.
8. Treat with melanin.
9. Air dry the preparation.
10. Bind the foil substrate to a SEM stub using sticky tab (E. F. Fullam, Inc.)
11. Use a stiff but tacky silver paint to assure that the foil is in conductive contact with the stub.
12. Incubate the preparation in a desiccator.
13. View the preparation in the SEM.

4.0 RESULTS AND DISCUSSION

Interactions between Melanin and Hemoprotein

Figure one shows that melanin reduces cytochrome C (cytC). Curve #1, which was derived from melanin treated cytC, shows an increased absorption at 550nm as compared with curve #2 which is for untreated cytC. Figure one was taken from the 1989 SFRP final report.

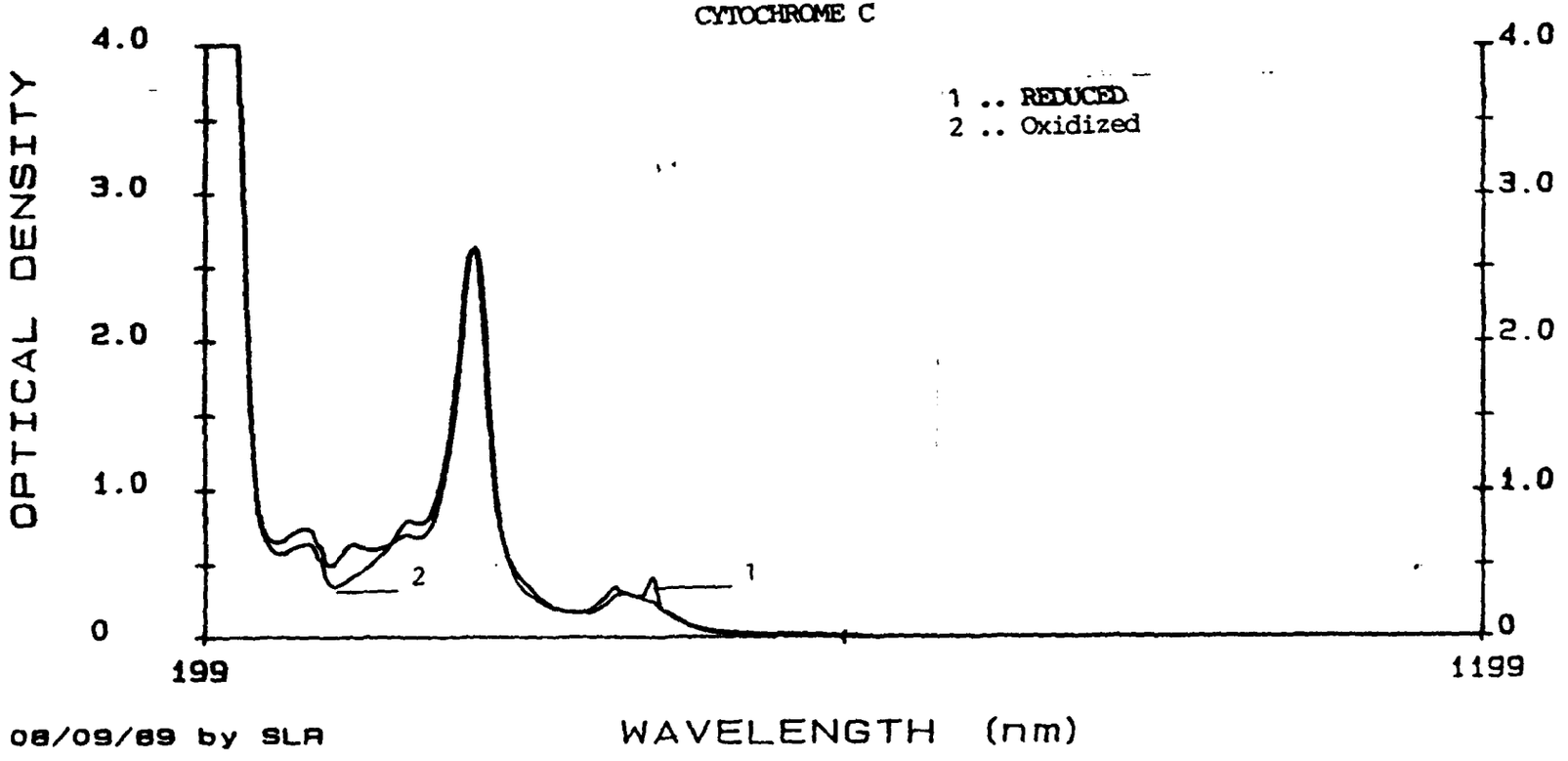
In figure two five hemoproteins are compared. They all have absorption maxima in the 408 to 402 nm range. The proteins: hemoglobin (hemogl), myoglobin (myogl), cytochrome C (cyto C), horseradish peroxidase (HRP), and catalase (cat), were used in the following concentrations respectively, 3.75uM, 15uM, 15uM, 7.5uM and 3.75um. It should be noted that the spectra for hemoglobin and myoglobin are congruent. The solvent used was pH 7.2 phosphate buffer. The spectra were obtained using a Varian DMS 100 spectrphotometer and Varian D515 data system. Figures 3 and 4 are obtained when the proteins are treated with melanin. A double beam configuration was used. For figure 3, the reference cell contained solvent equal in volume to that of the sample cell which contained melanin and hemoprotein. In obtaining figure 4, the reference cell contained solvent and melanin equal to that contained in the sample cell which contained the hemoprotein melanin mixture. In the difference spectra of figure 4, the absorbance due to unbound melanin is subtracted. The absorbance maxima are lower but they have not shifted. Suggesting that there is less hemoprotein absorbing, in each case. Since the electrophoresis experiment described below indicates that melanin binds to hemoprotein, the decreased absorbance is probably due to decreased concentration of free hemoprotein. One would therefore expect to detect evidence of the melanin hemoprotein complex as a new absorbance peak in a difference spectrum. This is seen in figure 5b which shows a titration of melanin against cytochrome C. The untreated cytochrome C exhibits the gamma Soret peak at 408 to 409 nm (figure 5a). It was shown in figure 4 that the gamma peak is reduced from 1.627, in figure 3, to 1.321 as seen in figure 4. In figure 5b the difference spectrum reveals a new peak at 416.9 nm.

Photographic plate #1 visualizes the ethanol extracted precipitate from a mixture of melanin and cytochrome C. It is of interest that the particles are conductive and do not require coating to facilitate SEM visualization.

Electrophoresis of melanin cytochrome C mixtures suggests that the two biopolymers form a relatively permanent complex which migrates to the positive pole at pH 8.6. Normally at that pH, cytochrome C (isoelectric point: 10.2) has a net positive charge and migrates to the negative pole. Data from an experiments comparing cytochrome C, hemoglobin, myoglobin are summarized in the following table, Table I.

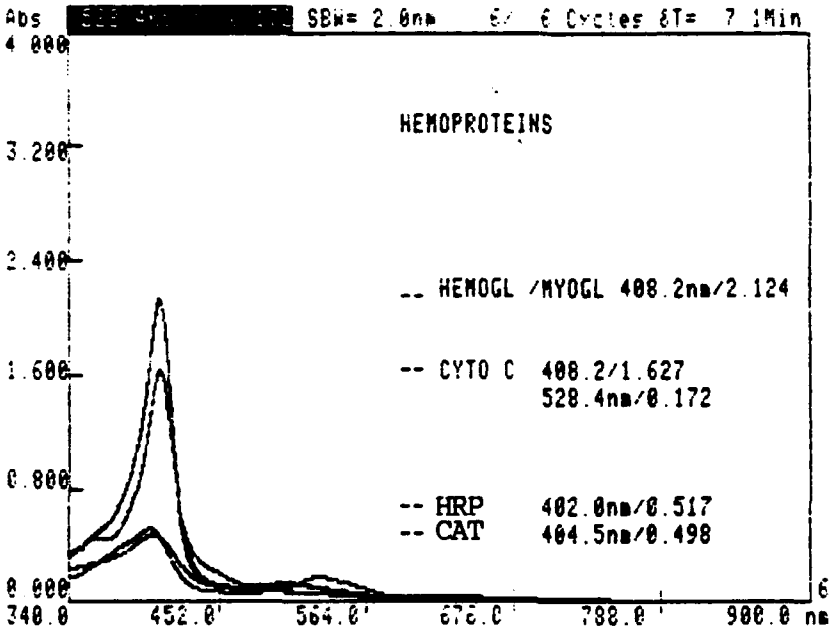
Figure 1

CYTOCHROME C



95-15

Figure 2


Abs
SBW= 2.0nm
6 Cycles
4.000
3.200
2.400
1.600
0.800
0.000
HEMOPROTEINS
-- HEMOGL /MYOGL 408.2nm/2.124
-- CYTO C 408.2/1.627
528.4nm/0.172
-- HRP 402.0nm/0.517
-- CAT 404.5nm/0.498
340.0
452.0
564.0
676.0
788.0
900.0 nm

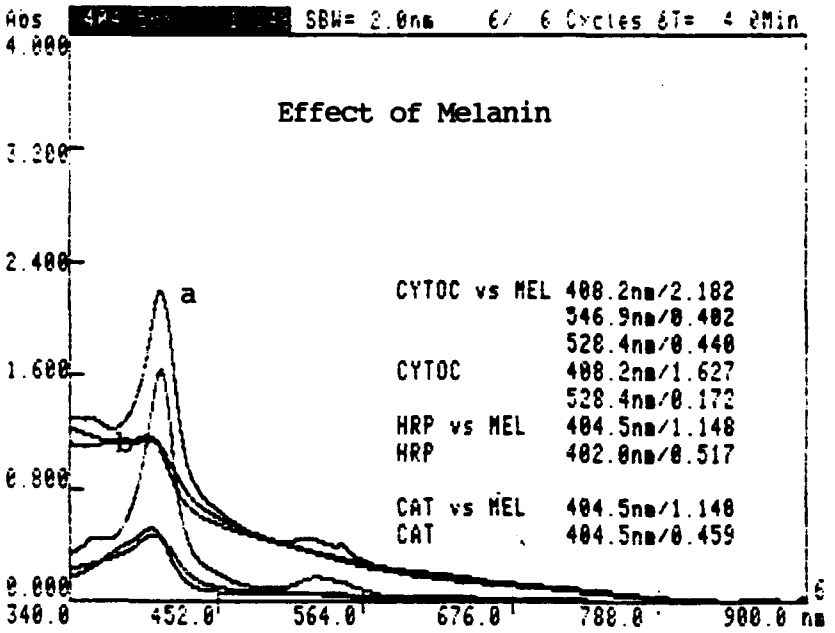


Figure 3

Figure 4

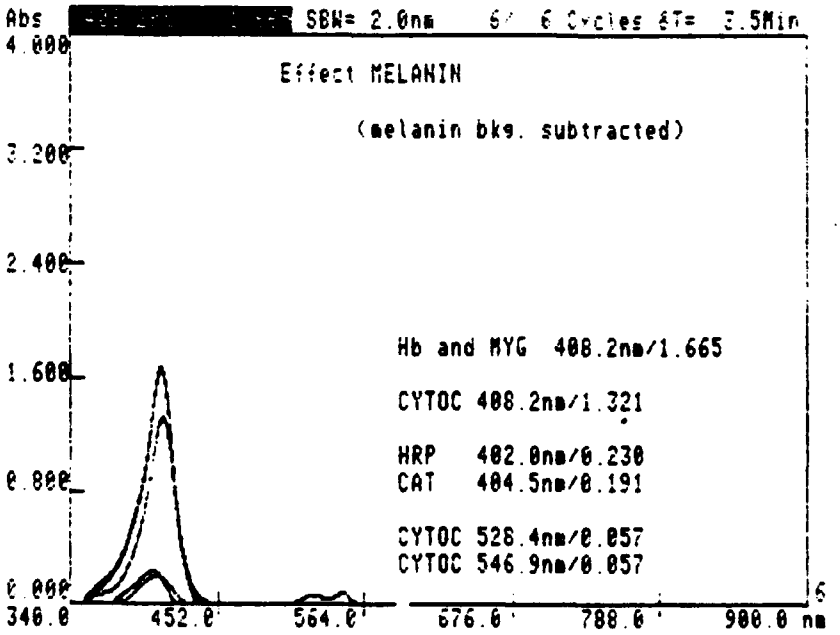


Figure 5a

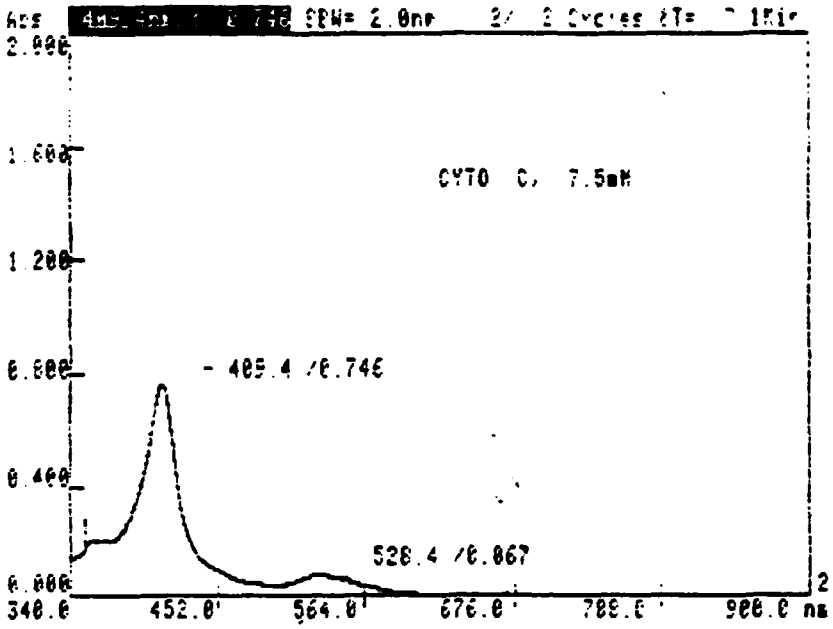


Figure 5b

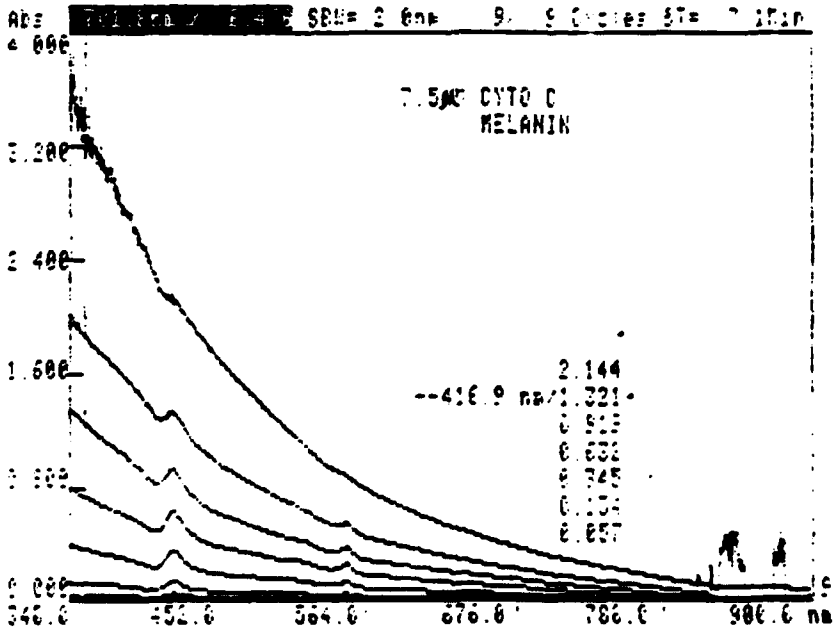


PLATE # 1

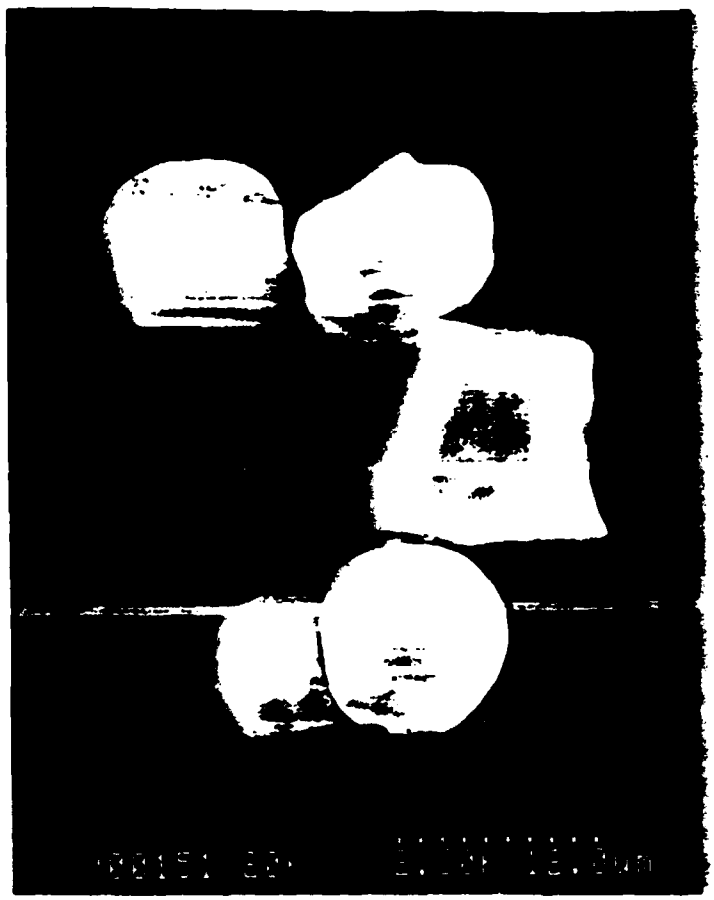


TABLE I
MIGRATION OF HEMOPROTEIN MELANIN COMPLEXES

Substance	Isoelectric point	Migration Distance
cytochrome C	10.2	migrates to negative pole (off gel)
myoglobin	7.2	1.8 cm
hemoglobin	6.9	4.0 cm
melanin		6.3 cm
*******		
cytochrome C/melanin		4.2 cm
myoglobin/melanin		5.2 cm
hemoglobin/melanin		5.4 cm

Cytochrome C, with an isoelectric point of 10.2, has a net positive charge in pH 8.6 buffer (Tris-glycine). It travels toward the negative pole and moves of the gel because of the orientation of the gel in the electric field. Myoglobin, isoelectric point: 7.2, and normal human hemoglobin, isoelectric point: 6.9, are negatively charged at pH 8.6, and they migrate toward the positive pole. The DOPA-melanin used in the experiment is considerably more negative than serum albumin which has an isoelectric point of 4.8. The isoelectric point of melanin was not determined. Melanin also binds to bovine serum albumin.

SEM Studies of Melanocytes in Tissue Culture

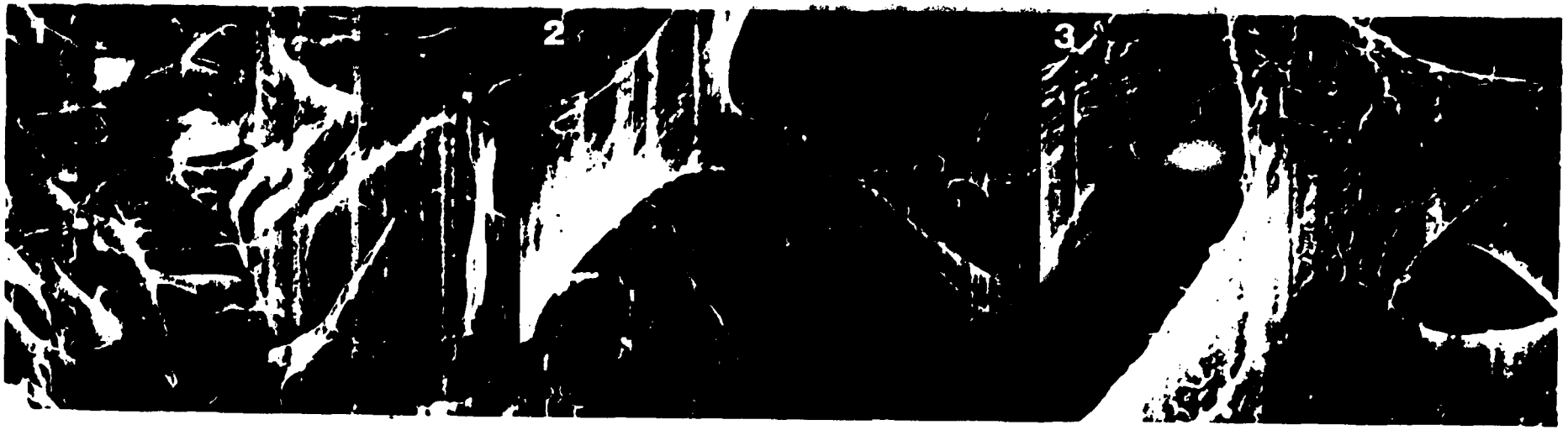
Melanocytes are large irregular dendritic cells. Even in very dense stationary cultures, confluent sheets were not formed. Rather densely packed "clonal foci" separated by moat like spaces predominate. The "clones" communicate by anastomoses of long processes. The "clones" grow vertically as well as horizontally. As a result of the "clonal" growth pattern, The mature culture resembles a three dimensional reticulum. Pigmentation begins when the foci attain dense vertical elevation. Thus a mature culture has the appearance of a brown reticulum.

Photo plate 2 shows four scenes exhibiting the "clonal" growth pattern. In panel 2-3 the vertical packing is very obvious. The scene in Panel 2-1 is magnified in plate 3. The cells have long processes which branch and interweave. The indicated scenes in panel 3-1 are shown magnified in panels 3-2 and 3-3. The processes vary in length and show branching complexity. The cell

PLATE # 2

95-21

PLATE # 3



95-22

surface exhibits decorations which might be a version of microvilli.

Panel 2-3 is shown enlarged in panel 1 of plate 4(panel 4-1). The dot and arrowhead indicate the subjects which are enlarged in panel 4-2 and 4-3 respectively..

Plate 5 shows a clone in an early stage of growth. Surface features and processes are visible. Panels 5-2 and 5-3 are higher magnifications of the the scene in panel 5-1.

Synthetic DOPA Melanin is a Conductive-coating Agent

Plate 6 shows a piece of melanin powder as it appears in the SEM. It is intrinsicly conductive, it does not require a metallic coat inorder to facilitate visualization in the SEM.

Cells may be visualized by SEM if they are first coated with melanin solutions. When DOPA melanin solutions, from which the powders are produced, are applied to fixed cells, the cells may be visualized in the SEM without the intervention of a metalic coat.

Plate 7 shows a sickle erythrocyte in an unusual configuration (Klein bottle ?). The cell was fixed, soaked in melanin and air dried prior to visualization.

Plate 8 shows telophase cells. The cell are shown at different magnifications. The cells in panels 8-2 and 8-3 are less conductive and the resolution is very poor. The cell in panel 8-1 is shown at higher magnification in plate 9. The resolutions is good and the conductivity is comparable to that of the cell in plate 10. The telophase cell in plate 10 received the conventional preparation described in the experimental design section, including coating with gold.

The methodology for producing optimum melanin coating is still evolving. The critical point drying regimen undermines the conductive coat. There is evidence that inclusion of materials to reduce surface tension during the postcoating air drying might be useful in improving the result.

PLATE # 4

95-24

PLATE # 5

95-25

PLATE # 6

PLATE # 7

PLATE # 8

PLATE # 9

PLATE # 10

Polymer Synthesis and Elongation

The effect of 3-AT on DOPA melanin synthesis was studied by following the protocol outlined in the methods section (page 10). The products have been obtained and stored for use in the future. The following summarizes the observation from that experiment. The reaction to produce the control specimen was prepared as usual. In making the T0 sample, one gram of 3-AT was added simultaneously with the DOPA. In making the T15 spemimen, 3-AT was added 15 minutes after stirring had begun. The control specimen was already dark after 15 minutes and black by 30 minutes. The T15 was also dark after 15 minutes, but took longer to turn black. The T0 sample was much retarded. It was a light orange at 30 minutes, subsequently turned orange brown, and turned black after an hour. The three samples were visibly identical after three days of stirring. It was necessary to check and readjusted the pH to 8.8. until it had stabilized by one hour after the start. The 3-AT samples required more NaOH to stabilize the pH at 8.8. The type III powdered melanins from the three samples were visually indistinguishable. The type II melanins were different in that it has not been possible to produce a powder from T0 and T15. T0 and T15 evaporate into a tacky tar-like material rather than a powder.

Prelininary NMR and IR Observations.

Figure two is the spectrum obtained for tyrosine-melanin (Sigma).The solvent used was chloroform and tetramethylsilane (TMS) was used to establish the reference peak. An attempt was also made to run a DOPA-melanin sample, but was not successful due to solubility problems. In the future dimethylsulfoxide will be used as solvent. No attempt is being made to analyze figure two. The main aim was to gain familiarity with the instrumentation and its output.

At this stage there is some indication that infrared spectroscopy will be very useful. Figure three is the IR spectrum of DOPA-melanin and Figure four is the IR spectrum of tyrosine-melanin. The spectra are quite different. The signature regions of the two spectra are distinctly different. In figure #3 the strong peaks are at 3420.23, 1616.55 and 1394.71. By comparison the significant peaks are at 3144.37, 2919.63, 1616.55. The only common peak is at 1616.55 which is tentatively identified a carbon/ carbon double bond in an aromatic ring. In figure #3, 3420.23 probably indicates hydroxyl groups as does the 3144.37 in figure #4.These preliminary observations are suggestive but no conclusion will be drawn at this point.

The NMR and IR experiments were subsequently discontinued due to equiptment unavailability.

5.0 CONCLUSIONS

The project objectives have been essentially accomplished.

The evidence indicates that melanin binds to all of the hemoprotein studied. Resulting in electronegative complexes which have absorption maxima at 416.9nm, 420nm, and [illegible]nm for cytochrome C, hemoglobin, and myoglobin respectively. The melanin-cytochrome C complex has been precipitated and visualized in the scanning electron microscope. The complex is conductive and does not require metallic coating for visualization by SEM.

The melanin not only binds to cytochrome C, it also changes its function because it reduces the molecule. It is not known whether melanin modifies the intrinsic functions of hemoglobin or myoglobin. Melanin suppresses horseradish peroxidase enzyme activity. A simple enzyme assay technique was used. An enzyme kinetics study might be useful in providing some insight into this effect. It is not known whether it is the melanin binding which inactivates the enzyme, or if it has some additional effect on the enzyme. Estimates of Kd (melanin-protein dissociation constant) for the reactions would also be useful.

The fact that melanin also binds to bovine serum albumin is a significant observation which was not pursued in these studies. It is possible that melanin binds to many proteins irrespective of net charge. If this speculation is true, the release of significant amounts of free melanin into the systemic circulation would be expected to cause pathological and diseased changes.

The SEM studies confirm the dendritic appearance of melanocytes. The reticulum like growth and anastomoses of the processes are propably new findings (of dubious importance). Typical microvilli were not detected although surface features were observed. The mitotic cells do "round-up", i.e., they do become spherical and are devoid of processes. As with other mitotic cells they are less tenaciously attached to the substrata.. This property was exploited to harvest mitotic cells for several experiments.

The most interesting result was the observation that melanin powders and films are conductive in the scanning electron microscope. This property was exploited and cells were made conductive by coating them with melanin. It is speculated that soluble conductive polymers may prove very useful in conventional scanning electron microscopy as well as scanning tunneling microscopy.

There is some evidence that 3-AT can influence the polymerization of chemically synthesized DOPA melanin. Products have been obtained. They have not, however, been analyzed.

6.0 REFERENCES

1. Friedmann,P.S. and B.A.Gilchrist, Ultraviolet Radiation Directly Induces Pigment Production by Cultured Human Melanocytes J. Cell. Physiol. 133: 88-94.

2. Kovacs,S.A, C.Radi and P.F.Agris. Transfer RNA Differences Between Swine Normal and Melanotic Tissue and Cell Strains" , in Proc. 10th Internatl. Pigment Cell Conference, Pigment Cell Biological Basis of Pigmentation (1977), 4: 79-86, Karger, Basal 1979.

3. Kiel, J.L., L. S. Wong and D. N. Erwin, Metabolic Effects of Microwave Radiation and Convection Heating on Human Mononuclear Leucocytes, Physiological Chemistry and Physics and Medical NMR, 18: 181-187, 1986.

4. Korytowski, W., B. Kalyanaraman, I. A. Menon, T. Sarna and R. C. Sealy, Reaction of Superoxide Anions with Melanins: Electron Spin Resonance and Spin Trapping Studies, Biochim. Biophy. Acta,882: 145-153, 1986.

5. Menon, I.A., S. Persad, H.F. Haberman, C.J. Kurian and P. K. Basu, Qualitative Study of the Melanins from Blue and Brown Human Eyes, Eye Res. 34: 531-537, 1982.

6. McGinness, J.E., R. Kono and W. D. Moorhead. "The Melanosome: Cytoprotective or Cytotoxic?", in Proc 10th International Pigment Cell Conference, Pigment Cell Biological Basis of Pigmentation (1977), 4: 270-276, S. Karger, Basal, 1979.

7. Pethig, R., Dielectric and Electronic Properties of Biological Materials, pp 342-344, John Wiley & Sons, N. Y., 1979.

8. Felix, C.C., J.S. Hyde, T. Sarna, R.C. Sealy, Melanin Photoreactions in Areated Media: Electron Spin Resonance Evidence for Production of Superoxide and Hydrogen Peroxide, Biochem. Biophys. Res. Comm., 84: 335-341, 1978.

9. Slawinska, D and Slawinski J., Electronically Excited Molecules in the Formation and Degradation of Melanins, Physiol. Chem. 14: 363-374, 1982.

10. Sealy, Roger, C., "Radicals in Melanin Biochemistry", in Methods in Enzymology, 105: 479-483, ed. Lester Packer, Academic Press, N. Y. 1984.

11. Sarna, Tadeusz, B. Pilas, E. J. Land, and T.G. Truscott, Interaction of radicals from water radiolysis with melanin, Biochim. Biophy. Acta, 883: 162-167, 1986.

12. Avers, Charlotte J., Molecular Cell Biology, Reading, MA, Addison-Wesley publishing Co., 319, 1985.

13. Crippa, P.R., A. Mazzini and D.Salmelli, Oxidation of NADH by Melanin: Effect of UV light and Copper Ions, Physiological Chemistry and Physics, 11: 491-499, 1979.

14. Slawinska, D., J. Slawinski and L. Ciesla, The Inhibition of Peroxyradical-induced Chemiluminescence by Melanins, Physiological Chemistry and Physics and Medical NMR, 15: 209-222, 1983.

15. Sarna, T., W. Korytowski and R. C. Sealy, Nitroxides as redox probes of melanins: darkk-induced and photoinduced changes in redox equilibria, Arch Biochem Biophys, 15:239(1) 226-233, 1985.

16. Worthington Enzyme manual,Enzymes and Related Biochemicals, ed. Charles C. Worthington, 207-209 Worthington Biochemical Corporation, Freehold, New Jersey, 1988.

17. Sheehan, D. C. and B. B. Hrapchak, Histotechnology St Louis, MO, The C. V. Mosby Co., 222, 1980.

18 Postek, M. T., K. S. Howard, A. H. Johnson, amd K. L. McMichaels, Scanning Electron Microscopy, Ladd Research Industries, Inc., 294, 1980.

Report # 96
760-0MG-110
Prof. Amir Karimi
Report Not Publishable

"ROBUST FILTERING OF BIOLOGICAL DATA"

1990 USAF-UES RIP

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

Prepared by:	Harold G. Longbotham, Ph. D.
Academic Rank:	Assistant Professor
Department and University:	Electrical Engineering University of Texas at San Antonio
Research Location:	USAFSAM/RZV Brooks AFB San Antonio, Texas 78235
USAF Researcher:	Captain Norman Barsalou

Date:	Oct. 3, 1990
Contract No.	F49620-88-C-0053/SB5881-0378

"ROBUST FILTERING OF BIOLOGICAL DATA"

by

Harold Longbotham

Abstract:

This report gives an update of our research to date in the application of nonlinear filters to signals with impulsive (or bursty) and statistical (independent and identically distributed) noise.

In the first part we report on our work in the development of an entire class of filters that may be designed for use in filtering piecewise constant signals in the presence of impulsive, bursty, or statistical noise. For our work at Brooks AFB these filters are used in the processing of VEP data. But since the sharpness of edges is one of the main criterion for a "good" image the filters will obviously have application in image processing. Our work in this area has produced more results than anticipated in this area.

In the second part we examine the Sweep VEP data and show how these same filters when used as prefilters eliminate the noise so that the point of inflection (point of acuity) may be easily inferred.

Finally we discuss what started out to be a tutorial but now is starting to look more like a book explaining these nonlinear filtering methods. We hope that this book (together with the demonstration package we are putting together) will explain these new techniques to the uninitiated scientist working in situations which call include impulsive noise.

Acknowledgements

First I would like to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsorship of this research, and Universal Energy Systems for their cheerful help in all the administrative details.

More specifically I would like to acknowledge the support I received at USAFSAM/RZV. Dr. Farr and Lt. Col. Cartledge offered me encouragement and support in all endeavors. Dr. Randolph Glickman provided the much needed expertise about VEP data and its historical analysis. Jim Brakefield (Krug International) and Captain Norman Barsalou provided the much needed support as to existing implementations, modern techniques that could be used, and support in signal processing in general.

I consider the time spent at Brooks AFB to be invaluable. The application of theories to real world situations is indeed the appropriate test bed for signal analysis.

I. Introduction

This research continued the research started in the summer of 1989 for the Laser Branch of the Radiation Science Division at Brooks Air Force Base (USAFSAM/RZV). Two conference papers were written and presented at the 1990 SPIE/SPSE conference in Sante Clara, California. A third paper was written and presented at the IEEE Circuits and Systems meeting in New Orleans, Louisiana. The first paper describe the LMS filter, a robust nonlinear filter to be used in the filtering of VEP (visual evoked potential) and sacadic eye movement. The second paper concerns the real time implementation of the Order Statistic (OS) filters in terms of stack filters.The third paper generalizes the use of OS filters in VEP data and demonstrate how it is useful anytime one is filtering a periodic signal in the presence of impulsive and/or independent and identically distributed (i.i.d.) noise. We include abstracts of these papers and detail how they are being expanded for future research. Lastly we are working on an exposition of these nonlinear filters and their use for the nonengineering (nonmathematical/statistical) community.

II Objectives of the Research Effort:

A. The LMS Filter, A Robust Alternative to OS Filtering

i. Conference paper

ii. Extensions of work

iii. Journal Paper

B. Stack Filters, The Real Time Implementation of OS and LMS Filters

i. Conference paper

ii. Extensions of work

iii. Journal paper

C. Various Approaches to Sweep VEP Data Analysis

i. Comparison of possible approaches

ii. Figure of results

iii. Conference paper

D. Appropriate Uses of Linear and Nonlinear Filtering

i. Expository paper written

ii. Analysis package

III. Results:

A. The LMS Filter, A Robust Alternative to OS Filtering

One conference paper was proposed to be written. Instead two papers were written in this area. These papers have been written, reviewed, accepted and presented. They are included in the appendix for review.

With regard to the first paper on the LMS filter, we have discovered how to generalize this into a class of filters we call the GOS/WMMR filters. In doing so we are presently developing a totally new field in nonlinear filtering. This class of filters have the following properties which make them optimal for many situations in biological and image processing.

(1) Optimal design procedures are being developed for impulsive noise.

(2) Optimal design procedures are being developed for I.I.d. noise that may be used simultaneously with (1).

(3) PICO (piecewise constant) signals are fixed points.

(4) Corrupted edges are enhanced upon filtering.

I deeply regret space does not exist for us to go into detail about this for I consider it to be a significant achievement.

B. Stack Filters, The Real Time Implementation of OS and LMS Filters

One conference paper was written in this area. This paper has been written, reviewed, accepted and presented. It is included in the appendix for review.

After talking to Dr. Ed Coyle (Purdue University) at the IEEE Circuits and Systems conference, we realize that the paper must be extended into a journal paper for it has considerable significance for his work in adaptive OS filters. However now is not the time for this. The algorithms for our work at Brooks AFB are still in the development stage and it would seem inappropriate to begin a detailed implementation program at this point.

C. Various Approaches to Sweep VEP Data Analysis

The following various approaches to Sweep VEP data were to be included:

(1) Current estimation technique used by P.I. of just "eyeballing" the data.
(2) Linear regression
(3) Robust linear regression
(4) Robust filtering in time followed by linear regression.
(5) Filtering across experiments

(1), (2), and (4) have been implemented in the lab. For (3) we have identified an approach called "least median of squares regression" that seems very promising. We hope to have this technique programmed within a month. The bootstrap and jackknife methods were also examined but rejected as being inappropriate. We must also program (5) at this point.

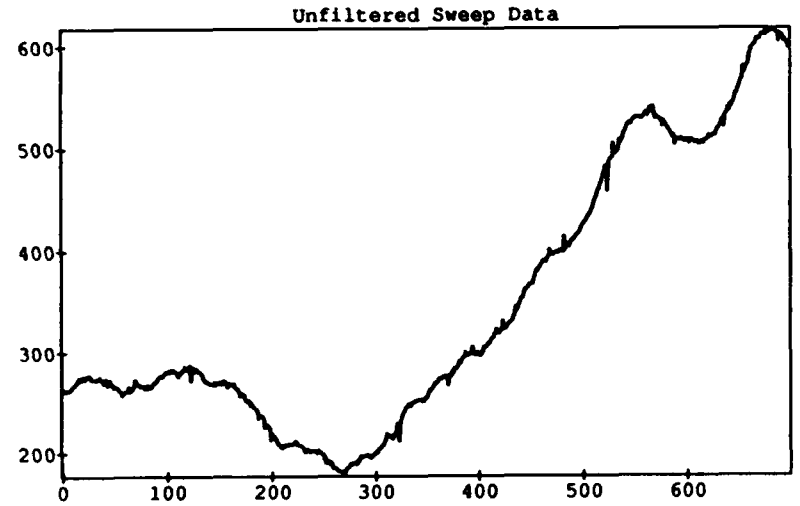
Method (4) seems to hold a lot of promise at this point as the results on the next page indicate.

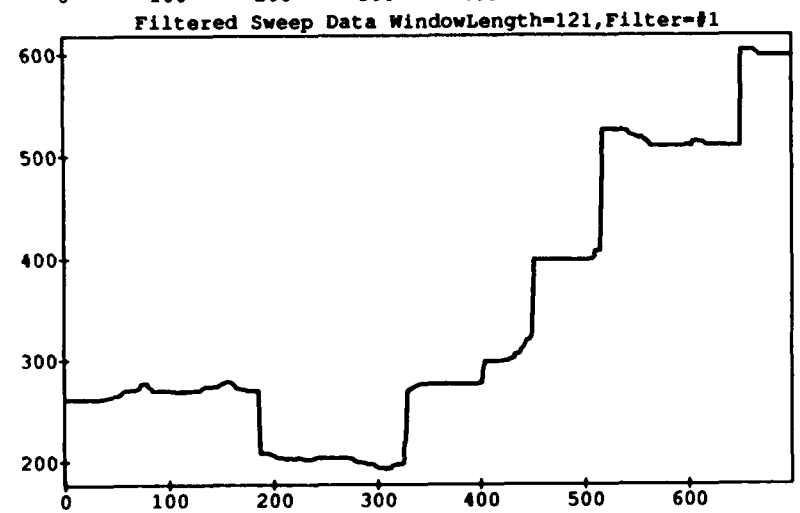
D. Appropriate Uses of Linear and Nonlinear Filtering

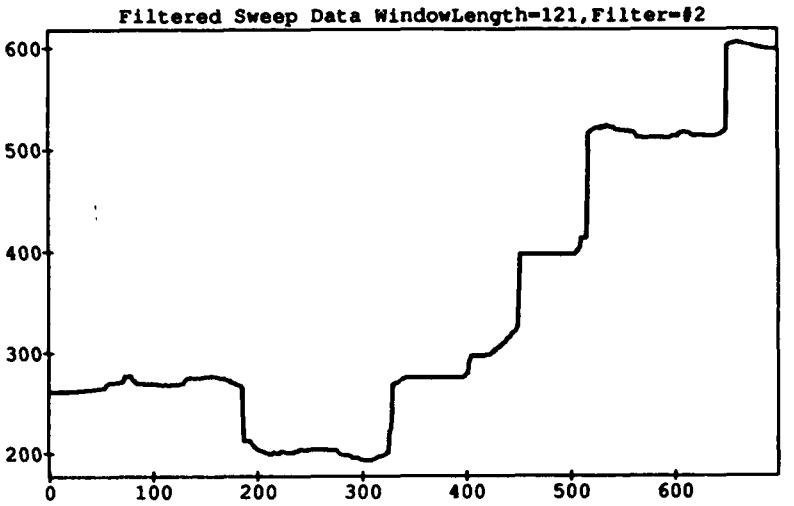
At this point we are still working on this manuel. We have yet to include all the new results on the GOS/WMMR filter. There is a demonstration package for the Steady State VEP data that is nearing completion.

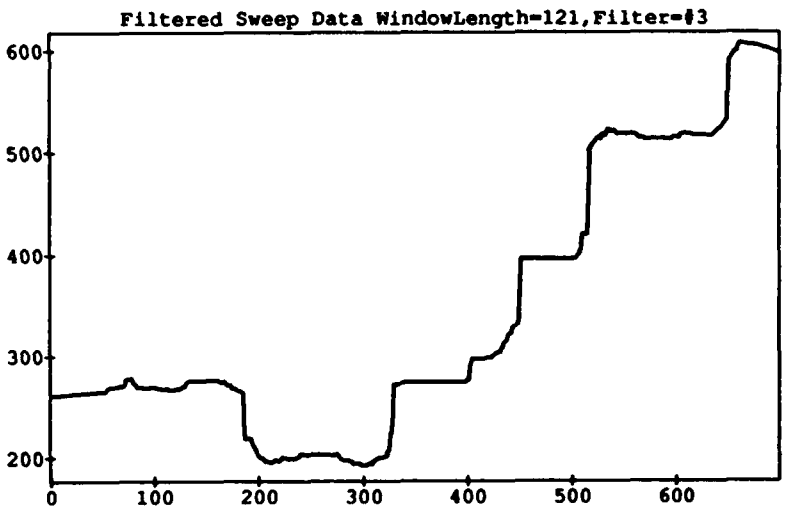
IV. Appendix

The appendix contains the three papers we have finished to date.


Unfiltered Sweep Data
600
500
400
300
200
0
100
200
300
400
500
600


Filtered Sweep Data WindowLength=121,Filter=#1
600
500
400
300
200
0
100
200
300
400
500
600


Filtered Sweep Data WindowLength=121,Filter=#2
600
500
400
300
200
0
100
200
300
400
500
600


Filtered Sweep Data WindowLength=121,Filter=#3
600
500
400
300
200
0
100
200
300
400
500
600

Unfiltered Sweep Data
Filtered Sweep Data WindowLength=121,Filter=#1
Filtered Sweep Data WindowLength=121,Filter=#2
Filtered Sweep Data WindowLength=121,Filter=#3
0
200
400
600
800

Unfiltered Sweep Data
Filtered Sweep Data Window Length=221,Filter=#1
Filtered Sweep Data Window Length=221,Filter=#2
Filtered Sweep Data Window Length=221,Filter=#3
100
200
300
400
500
600
0
200
400
600
800

The LMS, an adaptive optimal order statistic filter

Harold G. Longbotham
Norman Barsalou

Nonlinear Signal Processing Group
Department of Engineering
University of Texas at San Antonio
San Antonio, Texas
78285-0665

1. INTRODUCTION

The median filter has gained recognition as a filter that will edit impulses and retain edges. Order statistic (OS) filters such as the α-trimmed mean have also proven useful in providing smothers that also eliminate impulses. At first implementation was slow, but due to the introduction of the stack filter may now be implemented in real time.

These filters have been extended to the generalized order statistic filters (GOS) which include the least median of squares (LMS) filter. The LMS filter averages the N+1 closest grouped (algebraic range) values in a window of size 2N+1. We show the LMS eliminates impulses and preserves "perfect" edges exactly as the median. In addition the LMS is shown to enhance "nonperfect" edges.

We next introduce the breakdown point as measure of the robustness of a filter.The breakdown point of a filter is defined to be the percentage of aberrant values allowed in the window before the output can be made to take an arbitrarily large value. We note the breakdown point for the averager is 0%, for the median and LMS it is 50% (asymptotically), and for the α-trimmed mean can vary from 0% to 50%.

Presented at and Published in the Proceedings of
The 1990 SPIE/SPSE Symposium on
Electronic Imaging Science and Technology
February 1990, Santa Clara, California

Using the breakdown point we then look at the appropriateness of the distance measures currently in use and introduce a new distance measure. One object of filtering is to eliminate noise so that the input signal approaches a "nearest neighbor" w.r.t. some metric. The sum of squares (SS) is used in most linear applications where the noise is additive. We introduce a median of squares (MS) measure in its place and demonstrate its place when using "editing" GOS filters for removal of impulsive noise.

Some assumptions that are made herein are the use of odd length (2N+1) windows, output taken at the center of the filter, and the use of sliding overlapping windows. We do realize the output of the LMS filter may be undefined but since this only occurs with probability zero (all sequences are assumed to be defined on the reals), we exclude this case in the following considerations.

2. DEFINITIONS

The following definitions will be needed. It is assumed the definitions are w.r.t. a window of width 2N+1.

(1) segment: A finite number of consecutive values of a sequence.

(2) LOMO-(N): A digital signal is locally monotone of degree N or LOMO-(N) if each segment of length N is nondecreasing or nonincreasing.

(3) Padding: In this paper we consider all finite length signals to be padded with N values before (after) the first (last) value, each equal to the first (last) value.

(4) $\mathbf{x}_i$: The set of values $x_i, x_{i+1}, \ldots, x_{i+2N}$, for i any integer.

(5) fixed points: The sequence $\{x_i\}$ is a fixed point for filter T if $T[\mathbf{x}_i] = \mathbf{x}_i$, for every i.

(6) constant segment: A segment of N+1 or more equal values comprises a constant segment.

(7) Impulse: A nonconstant segment comprised of a segment of length N between two constant segments of the same amplitude.

(8) perfect edge: A segment encompassing a constant segment followed by another (nonequal) constant segment.

(9) nonperfect edge: A monotone segment encompassing two constant segments separated by a monotone nonincreasing (nondecreasing) segment of length less than N+1.

(10) outlier: value outside the normal realm of expectation. Usually defined w.r.t. a distribution but here we will be developing distribution free results.

(11) OS filter: The output is the weighted sum of the algebraically ordered windowed values.

(12) median filter: An OS filter that weights the central windowed value by one and all other values by zero.

(13) G.O.S. filter: The output is the weighted sum of the ordered windowed values [1].

3. FIXED POINTS OF THE LMS

In this section we show that for bivalued signals the LMS and the median have exactly the same fixed points. However, for multivalued signals the LMS has a passband similar to the median, but also "smooths" signals and "enhances" edges.

Lemma 1: The median and the LMS have exactly the same output for any bivalued input.

Proof: Assuming $\{x_i\}$ is a bivalued signal, the output of any window for the median filter is the value in the window that occurs at least N+1 times. With a bivalued input the output of the LMS filter is also the value that occurs N+1 or more times. Therefore the two filters have identically the same output for any input.

Corollary: The median and LMS have identically the same fixed points for bivalued signals, and for a finite length bivalued signal they converge to a fixed point at the same rate.

This corollary together with prior results for the median [2] show that if only padded finite length signals are considered, the fixed points of the LMS are all LOMO-(N+2), . In addition the corollary and prior results [3] show that if we

consider infinite length signals there are an infinite number of possible LOMO-(N+2) fixed points and a finite number of periodic fixed points possible. This demonstrates that since the median is computationally more efficient than the LMS, there is no reason to use the LMS when considering only bivalued signals, or if one considers thresholding signals and application of the filter at each level (as in stack filters).

When considering nonbivalued signals it is obvious the outputs may differ since the output of the LMS need not be a windowed value. Still, the similarities between the median and the LMS are striking with the LMS offering certain advantages as we outline in the following theorem. But first we need the following lemmas.

Lemma 2: Assume $\{x_1,x_2,....,x_{2N+2}\}$ is increasing (decreasing) monotonically, then LMS($\mathbf{x}_1$) is less than or equal to (greater than or equal to) LMS($\mathbf{x}_2$).

Proof: WLOG we assume $\{x_1,x_2,...,x_{2N+2}\}$ is increasing monotonically. Let $x^i = x_{N+i}-x_i$, for i from $1\le i\le N+2$. Let $X = \{x^i, 1\le i\le N+1\}$, x = MIN (X) and $Y = \{x^i, 2\le i\le N+2\}$, y = MIN (Y). Note that $X-Y = x^1 = x_{N+1}-x_1$, $Y-X = x^{N+2}=x_{2N+2}-x_{N+2}$ and LMS ($\mathbf{x}_1$) = AVE{N+1 values with range x}, LMS ($\mathbf{x}_2$) = AVE {N+1 values with range y}. LMS ($\mathbf{x}_1$) $\neq$ LMS ($\mathbf{x}_2$) iff we are not averaging the same N+1 values iff x_1 (x_{2N+2}) is used in calculating LMS ($\mathbf{x}_1$) (LMS ($\mathbf{x}_2$)). In either case LMS ($\mathbf{x}_1$) $\le$ LMS($\mathbf{x}_2$). Similar arguments hold if $\{x_1,x_2,...,x_{2N+2}\}$ is decreasing monotonically.

Corollary: A monotone increasing (decreasing) sequence is still monotone increasing (decreasing) after LMS filtering.

Lemma 3: For any LOMO-(N) signal, changes in monotonicity are separated by constant segments of length at least N-1.

Proof: [4]

Lemma 4: Assume a signal is LOMO-(N+2), then each segment of length 2N+1 has at least one of the following characteristics:

(a) It contains a constant segment of length N+1.

(b) The segment of length 2N+3 (the original segment extended by one point in each direction) is monotone.

Proof: WLOG consider the segment $(x_1,x_2,....,x_{2N+1})$ from a LOMO-(N+2) signal. Assume (b) is not fulfilled. Then there exist an i and j, $0 \le i,j \le 2N+2$, such that $x_i < x_{i+1}$, $x_j > x_{j+1}$. But for the signal to be LOMO-(N+2) there must be an N+1 constant segment separating x_i and x_j by the previous lemma. Therefore (a) must hold.

Theorem 1: For multivalued signals we have the following results upon LMS filtering:

(a) Perfect edges are fixed points.

(b) Impulses are eliminated upon one pass.

(c) Constant segments do not decrease in size.

(d) If the input is LOMO-(M) for M greater than or equal to N+2, then the output is LOMO-(M).

(e) Nonperfect edges become perfect edges upon iteration.

Proof: (a) If we are windowing a segment from a perfect edge, then at least N+1 consecutive values in the window are equivalent, therefore the output is the same as the central value.

(b) If the window contains an impulse, then there are at least N+1 equal values in the window (with a different value from that of the impulse) therefore the impulse is eliminated upon one pass.

(c) If the central value in the window is a value from a constant segment, there are at least N other equal values in the window. Therefore the output is the central value and a constant segment cannot decrease in size.

(d) WLOG assume the central value is x_{N+1}. If the input is LOMO-(M), $M \ge N+2$, then by the preceding lemma x_{N+1} is either in a constant segment or imbedded in the monotone segment $\{x_0,x_1,...,x_{2N+2}\}$. If x_{N+1} is in a constant segment then it is the output. If $\{x_0,x_1,...,x_{2N+2}\}$ is a monotone segment, then by a previous lemma the monotonicity is preserved by the output. In either case, if the input is LOMO-(M), the output is LOMO-(M).

(e) Assume the filter is in a nonperfect edge. If the central value is a value different from either of the constant regions then by (c) it is the output. If the central value is in between the constant regions then the output is closer to the closest constant region. Therefore the output will approach a perfect edge upon iterations of the filter. Note we have excluded intermediate values half way between the two constant regions since the LMS is indeterminate in that case.

4. THE "BREAKDOWN POINT CRITERION" AND ROBUSTNESS

Historically [5,6,7], the breakdown point has been used in the statistical literature as a measure of the robustness of an estimator or a minimization criterion. In this section we will describe its use as a measure of the robustness of an estimator in statistics, extend this to filters, then use it to compare the robustness of the median, LMS, averager, and α-trimmed mean filters. In the next section we will describe its use as a measure of the robustness of a minimization criterion.

Intuitively the breakdown point is the smallest amount of contaminants that may cause an estimator to take on an arbitrarily large aberrant value. The optimal value of the breakdown point is usually accepted to be 50% since if more than 50% of the values are perturbed they could be perturbed in such a way as to mask the original distribution/sample. We note the breakdown point of the mean is 0 and that of the median is 50%.

Hampel [6] used a Mote Carlo to study estimators of location in the presence of outliers. Six estimators were chosen, each of which rejected outliers by some criterion and then averaged the remaining values. He also calculated the breakdown point of each estimator and showed a simple summary and theoretical explanation of the results of the Monte Carlo study could be explained by the breakdown point. This and other results [7] have led us to expect the breakdown point to be in direct correlation with the robustness of an estimator.

In filtering we wish to define a similar concept. We note we cannot use the same concept since in an infinite length sequence we would need only perturb a segment of length $2N+1$ and if the filter depends at all on the signal values, at that point the value would be infinite. This would imply a breakdown of 0% for any filter. Therefore we will define the breakdown point of a filter to be the fraction of aberrant values in a particular window that will cause the filter to take on arbitrarily large values. We note the breakdown point for the median is 50% and for the averager is 0%. For the LMS the breakdown point is asymptotically 50%. If the α-trimmed mean is defined so as to (asymptotically) trim a total of α% of the values then it has a breakdown point of $(\alpha/2)$%.

This is one reason we call the LMS an optimal OS (actually a GOS [1]) filter. It is an smoother based on the rank order values that achieves the optimal breakdown point. To see the significance of this we note that if one compares the α-trimmed mean that averages N+1 values to the LMS he would expect them to have similar smoothing properties but that for the LMS to be capable of trimming twice as many outliers if the amplitude of the impulsive noise is from one tailed distributions such as the exponential.

5. SS AND MS, DISTANCE MEASURES

In this section we discuss the inappropriateness of the usual distance metric (sum of squares) when using GOS filters for impulsive noise. In its place we suggest a variant of the median of squares as introduced in the statistical literature. We then point out that the average of the N+1 closest clustered values in a sample of 2N+1 is the LMS (least median of squares) estimate yielding yet another reason why the LMS filter may be considered an optimal averager under certain noise conditions.

We normally wish to measure the distance of signals in several different instances. Three distinct times come to mind. We would like to use a distance measure to find the closest "desired" signal. After filtering we would like to measure the closeness of the output signal to the previously determined "desired" signal. And in adaptive filtering we would measure as part of the filtering process. Historically the SS (sum of squares) is used in all three instances, i.e. we take the sum of the squares of the differences of the signal and the desired signal. This measure has proven effective and tractable for linear filters and additive noise. Now I am proposing that we change our distance measure for "editing" filters such as the OS and GOS filters discussed here and elsewhere.

Take the example of an constant signal of length 2L and amplitude zero as the desired signal. If we perturb one value by L the SS distance from the new signal (call it signal 1) to the desired signal is L^2. If we then perturb every other point in the desired signal by 1 then the new signal (call it signal 2) is also a distance of L^2 from the desired signal. But our intuition tells us the desired signal is recoverable from signal 1 and not signal 2. To back up our intuition we might note there are any number of OS filters that one could use to recover the desired signal from signal one but I can think of no OS filter that would recover the desired signal from signal two. The point is that in using "editing" OS filters and in the presence of noise the SS measure is no longer meaningful.

A similar change is coming about in classical linear regression. Rousseeuw [7] notes "the method of LSE (least squares error) is being criticized more and more for its dramatic lack of robustness." He has suggested that instead we replace the mean with the median and use the LMS (least median of squares error). Indeed he uses the breakpoint to measure the robustness of the LMS and the LSE and show them to have breakpoints of 50% and 0% respectively. After the preceding section this and its implications should be obvious to the reader.

We do note that we cannot just accept the LMS as above for the same reasons we could nor accept the breakpoint as described in the statistical literature. In stead one might use the average of the MS of each segment of length 2N+1 when using the median or LMS filter, but this is still an area of current research. This method would have to be modified for editing filters with breakdown point less than 50%.

Now we come to another reason for calling the LMS filter an optimal filter. Rousseeuw has shown that just as the average is the least squares estimate, the average of the N+1 (out of a sample size 2N+1) closest clustered values is the least median of squares estimate. This estimate is also similar to "shorth" in the Princeton Monte Carlo study [A 1972]. Rousseuw [7] also suggest that in the case of multiple possible values for the LMS estimate you take the average. However, the case of multiple possible LMS estimates is assumed to have probability zero and therefore deleted from consideration (as we have in this paper).

6. IMPLEMENTATION OF THE LMS FILTER

We discuss the implementation of the LMS filter both as a factor of time and cost. The most obvious implementation of an LMS filter is using 2N+1 stack [8] filters (each of the rank orders) followed by a stack (min) filter. Therefore it may be implemented in real time. If C is the cost of one stack filter it is obvious the cost of a median filter is C, the cost of an a-trimmed mean that averages N+1 values is (N+1)C, and the cost of the LMS is (2N+2)C. We are presently working on modifications to stack filters that will generate the LMS at approximately the same cost as the median.

7. APPLICATIONS OF THE LMS FILTER

In discussing applications of the LMS we describe the application we first read that lead us to study the LMS, our own application to biological data, and its use in image processing.

C.M. Wang [9] used the LMS estimate is designing a algorithm for a robot wall follower. The problem was to design a robot that would follow corridor walls, turn into the first corridor possible (following the wall around the corner), and ignoring plants, doorways, water fountains, windows, etc.. He compared the LMS estimate to several other robust estimates and found it to preferred.

Visually evoked potential data [10] is normally in the millivolt range and at 1 to 10 hz. The usual data analysis [10] leads to sequences that would be constant except for the presence of noise. One type of noise is that of a muscle artifact which is bursty in nature and can be of length up to 10 samples (depending on the sampling rate). This burst may or may nct be all of the same sign. Therefore if one is to use an α-trimmed mean and hope to eliminate all of the burst he must make the α-trimmed mean filter twice as long as the corresponding LMS filter or median filter. As a general rule we use as small a window as possible and therefore would choose the median or the LMS. Historically the filter of choice has been the averager, therefore the filter of choice would be the LMS since it smooths and eliminates the same burst length as the median for a given window size.

The median filter has been successful in image processing due to its ability to eliminate impulses and preserve edges. The LMS has been shown (Theorem 1) to eliminate impulses, preserve perfect edges and enhance nonperfect edges. Therefore it to should have a place in image processing when one wishes to incorporate the properties of the median with smoothing and edge enhancement.

8. EXTENSIONS

In this section we note several possible extensions. The filter length can be even. One might wish to edit less and therefore enlarge the group with smallest range to be averaged. One could use a weighted average. We also note there are other GOS estimates that obtain the optimal breakdown point.

First we note there is no reason to assume there is an odd number of values in the window. The LMS could be defined as the average of the closest grouped majority of values. In the rest of this section we will assume the window length is N.

We note the LMS filter and measure of distance are perfect if one desires to use a smoother and expects to have approximately N/2 perturbed values in each window. But there will be occurances where one wishes to do more smoothing and expects less than N/2 perturbed values in each window. In these cases one might introduce the α-LMS, $N/2 \le \alpha \le N-1$, where the α closest values in the window are averaged. A corresponding measure of closeness should also be defined.

Just as the OS filter with different weights has been shown to be optimal for varying noise conditions [11], it is not inappropriate to assume the usefulness of the α-LMS may be improved by varying weights form those of the simple averager. Therefore given an OS with L consecutive zeros on each end, one would consider defining the "equivalent" LMS filter to be an (N-2L)-LMS filter where the closest grouped N-2L values are ordered and weighted by the same coefficients as the OS.

It should be noted that there are other parameter free statistical estimates (and therefore filters) with an optimal breakdown point. Most notably there are the Huber-type skipped mean [6] and the Shapiro Wilk [6] estimates. The LMS was chosen for this paper over these because of the obvious link to the α-trimmed mean and median, the usefulness of its set of fixed points, and the ease of implementation.

We call all GOS filters like the LMS and Huber-type skipped mean adaptive (in comparison to the standard OS filter) in that they are smoothers that adapt to edit bursty noise up to 1/2 the window in length, whereas the equivalent α-trimmed mean that averages N+1 values will only edit 1/2 of the burst (i.e. a burst only 1/4 of the window length).

9. ACKNOWLEDGEMENTS

I would like to thank Dr. Jerome Keating , Statistics Department, the University of Texas at San Antonio for introducing me to the LMS filter through

reference [9]. I would also like to thank Dr. Cartledge, Brooks AFB, for promoting my interest in applying OS filters to biological data. I would also like to thank the Air Force for sponsoring this research in part through contracts AFOSR-89-0490 and F49620-87-R-0004. The U.S. Government is authorized to reproduce reprints for government purposes, nothwithstanding any copyright notion hereon.

10. REFERENCES

[1] Longbotham, H.G., Bovik, A.C., and Restrepo, A.P., "Generalized order statistic filters," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, Scotland, 1989.

[2] Gallagher, N.C. and Wise, G.L., "A theoretical analysis of the properties of the median filter," *IEEE Trans. Acoust., Speech, Sig. Process.*, vol.ASSP-29, pp. 1136-1141, 1981.

[3] Eberly, D., Longbotham, H.G., and Aragon, J., "Complete classification of roots to 1-dimensional median and rank-order filters," submitted to *IEEE Trans. Acoust., Speech, Sig. Process*, Spring 1989.

[4] Longbotham, H.G., and Bovik, A.C., "Theory of order statistic filters and their relationship to FIR filters," *IEEE Trans. Acoust., Speech, Sig. Process.*, vol.ASSP-37, no. 2, pp. 275-287, 1989.

[5] Donoho D., and Huber, P.J., "The notion of breakdown point", **A Frestschrift for Erich L. Lehman**, P.J. Bickel, K. Doksum, and J.L. Hodges, eds., pp.157-184, Wadsworth, Belmont, Calif., 1983.

[6] Hampel, F.R., "The breakdown points of the mean combined with some rejection rules," *Technometrics*, 27, pp. 95-107, 1985.

[7] Rousseeuw, P.J., "Least median of squares regression," *Journal of the American Statistical Association*, 79, pp. 871-880, 1984.

[8] Wendt, P.D., Coyle, E.J., and Gallagher, N.C., "Stack filters," *IEEE Trans. Acoust., Speech, Sig. Process.*, vol.ASSP-34, no. 2, pp. 898-911, 1986.

[9] Wang, C.M., "A robust estimator for wall following," *Communications in Statistics- Theory and Methods*, 17(2), pp. 411-422, 1988.

[10] Regan, D., **Evoked Potentials in Psychology, Sensory Physiology and Clinical Medicine**, Chapman and Hall, London, England, 1972.

[11] Bovik, A.C., Huang, T.S., and Munson, D.C., "A generalization of median filtering using linear combinations of order statistics," *IEEE Trans. Acoust., Speech, Sig. Process.*, vol.ASSP-31, pp. 1342-1350, 1983.

ANALYSIS OF PERIODIC SIGNALS VIA ORDER STATISTIC FILTERS

Harold G. Longbotham[1], Thomas R. Arnow[1], Norman Barsalou[1],
Randolph D. Glickman[2], and Jerome Keating[1]

[1]The Nonlinear Signal Processing Group
The University of Texas at San Antonio
San Antonio, Texas 78285
[2] Department of Ophthalmology
The University of Texas Health Science Center
San Antonio, Texas 78284

Abstract

In this paper we will set forth an approach to the robust analysis of periodic digital data. This approach uses optimal order statistic (OS) filters for the filtering of non-Gaussian additive i.i.d. noise and OS and generalized order statistic (GOS) filters to factor out impulsive and/or bursty noise. The noise is not necessarily stationary and may vary with phase (amplitude) and /or time. The methodology is applied to visual evoked response data.

Introduction

FIR, OS, and GOS Filters

The finite impulse response (FIR) filter is used for elimination of frequency dependent noise. The design is well understood and will not be discussed here. We would like to describe the filter though because of its similarity to the OS filter. In an FIR filter of length N a data sequence is windowed by selecting N consecutive values. These N values are then multiplied (componentwise) by a sequence of length N of real values. These N multipliers are constant from window to window and it is the selection of the multipliers that allows the designer to retain certain frequency components and delete other frequency components.

The OS filter is similar to the FIR filter except that after windowing the N consecutive values, the windowed values are ordered from smallest to largest (according to algebraic magnitude) before multiplication by the chosen real numbers. Again it is the choice of the real numbers that make the OS filter optimal. These filters have been shown to be optimal (over FIR) filters for additive

Presented at the IEEE Symposium on Circuits and Systems
New Orleans, La.
May 1990

symmetric exponential noise that is nongaussian [1]. The OS filters may also be designed to be editors. For example if the first multiplier is zero we will eliminate the minimum value in each window. Another popular OS filter that will be used for comparison is the α-trimmed mean. The α-trimmed mean trims α% of the values from each window (it trims $\alpha/2$ % of the largest values and $\alpha/2$% of the smaller values) and averages the remaining values. The median is another OS filter we will utilize. The output of the median filter is simply the median value in each window. As an OS filter it would have symmetric exponential noise that is nongaussian [1]. The OS filters may also be designed to be editors. For example if the first multiplier is zero we will eliminate the minimum value in each window. Another popular OS filter that will be used for comparison is the α-trimmed mean. Themultipliers that are all zero except for a central one (note we will only consider odd length filters here). The median is also an α-trimmed mean with $\alpha=2N/(2N+1)$.

The GOS filters [2] are defined as filters with an output that is a function of the ordered (not necessarily algebraically) windowed values. Therefore the filter output could be a function of the ordered window values where the ordering is the distance of the value from the median in the window. The GOS filters we will be using are variants of the least median of squares (LMS) filter [3]. The LMS filter has as output the average of the N+1 closest grouped values in the window. This allows the filter to edit N outliers before averaging. Therefore we note the LMS should behave similarly to the (N/(2N+1))-trimmed mean for symmetrically distributed noise and have better editing properties for nonsymmetrical noise.

Fixed Points and Statistical Properties

Nonlinear filters such as the OS and GOS do not have passbands. Therefore their effect on signals is usually described in one of two ways [4]. We define a fixed point of a nonlinear filter as a sequence that passes through the filter unperturbed (usually finite shifts or time delays of the entire sequence are ignored).The only necessary fact about fixed points for this study is that the OS and the LMS have constant signals as a subset of their fixed points [3]. This is because the periodic signal to be filtered will be subsampled on intervals of one period, yielding sequences that would be constant in the absence of noise. Therefore if the filter is used under the "correct" noise conditions the resulting signal should reduce to a constant upon iteration.

The other common way to describe the effect of nonlinear filters is by specifying their statistical properties. For constant signals the effect of each filter on a window of values is usually called an location estimate. The FIR (linear estimator) properties are well known. The properties of OS and the LMS as estimators of location have been recently developed [1,3]. These will be used to infer the type of i.i.d. additive noise in a signal. For impulsive and bursty noise we will resort to newer techniques as we now outline.

The Breakdown Point

Historically [5,6,7], the breakdown point has been used in the statistical literature as a measure of the robustness of an estimator or a minimization criterion. In the next section we will describe its use as a measure of the robustness of a filter, then use it to compare the robustness of the median, LMS, averager, and α-trimmed mean filters. We will also describe how the breakpoint of the optimal filter will yield an indication of the impulsiveness and/or burstyness of the noise.

Application to Physiological Signals

These filters will be applied to the analysis of visually evoked bioelectric signals. This analysis is complicated due to the nonlinearity of the system, the low signal to noise ratio, and nonstationary, impulsive noise. It was the observation that conventional techniques for noise rejection were not completely successful that lead us to investigate the use of OS and GOS filters.

Periodic Signal Analysis

Assumptions on Signals To be Investigated

In the particular case studied a biological system was driven with a periodic stimulus and samples of the responses recorded. In the absence of any knowledge of the degree of linearity of the system we assume that after a short period of time the output sequence has reached a steady state condition and is periodic with the same period as the input. We do note it is possible that the stimulus period is a multiple of the response period due to the subjects response to every constant reversal. However we assume the period of the response is the same as the stimulus since response to every reversal implies a period the same as that of the stimulus. We note that if the system is linear the assumption is certainly valid [8] and that in other cases it is determined by the degree of linearity of the system.

We also assume there is an additive and/or impulsive noise source that will vary from signal to signal and within the signal as a function of input frequency, amplitude (phase) or time. Therefore the objective is to reduce the output to a periodic signal with the same period as the input.

Conversion of Periodic Signals to Constant Sequences

We assume the output is periodic with the same period as the input, allowing us to assume the interleaved samples one period apart form a constant sequence. Assume there are N sample points in one period of the response, yielding a sequence of points $\{x_i\}$, $1 \le i \le L$, where the number of complete periods of sample data is the largest integer in L/N. We also assume the jitter is constant, i.e. the response is sampled at exactly the same point in the period each time. Then we may assume each of the sequences $\{x_{J+jN}\}$ are constant, where J and j are nonegative integers such that $J+jN \le L$ and $1 \le J \le N$.

For example assume the input is periodic at 10 hz. and the sampling rate is 100 sps. Then in the absence of noise and assuming the system is linear, we may form 10 constant sequences. By selecting some nominal starting point, the first sequence is formed by selecting the first point in each period, the second sequence is formed by selecting the second point in each sequence, etc.

Measures for Comparison of Filtering Effects

In the presence of noise the resulting sequences will not be constant, although in what follows we will refer to them as constant sequences. Therefore we will apply filters to reduce the sequences to approximately constant sequences. One criterion that will guide us is to use the shortest filter possible since obviously any filter with length approximating that of the number of constant values in a constant sequence will yield a constant sequence.

There are two different measures for comparison of the effect of filters on the output. One obvious measure of the effect of a filter on the input is the average of the variance or the standard deviation of each of the "constant" sequences. We call the average of the standard deviation of the constant sequences the "measure of repeatability" of the signal. We note that if one does not average the standard deviations of the constant signals, but rather graphs the standard deviation as a function of the constant signals, the result will indicate the change in the noise as a function of phase and/or input amplitude. One may also graph the variance of consecutive segments of a constant sequence thus yielding an indication of the variability of the noise as a function of time (i.e. the stationarity). The second measure is to examine the power spectral density of the output assuming the output will have a major frequency component as indicated by the period of the input. In this study we will only be using the first measure.

For future study we plan to use various forms of the least median of squares measure [3] to estimate the impulsiveness of a signal both before and after filtering.

FIR, OS, and GOS Filters

For frequency dependent noise added to a constant signal there are well known FIR and IIR linear filter design criterion. Herein we will assume such noise has been eliminated by preprocessing.

If the noise is Gaussian (i.i.d.) it is well known the optimal filter is the averager. It has been shown [1] that for symmetric exponential nongaussian i.i.d. noise there are more optimal OS filters. Therefore if the noise is symmetric exponential i.i.d., a particular OS filter will be most optimal in reduction of the output to a periodic sequence. For nonsymmetric noise distributions, we believe nonsymmetric OS and GOS filters will prove to be optimal, this is a future area of investigation.

One must also take into account the possibility of impulsive and bursty noise. A measure recently introduced [5,6,7] to measure the robustness of a filter or of an location estimator is the breakdown point. Intuitively the breakdown point is the smallest number of contaminants that may cause an estimator to take on an arbitrarily large aberrant value. The optimal value of the breakdown point is usually accepted to be 50% since if more than 50% of the values are perturbed,

the original distribution of the sample could be masked. We note the breakdown point of the mean is 0 and that the median and LMS filters achieve the maximal breakdown point of 50% and will eliminate up to N impulses in a window of length 2N+1. Similarly it can shown that there are OS and GOS (derivatives of the LMS filters) filters with zero weightings for the outliers that have a breakdown point any where from 0 to 50% [3]. It is to be noted the GOS derivatives will be optimal (over the corresponding OS filter) for bursty noise that is not symmetrical within burst.

Hampel [6] used a Monte Carlo simulation to study estimators of location in the presence of outliers. Six estimators were chosen, each of which rejected outliers by some criterion and then averaged the remaining values. He also calculated the breakdown point of each estimator and presented a simple summary and theoretical explanation which explained the results of the Monte Carlo study by the breakdown point. This and other results [7] have led us to expect the breakdown point to be directly correlated with the robustness of an estimator.

In filtering we wish to define a similar concept. We note we cannot use the same concept since in an infinite length sequence we would need only perturb a segment of length 2N+1 to produce an infinite output value, if the filter depends at all on the signal values. This would imply a breakdown of 0% for any filter. Therefore we will define the breakdown point of a filter to be the fraction of aberrant values in a particular window that will cause the filter to take on arbitrarily large values. For the LMS the breakdown point is asymptotically 50%. If the α-trimmed mean is defined so as to (asymptotically) trim a total of α% of the values then it has a breakdown point of (α/2)%.

This is one reason we call the LMS an optimal OS (actually a GOS [2]) filter. It is a smoother based on the rank order values that achieves the optimal breakdown point. To emphasize the significance of this we note that if the α-trimmed mean that averages N+1 values is compared to the LMS, similar smoothing properties are obtained but the LMS is capable of trimming twice as many outliers if the amplitude of the impulsive noise is from a one tailed distribution such as the exponential.

Noise Analysis

It is to be noted that if one can assume the noise source is limited to the class of distributions for which the filter is optimal, then one has also identified the noise in identifying the optimal filter.

Application to VEP Data

The Visual Evoked Potential

The visual evoked potential (VEP) is a voltage potential usually recorded at or just beneath the scalp surface over the visual projection area of the brain. In the laboratory it is normally illicited by presenting the subject with varying spatial

patterns on a monitor. D. Regan [9] has defined a steady state VEP to be "scalp potentials that are locked in time (or in phase) to the sensory stimulus which evokes them." The system is normally assumed to be nonlinear e.g., " the fact that the VEP to sinusoidally modulated light contains higher harmonics indicates directly that the human VEP system is nonlinear" [10]. These statements have lead us to consider steady state VEP data collection as consisting of inputting a periodic signal into a noisy nonlinear system whose parameters are unknown, but with a response one could assume in the noiseless case to be of the same period as the stimulus.

Historical Signal Analysis and Results

Several approaches are generally used to reduce spurious noise in physiological recordings [9]. Differential recording eliminates common mode noise, e.g. 60 cycle power interference. The use of anesthetics and other pharmacological treatments tend to reduce spontaneous neural and muscle activity, although their use introduces the unwanted possibility of inducing non-physiological responses in the nervous system. Finally, signal processing the bioelectric signals themselves can remove random noise from the evoked response. Signal averaging is commonly used by physiologists to remove uncorrelated activity from the records, based on the assumption that the noise is Gaussian distributed about the zero potential level. If, however, the noise is not normally distributed, averaging will not be an optimal noise reduction method, and other filters, such as the OS and GOS may be more appropriate. Because previous studies have indicated the major component of noise in alert patients is impulsive (assumed to be due to muscle artifacts), and the system has demonstrated some degree of nonlinearity (as noted above), we were led to undertake the present study.

Laboratory Setting

Visual function was assessed by recording visual evoked potentials (VEPs) from rhesus monkeys. Recordings were made with gold cup electrodes placed on the animals scalp, roughly over the foveal projection area on each side of the head. The animals were anesthetized with nembutal during recording sessions, and their eyes were aligned with a high-resolution CRT display. VEPs were elicited by square or sine wave lumanance gratings generated on the CRT by a computer-based video frame buffer and counterphased in a square wave temporal pattern on the CRT. The mean luminance of the visual stimulus was 50 cd/m^2. The physiological signals, after being amplified, were digitized at 512 samp/sec and acquired by a Microvax II-based data aquisition system. The amplitude of the VEPs was typically 1-5 μV p-p with a background noise level of approximately 0.5 μV.

Results of VEP Data Analysis

The above technology was used to reproduce the following graphs. The first two graphs are the original VEP and the filtered data (median of length 7). The first two graphs on the next page contain raw data and filtered data taken from

the animal when in a less sedated state. The last two graphs show the results of processing the above data with seven different filters. Except in the case of the median, the labels by each curve indicate the noise distribution for which the OS coefficients used are optimal. We note that the normal (averaging) filter performed best when the subject was well sedated, but the results are less clear when the subject was in a less sedated state. Space limitations preclude presentation of the LMS results. We are currently investigating the effects of sedation on the noise distribution of the VEP.

Acknowledgements

Dr. Longbotham would like to thank Universal Energy Systems and the AFOSR for support in part of the presented research.

References

[1] Bovik, A.C., Huang, T.S., and Munson, D.C., "A generalization of median filtering using linear combinations of order statistics," *IEEE Trans. Acoust., Speech, Sig. Process.*, vol.ASSP-31, pp. 1342-1350, 1983.

[2] Longbotham, H.G., Bovik, A.C., and Restrepo, A.P., "Generalized order statistic filters," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, Scotland, 1989.

[3] Longbotham, H.G., Barsalou, N., "The LMS, an adaptive optimal order statistic filter", *Proceedings of the SPIE/SPSE Symposium on Electronic Imaging*, Santa Clara, California, Febuary, 1990.

[4] Coyle, E.J., Lin, J.H., and Gabbouj, M, "Optimal stack filtering and the estimation and structural approaches to image processing," *IEEE Trans. on Acoust. Speech, and Signal Process.*, vol. ASSP-37,no. 12, pp. 2037-2066.

[5] Donoho D., and Huber, P.J., "The notion of breakdown point", **A Frestschrift for Erich L. Lehman**, P.J. Bickel, K. Doksum, and J.L. Hodges, eds., pp.157-184, Wadsworth, Belmont, Calif., 1983.

[6] Hampel, F.R., "The breakdown points of the mean combined with some rejection rules," *Technometrics*, 27, pp. 95-107, 1985.

[7] Rousseeuw, P.J., "Least median of squares regression," *Journal of the American Statistical Association*, 79, pp. 871-880, 1984.

[8] Oppenheim, A.V., Schafer, R.W., **Discrete-Time Signal Processing**, Prentice Hall Signal Processing Series, 1989, first edition.

[9] Regan, D., **Evoked Potentials in Psychology, Sensory Physiology and Clinical Medicine**, Chapman and Hall, London, England, 1972.

[10] Desmedt, J.E., **Visual Evoked Potentials in Man**, Oxford University Press, Oxford, England, 1977.

FIXED POINT ANALYSIS OF STACK FILTERS

Parimal Patel
Harold Longbotham
Norman Barsalou

Division of Engineering
Nonlinear Signal Processing Group
University of Texas at San Antonio
San Antonio, Texas 78285

ABSTRACT

We extend the current theory of stack filters by determining fixed points for the infinite length signals using positive Boolean functions. Fixed point analysis is basically determining root structure. We describe the fixed point structure for three and four variable positive Boolean functions and describe how one may generate larger positive Boolean functions and simultaneously generate their fixed point structure. We give two explicit formulations for the stack filter that correspond to any positive Boolean function. We then discuss explicit algorithms for generalizing positive Boolean functions that will "stack" on each other. We define a generalized stack filter as the filter that has varying positive Boolean function at various levels. We provide sufficient conditions so that positive Boolean functions in a generalized stack filter can be varied at levels and still uphold threshold decomposition and stacking properties. We also describe a rule which must be followed in the stacking of rank order filters if the stacking property is to be satisfied. Design of stack filters is also presented.

2. INTRODUCTION

Signal detection and analysis problems occur because of the addition of unwanted noise, therefore the signal processing of greatest interest involve ways to filter noise. Some filters, such as linear filters, work well for some additive noise sources but blur sharp edges and are therefore undesirable in many instances. Nonlinear filters, such as the median filter, however, are proven edge preservers and remove unwanted impulses. One type of nonlinear filter is a set of filters known as stack filters. Stack filters may be implemented in real time and defined by threshold decomposition and stacking properties .

Stack filters are nonlinear digital filters that exhibit the property of threshold decomposition and the stacking property [1,2] as shown in Figure 1. Threshold decomposition is a limited superposition property. An M-valued input signal is decomposed into a set of M-1 binary signals. Each binary signal is obtained by thresholding the input signal at the level value k, for $1 \leq k \leq M-1$. The kth level binary signal takes on the value of 1 whenever the input signal is greater than or equal to the level k. Summing up these binary signals always produces the original input signal. The stacking property is an ordering property. After each level is filtered, if each output level is stacked on top of each other from level M-1 to 1, the output is always a

column of 1s with a column of 0s on top. The only desired output of the filter then is the number of the highest level that contains a 1. Positive Boolean functions are Boolean functions that contain only uncomplimentary variables. When identical positive Boolean functions are used as filters and are stacked, input signals obey threshold decomposition and output signals obey the stacking property. Filtering with positive Boolean functions involves no algebraic operations, therefore there are none of the arithmetic errors, quantization errors, etc., that occur with linear filters. Additionally, filtering each binary signal independently allows the operations to be done in parallel. Each filter is also trivial to implement since they are composed of positive Boolean functions only. Figure 1 also shows a filter with the positive Boolean function $X_1X_2+X_2X_3+X_1X_3$ producing the stacked output.

It has been shown that when the Boolean function at each level, B, is a rank order filter, the stack filter, S_B, is in turn the corresponding rank order filter. Fixed point structures for these specific filters have also been discussed [2] when the signal is finite and end effects are eliminated by padding. The importance of this past research can not be underestimated, in that it leads to a specific implementation of rank order filters with constant time delay.

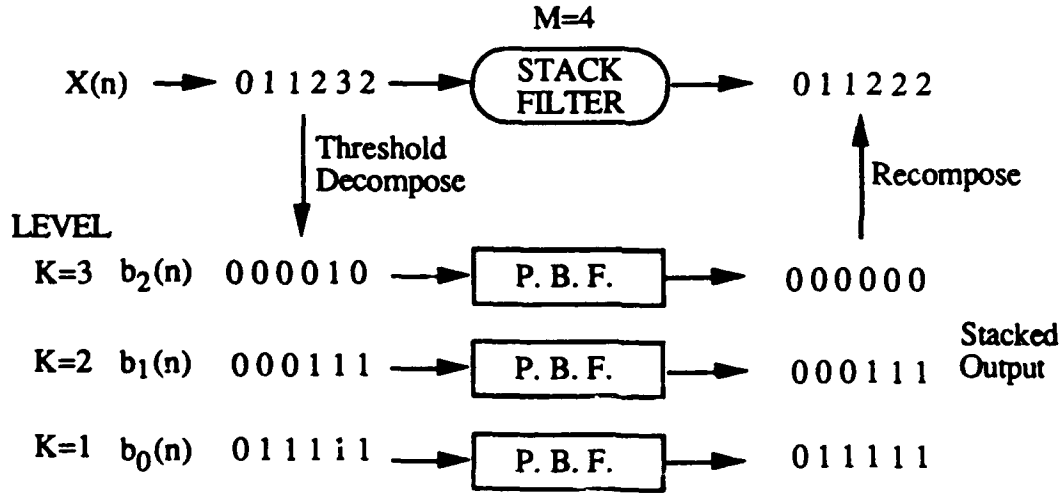


Figure 1: Digital filter illustrating threshold decomposition and stacking property, and positive Boolean functions at each level. The positive Boolean function used in the illustration is $X_1X_2+X_2X_3+X_1X_3$.

Maragos [3] has used mathematical morphology to demonstrate that stack filters are a finite max-min or min-max relation on the original quantized input values. We extend this by demonstrating (without the use of mathematical morphology) a simple one-to-one and onto function that maps the class of positive Boolean functions onto the class of filters, S_B, such that S_B is the MAX (MIN) of a set of MIN (MAX) operations, operating on subsets of the windowed values. This allows us to determine the classical stack filter, S_B, for any given positive Boolean filter B. Furthermore it serves to delineate what filters one may implement with a classical stack filter. We note that the above stated results imply the output of the classical stack filter must be

one of the values in the window. Therefore we cannot hope to implement any filter that assumes values other than those in the window with a single stack filter (examples of filters one cannot implement include FIR and OS (order statistic) filters).

Historically, a fixed point for a nonlinear digital filter is a signal that passes through the filter unchanged, i.e. a signal in the passband of the filter. Fixed point analysis is nothing more than determining fixed point structure. Fixed point for median and rank order filters when the class of signals is restricted to finite length signals with constantly padded endputs, has previously been discussed [4,5]. We [6] have extended these results to include two-sided infinite length signals, showing that the previous results actually eliminate other possible fixed points by requiring the finite length input to be padded on each end prior to filtering.

A generalization may be made by letting the Boolean functions vary from level to level. We provide necessary conditions on the variation of the Boolean filters from level to level in order for stacking to be upheld (threshold decomposition will be upheld as usual). Resulting possible stack filters are then discussed and it is shown that rank order filters may be stacked as long as a higher order occurs under a lower order.

In the next section we give the notations used, define the stack and generalized stack filter, and state and prove the supporting theorems. Section 4 discusses a specific implementation that is being currently carried out. Gate-level digital filter is designed and verified. In Section 5 we conclude by pointing out that work is in progress to design and implement a programmable generalized stack filter.

3. NOTATION, DEFINITIONS, AND THEORY

Notation:

B, B_i : positive Boolean functions

$\{x_i\}$: sequence of values from $\{0,1,...,L-1\}$

$\{\mathbf{x}_i\}$: sequence of windowed vector values of length N starting at i, i.e. $\mathbf{x}_i=(x_i,x_{i+1},...,x_{i+N-1})$

$\{t_{i,j}\}$: L sequences ($1\le j\le L$) of Boolean values obtained by thresholding x_i at level j,

i.e. $t_{i,j}=1 \Leftrightarrow x_i \ge j$

$\{\mathbf{t}_{i,j}\}$: sequence of windowed Boolean vector values on level j of length N starting at i,

i.e. $\mathbf{t}_{i,j}=(t_{i,j},t_{i+1,j},...,t_{i+N-1,j})$

LOMO : a sequence of data points is said to be LOMO-(N) if each consecutive set of N values is monotone.

When considering only one window the subscript i on $\mathbf{x}_i$ and $\mathbf{t}_{i,j}$ will be omitted, and if it does not matter which level we are filtering on, the subscript j will be omitted.

Definition: A stack filter, S_B, is one in which the binary filters B are identical on each level. A generalized stack filter, S, is one in which the binary filters may vary on each level, but still must maintain the stacking property.

The following theorem yields two isomorphisms between the positive Boolean filter, B, used at each level in a stack filter and the stack filter, S_B.

Theorem 1: Let S_B be a stack filter operating on the vector of windowed values, $x=(x_1,x_2,...,x_N)$ and B be the corresponding positive Boolean filter operating on the Boolean variable vector $t=(t_1,t_2,...t_N)$. B is in the form

$$B(t) = \sum_{i=1}^{M}\left(\prod_{k\in J_i} t_k\right) = \prod_{i=1}^{I}\left(\sum_{j\in Q_i} t_j\right)$$

where J_i, Q_i are subsets of $\{1,2,...,N\}$, $\Leftrightarrow$ $S_B = MAX\{MIN\{x_i: i\in J_1\}, MIN\{x_i: i\in J_2\}, ..., MIN\{x_i: i\in J_M\}\} = MIN\{MAX\{x_i: i\in Q_1\}, MAX\{x_i: i\in Q_2\},..., MAX\{x_i: i\in Q_I\}\}$.

Proof: The output of $S_B(x)=r \Leftrightarrow B(t_r)=1$ and $B(t_{r+1})=0 \Leftrightarrow \exists\ J_k$ in the hypothesis above $\ni t_{i,r}=1, \forall\ i\in J_k$ and $\forall\ J_i\ \exists\ t_{j_i} \ni t_{j_i,r+1}=0 \Leftrightarrow \exists\ J_k \ni x_i \geq r\ \forall\ i \in J_k$ and $\forall\ J_i\ \exists\ j_i \in J_i \ni x_{j_i} < r+1 \Leftrightarrow r=MAX\{MIN\{x_i: i\in J_1\}, MIN\{x_i: i\in J_2\}, ..., MIN\{x_i: i\in J_M\}\}$. Similarly, $S_B(x)=r \Leftrightarrow B(t_r)=1$ and $B(t_{r+1})=0 \Leftrightarrow \exists\ Q_i \ni \forall\ j \in Q_i$, $t_{j,r+1}=0$ and $\forall\ Q_i\ \exists\ j \in Q_i \ni t_{j,r}=1 \Leftrightarrow \exists\ Q_i \ni \forall\ j\in Q_i$, $x_j < r+1$ and $\forall\ Q_i\ \exists\ j \in Q_i \ni x_j \geq r \Leftrightarrow r = MIN\{MAX\{x_i: i\in Q_1\}, MAX\{x_i: i\in Q_2\},..., MAX\{x_i: i\in Q_I\}\}$.

Therefore we see that if the positive Boolean filter is given in the sum of products (product of sum) form, the corresponding classical stack filter is simply the maximum (minimum) of a subset of the original windowed data values, where each value in the subset is the minimum (maximum) of a set of windowed values that correspond to one of the sets of Boolean values in one term of the sum of products (product of sum) form. For example, if the window length is 5 and $B = t_1t_2+t_4$, then $S_B=MAX\{MIN\{x_1,x_2\},x_4\}$ and if $B = (t_1+t_4)(t_2+t_5)$, then S_B is $MIN\{MAX\{x_1,x_4\}, MAX\{x_2,x_5\}\}$. In either case the above shows that the output of any classical stack filter will be one of the windowed values. In an FIR or OS filter, with more than one nonzero coefficient, the output is a linear combination of the windowed values. Therefore, a stack filter cannot implement a general FIR or OS filter.

Some research has been done using finite length signals where the first and last points are repeated N times. We will consider infinite length signals, thereby deleting the constant segment introduced by padding. Infinite length fixed points are important for several reasons. In many real life situations [7,8] repeating the end points of a signal unnecessarily emphasizes them, and can have adverse affects in situations of

noisy data. Fixed points for the class of finite length signals (with padding) are a subset of the fixed points for the class of infinite length signals. Therefore in studying infinite length signals, we are including in our study the class of finite length signals.

It is well known [4,9] that when a signal has a monotone segment of length N+1, the only fixed points to a 2N+1 length median filter are LOMO-(N+2). For median filters it has also been shown [9] that if a signal does not have a monotone segment, then any other possible fixed points must be bi-valued. For rank-order filters other than the median it has been shown [5] that when considering only finite length signals (so that a monotone segment is introduced by padding) the only fixed points are constant signals. However, if one considers infinite length signals, rank order filters (including the median) are shown to have a finite number of periodic fixed points [6, 10]. One may expect similar results in the case of stack filters. We show later that if the requirement of a monotone segment of sufficient length is eliminated, periodic solutions are allowed, and if one compares our results to those in [2] the similarity is seen.

First we will consider positive Boolean filters consisting of a single term. We note that since we are considering infinite length signals, a signal that is only shifted (i.e. undergoes a constant time delay) is still considered to be a fixed point. For each of the following results on fixed points, let x_i denote the ith value from the left in the filter window $x = (x_1, x_2, ..., x_N)$, therefore below x_i is a Boolean variable.

Lemma 1: Any positive Boolean function has constant and monotone signals as fixed points.

Proof: It is obvious the output of any positive Boolean function with only zero (one) input is zero (one). If the signal is monotone increasing, the output is zero until the first one is encountered. Assume the Boolean function is in SOP form and the filter window is moving from left to right. The output will remain zero until there are enough ones in the window for one of the terms to be one. This term will remain one, therefore the output of the filter is monotone and equivalent to the input up to a shift. We have similar results for a decreasing signal or movement of the filter in the opposite direction.

Lemma 2: The positive Boolean filter $x_i x_{i+1}$ has no fixed points other than the constant and monotone fixed points.

Proof: If $\{x_i\}$ is not positive or monotone, it has constant sequence of M_1 ones with a zero on either end or a constant sequence of M_2 zeros with a one on either end. In the first case the ones are shortened to a length M_1-1 and in the second case the zeros are lengthened to a length of M_2+1.

Lemma 3: The positive Boolean filter $x_i x_j$ has as a fixed points any periodic signal with period j-i.

Proof: Since a Boolean value ANDed with itself is the same Boolean value, we have the desired result.

Example: The positive Boolean function x_1x_3 has as a fixed point the period 2 signal ...010101.... . The positive Boolean function x_1x_5 has as a fixed points ...000100010001... .

Theorem 2: Let $B= \Pi\, x_{j_i}$, where $1 \le j_i \le N$, and let $d_i = j_i - j_{i-1}$. Then B has as a fixed point of any signal with period p if p is a divisor of each d_i.

Proof: Similar to that of the lemma above.

As one would expect, there is a duality theorem that may be used to transfer the above results for a "multiplicative" term into similar results for an "additive" term. We now state and prove this duality theorem and list without proofs three corollaries which result.

Definition: If $B= (x_1,x_2,...,x_N,+,\bullet,')$, (where $+,\bullet,'$ are the Boolean operations of OR, AND, INVERSION), then its dual is defined as $B_d=(x_1,x_2,...,x_N,\bullet,+,')$.

Theorem 3: $\{x_i\}$ is a fixed point to the Boolean filter $B \Leftrightarrow \{x_i'\}$ is a fixed point to B_d.

Proof: $\forall\ x_i,\ B(x_i)=x_i \Leftrightarrow [B(x_i)]'=x_i' \Leftrightarrow B_d(x_i')]=x_i'$, where $x_i'=(x_1',x_2',...x_N')$.

Corollary 1: The positive Boolean function x_1+x_2 has only fixed points that are constant and monotone.

Corollary 2: The positive Boolean function $B=x_i+x_j$ has as a fixed point, any signal that is periodic with period j-i.

Corollary 3: Let $B = \Sigma\, x_{j_i}$, where $1 \le j_i \le N$, and let $d_i = j_i - j_{i-1}$. Then B has as a fixed point of any signal with period p if p is a divisor of each d_i.

Now we state and prove the following theorem to determine fixed points to a positive Boolean combination of two functions.

Theorem 4: Let B be the Boolean operator "OR" or "AND" and let $B_1\{x_i\} = \{x_{i+j_1}\}$ and $B_2\{x_i\} = \{x_{i+j_2}\}$. Then $B(B_1\{x_i\},B_2\{x_i\}) = B(\{x_{i+j_1}\},\{x_{i+j_2}\})$.

Proof: Obviously the proof is true for the two Boolean operators AND and OR.

We have derived fixed points for three and four variables positive Boolean functions, these are summarized in Table 1. The second through fourth columns are for a positive impulse sequence of length one, two, or three ones embedded in a string of zeros (minimum number of zeros on each side being N/2). Fifth through seventh columns are for a negative impulse sequence of length one, two, or three zeros

embedded in a string of ones (minimum number of ones on each side being N/2). We also have considered periodic functions of period two, three, and four (Columns eight through ten). First column of the table includes positive Boolean functions. These functions are grouped into three and four variables. Functions within the three variables group hold their validity under the four variables group. A "y" in the ith row and jth column indicates that the function *i* has a fixed points for input sequence *j*. "x" indicates that the function i does not have a fixed point for the input sequence j. "1", "2", and "3" in the column of positive and negative impulse indicates that output sequence is shifted by 1, 2, or 3 places left when the smallest variable $x_i = x_1$. In case of periodic input sequences, the sequence indicated implies that the function has a fixed point for that particular periodic input sequence. If the sequence has subscript then the subscript indicates number of places the output shifted left with respect to input. Note that for smallest variable x_i , $i \neq 1$ will result in additional (i-1) left shifts. Also note that for positive Boolean function X_iX_j ($j \neq i+1$), even though the table shows (row 3) fixed points for periodic sequence of period 2, we can generalize for any j for the same function to have a fixed point for j-i periodic sequence.

Boolean Functions	+ impulse of length			- impulse of length				period of		
	1	2	3	1	2	3		2	3	4
				3-variables						
X_i	y	y	y	y	y	y	y	y	y	
X_iX_{i+1}	x	x	x	x	x	x	x	x	x	
X_iX_j ($j \neq i+1$)	x	x	x	x	x	x	y	x	x	
$X_i + X_{i+1}X_{i+2}$	y	x	x	x	x	x	y	001	0001	
$X_{i+1} + X_iX_{i+2}$	1	1	1	x	1	1	x	001_1	$0001,0011_1$	
				4-variables						
$X_i + X_{i+2}X_{i+3}$	y	x	x	x	x	x	y	y	0001	
$X_i + X_{i+1}X_{i+3}$	y	y	x	x	x	x	x	y	0001,0011	
$X_{i+1} + X_iX_{i+3}$	1	1	1	x	x	1	1	x	0001_1	
$X_{i+2} + X_iX_{i+3}$	2	2	2	x	x	2	2	x	0001_2	
$X_iX_{i+1} + X_{i+3}$	3	x	x	x	x	x	3	001_3	0001_3	
$X_iX_{i+2} + X_{i+3}$	3	3	x	x	x	x	x	3	$0001,0011_3$	
$X_{i+1} + X_iX_{i+2}X_{i+3}$	1	1	1	x	x	x	1	001_1	$0001,0011_2$	
$X_{i+2} + X_iX_{i+1}X_{i+3}$	2	2	2	x	x	x	2	001_2	$0001,0011_2$	
$X_{i+1}(X_i + X_{i+2}X_{i+3})$	x	x	1	1	1	1	x	x	0111_1	
$X_i(X_{i+1} + X_{i+2}X_{i+3})$	x	x	x	y	x	x	x	011	0111	
$X_{i+2}(X_i + X_{i+1}X_{i+3})$	x	x	x	2	x	x	x	x	2	
$X_i(X_{i+2} + X_{i+1}X_{i+3})$	x	x	x	y	x	x	y	011	0111	
$X_i(X_{i+3} + X_{i+1}X_{i+2})$	x	x	x	y	x	x	x	y	0111	

Table 1: Fixed point analysis of positive Boolean functions of three and four variables.

In [2] it is stated that one may generalize stack filters by varying the Boolean filter from level to level, as long as they maintain the stacking property relative to each other. The following lemmas are results of this and the corollary generated allows us to understand how one may stack median and rank order filters.

Lemma 4: A generalized stack filter will result from stacking B_{i+1} on B_i if B_i is of the form $B_i = B_{i+1}+B$ for some positive Boolean function B.

Proof: Assume $B_{i+1} = 1$, then $B_i = 1$. Assume $B_i = 0$, then $B_{i+1}=0$. Since B_i and B_{i+1} are each positive Boolean functions , B_{i+1} stacks on B_i.

Lemma 5: Let P_i ,Q_i be subsets of {1,2,...,N}, then B_k stacks on B_{k-1} if

$$B_k = \sum_{i \in I} \prod_{p \in P_i} x_p$$

and

$$B_{k-1} = \sum_{j \in J} \prod_{p \in Q_j} x_p$$

where $\forall$ i,j $\exists$ P_i and Q_j ∋ Q_j is a subset of P_i.

Proof: $B_k = 1 \Leftrightarrow \exists P_i \ni x_p=1, \forall p \in P_i \Leftrightarrow \exists Q_j \ni x_p=1 \ \forall j \in Q_j \Leftrightarrow B_{k-1} =1$. $B_{k-1} = 0 \Leftrightarrow \forall j \in J \ \exists \ p \in Q_j \ \ni x_p = 0 \Leftrightarrow \forall i \in I \exists p \in P_i \ni x_p = 0 \Leftrightarrow B_{k-1} = 0$.

Example: If x_i represents Boolean variables we have, x_1 stacks on x_1+x_2, x_1x_2 stacks on x_1+x_3 (since $x_1x_2 + x_1+ x_3 = x_1 + x_3$), and $x_1+x_2x_3$ stacks on x_1+x_3 (since $x_1+x_3 = x_1 + x_2x_3 + x_3$).

Corollary 4: We may stack a rank order filter on another rank order filter if the lower filter is of higher order and obtain a generalized stack filter.

Proof: The i^{th} rank order filter of window size N is the sum of all terms with exactly i variables (all in uncomplimentary form of course). Therefore we may use lemma 5.

Example. We may stack the MIN filter on the Median filter, and the Median filter on the MAX filter and the result will be a generalized stack filter.

4. IMPLEMENTATION

A digital filter composed of positive Boolean stack filters has been designed and implemented. This filter has a 3- bit input, a 4-bit window, and a 3-bit output. As the input is 3-bit wide seven threshold levels are required. and can be broken down into the units shown in Figure 2. A modular design approach has been implemented. In order to realize the algorithm implemented by the stack filter composed of positive

Boolean filters, the system is subdivided into four major units. These four units are the Quantize Threshold Decomposition Unit, Windowing Unit, Positive Boolean Filter Unit, and the Threshold Recomposition Unit.

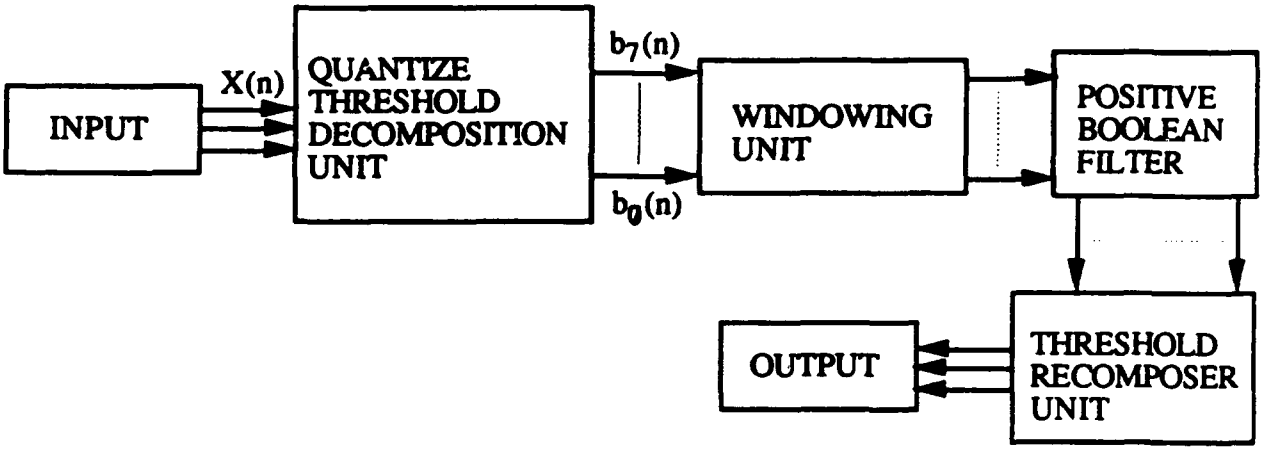


Figure 2: A block diagram of the digital filter.

The Decomposition Unit complies the properties of threshold decomposition and stacking which allow for an efficient VLSI implementation. The output is stacked upward from B_1 to B_7. A 7-bit output from the Threshold Decomposition Unit is feed to the Windowing Unit which provides for a 4-bit window at each level of the threshold decomposition unit output.

Each window level is composed of four master-slave D Flip-Flops which provide a valid 4-bit parallel output to a single level positive Boolean filter in the Filtering Unit on each rising edge of the clock. During this same clock pulse, a shift right operation was implemented on the falling edge of the clock. Consequently, the first three shift right operations of any new input sequence provide invalid data. Upon the fourth clock pulse, or shift right operation, a valid 4-bit parallel output would be provided to the Filter Unit. As the window width is 4-bit, fifteen different minterm combinations of the window variables are possible:

X_1	X_2	X_3	X_4
X_1X_2	X_1X_3	X_1X_4	X_2X_3
X_2X_4	X_3X_4	$X_1X_2X_3$	$X_1X_2X_4$
$X_1X_3X_4$	$X_2X_3X_4$	$X_1X_2X_3X_4$	

These 15 combinations are available to select the desired positive Boolean function filter. Many combinations of the minterms can be simplified, such as $X_1X_2+X_1X_2X_3 = X_1X_2$, therefore it is necessary to provide control logic to select only minimum sum of product expression filters. So, as shown in Table 2, if one chooses minterm X_1X_2, there are three filter terms that, if selected, would be redundant to X_1X_2. Table 2 shows all such combinations of "Chosen" filter terms and the terms that are then "Redundant."

Boolean filters, the system is subdivided into four major units. These four units are the Quantize Threshold Decomposition Unit, Windowing Unit, Positive Boolean Filter Unit, and the Threshold Recomposition Unit.

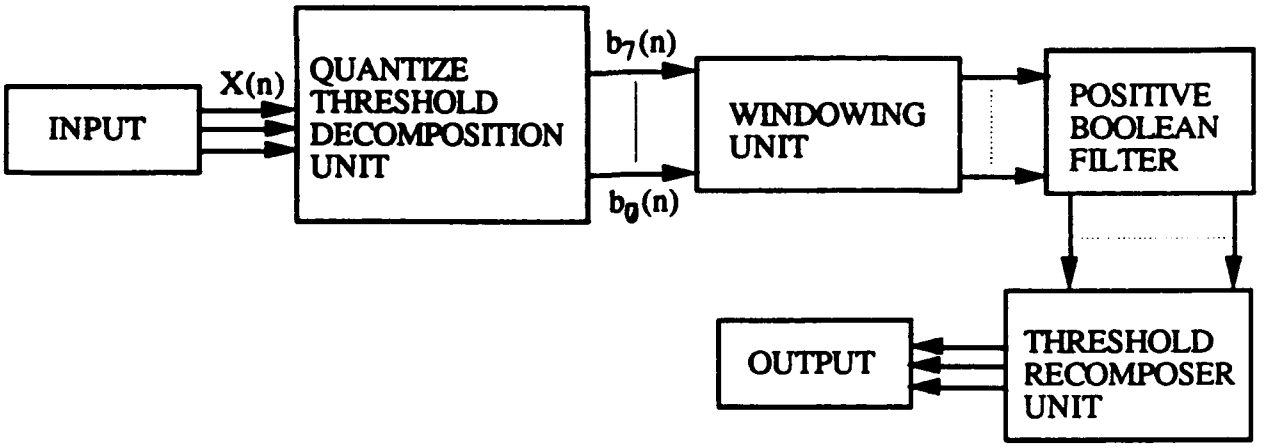


Figure 2: A block diagram of the digital filter.

The Decomposition Unit complies the properties of threshold decomposition and stacking which allow for an efficient VLSI implementation. The output is stacked upward from B_1 to B_7. A 7-bit output from the Threshold Decomposition Unit is feed to the Windowing Unit which provides for a 4-bit window at each level of the threshold decomposition unit output.

Each window level is composed of four master-slave D Flip-Flops which provide a valid 4-bit parallel output to a single level positive Boolean filter in the Filtering Unit on each rising edge of the clock. During this same clock pulse, a shift right operation was implemented on the falling edge of the clock. Consequently, the first three shift right operations of any new input sequence provide invalid data. Upon the fourth clock pulse, or shift right operation, a valid 4-bit parallel output would be provided to the Filter Unit. As the window width is 4-bit, fifteen different minterm combinations of the window variables are possible:

X_1	X_2	X_3	X_4
X_1X_2	X_1X_3	X_1X_4	X_2X_3
X_2X_4	X_3X_4	$X_1X_2X_3$	$X_1X_2X_4$
$X_1X_3X_4$	$X_2X_3X_4$	$X_1X_2X_3X_4$	

These 15 combinations are available to select the desired positive Boolean function filter. Many combinations of the minterms can be simplified, such as $X_1X_2+X_1X_2X_3 = X_1X_2$, therefore it is necessary to provide control logic to select only minimum sum of product expression filters. So, as shown in Table 2, if one chooses minterm X_1X_2, there are three filter terms that, if selected, would be redundant to X_1X_2. Table 2 shows all such combinations of "Chosen" filter terms and the terms that are then "Redundant."

Choose	Redundant
X_1	X_1X_2, X_1X_3, X_1X_4, $X_1X_2X_3$, $X_1X_2X_4$, $X_1X_3X_4$, $X_1X_2X_3X_4$
X_2	X_1X_2, X_2X_3, X_2X_4, $X_1X_2X_3$, $X_1X_2X_4$, $X_2X_3X_4$, $X_1X_2X_3X_4$
X_3	X_1X_3, X_2X_3, X_3X_4, $X_1X_2X_3$, $X_1X_3X_4$, $X_2X_3X_4$, $X_1X_2X_3X_4$
X_4	X_1X_4, X_2X_4, X_3X_4, $X_1X_2X_4$, $X_2X_3X_4$, $X_1X_3X_4$, $X_1X_2X_3X_4$
X_1X_2	$X_1X_2X_3$, $X_1X_2X_4$, $X_1X_2X_3X_4$
X_1X_3	$X_1X_2X_3$, $X_1X_3X_4$, $X_1X_2X_3X_4$
X_1X_4	$X_1X_2X_4$, $X_1X_3X_4$, $X_1X_2X_3X_4$
X_2X_3	$X_1X_2X_3$, $X_2X_3X_4$, $X_1X_2X_3X_4$
X_2X_4	$X_1X_2X_4$, $X_2X_3X_4$, $X_1X_2X_3X_4$
X_3X_4	$X_1X_3X_4$, $X_2X_3X_4$, $X_1X_2X_3X_4$
$X_1X_2X_3$	$X_1X_2X_3X_4$
$X_1X_2X_4$	$X_1X_2X_3X_4$
$X_2X_3X_4$	$X_1X_2X_3X_4$

Table 2: Essential and redundant minterms.

The recomposer is basically a unit that has 7 input lines from each filter level and produces the 3-bit solution for the current windowed data values. The output expressions can be determined from the truth table and are:

$$O_2 = L_4$$
$$O_1 = L_1L_3 + L_5$$
$$O_0 = L_6 + L_4L_5' + L_2L_3' + L_0L_1'$$

The gate level design of the stack filter composed of positive Boolean filters was done on SUN workstations using DAISY software. The design verification is carried out in DAISY CAE/CAD environment. Because of the large number of gates needed to realize the design, the nesting capability of DAISY was used in order to simulate and verify the valid operation of the stack filter. The input test data set shown in Table 3 is used in order to verify that design. The result is shown in Figure 3.

Filter Selected	Data Input	OUTPUT
$X_1X_2 + X_1X_4 + X_2X_3$	1 2 3 4 5 4 3 3 3 4 5 4 3 2 1	X X X 2 3 4 4 4 3 3 3 4 4 4 3
$X_1X_3 + X_2$	0 2 2 0 1 1 3 2 2 2 0	X X X 2 2 1 1 1 3 2 2
$X_1X_2 + X_1X_3 + X_2X_3$	1 1 0 2 3 3 1 2 2	X X X 1 1 2 3 3 2

Table 3: Test data to verify design.

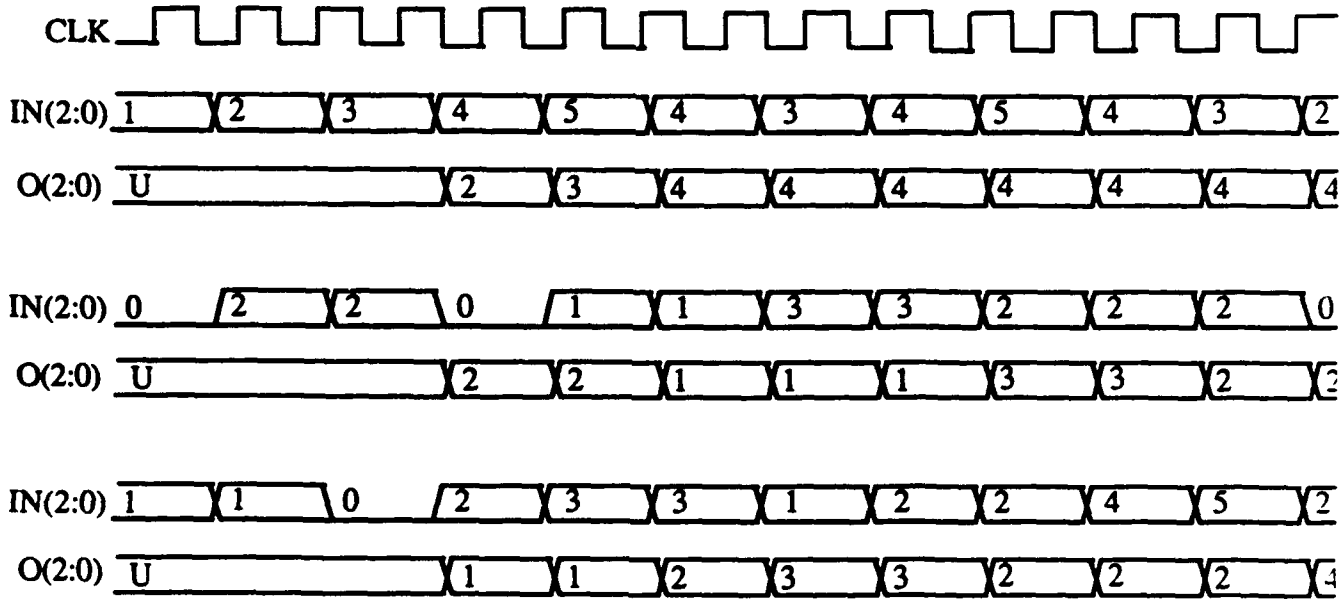


Figure 3: Simulation result of test data.

5. DISCUSSION AND CONCLUSION

Two explicit correspondences between stack filter operators and the corresponding Boolean operators have been demonstrated. The infinite length fixed points of some stack filters have been explored. A method of combining filters and generating fixed point structures simultaneously has been demonstrated. We also have shown explicitly how generalized stack filters may be generated and applied our result to show how one may stack rank order filters. To define the passband of stack filters and generalized stack filters, we must find all fixed points and show that signals that are not fixed points "approach" fixed points upon repetitive application of the filter. One does not really expect all input signal/filter combinations to approach a limit, but just as in the analog case we must define criterion for stability, convergence, and oscillations.

Gate-level design of hard-coded stack filter has been presented. The stack filter employs three variable positive Boolean functions and window size of 4. The input data is represented by 3-bit. As the precision increases (number of bit that are needed to represent data), the number of threshold levels increases in proportion to 2^n. This may create problem with higher precision. With higher precision, an additional processing in space domain is required before threshold decomposition and Boolean filtering is carried out. Work is in progress to design and implement programmable stack filter that will allow us to vary Boolean functions at various levels.

6. ACKNOWLEDGEMENT

This research is supported in part by NSF grant No. USE-8853283 and the Air Force Office of Scientific Research/AFSC, United States Air Force, under Contracts AFSOR-89-0490 and F 49620-87-R-0004. The Unites States government is authorized to reproduce reprints for government purposes, not withstanding any copyright notion hereon.

7. BIBLIOGRAPHY

[1] J. P. Fitch, E.J. Coyle, and N.C. Gallagher, "Median filtering by threshold decomposition" *IEEE Trans. on Acous. Speech and Signal Processing*, vol. ASSP-32, Dec. 84, pp.1183-1188.

[2] P.D. Wendt, E.J Coyle, and N.C. Gallagher, "Stack filters" *IEEE Trans. on Acous. Speech and Signal Processing*, vol. ASSP-34, Aug. 86, PP.898-911.

[3] P.T. Maragos, and R.W. Schafer, "Morphological filters-part II: Their relations to median, order statistic, and stack filters", *IEEE Trans. on Acous. Speech and Signal Processing*, vol. ASSP-35, Aug. 87, pp.1170-1184.

[4] N.C. Gallagher and G.L. Wise, "A theoretical analysis of the properties of median filters," *IEEE Trans. on Acous. Speech and Signal Processing*, vol. ASSP-30, Dec. 81, pp.1136-1141

[5] T.A. Nodes and N.C Gallagher,"Median filters: Some modifications and their properties,"*IEEE Trans. on Acous. Speech and Signal Processing*, ASSP-29, Oct.. 82, pp.739-746.

[6] H.G. Longbotham and A.C Bovik, "Theory of order statistic filters and their relationship to linear FIR filters," *IEEE Trans. on Acous. Speech and Signal Processing*,ASSP-37, Feb. 89, pp.275-287.

[7] Nieminen, A., Neuvo, Y., and Mitra,U., "Algorithms for real time trend detection," 1988 International Conference On Acoustics Speech And Signal Processing, vol. D, pp. 1150-1153.

[8] Longbotham, H.G., and Roberts, J. "Application of nonlinear filters to VEP data," 1988 Summer Faculty Research Program Technical Report, UES/AFOSR, Washington D.C., pp. 141-1 through 141-20.

[9] Tyan, S.G., "Median filtering: Deterministic properties," in Two Dimensional Signal Processing: Transforms and Median Filters, T.S. Huang, Ed., New York: Springer Verlag, 1981.

[10] Eberly, D., Longbotham, H. G., and Aragon, J., "Complete Classification of Roots to a 1-Dimensional Median and Rank-Order Filers," submitted to IEEE ASSP, Spring 1989.

Report # 98
760-0MG-101
Prof.James Mrotek
Report Not Publishable

FINAL REPORT

PROJECT NUMBER: S-760-7MG-091

PROJECT TITLE: ADENOSINE MODULATION OF NEUROTRANSMITTER RELEASE FROM HIPPOCAMPAL MOSSY FIBER SYNAPTOSOMES

PRINCIPLE INVESTIGATOR: Ralph I. Peters, Ph.D.
Department of Biological Sciences
The Wichita State University
Wichita, KS 67208

Specific Aim #1 - Validation of The Synaptosomal Preparation

Question:

Although this preparation had been shown to possess A1 adenosine receptors during my summer in the SFRP program, prior to this work it was not known if the preparation contained endogenous adenosine or if endogenous adenosine normally modulates release.

Results:

Several procedures were employed to test the hypothesis that this preparation is capable of producing adenosine, and that the adenosine produced normally inhibits the release of dynorphin B. Blockade of A1 receptors with 8-phenyltheophylline was found to reduce the spontaneous release of dynorphin B but not affect the release evoked by depolarizing concentrations of potassium ion (fig. 1). Blockade of adenosine uptake mechanisms with dipyridamole was found to have no effect on the spontaneous release of dynorphin B, but significantly reduced the evoked release (fig. 2). Augmenting adenosine degradation by the addition of adenosine deaminase to the superfusion buffers substantially elevated the spontaneous release from this preparation, and slightly elevated the evoked release (fig. 3). Taken together, these data strongly support the hypothesis that this preparation normally produces adenosine, and that this adenosine serves to reduce the release of dynorphin B.

Specific Aim #2 - Source of Adenosine

Question:

Since it seemed clear that the hippocampal mossy fiber synaptosomal preparation produced adenosine which reduced the release of dynorphin B, attention was directed towards the source of adenosine in the synaptosomes. The most likely source of adenosine seemed to be ATP, so attention was focused on the possibility that the preparation releases ATP and degrades the released ATP to adenosine.

Results:

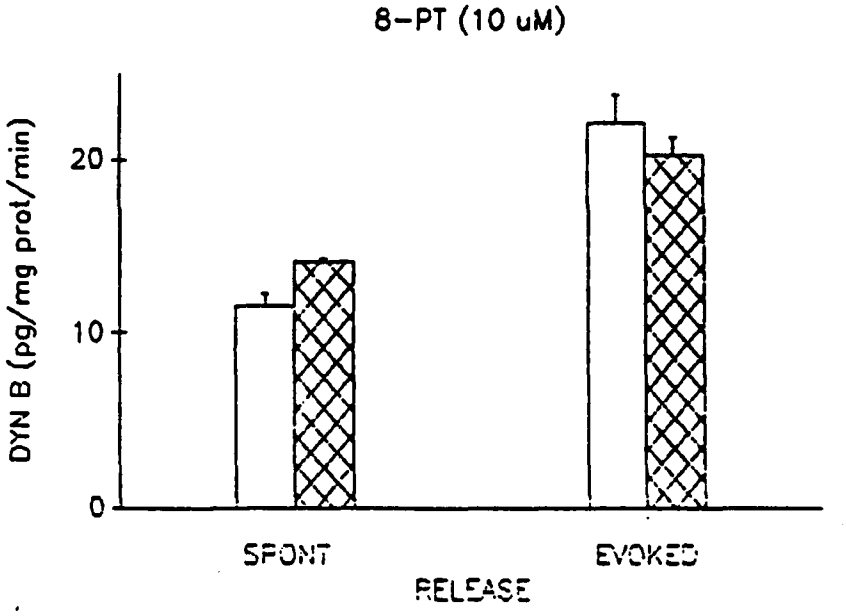
Direct evidence for the release of ATP from these synaptosomes was obtained by Dr. David Terrian's laboratory at the USAF School of Aerospace Medicine. They found that this synaptosomal preparation releases ATP in a depolarization dependent fashion, and that this release is calcium dependent (fig. 4). It is known that ATP may be enzymatically converted to adenosine by the actions of ATPase followed by 5'-nucleotidase. The nucleotidase may be inhibited by alpha, beta-methylene ADP, and this treatment would be expected to augment the release of dynorphin B. However, treatment with 10 micromolar concentrations of this compound had no effect on the spontaneous release of dynorphin B and actually reduced the evoked release (fig. 5). The unavailability of antiserum prevented further study of this phenomenon.

PRODUCT

Although none of the data from this project are involved, some of the ideas and preliminary results are contained in a manuscript submitted to The Journal of Neurochemistry. It is entitled "Adenosine modulation of the hippocampal mossy fiber synapse: ATP release, adenosine formation and inhibition of dynorphin and glutamic acid release" and authored by D.M. Terrian, P.G. Hernandez, M.A. Rea, and R.I. Peters.

The bulk of this work forms the basis for a Master of Science thesis written by Jonathan J. Lowen, a student in my laboratory. The thesis has been completed, and is scheduled for defense on 6 April, 1989. The work is publishable, and manuscript preparation will begin shortly. USAF-UES support will be acknowledged.

Figure 1


8-PT (10 uM)
DYN B (pg/mg prot/min)
20
10
0
SPONT
EVOKED
RELEASE

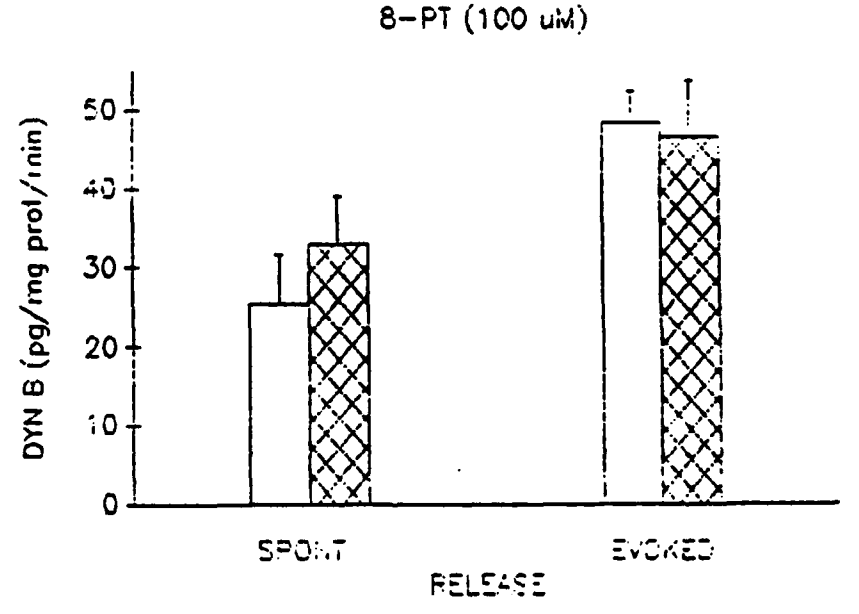

8-PT (100 uM)
DYN B (pg/mg prot/min)
50
40
30
20
10
0
SPONT
EVOKED
RELEASE

Figure 3

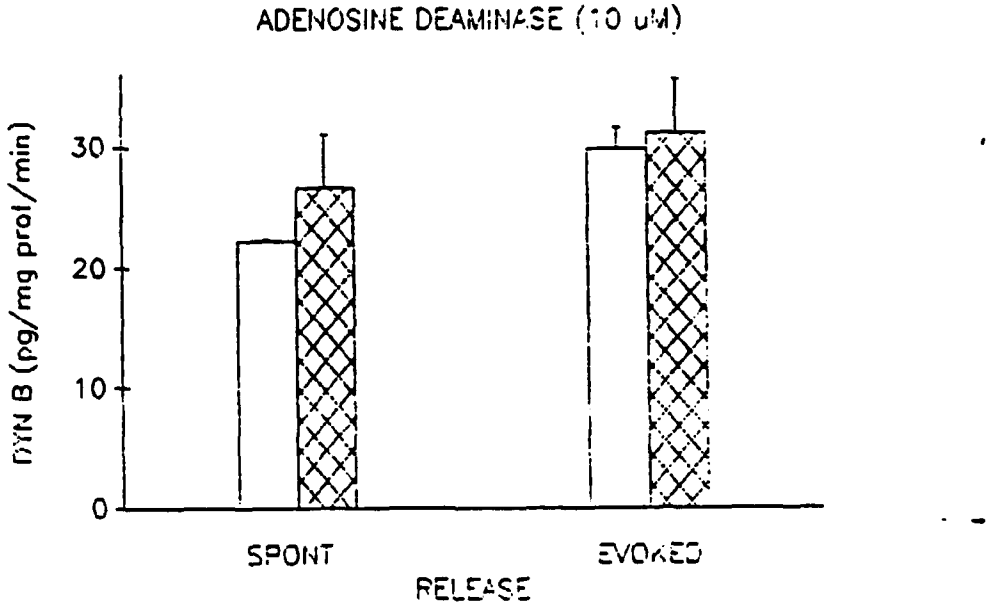


Figure 2

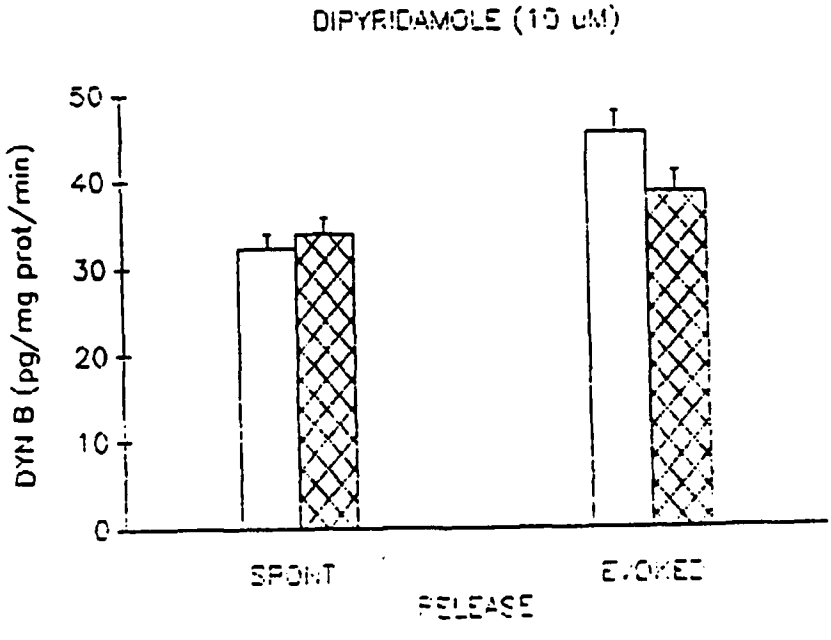
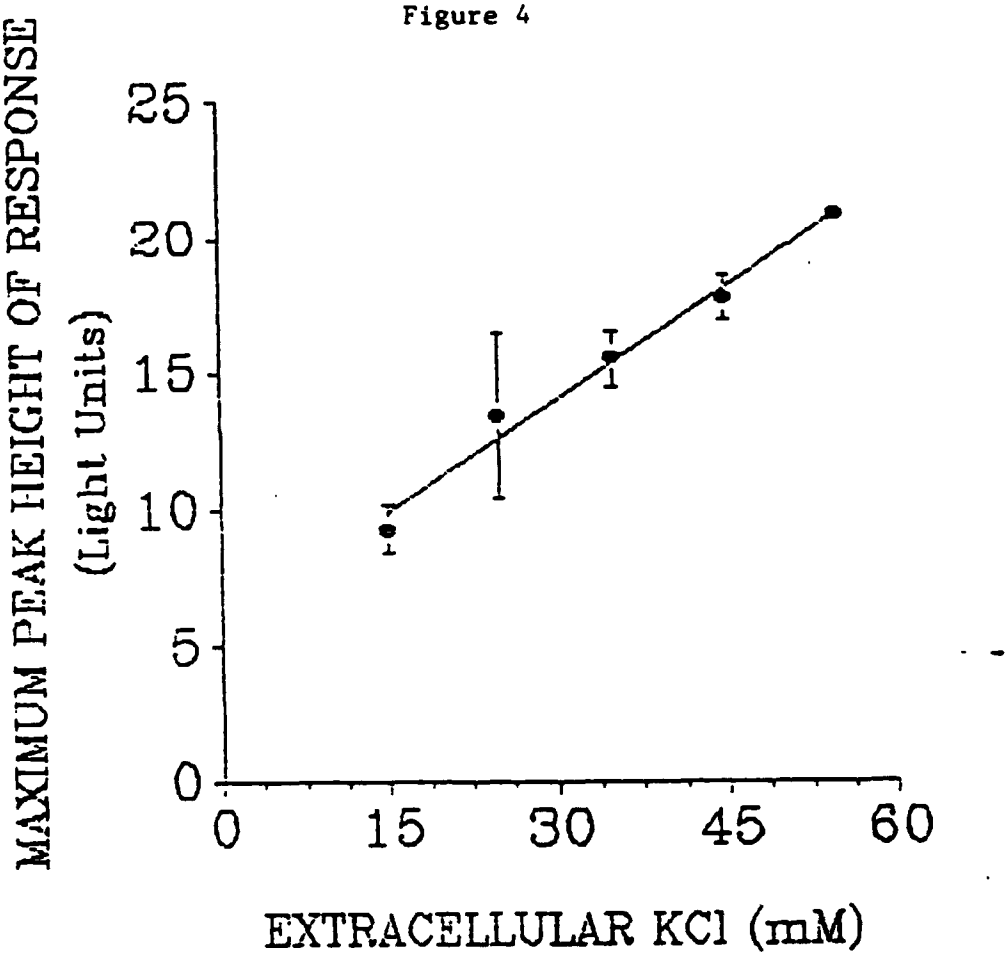


Figure 4


MAXIMUM PEAK HEIGHT OF RESPONSE
(Light Units)
25
20
15
10
5
0
0
15
30
45
60
EXTRACELLULAR KCl (mM)

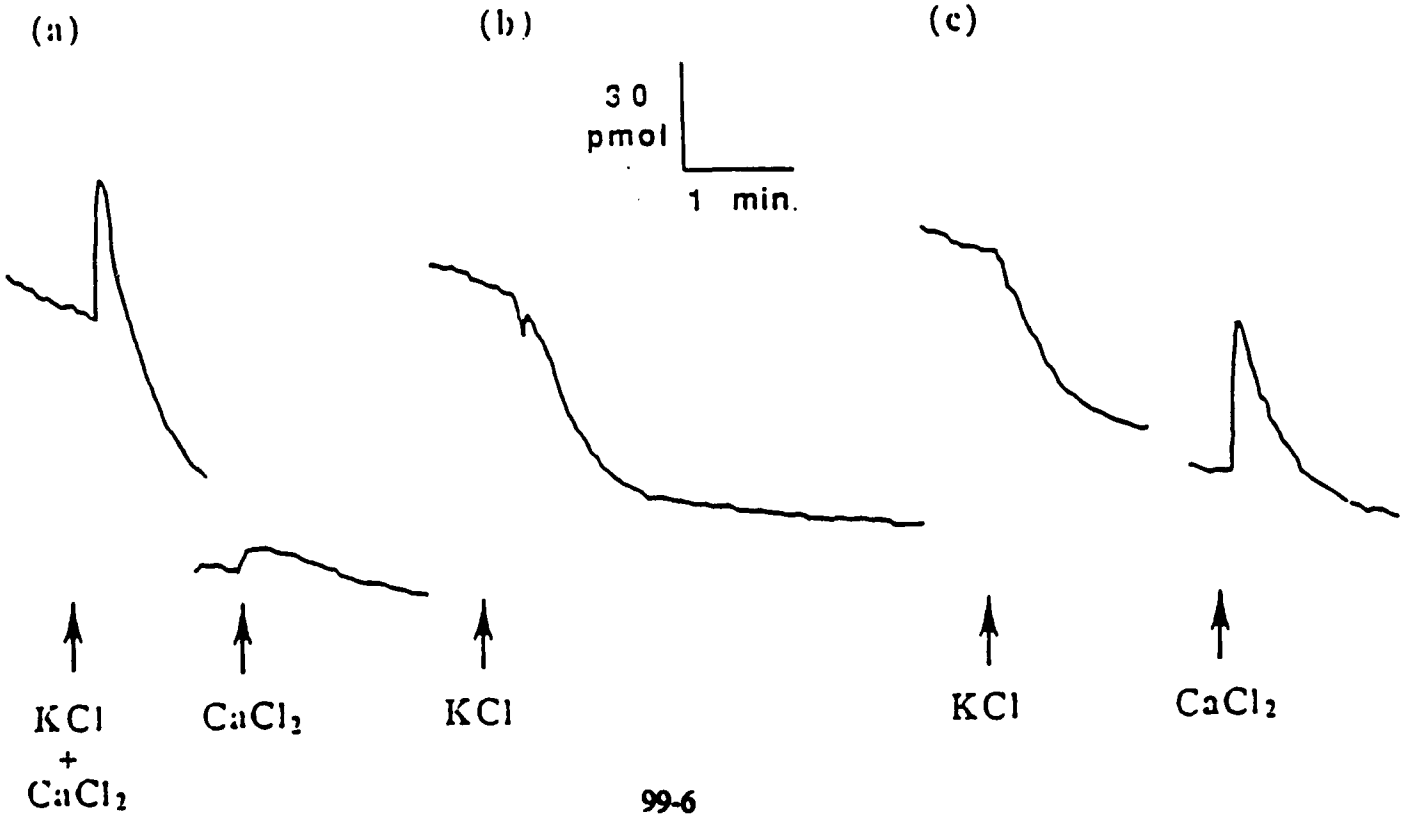
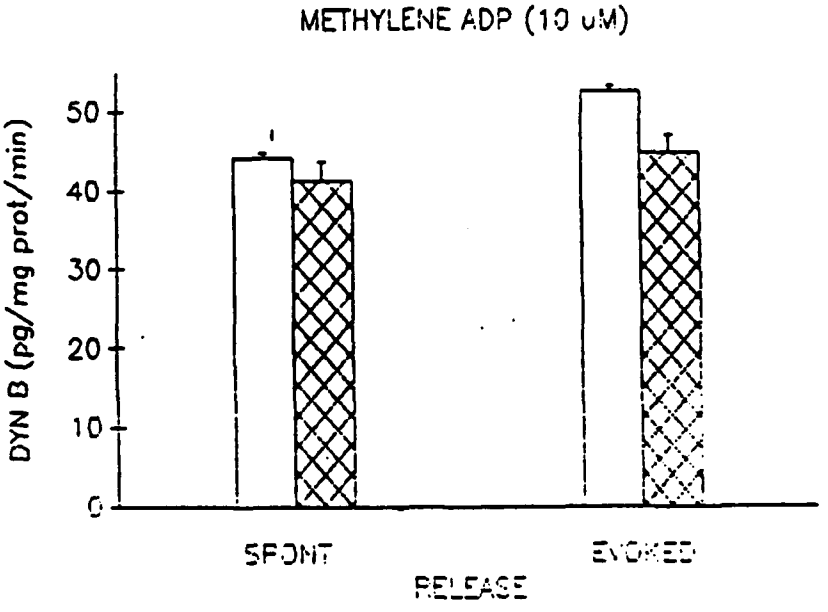

(a)
(b)
(c)
30
pmol
1 min.
KCl
+
CaCl2
CaCl2
KCl
KCl
CaCl2

Figure 5


METHYLENE ADP (10 uM)
DYN B (pg/mg prot/min)
50
40
30
20
10
0
SPONT
EVOKED
RELEASE

UNIVERSAL ENERGY SYSTEMS, INC.

1990 RESEARCH INITIATION PROGRAM

FINAL REPORT

Prepared by:	Raymond M. Quock, Ph.D.
Academic Rank:	Associate Professor of Pharmacology
Department:	Department of Biomedical Sciences
and University:	University of Illinois College of Medicine at Rockford
Research Location:	University of Illinois College of Medicine at Rockford
Date:	November 30, 1990
Contract No:	UES Project 210, S-210-10MG-035

Behavioral and Neurochemical Effects of Radiofrequency Electromagnetic Radiation

by

Raymond M. Quock

Abstract

Whole animal behavioral and neurochemical experiments were conducted to ascertain the influence of exposure to radiofrequency electromagnetic radiation (RFR) on opioid and benzodiazepine mechanisms in the central nervous system. In this research, we determined that exposure of male Swiss Webster mice to increasing specific absorption rates of RFR resulted in a thermal stress that activated endogenous opioid systems (as manifested by antinociception in the abdominal constriction test). Less conclusive effects of RFR were observed in the mouse tail flick test and staircase test. The observed antinociception in RFR-irradiated animals was sensitive to antagonism by relatively low doses of the opioid receptor blocker naltrexone, suggesting an involvement of μ-opioid receptors in mediation of the antinociceptive response to RFR. Neurochemical experiments still in progress indicate that exposure to RFR may alter binding of radioligands to opioid receptors in portions of RFR-exposed vs. sham-irradiated mice. These findings indicate that RFR does appear to cause an activation of endogenous opioid mechanisms, resulting in measurable neurochemical changes and overt behavioral responses. Based on the results of this investigation, recommendations are provided regarding future experiments and methodological concerns in continuation of this research.

Acknowledgements

I wish to acknowledge the excellent technical assistance of research specialists Janet Mueller and Elizabeth Ellenberger and Walter Rice Craig summer research fellow Roger H. Maillefer, University of Illinois College of Medicine at Rockford. I also wish to thank Dr. David G. Lange, Department of Anesthesiology and Critical Care Medicine, The Johns Hopkins University School of Medicine, and Dr. James Lin, Department of Bioengineering, University of Illinois at Chicago, for their generous assistance in this research. The RFR experiments conducted at exposure to 2450- and 185-MHz could not have been possible without Dr. Lange and Dr. Lin's loan of key components of the exposure systems and their continuing consultation. The collaboration of Dr. Linda K. Vaughn, Marquette University School of Dentistry, in the radioligand binding experiments was also critical to this project, and her experience, active participation and advice are gratefully appreciated. The counsel of Dr. Liang-fu Tseng, Medical College of Wisconsin, in setting up the radioimmunoassay procedures is also acknowledged.

I. INTRODUCTION

Increasing development, deployment and utilization of microwave-emitting radar and communications equipment has been accompanied by a heightened awareness and concern in possible influences of electromagnetic radiofrequency radiation (RFR). There is intensive ongoing research of potential health hazards of RFR to military personnel who operate this equipment and of how operation of RFR might impact upon performance of duties by military personnel. The ultimate objective of such investigation is to characterize the bioeffects of RFR and to elucidate the underlying mechanisms of the physiological responses to RFR exposure.

RFR is known to elevate tissue temperature and is suspected of inducing thermal stress. Stress per se represents the disruption of normal physiological equilibrium (homeostasis) by some external stimulus. Stress will provoke general adaptive changes that attempt to maintain physiological integrity and augment the resistance to stress (Selye, 1946). Based on research of non-RFR stressors, these bodily reactions are known to include activation of the sympathetic nervous system, the hypothalamic-pituitary-adrenal axis and endogenous opioid systems.

Michaelson suggested that high level RFR can act as a stressor and disrupt homeostasis (Michaelson et al. 1964). In a subsequent study, Michaelson and coworkers demonstrated that RFR exposure can increase both plasma corticosterone and colonic temperature in correlated fashion (Lotz and Michaelson, 1978). Other neuroendocrine patterns of reaction to RFR-induced stress include reduction in both thyroid and growth hormones (Lu et al., 1980).

There is also evidence that RFR exposure can influence activity of endogenous opioid systems. Lai and associates have demonstrated that acute exposure of rats to a pulsed RFR field (500 pps, 2-μsec) at 2450 MHz can induce a post-exposure hyperthermia that is sensitive to antagonism by opioid receptor blockade (Lai et al., 1984). In addition, they noted that this post-RFR hyperthermia was also blocked by serotonergic receptor antagonists and prostaglandin synthesis inhibitors (Lai et al., 1984, 1987a). This was remarkably reminiscent of the hyperthermic effect of the opioid peptide β-endorphin administered into the rat hypothalamus; this effect was also sensitive to antagonism by opioid receptor blockers, serotonergic receptor antagonists and prostaglandin synthesis inhibitors (Martin and Bacino, 1979). Based on these findings, Lai speculates that the hyperthermia that ensues pulsed RFR exposure is due to activation of endogenous opioid systems, probably involving β-endorphin, which then activates opioid receptors, possibly located in the thermoregulatory preoptic anterior hypothalamus, which, in turn, activate a serotonergic pathway resulting in hyperthermia.

In other research supporting the hypothesis of RFR activation of endogenous opioid systems, rats exposed to RFR exhibited enhanced catalepsy to morphine (Lai et al., 1983) and hyperthermia to amphetamine (Lai et al., 1986a); these potentiating effects of RFR were also reversed by naloxone. Narcotic antagonist drugs also prevented RFR-induced reduction in high-affinity choline uptake in the hippocampus of the rat (Lai et al., 1987b). Moreover, the naloxone precipitated abstinence syndrome in morphine dependent rats can be suppressed by RFR (Lai et al., 1986b).

Another consequence of stress-induced activation of endogenous opioid systems is antinociception (Amir et al., 1980; Bodnar et al., 1980; Chance, 1980). This phenomenon of stress-induced analgesia is important because it provides insight into the psychological and physiological factors that activate endogenous pain control and opiate systems. It is known that one physiological reaction of the body to stress is release of endogenous opioid peptides, resulting in suppression of pain via stimulation of opioid receptors in the medial brainstem (Madden et al., 1977). More recent research, though, has yielded sufficiently inconsistent findings to suggest that there may be multiple forms of stress-induced analgesia, some opioid in nature and others non-opioid in nature (Bodnar et al., 1980; Chance, 1980; Mayer and Liebeskind, 1974; Mayer and Price, 1976; Akil et al., 1976; Lewis et al., 1980). It is apparent that the nature of stress-induced analgesia is dependent upon the type of stressor as well as its duration and intensity (Grau et al., 1981).

The intention of research activities in our laboratory were to ascertain the interaction of RFR with endogenous opioid and benzodiazepine mechanisms. The plan was to conduct whole animal studies and correlate changes seen with neurochemical experiments. The plan was to determine whether RFR might induce sufficient thermal stress to activate endogenous opioid mechanisms to produce antinociception and to stimulate endogenous anxiomodulatory mechanisms to produce anxiety.

II. OBJECTIVES

As originally proposed, the objectives of this research were several fold: 1) identification of RFR exposure conditions that will evoke overt

behavioral responses (antinociception and anxiety); 2) determination of the sensitivity of such RFR-induced effects to antagonism by specific receptor blockers (opioid and benzodiazepine); and 3) identification of possible neurochemical correlates (radioligand binding to receptors and radioimmunoassay for endogenous neuropeptides) of such RFR-induced effects.

III. REPORT

A. RFR Exposures: RFR exposures were conducted at two different frequencies, 2450 MHz and 148 MHz. As a part of this research project, these exposure systems were borrowed from Dr. David G. Lange (Johns Hopkins University) and Dr. James C. Lin (University of Illinois at Chicago), respectively, with certain components (signal generators, power supplies and other accessories) purchased to complete the systems. Block diagrams of both RFR exposure systems are shown in Figs. 1 (2450 MHz system) and 2 (148 MHz system).

In one exposure system, a 2450-MHz Tappan microwave oven served as a source, and mice were exposed in a waveguide system attached by a coaxial cable to the microwave generator (Quock et al., 1986, 1987). The peak power density in the chamber was checked with a microwave leakage detector and adjusted with a double-stub tuner and line attenuator. Mice were housed in a styrofoam exposure chamber (9.0 cm L x 11.0 cm W x 6.5 cm H) in the waveguide. Since the power density in the exposure chamber was not uniform, a stiff plastic mesh was set within the chamber (9.0 cm L x 4.5 cm W x 6.5 H) and used to reduce animal movement and to make the power density to which each mouse was exposed more consistent.

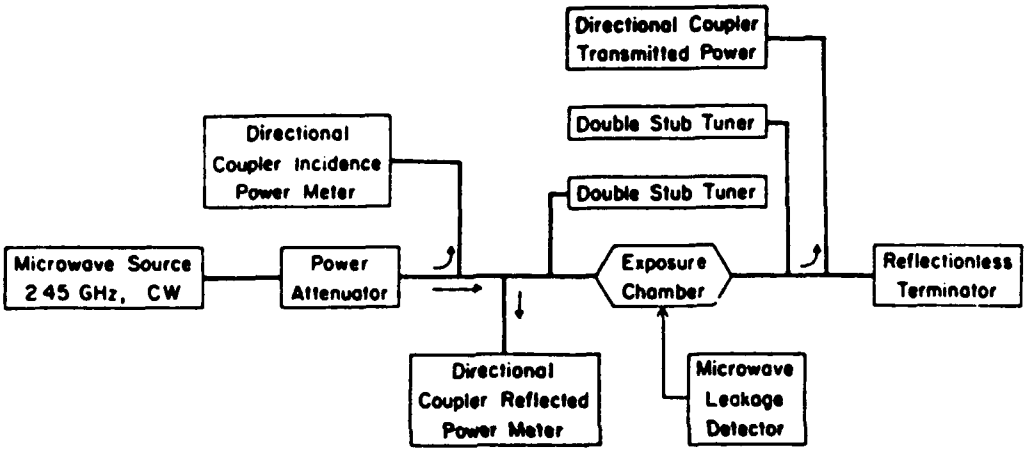

Directional Coupler Transmitted Power
Directional Coupler Incidence Power Meter
Double Stub Tuner
Double Stub Tuner
Microwave Source 2.45 GHz, CW
Power Attenuator
Exposure Chamber
Reflectionless Terminator
Directional Coupler Reflected Power Meter
Microwave Leakage Detector

Fig. 1

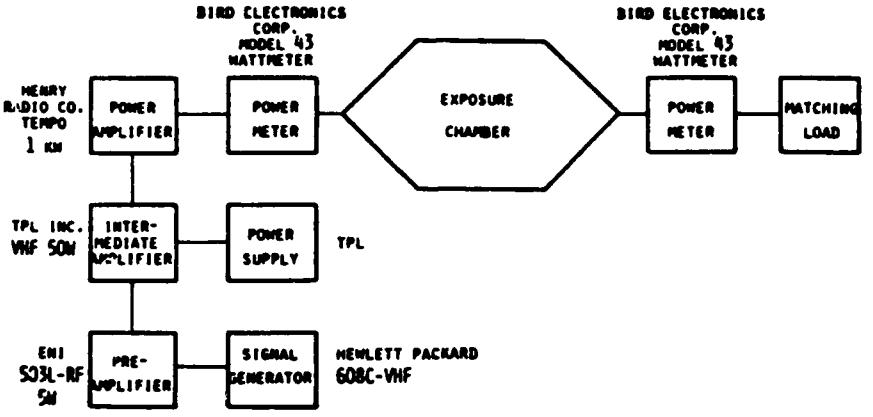

BIRD ELECTRONICS CORP. MODEL 43 WATTMETER
BIRD ELECTRONICS CORP. MODEL 43 WATTMETER
HENRY RADIO CO. TEMPO 1 KW
POWER AMPLIFIER
POWER METER
EXPOSURE CHAMBER
POWER METER
MATCHING LOAD
TPL INC. VHF 50W
INTER-MEDIATE AMPLIFIER
POWER SUPPLY
TPL
ENI 503L-RF 5W
PRE-AMPLIFIER
SIGNAL GENERATOR
HEWLETT PACKARD 608C-VHF

Fig. 2

In the other exposure system, a 10-480-MHz Hewlett Packard 608C signal generator was used as a source, and mice were exposed to 148 MHz RFR in a TEM exposure chamber attached by a coaxial cable to the generator (Lin et al., 1979). The power density in the chamber was checked with a power meter. Mice were housed in individual styrofoam containers inside a TEM exposure chamber (50 cm L x 50 cm W x 50 cm H) in the waveguide. The styrofoam containers produce minimal stress in the mice and also minimal distortion of the incident RFR field.

B. Antinociceptive Testing: Mice were exposed to varying durations of RFR at varying power densities. Experiments were initiated at an earlier date using the 2450-MHz exposure system; studies in the 148-MHz system were started only recently because of the delay in assembling the system. Antinociceptive testing utilized two nociceptive paradigms, the acetic acid abdominal constriction test and the hot plate test.

(1) Abdominal constriction test: In the 2450 MHz experiments, it was observed that if mice were tested for antinociception immediately following termination of RFR exposure, the responses were not as prominent as when animals were tested after a lag of 10-15 min. During this time, the hyperthermia had subsided but the antinociception was present. Preliminary experiments dealing with abdominal constriction antinociception as a test system showed that the optimal RFR exposure for antinociception was 10 min of 20 mW/cm^2.

Exposure of male, 20-25 g Swiss Webster mice to increasing specific absorption rates (SARs) of 2450-MHz RFR appeared to produce a progressive antinociceptive effect in the mice in both the abdominal constriction and tail flick paradigms. This antinociceptive effect was

apparent if the test was conducted 15 min following exposure but not immediately following exposure.

In the actual experiments for data, different groups of mice were exposed to peak power densities of 10, 15 and 20 mW/cm^2. Experiments at one additional power density, 30 mW/cm^2, were discontinued due to appreciable mortality. A specific absorption rate (SAR) of 23.7 W/kg at the nominal incidence influx of 10 mW/cm^2 was calculated in accordance with the Dewar-flask calorimetric technique (Blackman and Black, 1977; Durney et al., 1980). Accordingly, the SARs at the three peak power densities tested were 23.7 W/kg at 10 mW/cm^2, 34.6 W/kg at 15 mW/cm^2 and 45.5 W/kg at 20 mW/cm^2; the SAR at the lethal power density (30 mW/cm^2) was 71.1 W/kg.

Fig. 3 shows the analgesic effects in saline- and naltrexone-pretreated mice following sham exposure or exposure to increasing power densities of RFR. There was an apparent analgesic effect in sham-exposed mice, possibly associated with confinement stress; this analgesia was not sensitive to antagonism by naltrexone. On the other hand, there was an additional power density-dependent analgesia in RFR-exposed mice. Only the analgesia induced by exposure to 20 mW/cm^2 of RFR was significantly different from control (sham-exposed mice); this analgesia was markedly antagonized by naltrexone.

Fig. 4 shows the changes in body temperature in saline- and naltrexone-pretreated mice following sham exposure or exposure to increasing power densities of RFR. Neither sham exposure nor naltrexone pretreatment produced any appreciable effect on body temperature. However, there was a power density-dependent hyperthermia in RFR-exposed mice. This increase in body temperature was not sensitive to

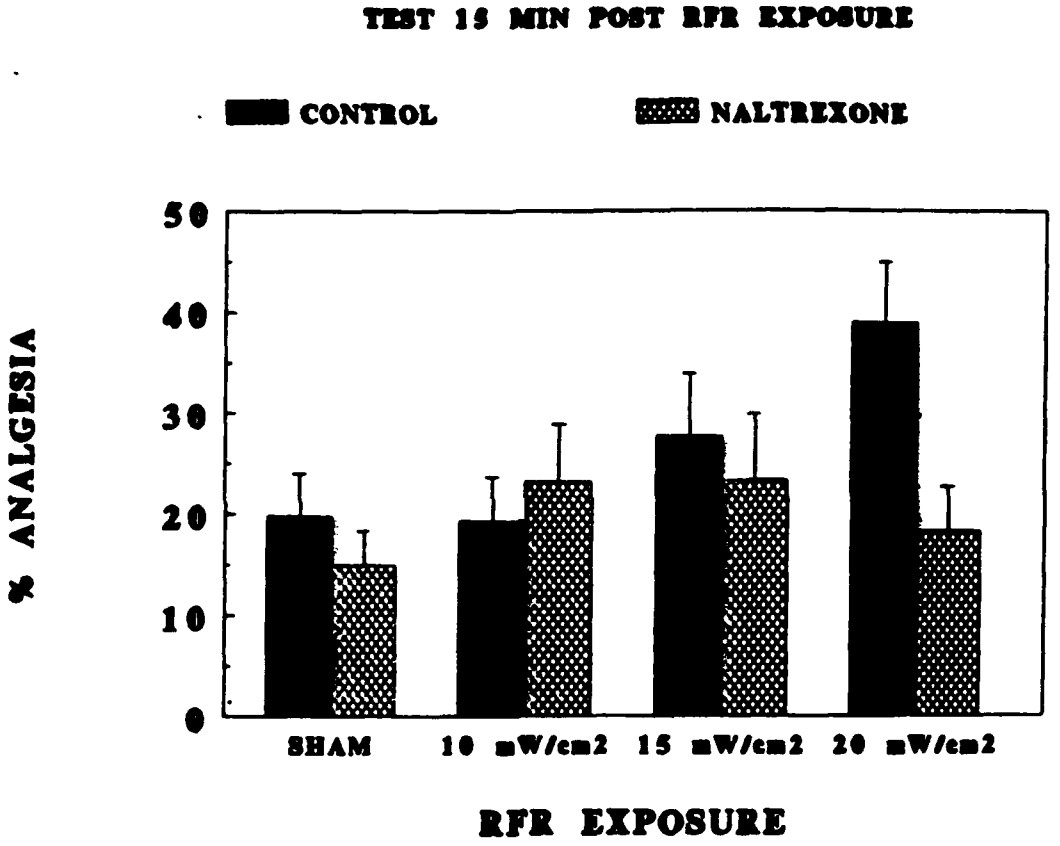

TEST 15 MIN POST RFR EXPOSURE
CONTROL
NALTREXONE
% ANALGESIA
50
40
30
20
10
0
SHAM
10 mW/cm2
15 mW/cm2
20 mW/cm2
RFR EXPOSURE

Fig. 3

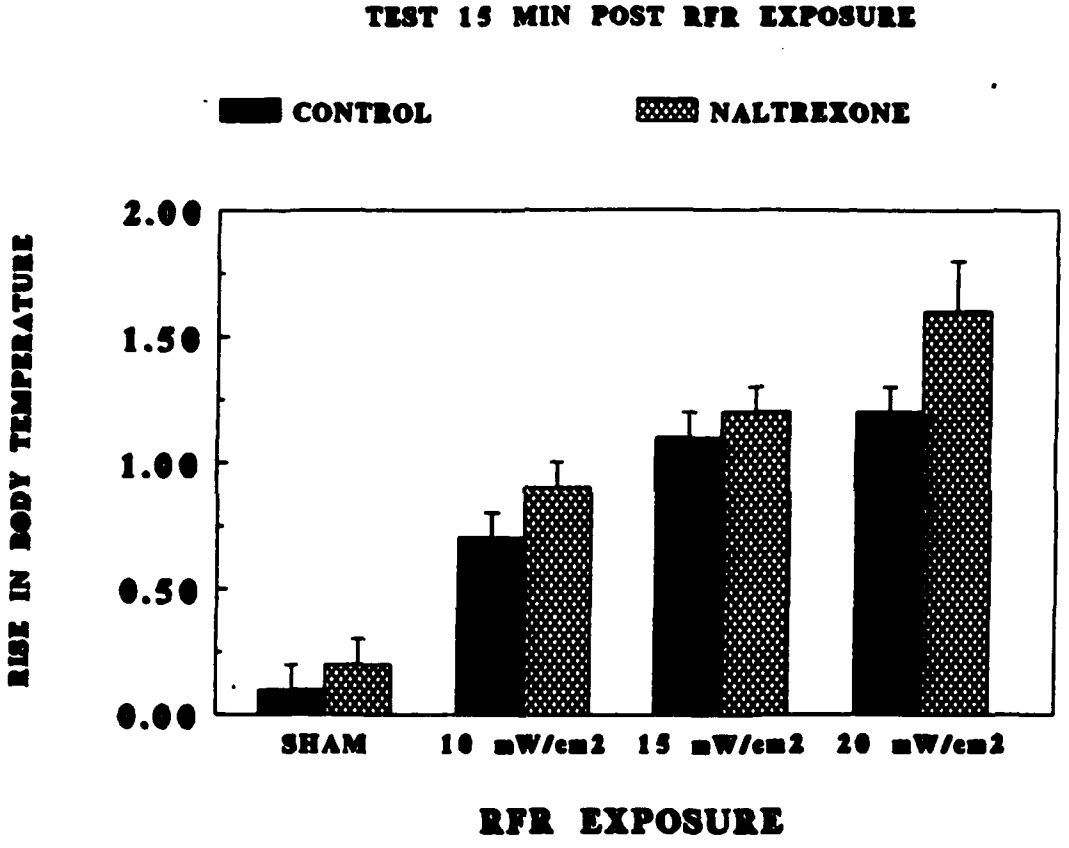

TEST 15 MIN POST RFR EXPOSURE
CONTROL
NALTREXONE
RISE IN BODY TEMPERATURE
2.00
1.50
1.00
0.50
0.00
SHAM
10 mW/cm2
15 mW/cm2
20 mW/cm2
RFR EXPOSURE

Fig. 4

antagonism by naltrexone. In fact, the hyperthermia induced by 20 mW/cm^2 RFR in naltrexone-pretreated mice was significantly greater than that in RFR-exposed, control mice.

While it has been demonstr ted that RFR can act as stressors, it has not been shown whether they can induce sufficient stress to activate endogenous opioid mechanisms. The purpose of this investigation was to determine whether exposure to RFR might induce a post-exposure state of analgesia and whether this analgesia was antagonized by an opioid receptor blocker.

The exposure system used in this study required confinement to a restricted area in order to make more consistent the RFR power density to which mice were exposed. This confinement apparently induced a degree of stress-associated analgesia which was unaffected by naltrexone treatment. When mice were so confined and exposed to RFR radiation, there was a further increase in analgesia (significant only at the highest power density), which was sensitive to antagonism by naltrexone. Based on the findings of this study, it would appear that there is a confinement stress-induced analgesia which is not opioid-mediated and a power density-(SAR-) dependent RFR-induced analgesia which is opioid-mediated.

(2) Tail flick test: In the 2450 MHz experiments, it was observed that there was an insignificant degree of antinociception in sham exposed mice following confinement in the waveguide. On the other hand, there was a significant antinociceptive effect following exposure to 10 min of 20 mW/cm^2 RFR. The antinociceptive effect was relatively stable at 10 and 20 min post RFR exposure; at 30 min following termination of the RFR exposure, the degree of antinociception had been reduced by almost 50%. In other

groups of mice exposed to 5, 10 and 15 mW/cm^2 RFR for 10 min, there was no appreciable antinociceptive effect at any time interval following RFR exposure for up to 30 min.

In other mice pretreated with 1.0 mg/kg naltrexone, there was a slight (and statistically insignificant) reduction in antinociception at 10 min and essentially no change in antinociception at 20 min following RFR exposure. These findings are shown in Fig. 5.

These findings suggest that at 2450 MHz, RFR exposure can induce antinociception detectable in the tail flick test. The failure of naltrexone to significantly antagonize this RFR-induced antinociception may suggest that either (1) the antinociception measurable in this paradigm is not opioid-mediated, or (2) the antinociception measurable in this paradigm may require greater doses of naltrexone (which may be indicative of mediation of an opioid receptor subtype somewhat more resistant to antagonism by naltrexone, e.g., possibly δ- or κ-opioid mechanisms). However, a concern here is the thermoregulatory behavior exhibited by the test animal following the thermal stress of the RFR exposure and whether or not such physiological consequences might influence response of the animal in the tail flick test, which utilizes a thermal stimulus.

B. Anxiety Testing: Different groups of mice were tested in the staircase test to determine whether RFR might induce anxiety following exposure. Animals were tested both immediately following exposure (Fig. 6) and also after a 10 min wait (Fig. 7) (to correspond to the period when RFR-induced antinociception was maximal). In neither case was there any overt increase in rearing, which according to the paradigm is indicative of anxiety.

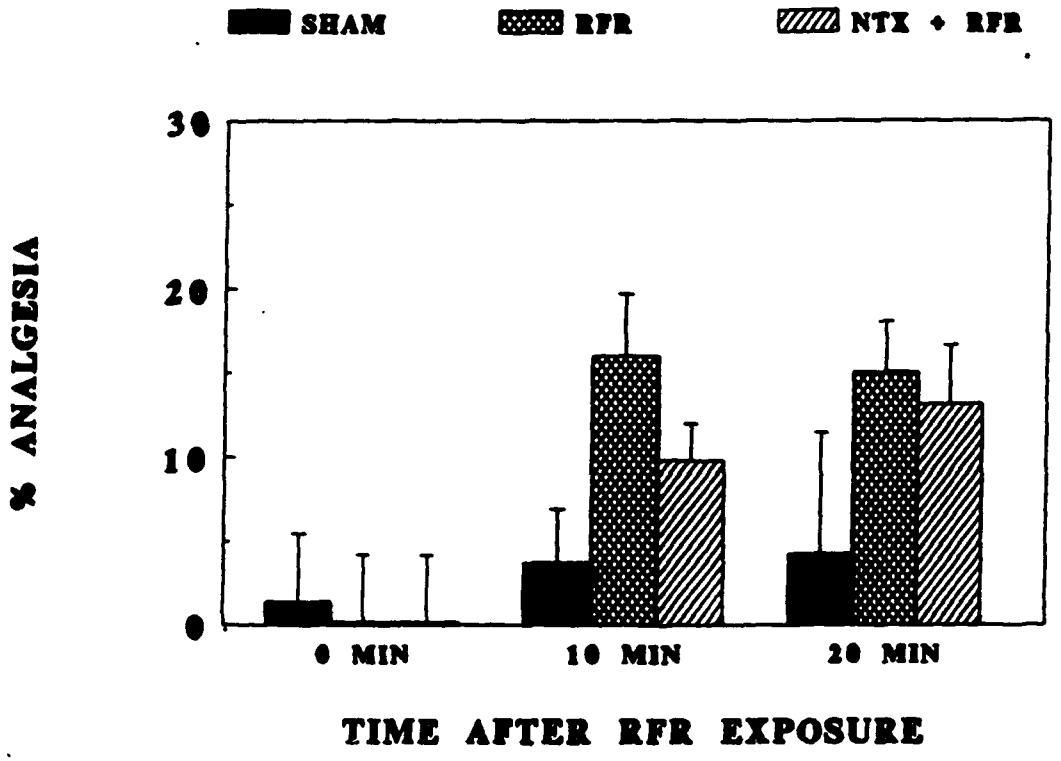

SHAM
RFR
NTX + RFR
% ANALGESIA
30
20
10
0
0 MIN
10 MIN
20 MIN
TIME AFTER RFR EXPOSURE

Fig. 5

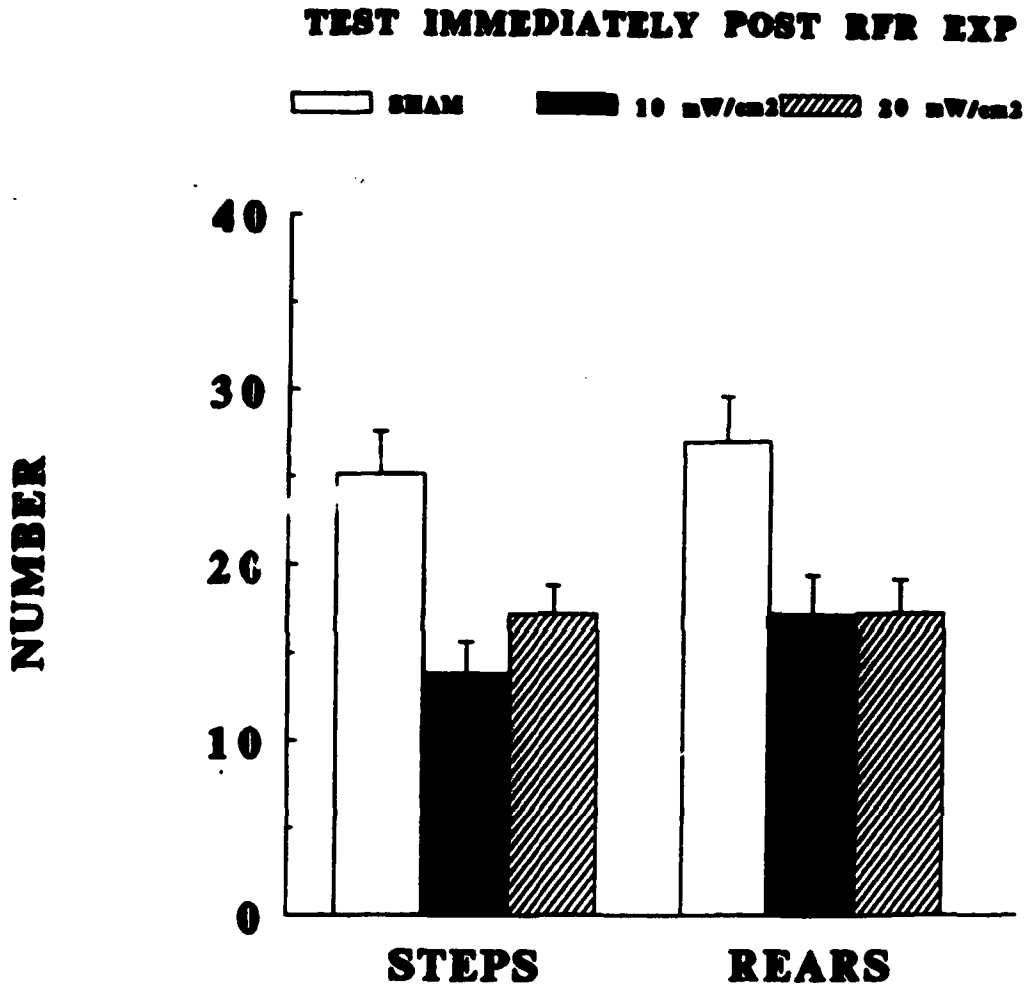

TEST IMMEDIATELY POST RFR EXP
SHAM
10 mW/cm2
20 mW/cm2
NUMBER
40
30
20
10
0
STEPS
REARS

Fig. 6

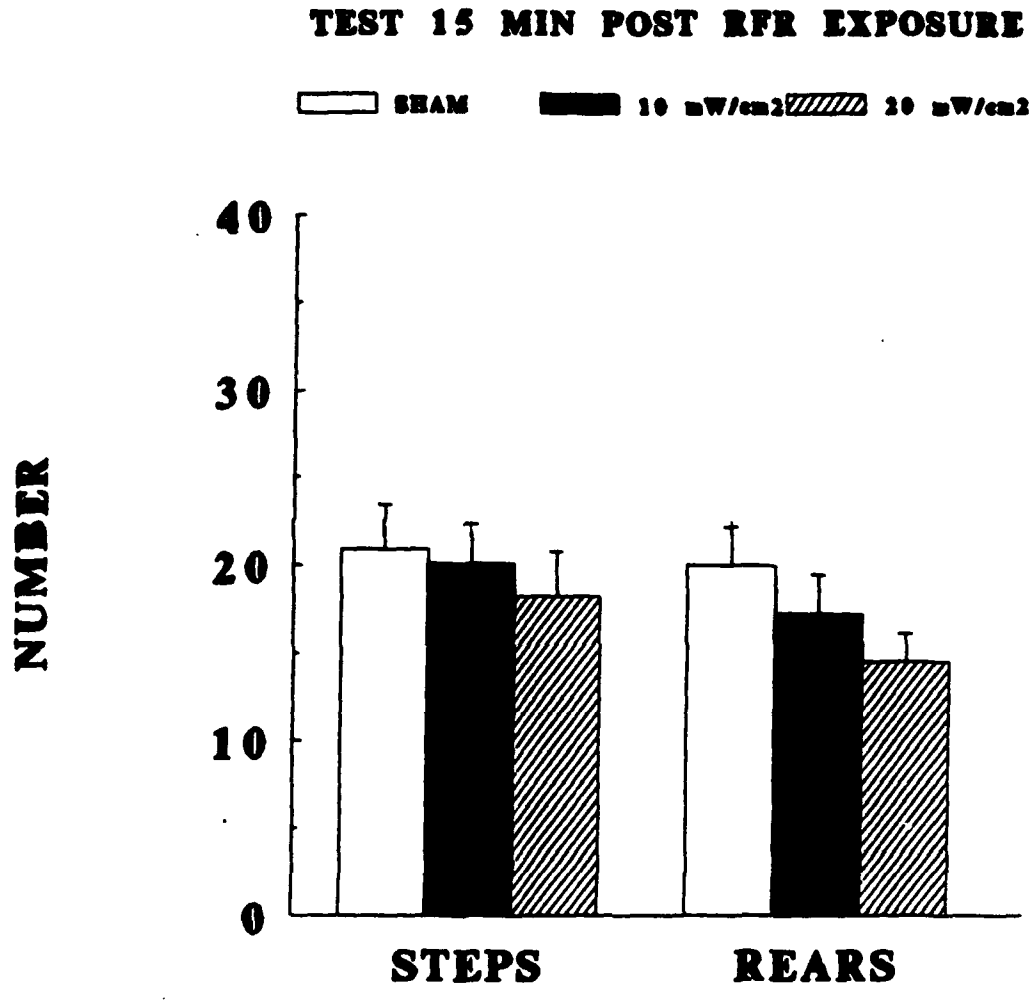

TEST 15 MIN POST RFR EXPOSURE
SHAM
10 mW/cm2
20 mW/cm2
NUMBER
40
30
20
10
0
STEPS
REARS

Fig. 7

In other experiments, we administered the benzodiazepine inverse agonist noreleagnine, an anxiogenic drug, or the benzodiazepine agonist chlordiazepoxide, an anxiolytic drug, to mice, then observed them in the staircase test. In accordance with the original paradigm, noreleagnine increased and chlordiazepoxide suppressed rearing activity, thus supporting the hypothesis that rearing represents anxious behavior. These findings are shown in Fig. 8.

The reduction of rearing behavior in RFR-exposed mice was somewhat unexpected but may be explainable based on the thermal stress imparted to the test animals, the reduction in step climbing and rearing being a behavioral thermoregulatory effect of the animal wishing to reduce its elevated body temperature. This may mean that anxiety testing dependent upon voluntary locomotor activity may be confounded by the temperature increase due to RFR exposure.

C. Neurochemistry Experiments: Two types of neurochemistry experiments have been initiated and continuing at this time, radioligand binding to endogenous opioid receptors and radioimmunoassay (RIA) quantification of endogenous opioid peptides.

(1) Opioid radioligand binding: Preliminary radioligand binding experiments were conducted, including the following: determination of optimal tissue treatment (fresh tissue vs. frozen tissue vs. frozen homogenates prepared in buffer vs. frozen homogenates prepared in sucrose solution), assessment of the effect of bovine serum albumin in the assay buffer, determination of inhibition of nonspecific binding by naltrexone, determination of association rate, and determination of tissue linearity.

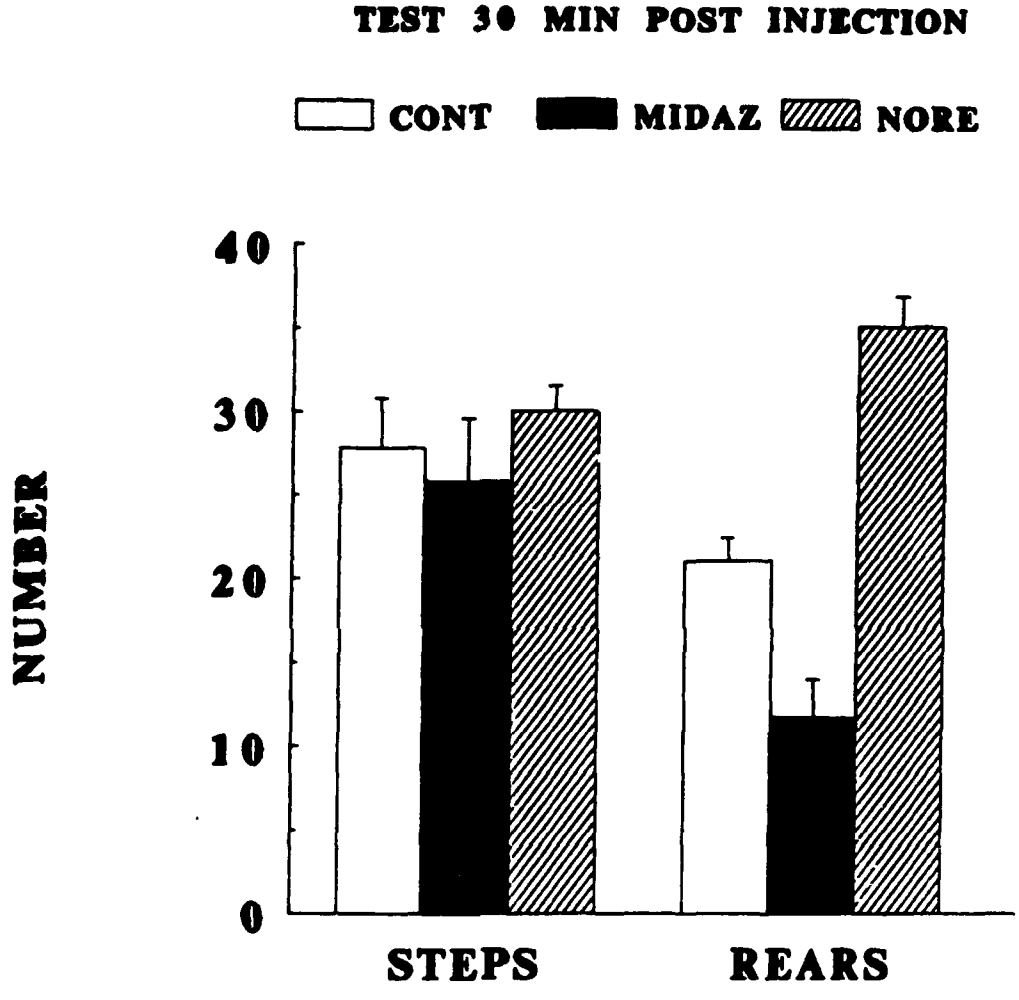

TEST 30 MIN POST INJECTION
CONT
MIDAZ
NORE
NUMBER
40
30
20
10
0
STEPS
REARS

Fig. 8

A large number of groups of mice have been exposed to 10 min sham radiation or 10, 15 and 20 mW/cm^2 2450-MHz RFR and sacrificed by cervical dislocation both immediately and 15 min following the period of irradiation. We are presently conducting radioligand binding experiments on sections of these brains and determining any microwave-induced changes in binding of [^{3}H]DAMGO, [^{3}H]p-Cl-DPDPE and [^{3}H]U-69,593 to μ-, δ- and κ-opioid receptors, respectively.

(2) Radioimmunoassay: Other groups of animals have also been exposed to 2450-MHz RFR, their brains removed and dissected either immediately or 15 min following the period of irradiation. The tissues have been lyophilized and stored at -70°C for RIA at a future date.

Because of the delayed assembly of our 148-MHz RFR exposure system, animals have yet to be exposed to 148 MHz RFR for neurochemistry experiments, however, these studies are planned.

IV. RECOMMENDATIONS

The preliminary findings of this investigation would imply that exposure to RFR does produce stress and that certain physiological responses, including activation of opioid mechanisms, might be triggered in response to this stress. Determinations of the influence of RFR on endogenous anxiomodulatory mechanisms were more equivocal. Neurochemical experiments are still in progress, supported by intramural funds of the university.

The following recommendations are made regarding future directions and methodological concerns of this research:

A. RFR Exposures

1. Additional power densities, durations of RFR exposure and post-RFR exposure intervals need to be investigated. This will be important in establishing optimal parameters of RFR exposure for identification of behavioral and neurochemical effects of RFR.
2. Different frequencies of RFR exposure need to be investigated, testing a range of sub- and supra-resonance frequencies as well as resonant frequency upon behavior and neurochemistry. It is probable that different frequencies may result in varying degrees of tissue heating in animals. The question of whether physiological responses to thermal stress are dependent upon core vs. peripheral heating remains to be investigated.
3. The present RFR exposure systems employed confinement of test animals to a restricted area to make the power density more consistent. This results in uneven distribution of heating in the test animal. Therefore, the consequences of this exposure may be highly variable. Another method of RFR exposure would necessitate restraining the animal to maintain a fixed orientation to the RFR field. A complication, however, would be restraint stress, unless the test animal is lightly anesthetized.

B. Antinociceptive Testing

1. A different nociceptive paradigm should be investigated for assessing RFR-induced antinociception. The abdominal constriction test used in this study tests μ- and κ-opioid analgesic activities only, and the tail flick test appeared to be affected, possibly by the RFR-induced hyperthermia. The hot plate test (which was not used in

this study) has the advantage of monitoring supraspinally organized response rather than a spinal reflex; however, it, too, utilizes a thermal noxious stimulus, and whether it, like the tail flick test, might be adversely affected by RFR-induced hyperthermia remains to be seen.

2. The possible activation of δ-opioid receptors by RFR needs to be investigated. This was not possible in the present study because of the different opioid mechanisms involved in antinociception monitored by the abdominal constriction paradigm.
3. Should an appropriate nociceptive paradigm be identified, the role of specific opioid receptor subtypes in RFR-induced antinociception needs to be investigated, especially with subtype-selective opioid receptor blockers, e.g. the μ-selective opioid antagonist β-funaltrexamine, the δ-opioid selective antagonist ICI-174,864 and the κ-selective opioid antagonist norbinaltorphimine.

C. Anxiety Testing

1. Another anxiety test is desirable for investigating the influence of RFR on anxiomodulatory systems. The staircase paradigm used in the present study is a simple and rapid screen for changes in level. However, it did appear that the thermal stress imposed by RFR might have suppressed some of the locomotor capability required in the staircase test for identification of central motor and anxiety functions. This may be problematic since most anxiety tests appear to involve a requirement for locomotor function, however, a test in which locomotion is required to a lesser degree may be useful.

- Interaction of benzodiazepine drugs and RFR needs to be studied. This was not accomplished in the present study because of the basic failure to detect increased anxiety in RFR-exposed animals. However, if RFR does induce anxiety as a result of stress, it will be important to determine whether pretreatment with an anxiolytic benzodiazepine or non-benzodiazepine (e.g., buspirone) might be prophylactic against the stress or anxiety.

D. Neurochemistry Experiments

1. One hypothesis under study is that RFR exposure can stimulate the neuronal release of endogenous opioid peptides. One approach would be to determine whether RFR exposure can protect opioid receptors from alkylation by β-chlornaltrexamine as a result of stimulated release of opioid peptides. This can be investigated neurochemically (via opioid radioligand binding) or pharmacologically (via analgesia testing following alkylation).
2. Another approach to determining whether RFR exposure stimulates neuronal release of endogenous opioid peptides would be to conduct experiments in centrally perfused rats, probably both ventricular-cisternal and intrathecal. Cerebrospinal fluid perfusate should be collected from animals before, during (if possible) and following RFR exposure with the perfusate fractions being subjected to radio-immunoassay determination of changes in the content of endogenous opioid peptides.
3. The method of radioligand binding is dependent the specificity of the radioligands in use. There is increasing evidence of multiplicity of opioid receptor subtypes, e.g., μ_1- vs. μ_2-opioid receptors and

possibly κ_1- vs. κ_2-opioid receptors. Therefore, there needs to be continual follow-up with newer and more specific opioid radioligands as they become available.

V. REFERENCES

Akil, H., D.J. Mayer and J.C. Liebeskind. Science 191: 961-962 (1976)

Amir, S., Z. Brown and Z. Amit. Neurosci. Biobehav. Rev. 4: 77-86 (1980)

Blackman, C.F. and J.A. Black. Radio Sci. 12: 9-12 (1977)

Bodnar, R.J., D.D. Kelly, M. Brutus, M. Glusman. Neurosci. Biobehav. Rev. 4: 87-100 (1980)

Chance, W.T. Neurosci. Biobehav. Rev. 4: 55-67 (1980)

Durney, C.H., M.K. Iskander, H. Massoudi, S.J. Allen and J.C. Mitchell. In: Radiofrequency Radiation Dosimetry Handbook, 3rd Ed., pp. 29-30. Report SAM-TR-80-32, USAF School of Aerospace Medicine, Brooks Air Force Base, TX (1980)

Grau, J.W., R.L. Hyson, S.F. Maier, J. Madden IV and J.D. Barchas. Science 213: 1410-1411 (1981)

Lai, H., A. Horita, C.K. Chou and A.W. Guy. Bioelectromag. 4: 205-214 (1983)

Lai, H., A. Horita, C.K. Chou and A.W. Guy. Bioelectromag. 5: 213-220 (1984)

Lai, H., A. Horita, C.K. Chou and A.W. Guy. Psychopharmacol. 88: 354-361 (1986a)

Lai, H., A. Horita, C.K. Chou and A.W. Guy. Pharmacol. Biochem. Behav. 24: 151-153 (1986b)

Lai, H., A. Horita, C.K. Chou and A.W. Guy. IEEE Eng. Med. Biol. pp. 31-36 (March 1987A)

Lai, H., A. Horita, C.K. Chou and A.W. Guy. J. Neurochem. 48: 40-45 (1987b)

Lewis, J.W., J.T. Cannon and J.C. Liebeskind. Science 208: 623-625 (1980)

Lin, J.C., J.C. Nelson and J. Ekstrom. Radio Sci. 14 (6S): 173-179 (1979)

Lotz, W.G. and S.M. Michaelson. J. Appl. Physiol.: Respirat. Environ. Exercise Physiol. 44: 438-445 (1978)

Lu, S.T., W.G. Lotz and S.M. Michaelson. Proc. IEEE 68: 73-77 (1980)

Madden, J. IV, H. Akil, R.L. Patrick and J.D. Barchas. Nature 265: 358-360 (1977)

Martin, G.E. and C.B. Bacino. Eur. J. Pharmacol. 59: 227-236 (1979)

Mayer, D.J. and J.C. Liebeskind. Brain Res. 68: 73-93 (1974)

Mayer, D.J. and D.D. Price. Pain 2: 379-404 (1976)

Michaelson, S.M., R.A.E. Thompson, M.Y. El Tamami, H.S. Seth and J.W. Howland. Aerospace Med. 35: 824-829 (1964)

Quock, R.M., J.M. Fujimoto, T.K. Ishii and D.G. Lange. Radiat. Res. 105: 328-340 (1986)

Quock, R.M., F.J. Kouchich, T.K. Ishii and D.G. Lange. Bioelectromag. 8: 45-55 (1987)

Selye, H. J. Clin. Endocrinol. 6: 117-130 (1946)

1990 USAF-UES RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

An Investigation of Dioxin Half-Life Estimation in Humans Based on Two or More Measurements Per Subject

Prepared by:	Ram C. Tripathi, Ph.D.
Academic Rank:	Professor
Department:	Mathematics, Computer Science and Statistics
University:	University of Texas at San Antonio
Research Location:	USAFSAM/EKB
	Brooks AFB
	San Antonio, Texas 78235
USAF Researcher:	Joel E. Michalek, Ph.D.

Date:	31 December 1990
Contract No:	F49620-88-C-0053/SB 5881-0378

Acknowledgements

I would like to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsoring this research. I would also like to thank the Universal Energy Systems, Inc. for administering and directing the Research Initiation Program.

My special thanks are due to Dr. Joel Michalek and his staff at the Biometrics branch of the Epidemiology Division for providing me with all the necessary facilities to carry out this research. I enjoyed working with Dr. Pushpa L. Gupta, University of Maine, and Dr. Samuel P. Caudill, Centers for Disease Control, on some problems related to this research. Finally, I also wish to thank Mr. Thomas White for his help on the computational aspects of this project.

An Investigation of Dioxin Half-Life Estimation in Humans Based on Two or More Measurements Per Subject

by

Ram C. Tripathi, Ph.D.

Abstract

In this report some estimates of half-life of 2, 3, 7, 8-tetrachlorodibenzo-p-dioxin (TCDD) in humans are developed by modeling the distribution of within-subject half-lives. Two models, conditional and unconditional, are developed based on two measurements per subject. The conditional model is based on the assumption that the dioxin level is decaying, therefore, the first measurement is larger than the second measurement. The unconditional model does not depend on such assumptions. Some extensions of these models are developed which utilize more than 2 measurements per subject. The probability distributions of the half-life are derived, which are utilized to obtain the maximum likelihood estimators of the related parameters. These in turn yield estimators and the confidence intervals for the median of these distributions which is the parameter of interest for estimating half-life. Another estimator of half-life is developed based on the repeated measures general linear model which incorporates covariates and permits the study of variability in half-life with respect to covariates, such as percent body fat or relative change in percent body fat.

1 Introduction

The estimation of half-life of TCDD, a contaminant in Agent Orange, in Vietnam veterans exposed to this substance, has been of considerable interest to the U. S. Air Force. The researchers at the Biometrics Branch of the Epidemiology Division, Brooks AFB, have been conducting investigations to estimate half-life of TCDD in humans based on two measurements per subject (Michalek et. al. 1989). Recently, due to a breakthrough in chemistry, the scientists are able to measure the levels of TCDD in human serum more economically and accurately. Consequently, it is expected that more measurements on the levels of TCDD in the Vietnam veterans will soon be available. Hence, the research has been expanded to develop estimates based on multiple repeated measurements per subject, and also to examine the variability of half-life with respect to some relevant covariates.

There are three methods available for estimating the half-life of TCDD in humans when the initial dose is unknown: 1) a non-parametric estimate of Phillips (1989), 2) a parametric estimate of Michalek et al (1989), 3) an estimate based on mixed linear models with repeated measurements, Tripathi (1989).

In this report, the estimate based on the mixed linear models with repeated measurements is further examined. This estimate is updated to utilize fixed effect linear models and is now obtainable by using the PROC GLM of the Statistical Analysis Systems (SAS). This allows incorporating the covariates in the model to study the possible heterogeneity of half-life with respect to these covariates. This model is applied to the Ranch Hand data and the results appear in Section 3. As a check, this model is also applied to estimate half-life on simulated data with multiple repeated measurements per subject. The results compare favorably with the known half-life which is used to generate the data.

Some new estimates of half-life are proposed here by modeling the probability

distribution of within-subject half-life. The first estimate is based on a model called "conditional model" which assumes that the level of TCDD in humans is decaying over time, therefore, the first measurement is larger than the second measurement for each subject. The second estimate is based on a model called "unconditional model" which does not require such an assumption. The second model is more realistic as there are cases in the data with the second measurement larger than the first. For both the models, the median of the distribution is regarded as the parameter of interest representing the half-life of TCDD in Vietnam veterans. The method of maximum likelihood is utilized to obtain estimates and confidence intervals for the half-life based on these models. The results for the Ranch Hand data are presented in Section 3.

Finally, some extensions of the unconditioned model are presented to incorporate more than two measurements per subject. These extended models are used on the simulated data mentioned earlier, the results appear in Section 4, and compare favorably with the known half-life which was used to generate the data. Finally, some recommendations are made to refine these models, compare the estimates from various models, and examine some properties of these estimates such as bias, asymptotic relative efficiency. It is also recommended to assess goodness of fit of the models proposed.

2 Objectives

The objectives of this research were to examine and refine the existing methods for estimating half-life of toxic compounds in humans, and to develop new methods which can accommodate multiple repeated measurements per subject along with covariate-information. This involved: 1) a further analysis of the mixed linear models (Tripathi 1989) on some generated data to validate the results, 2) develop new

models and methods for estimating the half-life based on two measurements per subject and compare the numerical results, 3) extend the models in 2) to multiple measurements.

3 Statistical Methodology

Consider the first-order kinetics model

$$C_T = C_0 e^{-\lambda T} \tag{1}$$

where C_T is the concentration T years after exposure, C_0 is the initial exposure, and λ is an unknown constant-decay rate. Based on (1), the true population half life is

$$T_{1/2} = \frac{\ell n\ 2}{\lambda}. \tag{2}$$

Therefore, in order to estimate $T_{1/2}$, we need to estimate λ. Various approaches for estimating λ, and hence $T_{1/2}$ will now be presented.

3.1 Linear Models Approach

In (1), if we take natural logarithm of both sides, and let $y_T = \ell n C_T, \alpha = \ell n C_0, \beta = -\lambda$, we get

$$y_T = \alpha + \beta T. \tag{3}$$

Since α is the natural logarithm of initial exposure, it is unobservable, and hence random for every subject, whereas β is an unknown constant. Tripathi (1989) took (3) as a motivating equation for a model which can accommodate multiple measurements per subject as well as covariates, and considered the following mixed-effects linear model with repeated measures:

$$\mathbf{y_i} = \mathbf{X_i}\beta + \mathbf{Z_i}\alpha_i + \epsilon_i \tag{4}$$

where $\mathbf{y_i} = (y_{1i}, y_{2i}, ..., y_{Ji})$ is a vector of J measurements of log concentrations on the i^{th} individual, $i = 1, 2, ..., N$, $\mathbf{X_i}$ and $\mathbf{Z_i}$ are known $J \times (p+1)$ and $J \times q$ design matrices. β is a vector of $p+1$ fixed effects to be estimated, and α_i, ϵ_i are random vectors of dimension q and J such that

$$\alpha_i \sim N(0, \sigma^2 D)$$
$$\epsilon_i \sim N(0, \sigma^2 I_J).$$

Model (4) was tried on the Ranch Hand data with 2 measurements per subject using BMDP5V and gave reasonable results. It was decided to check whether the above model gave satisfactory results when there are more than two repeated measures per subject. A data set was simulated with known λ, and hence half-life, which has 5 repeated measurements per subject. When this model was fit to the simulated data using $BMDP5V$, the results were not encouraging.

Then it was decided to try some other version of linear models. Let T_{ij} denote the number of years after exposure for individual i at the time of his j^{th} TCDD measurement with $j = 0, 1, 2, ..., J$, with $T_{io} = 0, i = 1, 2, ..., n$. Then, from (1), we have

$$\ell n(C_{Tij}) = \ell n(C_{io}) - \lambda T_{ij} \tag{5}$$

Since $\ell n(C_{Tij})$ differs for each individual, equation (5) suggests the use of a repeated measures analysis of covariance model such as

$$y_{ij} = \beta_0 + \tau_i + \beta_1 T_{ij} + \epsilon_{ij} \tag{6}$$

where y_{ij} is the natural logarithm of the j^{th} TCDD background- corrected measurement on the i^{th} individual, T_{ij} is the number of years between the j^{th} and first TCDD measurements, $-\beta_1$ represents the common decay rate λ, $\beta_0 + \tau_i$ is the intercept for the i^{th} individual, and ϵ_{ij} is the residual error term for measurement y_{ij}. The ϵ_{ij} are assumed to be $N(0, \sigma^2)$ where σ^2 is the analytical variance associated with a single y_{ij} measurement and is assumed to be constant across y_{ij} levels. In (6), y_{i0}

is not observed because the initial dose C_{i0} is assumed unknown for all subjects. Correction for background was accomplished by subtracting 4 *ppt* before taking the logarithm (Pirkle et al. 1989).

The model defined by equation (6) is easily extended to include additional classification variables and covariates. Thus, to examine the set of data for 36 Ranch Hand participants with 1982 and 1987 TCDD values greater than 10 *ppt*, we consider the following model:

$$y_{ij} = \beta_0 + \tau_i + \beta_1 T_{ij} + \beta_2 x_{ij} + \beta_3 T_{ij} x_{ij} + \epsilon_{ij} \tag{7}$$

where $y_{ij}, \beta_0, \tau_i, \beta_1, T_{ij}$ and ϵ_{ij} are defined as before and x_{ij} represents percent body fat of individual i at the time of his j^{th} TCDD measurement, β_2 is the parameter corresponding to x_{ij}, and β_3 corresponds to the interaction between x_{ij} and T_{ij}. With this model, a significant change in half-life with the covariate x_{ij} is produced by β_3 being significantly different from zero. Models (6) and (7) can be fit to the data using the SAS GLM procedure.

This model is now applied to the Ranch Hand data. We denote $y_{ij} = \ell n(C_{T_{ij}} - 4), i = 1, 2, ..., n$; where C_{ij} is the TCDD concentration for subject i at $j = 1$ (1982) and 2(1987). Percent body fat (PBF_{ij}) was computed from the formula

$$PBF_{ij} = 1.264 w_{ij}/H_{ij}^2 - 13.305$$

where w_{ij} is the weight in kilogram and H_{ij} is the height in meters for subject i at time j. We also considered the relative change in percent body fat from 1982 to 1987 defined by

$$RPBF_{i12} = (PBF_{i2} - PBF_{i1})/PBF_{i1}$$

and the relative change from tour to 1987

$$RPBF)_{i02} = (PBF_{i2} - PBF_{i0})/PBF_{i0}$$

where PBF_{i0} is the percent body fat within 5 years of tour calculated using weight within 5 years of the end of tour and height determined in 1982.

The unadjusted repeated measures model (6) was applied; the least- squares estimate of β_1 was $-.090305$ and its standard error was .011059. The corresponding unadjusted half-life estimate was 7.68 years and the 95% confidence interval for half-life was 6.2 to 10.1 years. The R-square value for this analysis was .97.

An application of the adjusted model (7) with percent body fat entered as a time-dependent covariate was carried out. The results are given below:

TABLE 1

Repeated Measures Analysis of TCDD Half-Life with Adjustment for Percent Body Fat in 36 Ranch Hands

Coefficient	Estimate	Standard Error	P-value
Time	-0.155302	0.044407	0.002
Percent body fat	-0.111670	0.048388	0.03
Time by percent body fat	0.003590	0.002064	0.09

The R-square value for this analysis was 0.98. The change in half-life with percent body fat is marginally significant ($p = 0.09$). To visualize the influence of percent body fat on the half-life estimate, the subjects were ranked with respect to the average of PBF_1 and PBF_2, denoted as $AVPBF$, and half-life and its 95% confidence interval were determined by fitting model (7). The TCDD half-life of the subject with an $AVPBF$ of 71.5 (the first quartile) was 7.7 years, 95% CI 5.9 – 11.0. The TCDD half-life of the subject with an $AVPBF$ of 24.7 (the third quartile) was 10.8 years, 95% CI 7.4 – 19.8.

The relative change in percent body fat from 1982 to 1987, $RBPF_{12}$, was also assessed with regard to changes in half-life using the adjusted repeated measures model (7). The results are shown in Table 2 below. The R-square for this analysis

was 0.98.

TABLE 2

Repeated Measures Analysis of TCDD Half-Life with Adjustment for Relative Change in Percent Body Fat in 36 Ranch Hand Between 1982 and 1987

Coefficient	Estimate	Standard Error	P-value
Time	-0.07586	0.01244	0.001
$RPBF_{12}$	2.316190	6.21403	0.710
Time by $RPBF_{12}$	-0.15771	0.30158	0.60

There was no significant change in half-life with relative changes in percent body fat from 1982 to 1987 ($p = .60$). Removal of the time-by-$PRBF_{12}$ interaction term resulted in a p-value of 0.025 for $RPBF_{12}$. Therefore, a model containing the covariate PBF and the interaction was fit with a stepwise introduction of $RPBF_{12}$ and each pairwise interaction, no new significant results were observed.

Finally, the relative change in percent body fat from tour to 1987 ($RPBF_{02}$) was assessed with regard to changes in TCDD half-life. The R-square for this analysis was 0.96. The results are shown in Table 3 below:

TABLE 3

Repeated Measures Analysis of TCDD Half-Life with Adjustment for Relative Change in Percent Body Fat From Tour to 1987 in 29 Ranch Hands.

Coefficient	Estimate	Standard Error	P-value
Time	-0.074495	0.229177	< 0.001
$RPBF_{12}$	-1.356103	2.195609	0.54
Time by $RPBF_{12}$	0.025878	0.135731	0.85

There was no significant change in TCDD half-life with relative change in percent body fat from tour to 1987 ($p = 0.85$). A model containing the PBF and the covariate $RPBF_{02}$ and all pairwise interactions was also fit; no new significant results were noted.

Another way to assess the half-life heterogeneity with respect to PBF and $RPBF_{12}$ is to perform the analysis using model (6) after stratifying the data on the basis of percent body fat in 1982 and relative change in percent body fat from 1982 to 1987. We performed this analysis on the 36 Ranch Hand cases by cross classifying the data into four cells based on PBF_1 and $RPBF_{12}$, each dichotonized at the median. The results of this analysis, presented in Table 4 below, suggest that both PBF and relative change in PBF may influence the observed half-life. With a larger data set we should be able to determine whether these results are generally true.

TABLE 4

Estimated TCDD Half-Life by Stratum of Dichotomized 1982 Percent Body Fat (PBF_1) and Relative Change in Percent Body Fat From 1982 to 1987 ($RPBF_{12}$)

	$PPF_1 \leq 0.2$		$PBF_1 > 0.2$		Total	
	Half-Life(n)		Half-Life(n)		Half-Life(n)	
$RPBF_{12} \leq 0.06$	9.15	(7)	11.81	(11)	10.64	(18)
$RPBF_{12} > 0.06$	5.26	(11)	7.69	(7)	5.98	(18)
Total	6.29	(18)	9.82	(18)	7.68	(36)

3.2 Approaches Modeling the Half-Life

Under the assumption of the first-order kinetics, as seen earlier, half-life is given by equation (2). We will denote the j^{th} TCDD concentration of the i^{th} case

by $C_{ij}, j = 1,2,...,J, i = 1,2,...,n$. When there are only two measurements per subject, we can write the half-life of the i^{th} individual as

$$Y_i = (\Delta T_i)(\ell n2)/\ell n(C_{i1}/C_{i2}), i = 1,2,...,n \tag{8}$$

where $\Delta T_i = T_{i2} - T_{i1} > 0$. Although, ΔT_i varies from subject to subject, in what follows, we have assumed $\Delta T_i = \Delta T, i = 1,2,...,n$, and denote $\Delta T(\ell n2) = d$ (a known constant). If we let $X_i = \ell n(C_{i1}/C_{i2})$, since $\ell n(C_{ij}) \sim N(\mu_j, \sigma_j^2), j = 1,2; X_i \sim N(\mu, \sigma^2)$, where $\mu = \mu_1 - \mu_2, \sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$ with $\sigma_{12} = cov(\ell nC_{i1}, \ell nC_{i2})$. It will be assumed that $X_1, X_2, ..., X_n$ are independent identically distributed (i.i.d.) random variables (r.v.) with the common $N(\mu, \sigma^2)$ distribution. Similarly $Y_i = d/X_i, i = 1,2,...,n$ are also i.i.d. with some common distribution. We will assume that the half-life of TCDD is a random variable denoted by Y with this common distribution. The median of the distribution of Y, denoted by M, is the true parameter of interest representing half-life that we wish to estimate based on a random sample $Y_1, Y_2, ..., Y_n$ from this distribution. In what follows, we denote the distribution function and the probability density function (p.d.f.) of the standard normal distribution by $\Phi(\cdot)$ and $\phi(\cdot)$ respectively. We consider two models

i) $C_{i1} > C_{i2}$, called the conditional model,

ii) unconditional model, without any condition.

3.2.1 Conditional Model

The p.d.f. of Y under the condition $C_{i1} > C_{i2}$ is

$$f_Y(y) = \frac{d}{\sqrt{2\pi}\sigma(1 - \Phi(-\mu/\sigma))} \exp[-(1/2\sigma^2)(d/y - \mu)^2], y > 0. \tag{9}$$

Then, the median of (9), the half-life parameter of interest is

$$M = \frac{d}{\mu + \sigma\Phi^{-1}(.5 + .5\Phi(-\mu/\sigma))}. \tag{10}$$

Assuming that $Y_1, Y_2, ..., Y_n$ is a random sample from (9), we shall now obtain the maximum likelihood estimator (m.l.e.) of M. The likelihood function of the sample is

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{d_i}{\sqrt{2\pi}\sigma(1 - \Phi(-\mu/\sigma))} \exp[-(1/2\sigma^2)(\frac{d_i}{y_i} - \mu)^2] \tag{11}$$

Taking the natural logarithm of (11) and differentiating with respect to μ and σ^2 gives the likelihood equations

$$\frac{n\phi(-\mu/\sigma)}{\sigma[1 - \Phi(-\mu/\sigma)]} + \frac{1}{\sigma^2} \sum_{i=1}^{n} (\frac{d_i}{y_i} - \mu) = 0 \tag{12}$$

and

$$-\frac{n}{\sigma} + \frac{n\mu\phi(-\mu/\sigma)}{\sigma^2[1 - \Phi(-\mu/\sigma)]} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (\frac{d_i}{y_i} - \mu)^2 = 0. \tag{13}$$

On solving (12) and (13) we obtain

$$\sigma^2 = \frac{1}{n} \left[\sum_{i=1}^{n} (\frac{d_i}{y_i})^2 - \mu \sum_{i=1}^{n} (\frac{d_i}{y_i}) \right]$$

and

$$\mu + \sigma \frac{\phi(-\mu/\sigma)}{1 - \Phi(-\mu/\sigma)} = \frac{1}{n} \sum_{i=1}^{n} (\frac{d_i}{y_i}).$$

These equations can be solved by Newton-Raphson's method to give the m.l.e. $\hat{\mu}, \hat{\sigma}$. The m.l.e. of M can now be obtained as

$$\hat{M} = \frac{d}{\hat{\mu} + \hat{\sigma}\Phi^{-1}(.5 + .5\Phi(-\hat{\mu}/\hat{\sigma}))}. \tag{14}$$

In (14), we can replace d by $\overline{\Delta T}(\ell n 2)$, where $\overline{\Delta T}$ is the mean of $\Delta T_i, i = 1, 2, ..., n$. The asymptotic distribution of $\hat{M}$ is $N(M, \sigma^2_{\hat{M}})$ if n is large, where $\sigma^2_{\hat{M}}$ can be obtained from the delta method, and from the variance covariance matrix of $(\hat{\mu}, \hat{\sigma}^2)$ which is the inverse of the information matrix. Then $100(1 - \alpha)\%$ CI for M can be obtained as $\hat{M} \pm Z_{\alpha/2}\hat{\sigma}_{\hat{M}}$ where $\hat{\sigma}_{\hat{M}}$ is an estimate of $\sigma_{\hat{M}}$.

3.2.2 Uncondition Model (Full Model)

In this section, we give the p.d.f. of Y without any restriction. This is given by

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma}\frac{d}{y^2}\exp[-\frac{1}{2\sigma^2}(\frac{d}{y}-M)^2] & y \neq 0 \\ 0 & y = 0 \end{cases} \tag{15}$$

The median of this distribution is given by

$$M = \frac{d}{\mu + \sigma\Phi^{-1}(a + \Phi(-\mu/\sigma))} \tag{16}$$

where $a = \pm.5$ according as $M \gtrless 0$. The likelihood function of the sample is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}(\frac{d_i}{y_i^2})\exp[-\frac{1}{2\sigma^2}(\frac{d_i}{y_i}-\mu)^2] \tag{17}$$

The m.l.e. of μ and σ^2 can be obtained from (17) and are given by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}(\frac{d_i}{y_i}), \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\frac{d_i}{y_i}-\hat{\mu})^2.$$

The m.l.e. of M is

$$\hat{M} = \frac{d}{\hat{\mu} + \hat{\sigma}\Phi(a + \Phi^{-1}(-\hat{\mu}/\hat{\sigma}))}. \tag{18}$$

Again, using the large sample properties of $\hat{M}$, $100(1-\alpha)\%$ approximate CI for M is $\hat{M} \pm Z_{\alpha/2}\hat{\sigma}_{\hat{M}}$ where $\hat{\sigma}_{\hat{M}}$ is an estimate of $\sigma_{\hat{M}}$ which can be obtained from the delta method and from the covariance matrix of $(\hat{\mu}, \hat{\sigma}^2)$. The latter can be obtained from the inverse of the information matrix.

3.2.3 Analysis of Ranch Hand Data Based on Modeled Half-Life.

In Table 5 below, we present analyses of Ranch Hand data based on these two models.

TABLE 5

Estimates of TCDD Half-Life for Ranch Hand Data

Based on the Conditional and Unconditional Models

Model	n	$\hat{\mu}$	$\hat{\sigma}$	$\hat{M}$	$\hat{\sigma}_{\hat{M}}$	95% CE for M	Width of CI
Conditional	32	0.4925	0.2688	6.6855	0.1898	(6.3135, 7.0574)	.7439
Model	134	0.2154	0.4821	8.5231	0.4292	(7.6818, 9.3644)	1.6826
Unconditional	32	0.4402	0.3149	6.6796	0.3149	(5.8691, 7.4902)	1.6211
Model	155	0.3745	0.4030	6.3677	0.2987	(5.7823, 6.9531)	1.1708

The confidence intervals obtained here are narrower than the one obtained by Pirkle et.al. (1989). Pirkle et.al. estimated the half-life of TCDD to be 7.1 years by taking the median of within subject half-lives $\frac{\Delta T_i \ell n 2}{\ell n(C_{i1}/C_{i2})}, i = 1, 2, ..., n$, with a 90% CI (5.8, 9.6). These results are not surprising because the estimates in Table 5 are based on parametric procedures while Pirkle et.al's estimate is nonparametric.

4 Extensions of Modeled Half-Life to Multiple Repeated Measurements

The two models considered in Section 3 for half-life accommodate only two measurements. In this section, we present three models based on a similar idea which accommodates multiple repeated measurements per subject. In the notation of Section 3, suppose C_{ij} is the j^{th} TCDD concentration measurement on the i^{th} individual at time T_{ij}; and denote $\Delta T_{ij} = T_{ij+1} - T_{ij}, j = 1, 2, ..., J-1, i = 1, 2, ..., n$.

Under the assumption of first-order kinetics, the half-life is given by $Y = \frac{\ell n 2}{\lambda}$. We can get values of λ for every subject as follows:

Model I: From equation (1), for every individual, we have

$$\ell n(C_{ij}/C_{ij+1}) = \lambda \Delta T_{ij}, j = 1, 2, ...J-1 \tag{19}$$

hence,

$$\lambda = \frac{1}{\Delta T_{ij}} \ell n(C_{ij}/C_{ij+1}). \tag{20}$$

On summing (20) both sides with respect to j, we get

$$\lambda = \frac{1}{J-1} \sum_{j=1}^{J-1} \frac{1}{\Delta T_{ij}} \ell n(C_{ij}/C_{ij+1}).$$

On putting λ in (2) and denoting the resulting half-life for subject i by Y_i, we have

$$Y_i = \frac{(J-1)\ell n2}{\sum_{j=1}^{J-1} \Delta T_{ij} \ell n(C_{ij}/C_{ij+1})}, i = 1, 2, ..., n$$

Since $Y_1, Y_2, ..., Y_n$ are i.i.d.; these can be regarded as a random sample from a distribution corresponding to a r.v. $Y = d/X$, with $d = (J-1)\ell n2$ and X denoting the denominator of Y_i disregarding i. The p.d.f. of Y is the same as in (15) with Y as defined here.

The m.l.e. of μ and σ for the distribution of Y and of its median M can be obtained as in Section 3.2.2. The variance $\sigma^2_{\hat{M}}$ of $\hat{M}$ can also be obtained by the delta method. Then, an approximate $100(1-\alpha)\%$ CI for M is $\hat{M} \pm Z_{\alpha/2}\hat{\sigma}_{\hat{M}}$.

Model II: From equation (1), we have

$$\ell n(C_{i1}/C_{ij}) = \lambda \Delta_1 T_{ij}, \qquad j = 2, 3, ...J, i = 1, 2, ..., n \tag{21}$$

where $\Delta_1 T_{ij} = T_{i1} - T_{ij}$. On summing both sides on (21) and solving for λ, we get

$$\lambda = \frac{\sum_{j=2}^{J} \ell n(C_{i1}/C_{ij})}{\sum_{j=2}^{J} \Delta_1 T_{ij}}.$$

On putting λ in (2) and denoting the resulting half-life for subject i by Y_i, we have

$$Y_i = \frac{(\ell n 2)\sum_{j=2}^{J} \Delta_1 T_{ij}}{\sum_{j=2}^{J} \ell n(C_{i1}/C_{ij})} \qquad i = 1, 2, ..., n.$$

As for the Model I above, $Y_1, Y_2, ..., Y_n$ can be regarded as a random sample from a distribution corresponding to a r.v. $Y = \frac{d}{X}$, with $d = (\ell n 2)\overline{\Delta_1 T}$, $\overline{\Delta_1 T} = \sum_{i=1}^{n}\sum_{j=2}^{J} \Delta_1 T_{ij}$, and X denoting the denominator of Y_i disregarding i. The p.d.f. of Y is the same as in (15) with Y as defined here. The m.l.e. $\hat{M}$ of M, the median of interest and the estimate $\hat{\sigma}_{\hat{M}}$ of the standard deviation of $\hat{M}$ can be obtained as mentioned earlier. These yield the $100(1-\alpha)\%$ CI for M.

Model III: From (21),

$$\lambda = \frac{\ell n(C_{i1}/C_{ij})}{\Delta_1 T_{ij}} \qquad j = 2, 3, ..., J.$$

Summing both sides over j and solving for λ yields

$$\lambda = \frac{1}{J-1}\sum_{j-2}^{J} \frac{\ell n(C_{i1}/C_{ij})}{\Delta_1 T_{ij}}.$$

On substituting λ in (2) and denoting the resulting half-life for subject i by Y_i, we get

$$Y_i = \frac{(J-1)\ell n 2}{\sum_{j=2}^{J} \frac{\ell n(C_{i1}/C_{ij})}{\Delta_1 T_{ij}}} \qquad i = 1, 2, ..., n.$$

As in Models I and II, $Y_1, Y_2, ..., Y_n$ can be regarded as a random sample from a distribution corresponding to a r.v. $Y = d/x$ with $d = (J-1)\ell n 2$, and X denoting the denominator of Y_i disregarding i. The p.d.f. of Y is the same as in (15) with Y as defined here. The m.l.e. of M and $\sigma_{\hat{M}}$ can be obtained as usual, hence also the $100(1-\alpha)\%$ CI for M.

These models were applied to simulated data with known half-life of 7.5424 years. The data was generated on 100 subjects with 5 measurements. Four analyses were carried out on each of the three models with 2, 3, 4, and 5 measurements per subject to see the effect of the number of measurements on the estimate. From the table it appears that the estimates from Model I are relatively insensitive to the number of measurements, while those from Models II and III seem to get more precise as the number of measurements increase, with Model II giving better results.

It is interesting to note that Models I and III are identical when there are only two measurements per subject. Also, the unconditional full model of Section 3.2.2 is equivalent to Model II in the case of two measurements. The conditional model of Section 3.2.1 was also applied to this data set with only two measurements, the results compare with those of the other models based on 4 or 5 measurements.

For comparison, the linear model (7) with repeated measurements was applied to the same simulated data with five measurements per subject. The estimates of λ and its standard error were $\hat{\lambda} = 0.087712, \hat{\sigma}_{\hat{\lambda}} = 0.00146$. This gives an estimate of half-life of 7.9025 years with estimate of its standard error 0.1311. The 95% CI for the half-life is (7.6537, 8.1681) with a width of 0.5144 which compares favorably with those obtained for Models II and III.

5 Recommendations:

1. It is important to explore why the mixed linear model approach using BMDP5V that works when there are two measurements per subject does not produce meaningful results on simulated data with multiple measurements. The results of mixed-model approach are necessary to offer a check for the results based on linear-model approach with repeated measurements proposed here.

2. It is also important to investigate the properties of the estimators obtained by

the repeated measures approach. These properties should include bias, and asymptotic relative efficiency of the estimators.

3. The conditional and the unconditional models proposed in Section 3 should be extended to accommodate covariates. Similar studies should also be carried out for the extended models Model I, Model II and Model III.

4. Models I, II and III are very similar and differ only in how we group the time intervals and TCDD concentration measurements, and how we solve for λ. The estimators obtained from these models should be assessed for bias and efficiency. Recommendation for a specific model can only be made after such a comparison.

TABLE 6

Estimates TCDD Half-Life for Simulated Data Based on Models I, II, and III with Multiple Measurements Per Subject for 100 Subjects

Model	No. of Measurements	$\hat{\mu}$	$\hat{\sigma}$	$\hat{M}$	$\hat{\sigma}_{\hat{M}}$	95% CI	Width of 95% CI
I	2	0.0943	0.0648	6.5271	0.3234	(5.8932, 7.1610)	1.2678
	3	0.0912	0.0296	7.5977	0.3346	(6.9419, 8.2534)	1.3115
	4	0.0882	0.0222	7.8592	0.3441	(7.1847, 8.5337)	1.3490
	5	0.0884	0.0175	7.8419	0.3417	(7.1721, 8.5116)	1.3395
II	2	0.0947	0.0662	6.4499	0.3220	(5.8189, 7.0810)	1.2621
	3	0.0924	0.0369	7.4579	0.2834	(6.9024, 8.0134)	1.1110
	4	0.0903	0.0267	7.6738	0.2253	(7.2321, 8.1154)	0.8833
	5	0.0895	0.0207	7.7426	0.1781	(7.3936, 8.0916)	0.6980
III	2	0.0943	0.0648	6.5271	0.3234	(5.8932, 7.1610)	1.2678
	3	0.0927	0.0423	7.3567	0.3044	(6.7601, 7.9532)	1.1931
	4	0.0912	0.0331	7.5783	0.2671	(7.0549, 8.1017)	1.0468
	5	0.0905	0.0272	7.6557	0.2281	(7.2086, 8.1028)	0.8942
*Model in Section 3.2.1	2	0.4234	0.3400	7.6187	0.2224	(7.1689, 8.0684)	0.8995

*Out of the 100 cases, 5 cases were thrown out due to $C_{i2} > C_{i1}$.

REFERENCES

Gupta, P. L. (1990). Dioxin Half-Life Estimation in Veterans of Project Ranch Hand. UES Report.

Michalek, J. E., Tripathi, R. C., Caudill, S. P. and Pirkle, J. L. (1990). An Investigation of TCDD Half-Life Heterogeneity in Veterans of Operation Ranch Hand, A manuscript underpreparation.

Michalek, J. E., Mihalko, D., White, T., and Patterson, Jr., D. G. (1989). Maximum Likelihood Estimation of Half-life Based on Two Measurements per Subject. (submitted)

Phillips, D. L. (1989). Propagation of Error and Bias in Half- Life Estimates Based on Two Measurements. *Archives of Environmental Contamination and Toxicology* (In Press).

Pirkle, J. L., Miner, J. C., Peterson, M. R., and Phillips, D. G. (1989). Estimates of the Half-life of 2, 3, 7, 8-tetrachlorodibenzo-p-dioxin in Vietnam veterans of Operation Ranch Hand. *Journal of Toxicology and Environmental Health*, 27, 165-171.

Tripathi, R. C. (1989). An Investigation of Dioxin Half-Life Heterogeneity in Humans Based on Two Measurements Per Subject. UES Report.

Report # 102
760-0MG-049
Prof. John Flach
Report Not Publishable

Report of Research Initiation Proposal: Degradation of the renal peritubular basement membrane in relation to toxic nephropathy from compounds of military interest

Submitted to: Universal Energy Systems, Dayton, OH
Administrators of The U.S. Air Force Summer Faculty Research Program under contract to the U.S. Air Force

Author: Thomas D. Lockwood, Ph.D.
Guest Scientist 1989 Summer Faculty Research Program
Armstrong Aerospace Medical Research Laboratory
Wright Patterson Air Force Base
Associate Professor
Department of Pharmacology and Toxicology
School of Medicine
Wright State University
Dayton, OH 45435

Administrative

The present report describes progress on a small research initiation seed grant of $20,000 (i.e., Minigrant) awarded by the Air Force as a subcontract administered by United Energy Services, Inc. Portions of the present research were contributed informally by personnel of the Toxic Hazards Division of the AAMRL and Department of Anatomy of the Wright State University School of Medicine. The proposed research has been successful to the present stage of completion. Completed results partially described herein can yield a short collaborative publication; however, additional effort will yield a full length manuscript publishable in a reputable peer reviewed journal. Thus, we elect to continue this project initiation to the point of completion of a full length paper of substantive significance to the field.

Summary

The degradation of the renal peritubular basal lamina (formerly called the basement membrane) was investigated in relation to tubular injury from toxic compounds of military interest. The ultrastructure of the lesion was characterized by transmission and scanning electron microscopy. The enzymatic biochemical basis for degradation of the peritubular basal lamina was studied by characterizing extracell release of degradative enzymes including proteases and glycosidases from the functional isolated perfused kidney during injury.

The kidney is normally a degradative organ with a very high content of lysosomal acid hydrolases that degrade filtered endocytosed plasma protein and glycoprotein. We hypothesized that after injury, these enzymes are released extracellularly and play a role in degradation of the peritubular basal lamina. The maintenance of the peritubular basal lamina is critical to the regeneration or degeneration of the conduit structure of the nephron following toxic or ischemic injury.

1. Although the degradation of the basal lamina can be observed ultrastructurally only at one day following injury the enzymatic basis for this degradative lesion was found to begin within minutes of severe tubular cell injury.

2. Lysosomal degradative enzymes are normally released in large amounts from the apical surface of uninjured tubular epithelial cells into the tubular lumen and eventually urine. We found that these acid hydrolases are not normally released in appreciable amounts from the basal or contralumenal surface of the uninjured tubule into the renal interstitial space and lymphatic system.

3. Within minutes of severe tubular injury lysosomal acid hydrolases and neutral proteases were observed to be released in large amounts into the tubular lumen and urine. Within 15 min of tubular cell injury acid hydrolases were released in large amounts from the basal contralumenal surface of the tubular cells, across the peritubular basal lamina, into the renal interstitium, and lymphatic system.

4. Contralumenal release of the acid hydrolases into the renal interstitium was correlated with the time course of functional decline of the preparation at 15-30 min after severe injury from nutrient deprivation.

5. At 24 hr after injury ultrastructural evidence of concomitant frank degradation of the peritubular basal lamina was observed in correlation with fusion of lysosomal vesicles with the basal plasma membrane. Other ultrastructural aspects of the cell lesion are described.

Introduction and Background

Due to its unique structure and function the kidney is a critical target organ of great sensitivity to a wide variety of toxic substances including occupational-environmental substances of military interest. The Air Force and Navy have studied the nephrotoxic hazard of fuels, fire retardants and hydraulic fluids for more than 10 years using generally accepted routine procedures (Bruner, 1984; Bruner & Pitts, 1983; Kinkeod et al., 1987; Mattie et al, 1986, Norton et al., 1985; Norton & Mattie, 1987). Existing methods of renal toxicology and pathology in animals and humans are insensitive and far from definitive (see below). Moreover, basic nephrotoxic mechanisms are not well understood. Current in vivo methods in toxicology suffer from a variety of well known scientific, economic and administrative problems. The isolated perfused kidney offers some advantages in the study of pathogenic mechanisms as described below.

Renal pathology can be subdivided into 4 related areas: the glomerulus, the tubule, the vasculature, and the interstitium (immune response). The present research focused on the tubule. The tubular epithelium is frequently the limiting organ of exposure to agents which only slightly damage other tissues. Considerable gaps remain in present understanding of two aspects of tubular injury: (a) What is the biochemical-cellular basis for the irreversible disruption of tubular conduit structure and resultant permanent loss of the nephron? (b) How can this process be recognized as a predictive sublethal ultrastructural lesion in animal models, and how does tubular pathogenesis correlate with customary functional indicators?

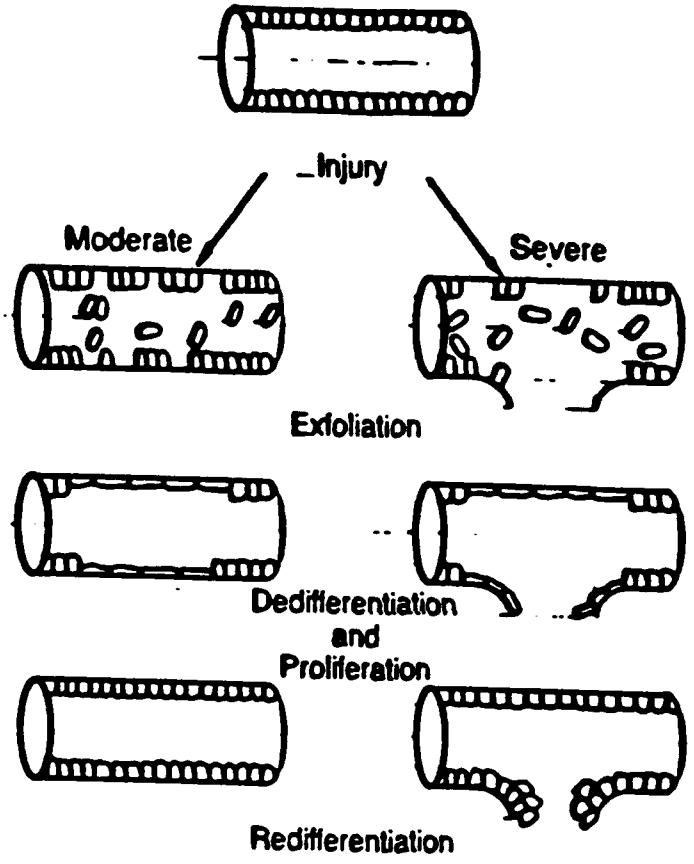


Fig. 1. A. Exfoliation and regeneration of a tubular epithelial cell population after moderate injury without disruption of the peritubular basal lamina. B. Severe injury accompanied by disruption of peritubular basal lamina and loss of conduit integrity (see text and Norton and Mattie, 1987).

Particular sensitivity of the tubular epithelial cells to damage from diverse toxic biochemical mechanisms. Several factors conspire to render the tubular epithelial cell population extremely sensitive to a wide variety of therapeutic and environmental-occupational agents. The transport function of the kidney epithelium requires very large amounts of energy as evidenced by numerous prominent mitochondria; thus the kidney is particularly sensitive to metabolic disruption of glycolysis or respiration. The kidney contains a high activity of drug metabolizing enzyme systems; these can produce reactive intermediates which cause macromolecular damage in various cell structures including membranes. Excessive filtered plasma proteins can exceed the normal ability of the tubular epithelium to degrade and clear them. The kidney receives 20% of the cardiac output which delivers large amounts of blood-borne toxicants. However, the most significant contribution to renal sensitivity is the result of the very function of the kidney: concentration. Because 99.5% of the water is reabsorbed, the potential concentrating effect or exposure to nonreabsorbed toxicants is 200 fold.

Critical role of the peritubular basal lamina in regeneration or degeneration of the nephron following toxic injury. The tubule is a conduit that must remain sealed to filtrate in order to regulate the volume and composition of the body fluids. If the tubular epithelium is disrupted at any point along the long course of the nephron, then unregulated tubular fluid escapes and is returned to the body.

The epithelial cells of the nephron are anchored at their basal side to the peritubular basal lamina. This cylindrical "sleeve" around the entire length of the nephron provides mechanical support in order to withstand a transmural pressure gradient. The basal lamina is composed of several proteins, polysacharides and glycoproteins, including collagen, fibronectin, heparin sulfate, etc. This basal lamina is replaced very slowly although the exact kinetics and mechanisms of turnover are obscure. All of the enzymes necessary to degrade basement membrane constituents are found in lysosomes, i.e., various proteases and various glycosidases. After tubular injury the peritubular basement membrane plays a critical role in facilitating the regeneration of the damaged and disrupted epithelial cell population (Fig. 1). If damage is moderate, the peritubular basal lamina remains intact although some epithelial cells exfoliate and appear in urine. Cells dedifferentiate and proliferate over the cylindrical framework retained by the intact basal lamina (Fig. 1). However, if cell damage is severe, then the basement membrane is disrupted. Regenerating cell proliferation is unguided and the conduit is lost (Fig. 1). Importantly, the nephron is a very long structure, and loss can result from disruption only at a single point. Thus the critical determinant of nephron regeneration vs. degeneration is the loss of integrity of the peritubular basal lamina framework. The enzymatic and ultrastructural features of normal basal membrane turnover and accelerated degradation after injury have not previously been described.

Normal degradative function of the kidney and pathogenic role of lysosomal degradative enzymes in tubular injury following toxic exposure. It is seldomly appreciated that the normal kidney is a degradative organ that hydrolyzes considerable amounts of plasma protein and glycoprotein that is filtered through the normal glomerulus. Filtered protein is reabsorbed by endocytosis in all segments but largely in the proximal tubule. The kidney has the highest concentration of lysosomal enzymes in the body and a morphologically prominent lysosomal-vacuolar system. Kidney lysosomal enzymes are normally secreted from live cells in large amounts and are excreted in urine (Lockwood and Bosmann, 1979). Exocytotic vesicles containing undegraded cell debris can be readily seen being released through the brush border into the tubular lumen (see below micrographs). These exocytotic vesicles have been shown to stain histologically for lysosomal acid hydrolases. Indeed when the glomerulus is damaged or when cell protein is released into the bloodstream from crushed or injured tissues then acute renal failure commonly results from an overload of undigested protein within endosomes of the proximal tubule (see below micrographs). The extremely high renal content of lysosomal enzymes can also cause massive damage to intra and extracell structures of the kidney when their containment is disrupted by cell damage (see below micrographs).

Present Studies

Against the above background, studies presently completed describe intra- and extracellular ultrastructural features of the disruption of the peritubular basal lamina following in vivo injury by nephrotoxic compounds. Second, the biochemical basis for disruption of the peritubular basement was studied by correlating the simultaneous release of degradative enzymes into tubular lumen and across the basement membrane into the renal interstitium.

Methods

Release of renal degradative enzymes from the isolated perfused kidney. The perfused kidney preparation was modified for the correlation of simultaneous enzyme release into the tubular lumen (urine) and renal interstitium (lymphatic) (see Fig. 2). General methods of perfusion with physiologic perfusate have been previously described (Lockwood and Bosmann, 1979). The renal artery is cannulated in situ without interrupting the flow of blood. Urine is collected by cannulating the ureter. The entire output of capsular and hilar lymphatics was collected separately by draining into a funnel (Fig. 2). This kidney preparation is functional for more than 1 hr as evidenced by reabsorption of sodium and water (see Lockwood and Bosmann, 1979). Lysosomal degradative enzymes were assayed as previously described (Lockwood and Bosmann, 1979).

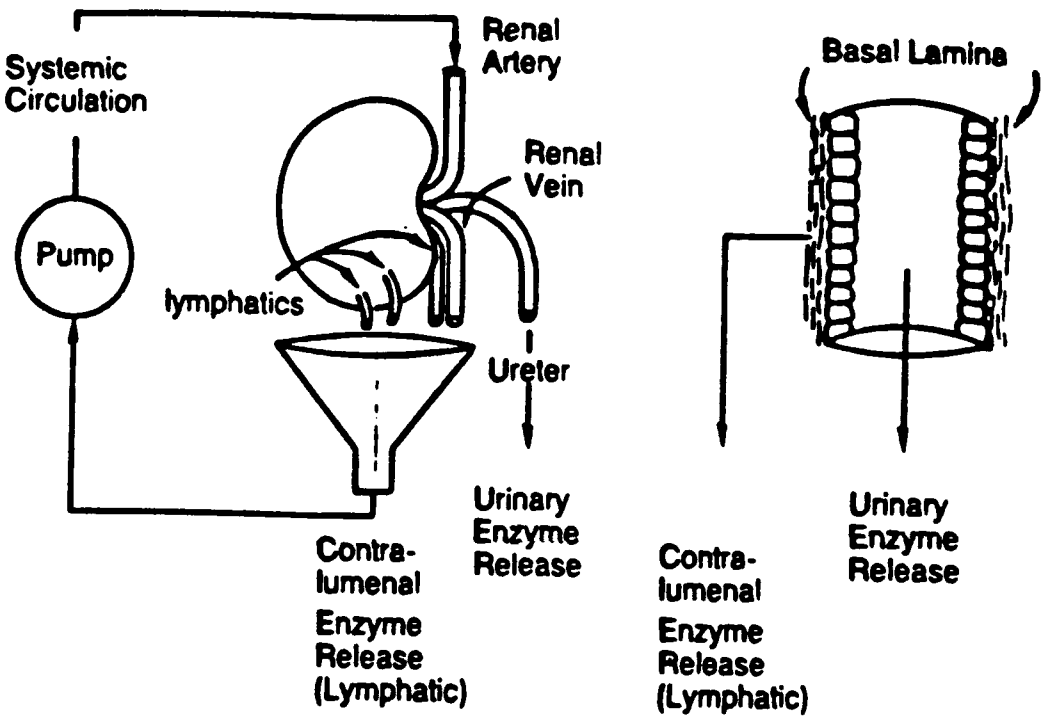


Fig. 2. Simultaneous measurement of release of degradative enzymes from the tubular epithelial cells (a) into the tubular lumen and urine and (b) across the peritubular basal lamina into the interstitium and lymphatic system (see text) (Lockwood, unpublished).

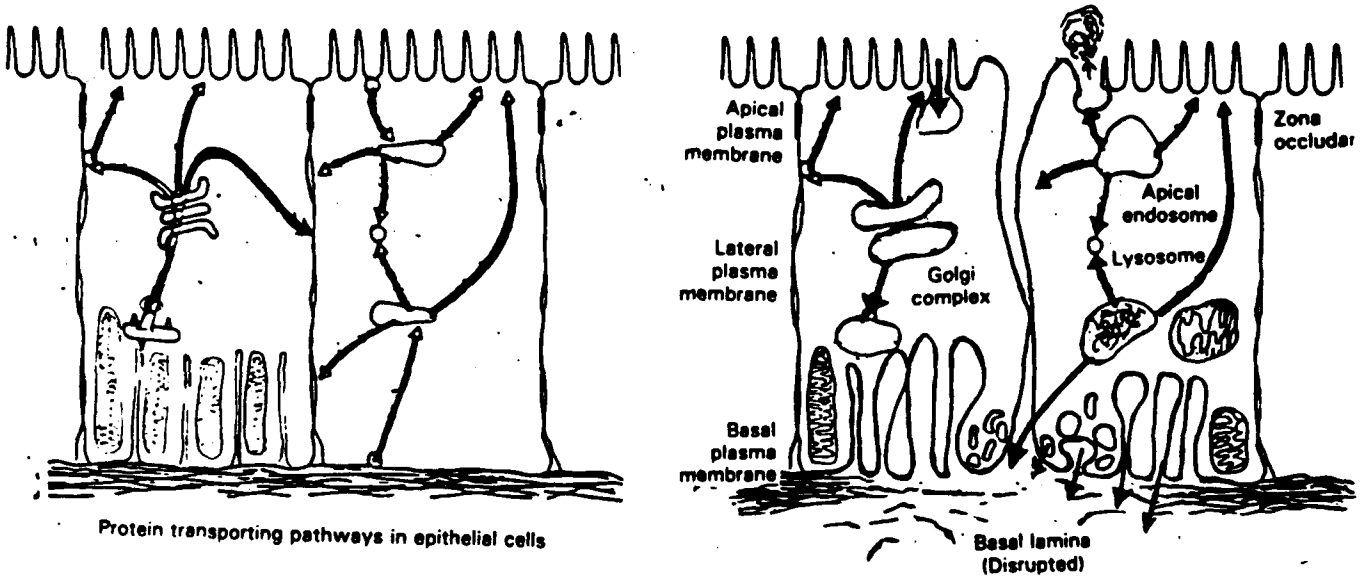
Exposure of animals to toxic compounds. Rats were administered various indicated toxicants by oral gavage as previously described

Electron Microscopy. Methods of transmission and scanning microscopy were as previously described (Churg et al., 1980; Brunner, 1984; Mattie et al., 1986). Kidneys were perfused with fixative in situ in anesthetized rats.

Results

Ultrastructural observations of the lysosomal-vascular system and peritubular basal lamina. Findings revealed by transmission and scanning microscopy are described in the legends of Fig.3 through 6.

Extracell release of lysosomal degradative enzymes from the lumenal and contralumenal surfaces of the tubular epithelial cell during cell injury. In order to produce a controlled progressive injury to the tubular epithelium the kidney was perfused without glucose or any other nutrients in a balanced salt solution. This procedure leads to a depletion of ATP, which is a final common pathway of damage from a wide variety of toxicants. In the nourished kidney it was found that normally, lysosomal degradative enzymes are released into the tubular lumen and urine only. However, within less than 30 min of injury from nutrient deprivation, the acid hydrolases are released in massive amounts unto the peritubular fluid and lymphatic system. The contralumenal release is correlated with the functional decline of the preparation. The details of acid hydrolase release are described in the legend of Figs. 7, 8. This acid hydrolase release is also indicative of several other glycosidases, acid phosphatase, β-glucuronidase and protease activity measured simultaneously (data not shown).



Protein transporting pathways in epithelial cells

Fig. 3. The structure and function of the "polarized" epithelial cell requires maintenance of adhesion to the supporting basal lamina. A description of the function of various components of kidney epithelial cells is found in the review from which the diagrams were redrawn (Rodriguez-Boulan and Nelson, Science 245:718-725 (1989). This information is not repeated here.

We and others have reported that lysosomal acid hydrolases are normally secreted into the tubular lumen from the apical side of the polarized epithelial cell. However, the preliminary unpublished results presented here indicate that these degradative enzymes are not normally released from the basal side of the cell across the peritubular basement membrane. Only after tubular injury are the degradative enzymes released from the isolated perfused kidney into the renal interstitium and systemic circulation (see below).

There are two distinct lysosomal systems in kidney epithelial cells: the apical and basal (reviewed by Rodriguez-Boulan and Nelson, 1989). Our unpublished data from the isolated perfused kidney (shown below) indicated that lysosomal degradative enzymes are normally released from the uninjured kidney only into the tubular lumen and urine. They are not normally released across the basal lamina. The extensive tubular lysosomal system and very high amount of lysosomal degradative enzymes evolved to serve the digestive function of the kidney in degrading endocytosed filtered proteins (see text). After tubular cell injury we hypothesize that the system might work in reverse: the massive extracell enzyme release might explain proteolytic and glycosidolytic degradation of the peritubular basal lamina to which cells normally adhere. If severe, such degradation prevents regeneration of a continuous cell monolayer after remaining cells proliferate following tubular injury. When the kidney becomes injured, tubular epithelial cells release their high amount of degradative enzymes from the basal portion of the cell where they can be rapidly detected in the circulating perfusate of the isolated perfused kidney preparation. This enzyme release presumably explains the degradation of the peritubular basement membrane observed approximately 1 day after tubular injury.

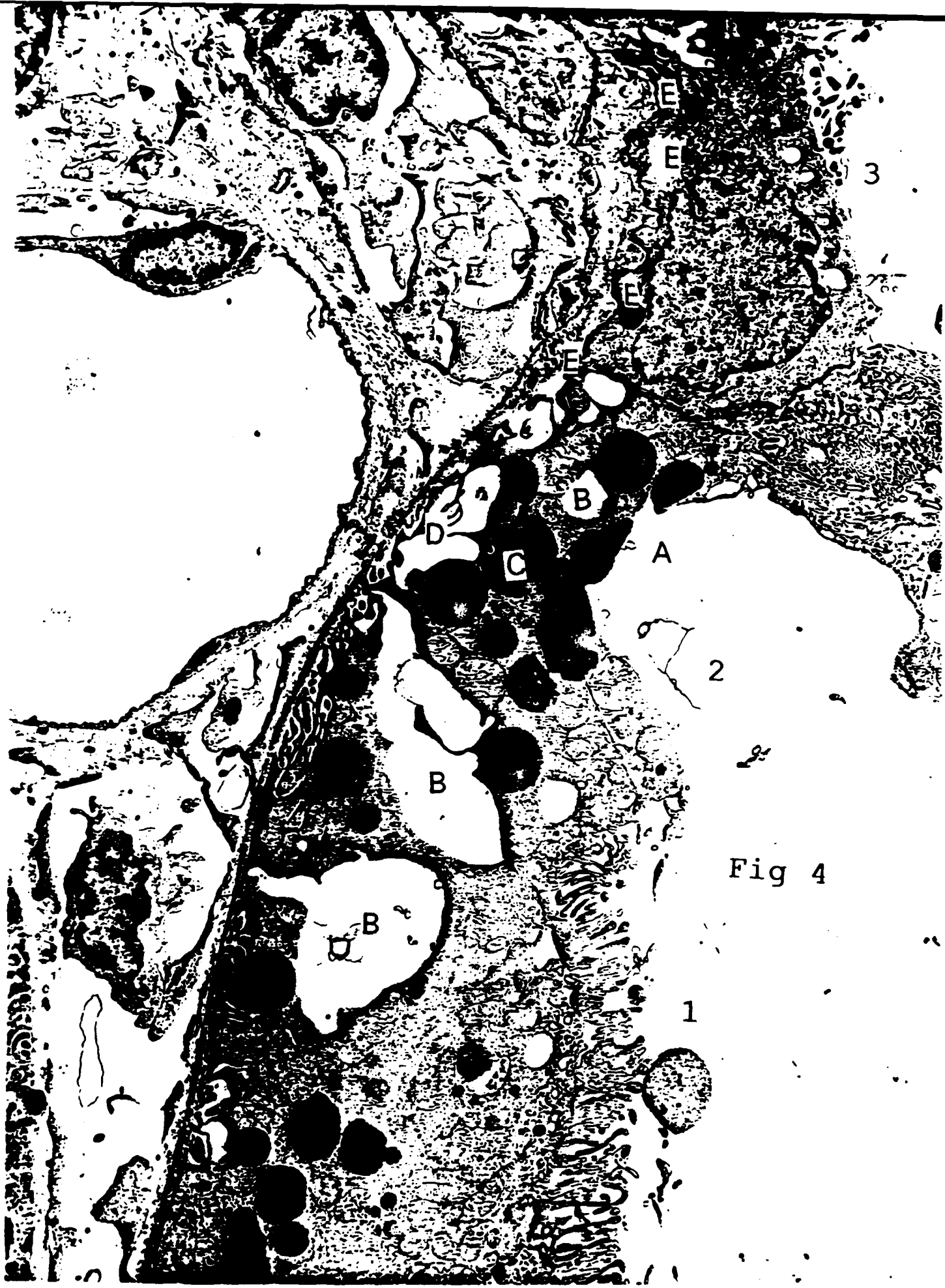


Fig 4

Fig. 4. Injured tubular epithelium after trimethyl pentane. This transmission micrograph shows tubular epithelial cells 24 hr after injury by oral administration of the nephrotoxicant trimethylpentane (2 mL/kg). Three cells are shown with a gradation of injury; cell #1 is moderately injured. Although the brush border of Cell #1 is abnormal, it remains intact. Cell #2 is sublethally injured severely, and has undergone some dedifferentiation in preparation for regenerative proliferation. The brush border (A) of proliferating cell #1 is absent. Cell #3 is severely and perhaps lethally injured. In all cells, the lysosomal system (B) can be seen to discharge their contents into the peritubular space in contact with the peritubular basal lamina. The dark vesicles (C) are known to be endocytosed protein that was filtered through damaged glomeruli. The peritubular basal lamina exhibits a corresponding gradation of injury increasing in severity from bottom to top. As shown by the segment delineated by the "E" is the marked degradation of the basal lamina corresponds to the basal boundary of the severely injured Cell #3.

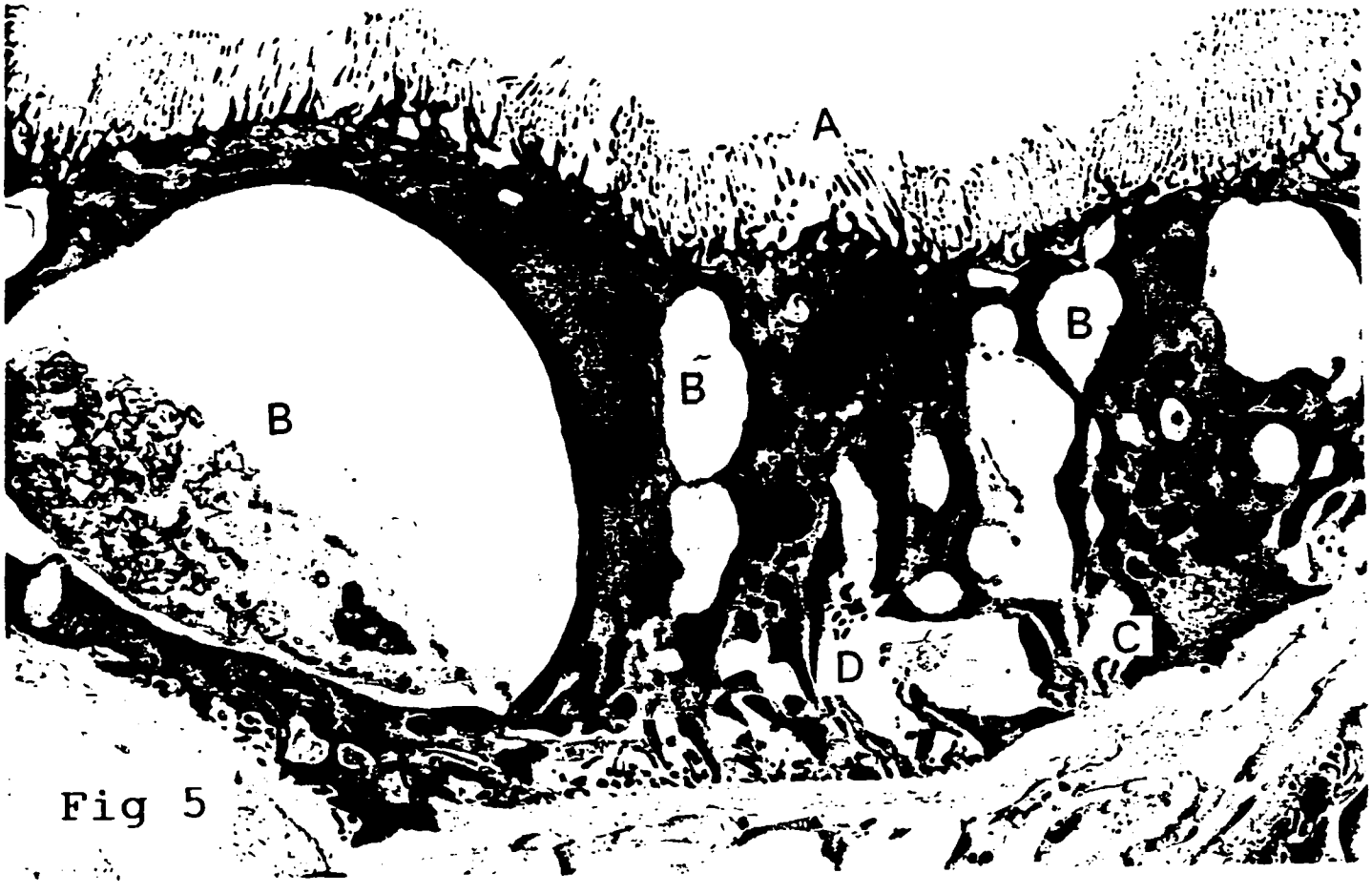


Fig. 5. Enlargement of the lysosomal vacuolar system in association with tubular cell injury. A large autophagic vacuoles (B, left) contains undegraded cell debris. Under uninjured conditions a lesser and slower amount of degradation is a normal part of the turnover of cell constituents. In association with cell injury normal degradation is dangerously accelerated. In this cell basal vesicles can be seen to discharge their contents of lysosomal enzymes by fusion with the basal plasma membrane. Some amount of this basal vesicular fusion can also be observed in normal uninjured cells. This micrograph is shown in order to illustrate that there is no simple correlation of frank observable degradation of the basal lamina with sublethal cell injury; however disruption of the basal lamina is a general finding in severely injured areas.

Fig. 6A. Scanning electron micrograph of disruption of peritubular basement membrane 48 hr after injury with jet fuel ML-H-5606 (mg/kg). Jet fuel 5606 is a mixture of aliphatic and aromatic hydrocarbons including tricresyl phosphate which is extremely nephrotoxic. A frank disruption of the peritubular basement membrane can be observed at the center of the photograph (1,250 x magnification).

Fig. 6B. In the same group of animals as shown in the separate kidney of Fig. 6A. A frank disruption of the peritubular basal lamina can be seen bounded by the "A's". The collagen fibrills surrounding this region are absent from the area of disruption. The size of the disrupted area appears to be the approximate size of one tubular epithelial cell perhaps lethally injured (10,000 X magnification).

Fig. 6C. In the same group of rats a disruption of the basal lamina can be seen within the area bounded by the letters "A". Collagen fibrills are absent over this area. A red blood cell (10 μM) can be observed.

Fig 6B

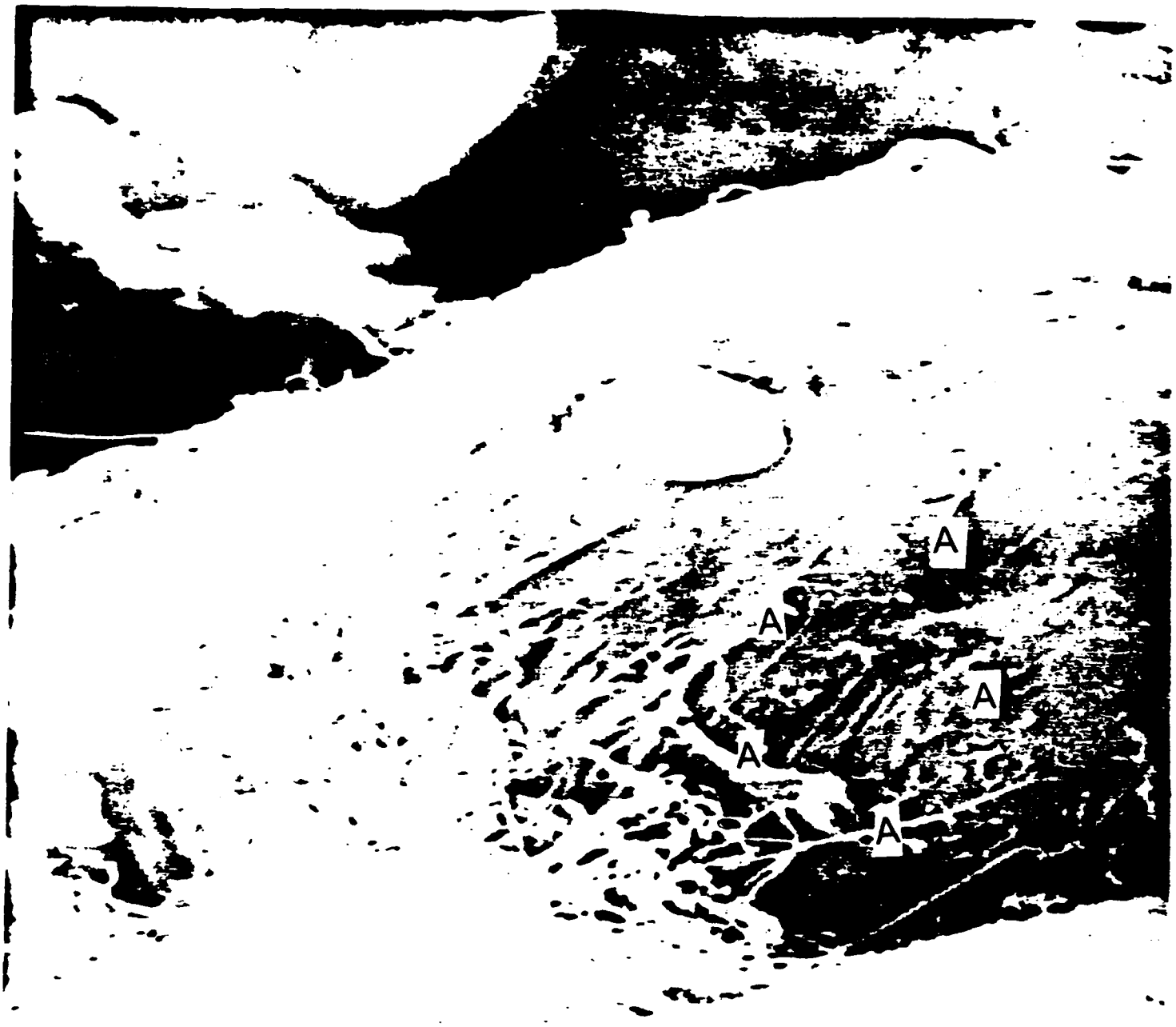


Fig 6 c

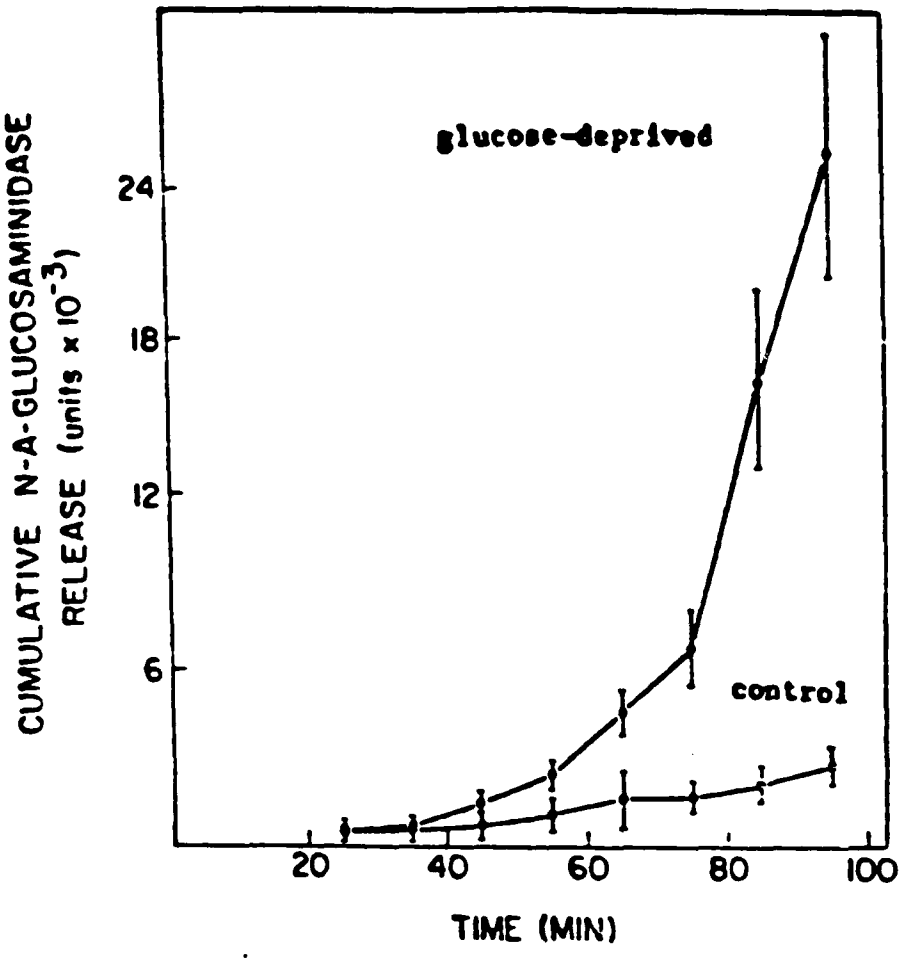


Fig. 7. Cumulative release of renal tubular N-acetyl-β-glucosaminidase into renal interstitium under perfusion with and without glucose and amino acids. Five kidneys were injuriously glucose-deprived by perfusion with glucose and amino acids in the perfusate, and six were perfused with only a balanced ionic salt solution containing 6% bovine serum albumin, as described in the text. Contralumenal release of N-acetyl-β-glucosaminidase through the peritubular basement membrane into renal interstitium and ultimately into circulating perfusate via renal lymphatics is shown for both conditions. The initial rate of release was difficult to determine due to low concentration of activity; however, it was either extremely small or--more probably--zero. The in-vitro longevity of the preparation perfused with glucose and amino acids is less than 2 hr. At 55-65 min, the control kidneys perfused with glucose and amino acids began a slow, gradual increase in enzyme release (bottom trace). In sharp contrast, those initially deprived of these substrates began a precipitous interstitial release at 45 min. By one hour, the mean cumulative contralumenal enzyme release into renal interstitium was one third of the normal basal release into tubular lumen and urine, whereas initially it was negligible. Most of the variability in the means shown is due to differences in the time course of the increase. In three experiments, acid phosphatase, β-galactosidase, and β-fucosidase and acid protease were also measured and exhibited a similar pattern of release. The differences are highly significant. All data are shown as means ± S.E.M.

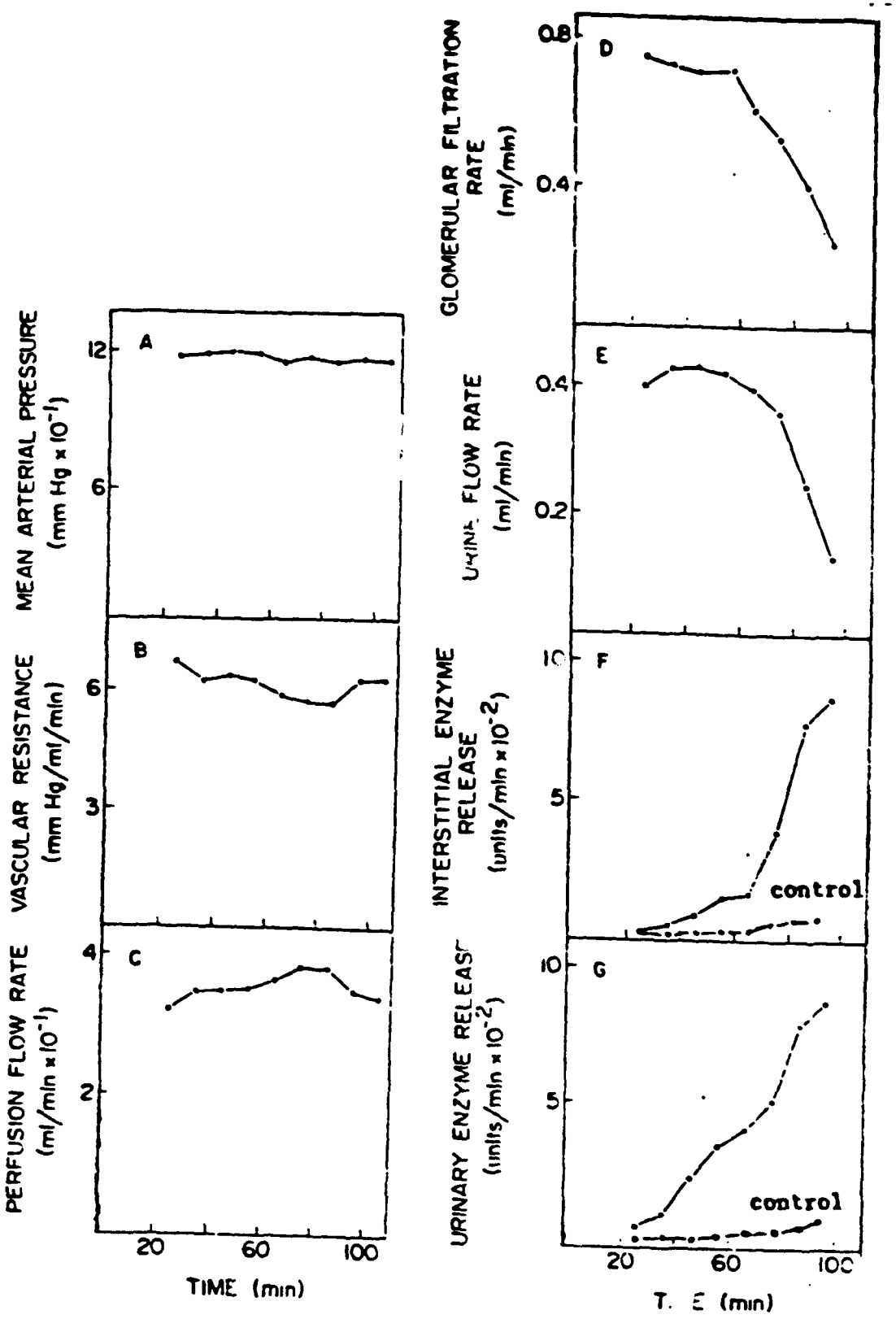


Fig. 8. Correlation between decline in functional parameters and release of lysosomal degradative enzyme into (a) tubular lumen and urine and (b) interstitial space and systemic circulation under progressive tubular cell injury from glucose deprivation. The kidney was perfused and release of

lysosomal enzymes was measured as described in the text. The perfusate normally contains glucose and amino acids, and a balanced Krebs salt solution with 95% O_2/5% CO_2 (37°C). In this preparation a gradual tubular cell injury was induced by omitting all metabolic substrates from the perfusate. Glucose deprivation was found to produce a gradual injury with convenient time course in comparison with the in vitro longevity of the preparation. Initial sodium reabsorption was 30% less than controls; initial urine flow was correspondingly 30% greater than controls (not shown).

A. Mean arterial pressure. Mean arterial pressure is the constantly regulated variable; however, it characteristically varies little and does not require appreciable adjustment as demonstrated by the constancy of resistance.

B. Vascular resistance. Vascular resistance is the ratio of mean arterial pressure to perfusate flow rate. In repeated experiments the fluctuations shown are random and do not characterize a uniform pattern.

C. Perfusate flow rate. Perfusate flow rate varies as a function of vascular resistance since mean arterial pressure is regulated experimentally constant. The fluctuations shown are random in repeated experiments.

D. Glomerular filtration rate. Glomerular filtration rate was measured by the clearance of inulin and assumes that the tubule remains impermeable and all filtered inulin is collected in urine. If the tubule becomes permeable to inulin and part of the filtrate leaks back into the renal interstitium and circulating perfusate, the actual decline in glomerular filtration rate could be considerably less.

E. Urine flow rate. Total urine output was collected at 10 min intervals.

F. Rate of release of tubular N-acetyl-β-glucosaminidase and other degradative enzymes into circulating perfusate (interstitial enzyme release). Tubular release of the lysosomal glycosidase enzyme into circulating perfusate represents contralumenal release through the peritubular basement membrane into renal interstitium and movement of the enzyme through capsular and hilar lymphatic vessels out of the kidney into the circulating perfusate (see diagram, Fig. 2). The average initial rate of contralumenal release is zero, whereas a large basal urinary release does exist. By 90 minutes the rate of contralumenal degradative enzyme release was approximately equal to the rate of release into tubular lumen and subsequently urine. Speculatively, this probably indicates a disruption of the peritubular basement membrane of unknown characteristics. The other acid hydrolases, β-fucosidase, β-galactosidase, acid phosphate and acid protease followed a similar pattern of release. The control enzyme release from uninjured preparations with glucose is indicated.

G. Rate of release of N-acetyl-β-glucosaminidase into urine. The basal rate of release from both in-situ kidney and the isolated perfused kidneys provided with glucose and amino acids was similarly 600 ± 150 units, indicating that the urinary enzyme is derived entirely from the kidney. In this experiment the rate of release after 90 min of substrate deprivation is 12 times normal. In addition to the other acid hydrolases listed above, acid protease activity was assayed in urine and all of these followed a similar pattern of release.

Abbreviated Discussion

Biochemical etiology of tubular conduit degeneration following tubular cell injury. Present results provide a novel description of the enzymatic basis for the degeneration or regeneration of the nephron following toxic injury to the tubular epithelial cells. Maintenance of the peritubular basal lamina is critical to the regeneration of the conduit structure of the nephron. The exact kinetics of normal turnover of the peritubular basal lamina constituents are unknown. However it is likely that low level extracellular release of tubular degradative enzymes is responsible for this turnover of the basal lamina. When injury causes massive elevation of extracell release of degradative enzymes the normal degradative process is pathogenically accelerated. Although ultrastructural evidence of degradation of the basal lamina cannot be clearly observed for approximately 12-24 hr, the involved degradative enzymes are released in massive amounts across the basal lamina into lymphatics almost simultaneously with the onset of injury. Thus the novel molecular etiology we describe precedes the subsequent ultrastructural evidence. Acute renal failure can be established within 1-2 hr of an acute toxic injury to the kidney.

Ultrastructural correlates of the release of degradative enzymes from the injured tubular epithelium. The ultrastructure of the peritubular basal lamina after injury has not previously been studied. It is presently unknown whether the absence of contralumenal release of lysosomal enzymes is due to the impermeability of the surrounding basal lamina to the normal amount of enzymes released from uninjured cells or alternatively to a true absence of release from the basal cell surface. After injury, enzymes released from the apical portion of the cell might also be found in the lymphatic system of the perfused kidney if the tubule is frankly disrupted. Following injury the peritubular basal lamina is probably degraded partially long before ultrastructural evidence can be seen by microscopy.

Present findings integrate biochemical functional and ultrastructural descriptions of the critical lesion determining degeneration vs regeneration of the nephron after injury.

Acknowledgments

David Mattie, Ph.D., and other personnel of the Toxic Hazards Division of the AAMRL contributed the transmission scanning micrographs. Dr. Christopher Jones, Ph.D., contributed kidney tissue Dr. Ann Taylor, Ph.D., Department of Anatomy and Alan Rice, Department of Pharmacology contributed scanning electron micrographs.

Abbreviated References

Bruner, R.H. Pathologic findings in laboratory animals exposed to hydrocarbon fuels of military interest. In: Renal Effects of petroleum hydrocarbons MA Mehlman ed. Princeton-Scientific Publishers Princeton. pp. 130-140 (1984).

Bruner, R.H., and Pitts, U. Nephrotoxicity of hydrocarbon propellants to male Fisher-344 rats. Proc. 13th Conference in Environmental Toxicol. J.A. Hurst ed. Air Force Aerospace Medical Res. Laboratory, Wright-Patterson Air Force Base. pp. 337-349 (1983).

Buckley, L.A., Clayton, J.W., Bagle, R.B., and Gandotti, A.J. Chlorotriflouroethylene nephrotoxicity in rats: A subacute study. Fund. Applied Toxicol. 2:1810186 (1982).

Churg, J., Spargo, B.H., Jone, D.B. Diagnostic electron microscopy of renal diseases. In: Diagnostic electron microscopy. Ed. B.F. Trump, and R.T. Jones. John Wiley & Sons. New York. pp. 203-315 (1980).

Clayton, J.W. Toxicology of the Flouroalkenes: Review and research needs. Environ. Health Perspective 21:255-267.

Cook, E.W., and Pierce, J.A. Toxicology of Flouro-oletins. Nature 242:337-338 (1973).

Gad, S.C., Rusch, G.M., Rugle, K.S., Darr, R.W., Hoffman, G.M., Peckham, J.C., and Schardein, J.C. J. Am. College of Toxicol. 7:663-674 (1988).

Kanerva, R.L., Ridder, G.M., Stone, L.C. and Alden, C.L. Characterization of spontaneous and decalin-induced hyaline droplets in kidneys of adult male rats. Food Chem. Toxicol. 25:63-82 (1987).

Ingber, D.E., Madri, J.A., and Jamieson, J.D. Basement membrane as a spatial of polarized epithelia. Am. J. Pathol. 122:129-139 (1986).

Kinkeod, E.R., Culpepper, B.T., Henry, S.S., Szolak, P.S., Flemming, C.D., Kutzman, R.S., Brunner, P.H., Wyman, J.F., and Mattie, D.R. The determination of the acute and repeated oral toxicity of halocarbon oil series 27-S. Contract report AAMRL-TR-89-007, NMRI 88-1b.

Little, J.R., and Cohen, J.J. Effect of albumin concentration on function of the isolated perfused kidney. Am. J. Physiol. 226:512-517 (1974).

Mattie, D.R. The comparative toxicity of operational Air Force hydraulic fluids (Personal Communication, 1989).

Mattie, D.R., Chase, M.R., and Hobson, P.W. The effect of 2,3,4 trimethylpentase on the ultrastructure of kidneys from normal vs castrated male rats. Proc. 44th Int. Meeting of the Electron Microscopy Soc. America p. 356-357 (1986).

Mehlman, M.A., Hemstrut, G.P. Thorpe, J.J., Weaver, N.N. eds. Renal effects of petroleum hydrocarbons. Princeton Scientific Publisher, Princeton (1984).

Norton W.N., and Mattie, D.R. The cytotoxic effects of trimethylpentase on rat renal tissue. Scanning Microscopy 1:783-790 (1987).

Norton, W.N., Mattie, D., and Kearns, C.L. Cytopathologic effects of specific aromatic hydrocarbons. Am. J. Pathology 118:387-397 (1985).

Plaa, G.L. Experimental evaluation of haloalkanes and liver injury. Fund. and Applied Tox. 10:563-570 (1973).

Potter, C.L., Gandolfi, A.J., Nagle, R.B., and Clayton, J.W. Effects of inhaled chlorotriflouroethylene and hexaflouropropene on the rat kidney. Toxicol. Appl. Pharmacol. 2:181-186 (1981).

Rodriguez-Boulan, E., and Nelson, W.J. Morphogenesis of the polarized epithelial cell phenotype. Science 245:718-725 (1989).

1990 RESEARCH INITIATION GRANT

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

Parametric Studies of the Breakdown of Total Information Processing Time into During-Display and Post-Display Components

Prepared by:	Ethel Matin, Ph.D.
Academic Rank:	Professor
University:	Long Island University
Research Location:	Department of Psychology, Post Campus, LIU

Date: 27 December, 1990

Parametric Studies of the Breakdown of Total Information Processing Time into During-Display and Post-Display Components

Ethel Matin

Abstract

Information processing speed for data frames accessed with and without saccadic eye movements was studied in three factorially designed experiments using performance measures that permit the separation of total processing time into during-display and post-display components. For all subjects in all experiments, the results showed a significant reduction in total time when the need for saccadic eye movements was eliminated by presenting the data in temporal succession in one spatial location. The time saved with this serial presentation method increased with the number of data frames and decreased with the amount of information per frame. For a range of frame durations that depended on the amount of information per frame, the results showed a partial tradeoff between the during-display and post-display components of total time, with decreases in post-display time to compensate for increases in during display time. Subjects generated a specifically predicted "waste time" function for frame durations that extended beyond the tradeoff region. Within the tradeoff region there was an optimal frame duration that was usually longer than the minimum required for performance at the 85% criterion employed in these studies.

Acknowledgements

This research was supported by a Research Initiation Grant from the Air Force Office of Scientific Research to Long Island University. I thank Frank Ancona, Laura Cohen, and Gregory Viscovitch, who participated in the studies as candidates for M.A. degrees in Experimental Psychology. Working with them was fun and I appreciate their many thoughtful contributions. Finally, thanks to my Air Force colleague, Dr. Kenneth Boff. We shared many exciting discussions during our collaboration on the "Rap-Com" effort.

I. INTRODUCTION

The studies described in this report were designed to address some of the questions raised by an experiment on serial visual displays that I performed as an AFOSR Faculty Research Fellow in the summer of 1989. Accordingly, I will introduce the new research with brief descriptions of the serial display concept and of the earlier "fellowship experiment".

A human-computer interface technology based on the serial presentation of visual information is being developed in the DEfTech Laboratory at Wright-Patterson Base under the direction Dr. Kenneth Boff, my Air Force colleague. In addition to exploring the scientific basis for the technology, Dr. Boff and I applied for a patent, which was awarded to the Air Force last year (Matin & Boff, 1989). In the summer of 1990 this work was reviewed by the Technology Transfer Project team at the University of Texas, San Antonio. They recommended the development of a commercialization plan for the technology, which we have dubbed RAP-COM (for Rapid Communication Display).

Essentially, a serial display is characterized by the presentation of single "eye-fulls" of data in temporal succession in one spatial location. It offers three potential advantages as a design option:

1. Reduced display space requirements. A serial display with N frames requires only 1/N of the space that would be

needed with a conventional, spatially distributed format.

2. Increased rate of information transfer. An increased rate is possible for some types of information because the serial format eliminates the saccadic eye movements normally used to scan an array of displays.

3. Designer control over information sampling sequences. In nuclear power plants and other process control environments, operators monitor large numbers of displays to determine when human intervention is required. All of these displays must be sampled periodically, some more frequently than others, depending on the bandwidth of the information. With conventional displays, the designer's task is to locate them in ways that will encourage optimal monitoring sequences. With a serial display, the optimal sequence can be built into the display itself.

In the course of our research on the serial format, Dr. Boff and I developed a measure called the "frame duration threshold" (Matin & Boff, 1990). Briefly, the duration threshold is the time that a frame must be presented for a given level of performance (e.g., 85% correct responding). By computing the difference between the threshold for information presented as a spatially distributed array of frames and the threshold for an otherwise identical serially presented array, we obtained a measure of the overhead that can be attributed to saccadic eye movements.

We used the duration threshold measure to study a variety of information processing tasks and found that a

substantial amount of time (approximately 100 ms/saccade) can be saved with serial displays for information that can be processed at rates of approximately four items per second or greater (Matin & Boff, 1988; Matin & Boff, submitted; Matin, Boff, & Donovan, 1987; Osgood, Boff, and Donovan, 1988). However, it seemed possible that this measure tapped the perceptual/ encoding/ memory aspects of information processing but not the entire cognitive processing, which could have continued after the last data frame had disappeared. I addressed this question in the fellowship experiment by comparing total processing time (onset of display to response) in serial and spatially distributed visual displays with three frames per display and a single digit (2-9) in each frame. The subject was required to process that information by counting the number of odd digits that appeared and then making a binary response on a computer keyboard (one key for ODD, another for EVEN, depending on whether the number of odd digits was odd or even - e.g., for the three digits 3, 2, 4, there is one odd digit and the correct answer is ODD). The data were collected in a way that permitted calculation of the total processing time for the three frame display and also allowed for the division of total time into two components, during-display time and post-display time. To get a better understanding of this breakdown, I examined the effect of the subject's strategy by studying performance with two different operating modes - emphasis on making the display time as short as possible or emphasis on responding as rapidly as possible after the display disappeared. For both modes, subjects were asked to

remember the overall goal: keep the processing time as short as possible for both the serial and the spatially distributed formats.

For all subjects and for both operating modes, the results of the experiment provided an unequivocal answer to our primary research question: for each subject, total information processing time was significantly less with the serial format. In addition, with both serial and spatially distributed displays, subjects were able to exercise some control over the total time, allotting a relatively long during-display time and a correspondingly short post-display time with operating mode 1 (minimize during-display time) and the converse with mode 2 (minimize post-display time). A high negative correlation, accounting for approximately 80% of response variance, was found between the two processing components, as would be expected if there is a tradeoff between them in accomplishing the total task. Some subjects were more efficient (shorter total processing time) with the mode 1 emphasis, while others were more efficient with mode 2 (Matin, 1989; Matin & Boff, 1990).

II. OBJECTIVES OF THE RESEARCH EFFORT

The first objective of the present research was to test the generality of the above findings with digit processing in a three frame display by considering tasks other than digit processing and displays with various numbers of frames. This objective was implemented in Experiments 1 and 2, for which we used the methodology

developed for the fellowship experiment. The experiments were designed to maximize the basic science yield. However, information about these variables should be useful for designers, who will have to decide whether and when to use a serial display, and how many frames to incorporate in it.

The second objective was to study the effect of frame duration on processing efficiency. This objective was implemented in Experiment 3, for which we developed a new methodology. Both the methodology and the information that we expect to obtain with it are interesting from a scientific perspective. Moreover, we expect the results to be useful in applied settings when decisions are made about frame durations for serial displays.

III. METHOD AND RESULTS

All experiments were run with an IBM XT microcomputer with an enhanced graphics adapter card, an IBM data acquisition card, a standard IBM keyboard, and an enhanced color display (EGA). The computer presented stimuli, recorded responses, and stored the data on disk for later statistical analysis. It was programmed in BASIC with 8088 Assembly language routines for display timing, synchronization with the 60 Hz raster, and measurement of the post-display response time.

In all experiments, subjects ran in a serial mode and a spatial mode. For the serial mode, the data frames appeared in temporal succession in a single window,

obviating the necessity for saccadic eye movements. For the spatial mode, the frames were accessed by making saccades between two windows centered vertically on the screen and separated horizontally by 11 deg (visual angle). Except for the number of windows (and the consequent need for saccadic eye movements in the spatial condition), the serial and spatial formats were identical. This ensured that the results could lead to a clear interpretation of the saccadic overhead, uncomplicated by other factors such as data access through peripheral vision or regressive saccades at long frame durations.

Experiment 1

Purpose. As one approach to extending the generality of the findings of the fellowship experiment, we studied the effect of number of frames on the relative efficiency of serial and spatially distributed displays and on the division of total information processing time into during display and post-display components.

Method. For both the serial and the spatial conditions the task was essentially identical to the digit processing task with binary response employed in the fellowship experiment. However, rather than viewing three frames/trial throughout the experiment, subjects viewed two frames in some runs, four in others, and six in still others. In all cases, each frame contained one digit (in the range 2-9). Regardless of the number of frames, the task was to count the number of odd digits and to respond

ODD if this count yielded an odd number, EVEN otherwise. The keys V and M on the IBM keyboard were labelled EVEN and ODD and used to enter the subject's responses. For all conditions, subjects were instructed to attempt to minimize total processing time. To initiate a trial they hit one of the two response keys while looking at a small window centered on the monitor. They kept their two index fingers on the keys, ready for the binary ODD/EVEN response at the end of the trial.

For the serial condition, which eliminated the need for saccadic eye movements, the digits were presented successively as individual frames in one window (2, 4 or 6 frames, depending on the level of the number of frames factor for the run). Each frame for a trial was presented for the same duration, which was chosen by the procedure described in the last paragraph of this method section. Feedback - the subject's response (O or E) paired with the correct response (O or E) - was presented for 800 ms just above the viewing window immediately after the subject entered the response. Subjects were free to initiate the next trial as soon as the feedback disappeared.

For the spatial condition, there were two windows, centered vertically on the screen and separated horizontally by 11 deg (visual angle). The first digit appeared in the left window and the last digit (as well as the feedback) appeared in the right window. Depending on the number of frames for the run, the subjects made 1, 3, or 5 saccades to access the information (e.g., for the

4 frames per trial condition, the first digit appeared in the left (fixation) window and was accessed without an eye movement; subjects then made a saccade to the right window to get the second digit, a second saccade to the left to get the third digit, and finally a third saccade to the right to get the fourth digit).

At the beginning of each run, subjects were informed about the viewing condition (serial or spatial) and about the number of frames (2, 4, or 6). In the first half of the run, the computer "tracked" the frame duration threshold for the conditions of the run (it measured the frame duration needed for 85% correct performance - see Matin & Boff, 1990, for details). This usually required 50-60 trials. Immediately thereafter, the subject received another 40 trials, for all of which the frame duration was fixed at the threshold. Post-display response time (which was negative if the response occurred before the last frame had disappeared) was measured on each trial in both parts of the run. However this measure was only analyzed for the 40 trials at the fixed, computer-selected threshold duration. In short, for the purposes of the present analysis, each experimental run yielded three numbers: the duration threshold, the mean post-display time for trials run at the threshold (based on trials with correct responses only), and the % correct responding for these 40 trials. Each run took approximately 5-7 minutes and subjects relaxed a few minutes between the four runs that constituted an experimental session.

Design. Three subjects ran in a completely balanced 2 X 3 factorial experiment. Factor 1 was the display condition (serial or spatial) and Factor 2 was the number of frames (2, 4, or 6). The dependent measures were total processing time and its two components, during-display time and post-display time. One value of the number of frames variable was used throughout an experimental session, which consisted of two serial and two spatial runs. If any of the four scheduled runs yielded a % correct less than 80% or greater than 90%, the run was repeated. Over the course of the 12 experimental sessions, serial and spatial runs were counterbalanced to control for position within the four-run session. To control for possible learning effects during the 12 session experiment, each value of the frames per trial variable appeared three times, once during sessions 1-3, once during sessions 4-6, etc. In all, there were eight estimates of the values of the dependent variables for each of the six experimental treatments. Subjects practiced extensively to stabilize performance before the formal data collection began. With few exceptions, there was only one experimental session per day.

Results and Discussion. Figure 1 on the next page shows the total processing time as a function of number of frames for the three subjects for the serial and spatial conditions, respectively. (For this figure and throughout the remainder of the paper, the subject's identification code will consist of a number for the experiment, and a letter for the particular subject within the experiment - i.e., 1A, 1B, and 1C for the three subjects in Experiment

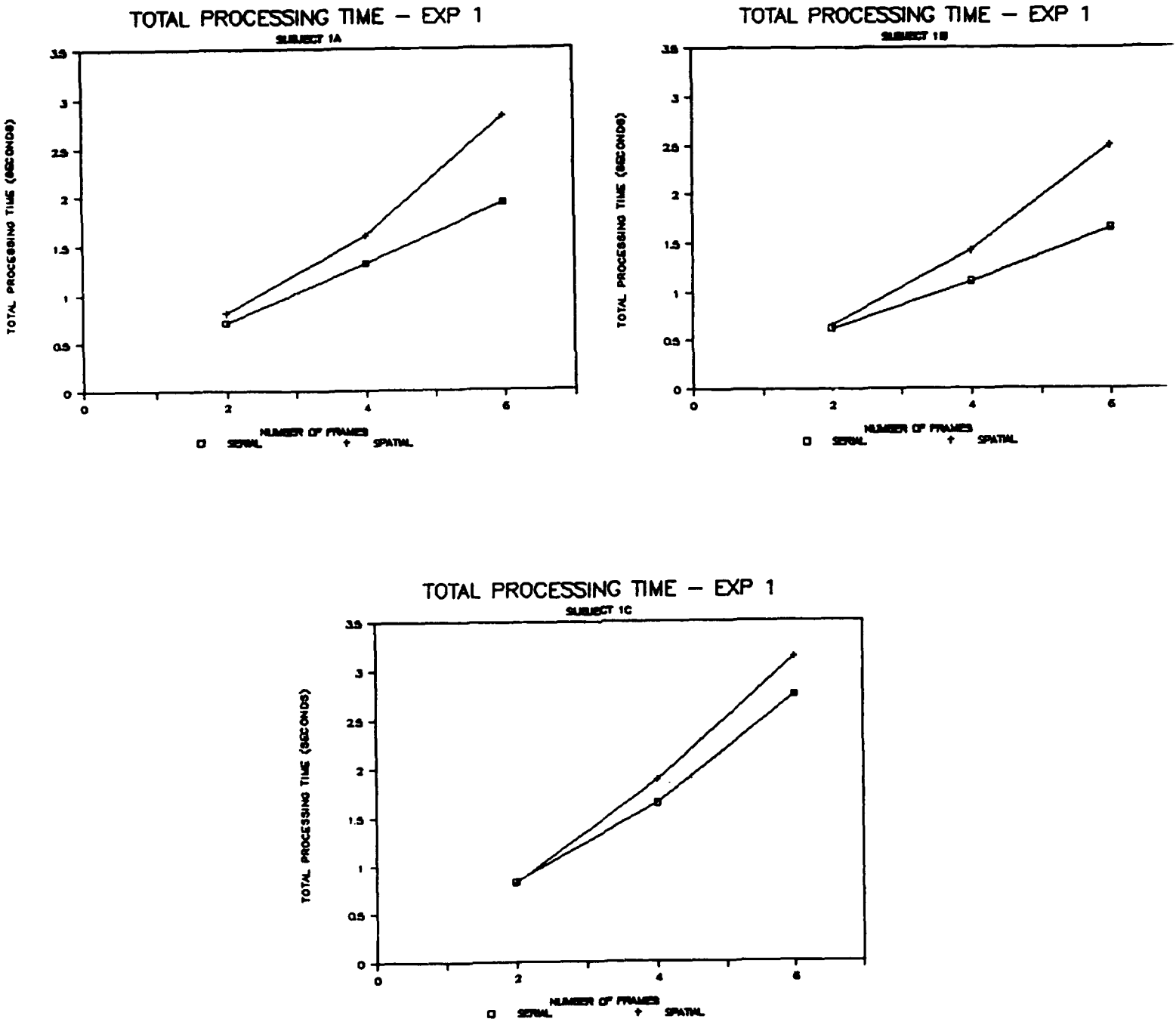


Figure 1. The results of Experiment 1. Total Information processing time is plotted as a function of number of frames for the serial and spatial formats.

1).

For all subjects, these results show an increase in the serial format's speed advantage as the number of frames increases (statistically, an interaction between the serial and spatial functions - see Table 1).

Table 1

Analysis of Variance for the Total Processing Time - Exp 1

	Subject		
Source	1A	1B	1C
Display Format			
F	195.42	292.89	23.10
p	<.000	<.000	<.000
Number Frames			
F	944.74	1291.46	819.40
p	<.000	<.000	<.000
Interaction			
F	61.85	104.93	7.41
p	<.000	<.000	<.002

Qualitatively, this result was expected: In effect, the part of total processing that can be attributed to time for the odd/even decision plus motor response should be a diminishing proportion of total time as the number of frames increases. See Figure 2 on the next page, which shows that the expected trend t ward a diminishing proportion for the post-display time was found in the data.

For subjects 1A and 1B the magnitude of the increase in the difference between the serial and the spatial functions wa

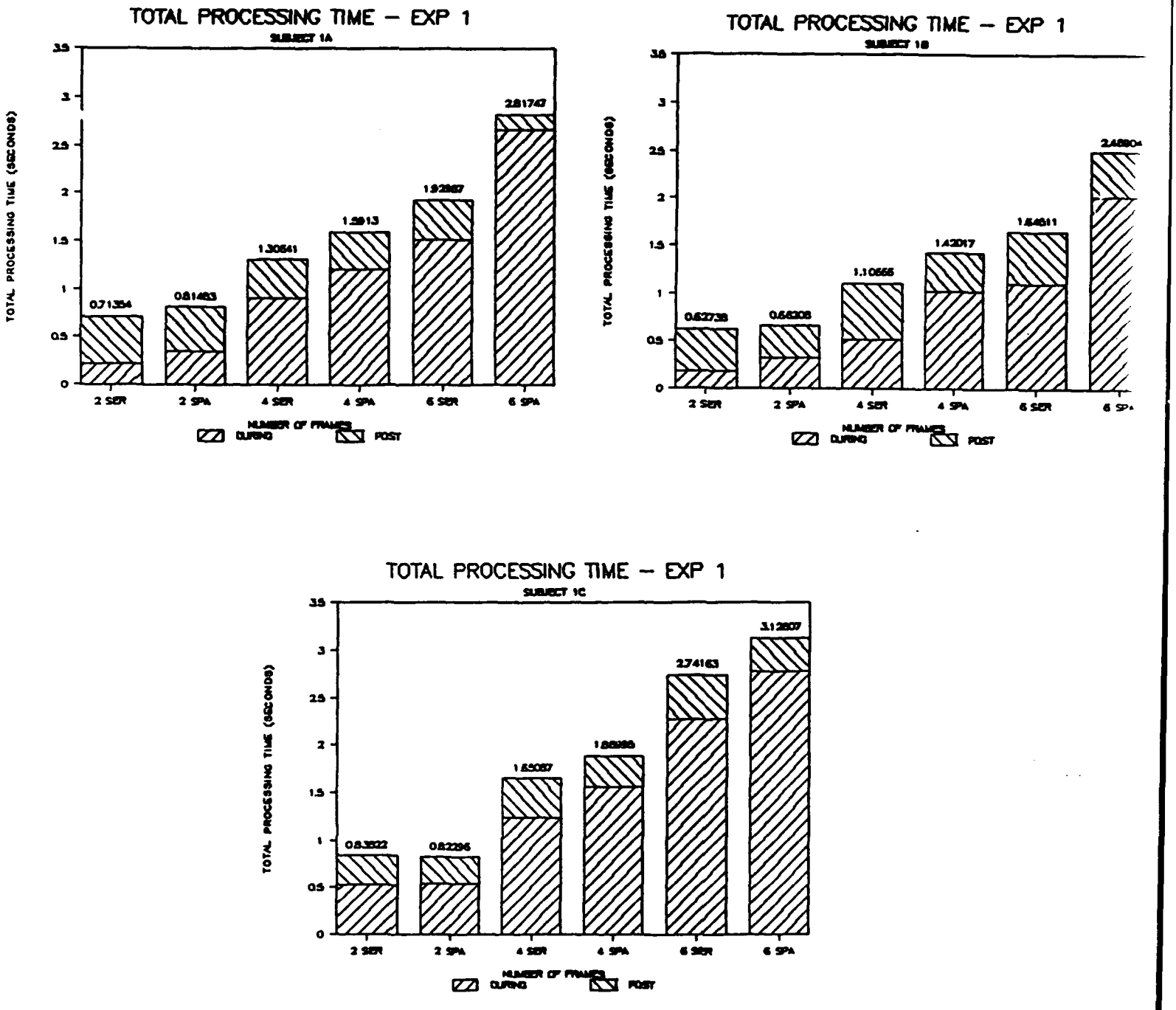


Figure 2. The results of Experiment 1. Total processing time for the serial and spatial conditions is broken down into during-display and post-display components.

considerably larger than we anticipated. Our expectations were conditioned by the following considerations: For the serial display we predicted a linear increase in total time as the number of frames increased. The slope of this function could be interpreted as the processing time/frame, and its intercept as the decision/response component of total processing time. For the spatial function, we expected a linear increase in total time with number of frames that reflected two components: time to process the frames plus time to make the saccades. The first component - time to process the frames - was presumable similar in the serial and spatial functions. Accordingly, the difference between the two slopes should reflect the time to make the saccades, which increases with the number of frames. However, the saccadic overhead - (spatial time - serial time)/number of saccades) was expected to be a constant of about 100 ms/saccade. In fact, the overhead increased with the number of saccades. This trend can be seen in Figure 1. To make it more obvious, however, we plotted the saccadic overhead as a function of the number of saccades (with 1, 3, and 5 saccades for the 2, 4, and 6 frame conditions - the first frame in the spatial condition appeared in the fixation window and required no saccade). This plot (Figure 3 on the next page) shows a 25 ms increase in overhead per saccade, not the constant overhead that was expected.

Experiment 2

Purpose. To extend the generality of our previous findings with the odd/even digit processing task we used a monitoring task with a binary "within-bounds", or "out-of-bounds"

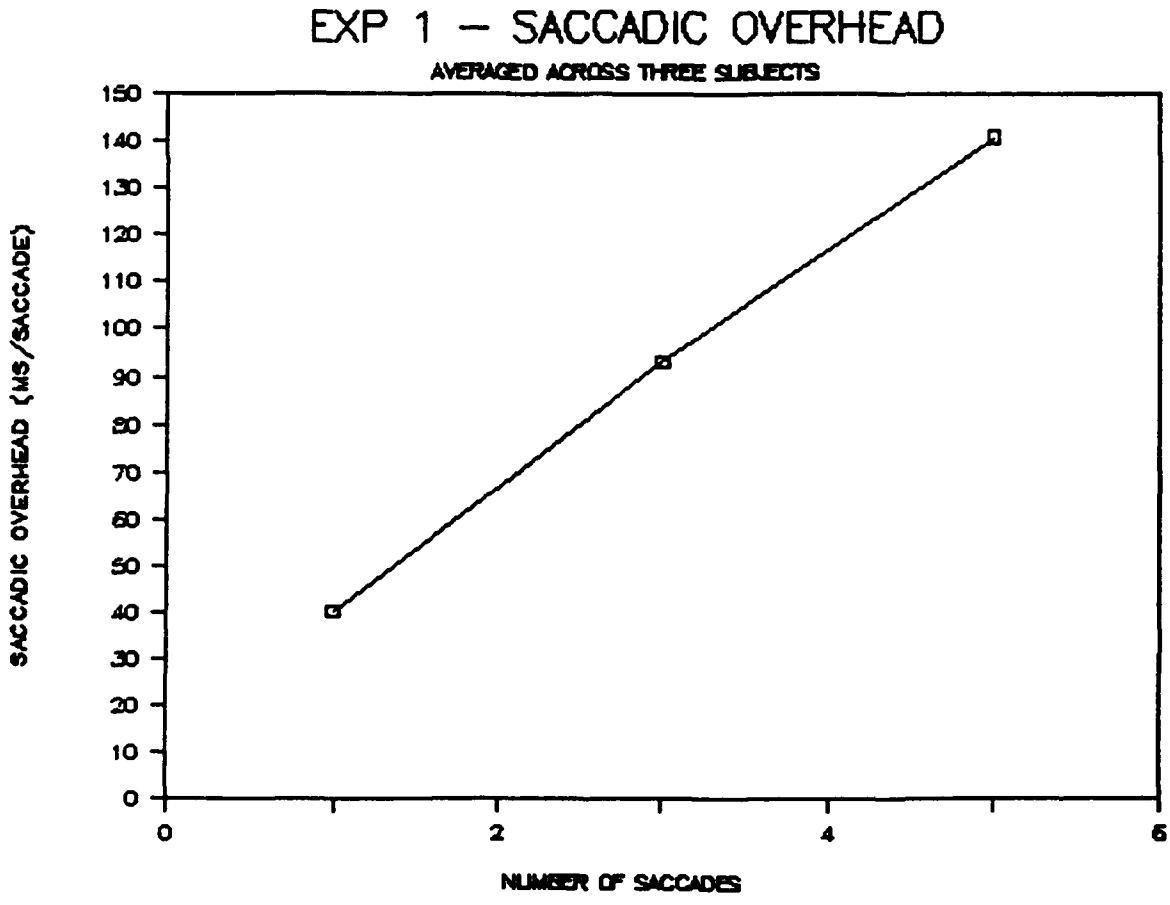


Figure 3. The results of Experiment 1, plotted to show the increase in saccadic overhead (ms/saccade) with increase in number of saccades. The data were averaged over all subjects for this plot.

response depending on the information presented in a sequence of frames. Two levels of task difficulty were employed to clarify some earlier findings which suggested that the serial display's speed advantage is limited to information frames that can be processed rapidly (Matin & Boff, submitted).

Method. For both the serial and the spatial conditions, subjects viewed three successively presented frames. The first frame contained a randomly selected two-digit random number in the range 30 through 50. Another 2-digit random number between 20 and 40 was presented in the second frame. The third frame contained the sum of the numbers in the first two frames plus or minus 3 for "in-bounds" trials, and plus or minus 7 for "out-of-bounds" trials. The subject was instructed to compare the sum of the numbers in the first two frames to the number in the third frame, and then to respond on one key for in-bounds trials and on another key for out-of-bounds trials. Two levels of task difficulty were employed, easy and difficult. For the easy condition, at least one of the two digit numbers ended in a zero (e.g., 20 + 15). For the difficult condition, neither number had a zero as its second digit (e.g., 32 + 24). Accordingly, subjects were required to add the digits in the tens place and also in the ones place for the difficult condition, while they only had to add the digits in the tens place for the easy condition.

With the exception of the task itself, the procedure in Experiment 2 was otherwise identical to the procedure employed in Experiment 1. All subjects practiced extensively

before formal data collection began.

Design. Three subjects ran in a completely balanced 2 X 2 factorial experiment. Factor 1 was the display condition (serial or spatial) and factor 2 was task difficulty (easy or difficult). Total processing time and its two components, during-display time and post-display time were the dependent measures. Every treatment (spatial easy, serial easy, spatial difficult, serial difficult) was run as a block in each of the eight experimental sessions to control for possible improvements in performance as the experiment progressed. In addition, the position of a given condition within the four-block session was counterbalanced across sessions to control for possible performance differences due to block position within a session. In total, eight samples of the dependent variables were obtained for each of the four treatments. With few exceptions, only one session was run per experimental day.

Results and Discussion. The results are presented graphically in Figure 4 on the next page. For each of the three subjects, the serial method resulted in a significantly faster total processing time for both the easy and the difficult task. Moreover, the effect of task difficulty was significant in all subjects. Averaged across all subjects and all treatments, the time saving with the serial format was 179 ms, for a saccadic overhead of 89.5 ms/saccade (179/2 because the three frames in the spatial condition were accessed with 2 saccades).

For each subject, a two-way ANOVA showed highly significant effects for both experimental factors ($p <= .001$).

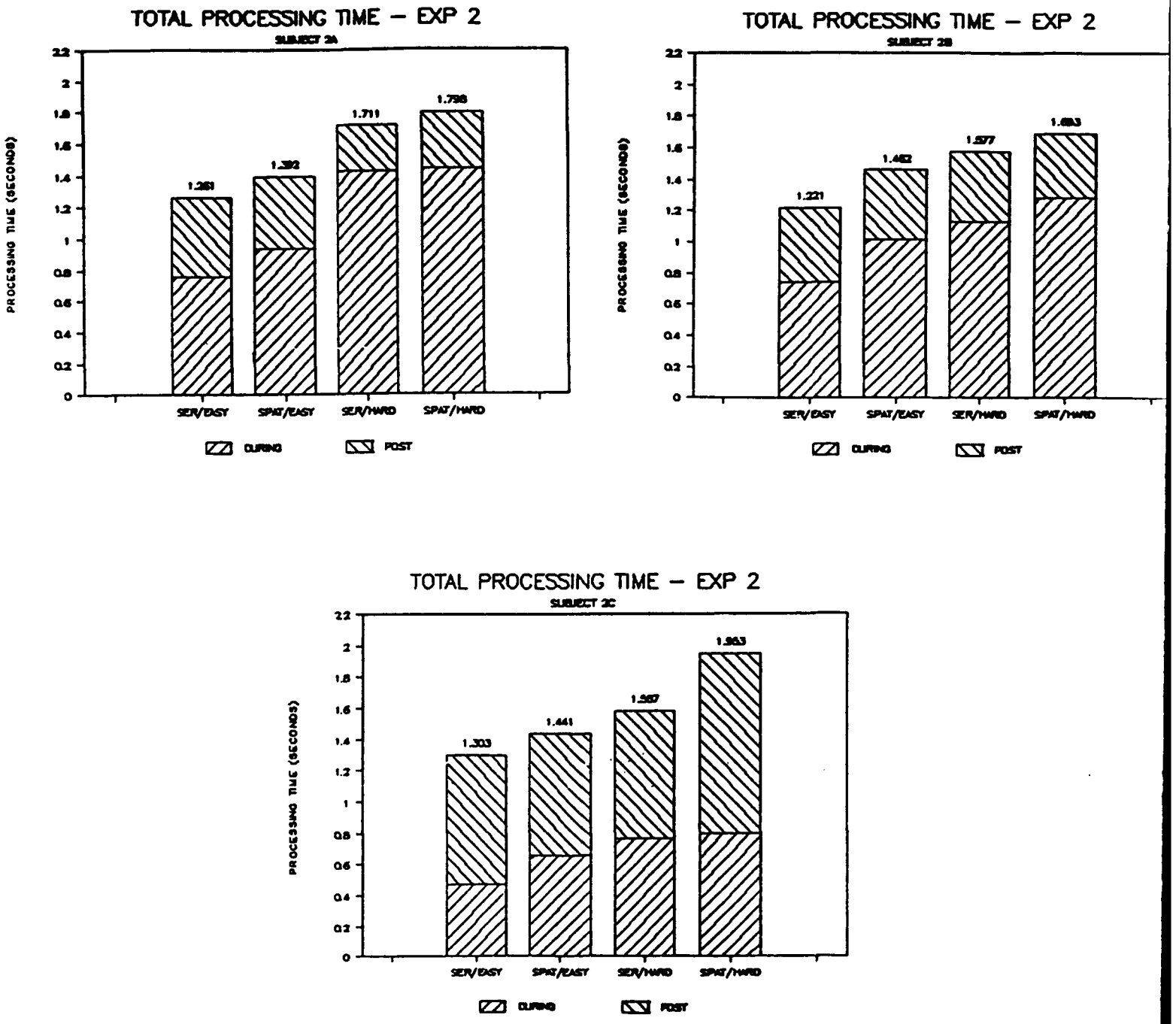


Figure 4. The results of Experiment 2. Total information processing time and its two components, during-display time and post-display tim for the serial and spatial conditions for both levels of task difficulty.

Moreover, there were no significant interactions (see Table 2). Two of the three subjects showed the expected decrease in saccadic overhead for the more difficult task (the overhead was 69.9 ms/saccade and 125.5 ms/saccade for the easy task and 38.0 ms/saccade and 61.0 ms/saccade for the hard task for subjects 2A and 2B, respectively). However, subject 2C's overhead for the hard task was greater, not less (69.6 ms/saccade for the easy task versus 174.2 ms/saccade for the hard task). A possible explanation for this departure from expectation will be considered in the General Discussion after the results of Experiment 3 have been described.

Table 2

Analysis of Variance for the Total Processing Time - Exp 2

	Subject		
Source	2A	2B	2C
Display Format			
F	97.60	50.82	29.97
p	<.000	<.000	<.000
Difficulty			
F	6.532	20.84	13.58
p	<.000	<.000	<.000
Interaction			
F	.580	2.487	2.498
p	.453	.126	.125

Experiment 3

Purpose. The primary purposes of the experiment were measurement of the optimal frame duration for maximizing processing efficiency and further study of the tradeoff between the during-display and post-display components of total processing time. To implement these objectives, we developed a new variant of the classical choice reaction time procedure. This technique was used to measure total processing time as a function of frame duration at a fixed performance level (85% correct).

Procedure. As in Experiment 1, subjects performed the odd/even digital processing task (i.e., they counted the number of odd digits that appeared in a sequence of frames, responding ODD for an odd number of odd digits, EVEN otherwise). However, rather than measuring their performance at the 85% duration threshold, we asked subjects to respond at a fixed level (85% correct) independent of frame duration, which was placed under experimental control. With practice, all subjects were able to implement these instructions by adjusting the response time at the end of the trial. In effect, they made it shorter for longer frame durations, thereby counteracting the tendency for improved performance with longer frame durations that would otherwise occur.

Initially, we ran three subjects using a three frame per trial sequence with one digit per frame. Two of them were able to commit the time for an extended experiment. For the latter, we varied task difficulty by using frames with 1, 2,

or 3 digits (3 frames per trial in all cases). The design and results described in the following paragraphs refer to the extended experiment, which was preceded by many hours of practice with the experimental task.

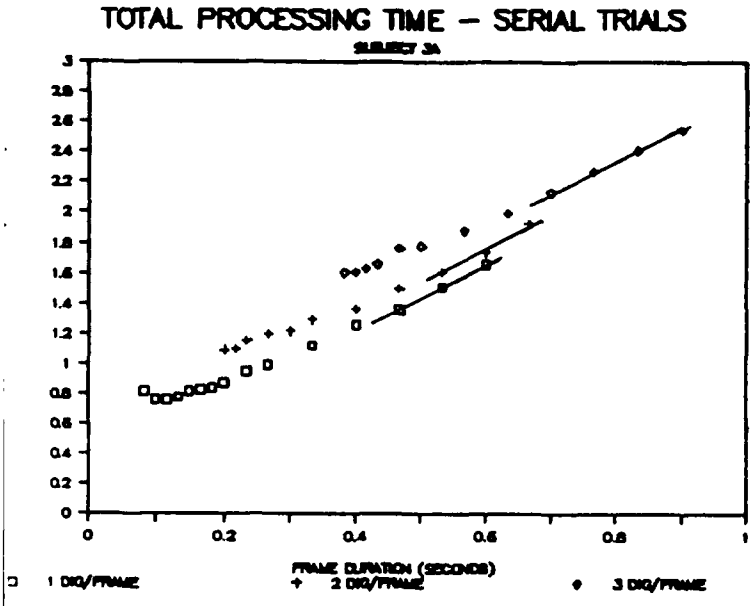
Design. Each subject ran in a completely balanced experiment with three factors (display condition, serial or spatial; number of digits per frame - 1, 2, or 3; and frame duration, with 12 to 15 values, depending on the subject and the levels of the two other factors). For each subject, we selected a range of frame durations that extended from the shortest duration for which it was possible to obtain 85% correct performance to frame durations that were long enough to create "negative" post-display times (i.e., subjects responded before the last frame disappeared). Two runs occurred on each experimental day, one serial and one spatial. In both cases, the value of the digits/frame variable was held constant and all values of the frame duration variable were presented in ascending order. At each duration, subjects received a block of 20 trials with brief rest periods (1 - 2 minutes) between blocks. If the percentage of correct responses for a block was less than 80% or more than 90%, the block was repeated (this happened on approximately 10% of blocks). Each run took about 30 - 40 minutes and a period of several hours occurred between the two runs.

Over the course of the experiment, six 20-trial blocks were run for each experimental treatment, with position of the serial and spatial runs within the session counterbalanced across sessions. Moreover, a three-day cycle was used for

the digits per frame variable to control for possible improvements in performance as the experiment progressed. In all, 18 experimental days were required to obtain the six samples for each treatment that were averaged in analyzing the results.

Results. Figure 5 (next page) shows the results of the serial runs for subjects 3A and 3B. The corresponding results for the spatial runs are shown in Figure 6 (also on the next page). In all cases, the independent variable is frame duration and the dependent variable is total processing time (during-display time + post-display time at the 85% criterion). The average % correct responding over the six samples at each data point was within .017% of the 85% criterion for subject 3A and within .01% of the criterion for subject 3B. The purpose of the lines that are superimposed over the data at long frame durations will be explained in the following paragraph.

With the plotting method used in Figures 5 and 6, we expected to find functions with two branches, a "tradeoff" branch for the relatively short-duration frames and a "waste-time" branch for the remaining, longer frame durations. For the former, a slope of zero would be found if the tradeoff between the during-display and post-display components of processing time was perfect (constant total time, independent of frame duration). For the latter, we anticipated a slope of two for the following reasons: When the frame duration exceeds a critical value (the time needed to access the data in the frame and process it completely) the subject would necessarily waste time for the first two



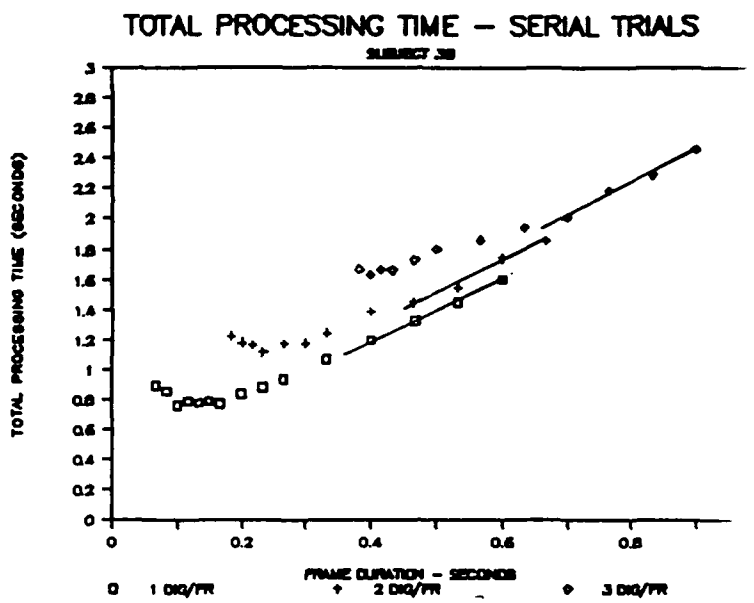
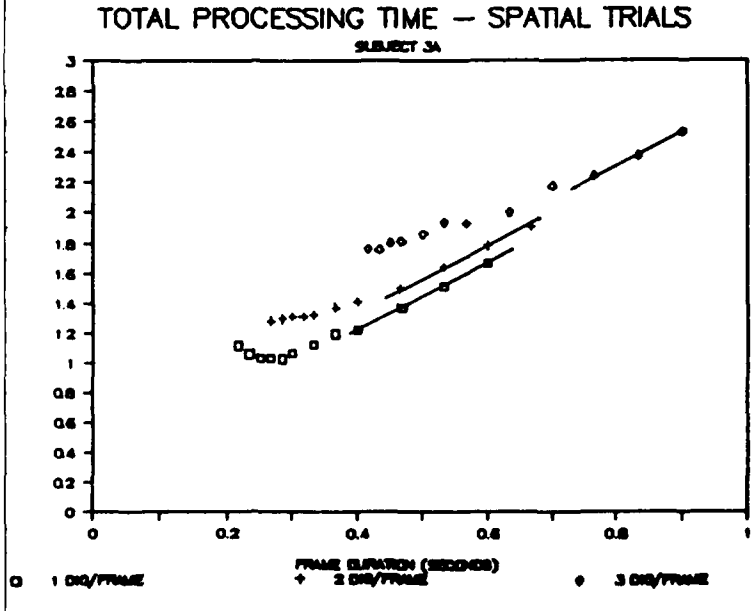


Figure 5. The results of Experiment 3, serial condition. Total information processing time is plotted as a function of frame duration. The line with slope 2 that is superimposed over the data represents the "waste time" branch of the function. See the test for an explanation.



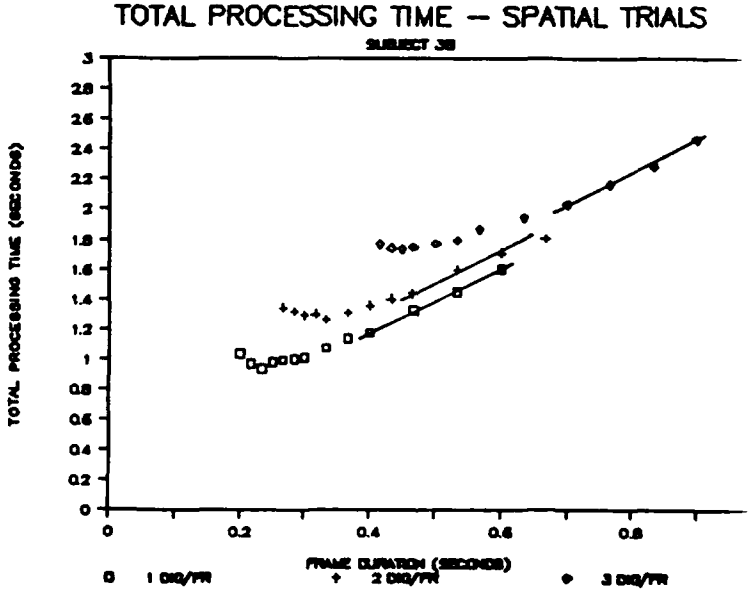


Figure 6. The results of Experiment 3, spatial condition. The plotting method is identical to the one used in Figure 5.

of the three frames. Specifically, one ms would be wasted for every millisecond of frame duration beyond the critical value as the subject waits for the first frame to disappear so the second frame can be accessed. Similarly, one ms would be wasted for every millisecond of frame duration longer than the critical value as the subject waits for the second frame to disappear so the third frame can be accessed. Because the response could be made as soon as the subject was ready, there was no reason to waste time on the third frame, for which reductions in the post-display time could continue to compensate for increases in frame duration. Accordingly, we expected to find a two ms increase in total processing time for every millisecond of increase in frame duration beyond the critical value (i.e., a slope of two).

The two-branch function described in the preceding paragraph was predicted for both the serial and spatial cases because the basic procedures for the two conditions were exactly the same except for the presence or absence of saccades. However, the waste time branch for the spatial function was expected to begin later because of the time needed for the two saccades. In addition, the minimum frame duration for 85% correct responding was expected to be less for the serial condition for all values of task difficulty, with a diminishing serial/spatial difference with increases in difficulty (number of digits per frame).

To simplify the description of the results, we imposed linear functions with a slope of two over the data at the long frame durations, where the waste-time function was expected. As can be seen in Figures 5 and 6, the waste-time

lines provide remarkably good fits. However, the data at the short frame durations show systematic departures from a perfect tradeoff between during-display and post-display time. In each of the twelve functions (three serial and three spatial for each subject) there is an optimal frame duration for minimizing total processing time. This optimal duration is greater than the shortest duration used in the experiment in eight of the twelve functions. For shorter or longer frame durations, the processing can be accomplished at the criterion level, but only with added costs. In short, for frame durations prior to the onset of the 2 ms waste-time branch, there is a partial tradeoff, with decreases in post-display time that do not completely compensate for the during-display time increases. Because it seems likely that the subject would need more time for the second frame than for the first (the time needed to remember the odd-digit count from the first frame, and then to increment it after the odd digits in the second frame were counted), we think that there is probably a 1 ms waste-time branch before the 2 ms branch appears. However, more data are needed at the shorter frame durations to clarify this possibility.

To simplify the comparison of the serial and spatial functions, they were plotted together on one graph for each value of the digits/frame variable for each subject. These graphs are shown in Figures 7, 8, and 9 for the 1, 2, and 3 digits/frame conditions, respectively. As was expected, there were no differences between the serial and spatial functions at the long durations, for which both functions had entered the waste-time branch of the data. However, for both subjects and for all values of the digits/frame

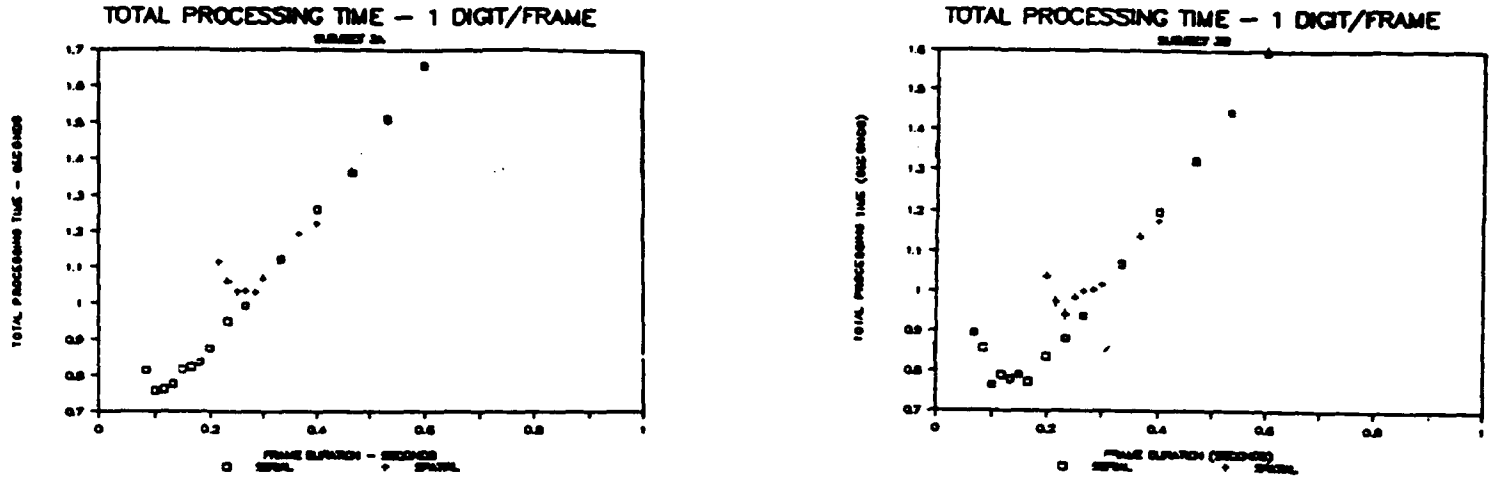


Figure 7. The results of Experiment 3, serial/spatial comparison for the 1 digit/frame condition. Total processing time is plotted as a function of frame duration.

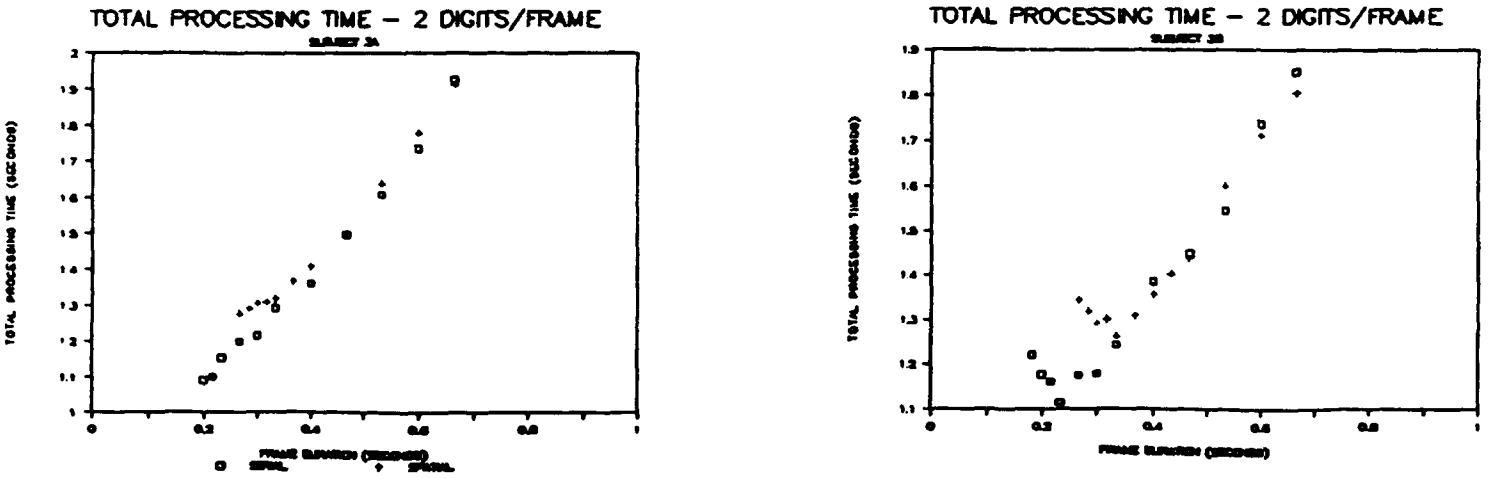


Figure 8. The results of Experiment 3, serial/spatial comparison for the 2 digits/frame condition. Total processing time is plotted as a function of frame duration.

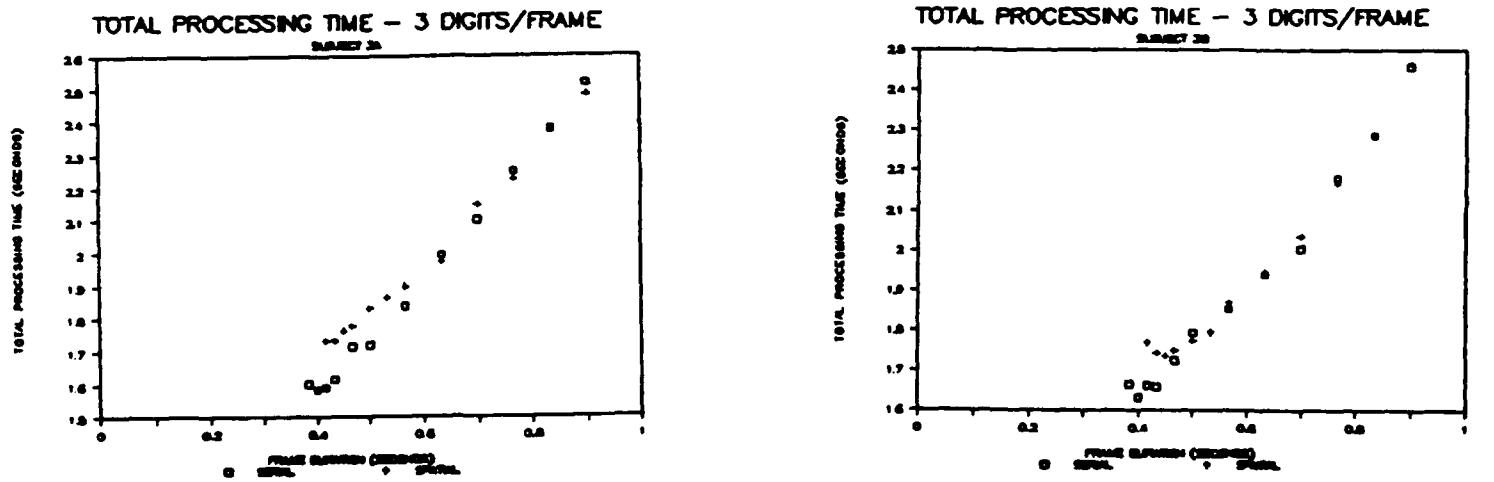


Figure 9. The results of Experiment 3, serial/spatial comparison for the 3 digits/frame condition. Total processing time is plotted as a function of frame duration.

variable, the results showed shorter total processing times with the serial display during the tradeoff period. The magnitude of this difference decreased with the difficulty of the task, as was expected from earlier research and from the results for subjects 2A and 2B of Experiment 2 in this report. This decrease can be seen graphically in Figures 7-9. Quantitative details are provided in Table 3, which shows the minimum total processing time for the spatial and serial condition and the difference between them for each value of the digits/frame variable.

Table 3

Minimum Values of Total Processing Time (MS) - Exp 3

	Subject	
Digits/Frame	A	B
1		
Spat	1028.5	939.9
Ser	759.2	764.2
Diff	269.3	175.7
2		
Spat	1279.2	1292.7
Ser	1092.1	1115.7
Diff	187.1	177.0
3		
Spat	1756.8	1736.8
Ser	1603.4	1631.9
Diff	153.4	104.9

IV. GENERAL DISCUSSION

As was explained in the discussion of Experiment 1 and illustrated in Figure 3, there was an unexpected increase in saccadic overhead with increase in number of frames. (Please note carefully that the difference between the spatial and serial processing times was expected to increase; however, the saccadic overhead - difference per saccade - was not). This finding suggests that the serial display's speed advantage was underestimated by our previous research, which was limited to serial/spatial comparisons in displays with three frames. We assumed that the total time saved with the serial format could be estimated by a simple engineering rule of thumb - multiply the overhead for the type of information in the display by the number of eliminated saccades.

To explore possible explanations for the unexpected increase in overhead, a careful examination of the effect of number of frames using the technique developed for Experiment 3 would be useful. This technique shows performance at a fixed operating level for a large range of frame durations and would thus allow us to see whether shifts in the subject's operating point on the continuum of frame durations could have produced the loss of efficiency with the spatial format as the number of frames increased. Similarly, such shifts might account for some of the individual differences in Experiments 1 and 2. For example, a shift to a less favorable point on the frame duration continuum with the spatial format might account for the unexpected increase in saccadic overhead with task difficulty in Subject 2C.

In the present experiments, as in all other studies in the program of research that Dr. Boff and I have been implementing, we emphasized the time-saving features of the serial format. However, full exploitation of this display's potential also requires studies of the space-savings aspect of the technology. To see the possibilities, consider the stylized version of a conventional check-reading display in Figure 10 on the next page. Displays of this type are commonly used when large amounts of information about the status of an overall system and its various sub-systems must be supplied to a human operator, with no need for precise details because the operator simply checks for indications of malfunctioning in a largely computer-controlled process. Usually, all processes are operating within prescribed limits and all the gauges are lined up in the same direction. Signs of a malfunction (as in systems A3 and C6 in Figure 10) are easily detected as offsets from a normal setting.

While this ergonomically designed method has many advantages, it shares with other conventional displays the disadvantage of requiring large amounts of space - one dial for each variable presented. With a serialized version, it would be possible to retain the advantages of the check-reading concept without incurring large "real-estate" costs. As an example, consider Figure 11, which shows a possible serial counterpart of the bank of dials in Figure 10. In Figure 11 (next page), the 40 variables from Figure 10 are displayed in a single window. The first row of ten variables from Figure 10 appears in Figure 11 as a sequence of 10 rapidly presented frames along the 0° axis when all

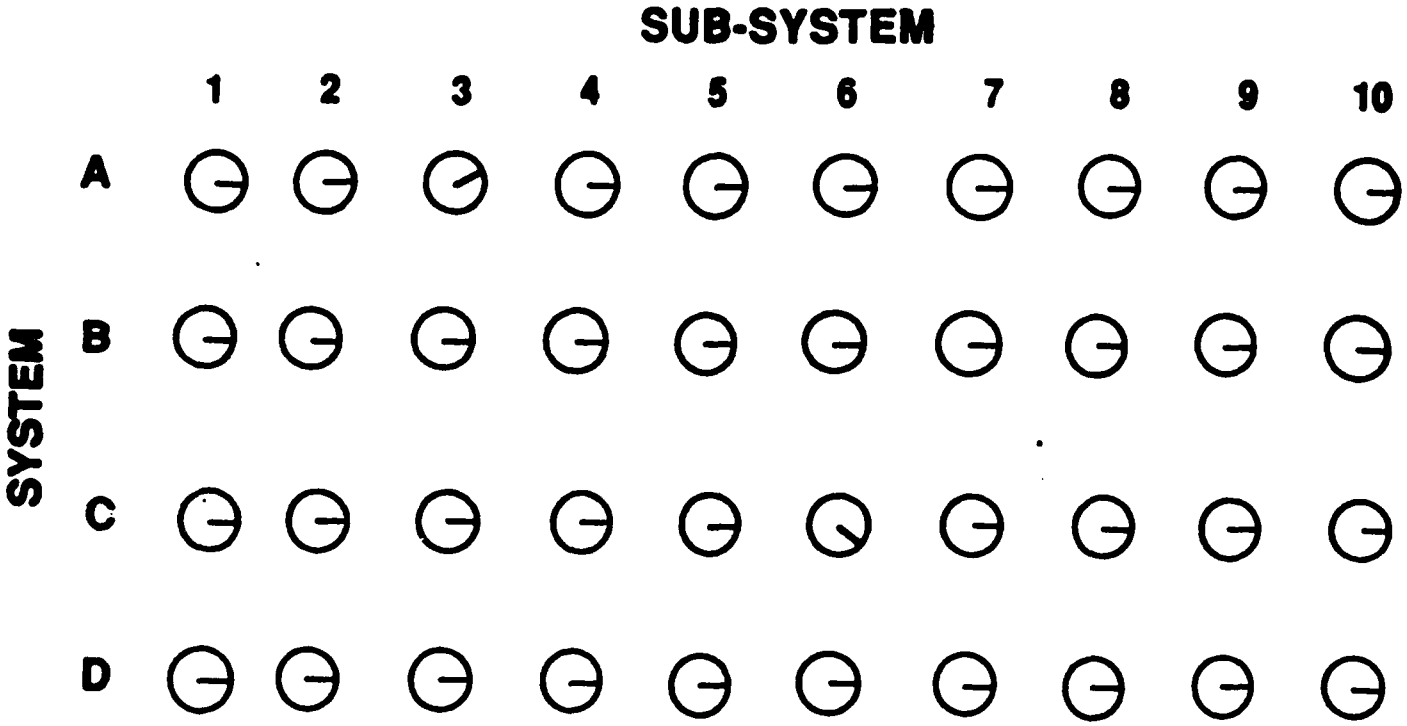


Figure 10. The 40 dials in the illustration are communicating information about the status of 4 systems (A,B,C,D), each with 10 sub-systems (1-10). In the example, 38 of the 40 sub-systems are operating within tolerance, while two (A3 and C6) are out of bounds.

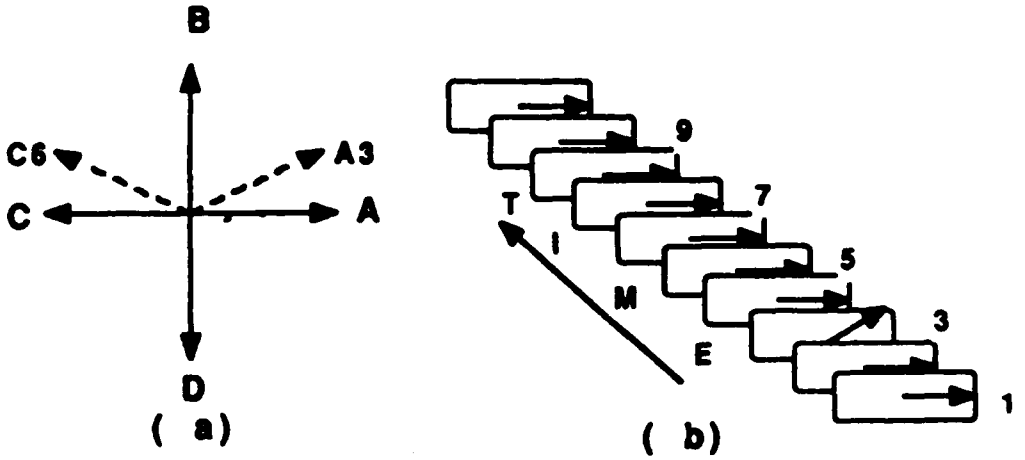


Figure 11. A possible serial equivalent of the Figure 10 check-reading display. The status of the 10 sub-systems of System A is displayed as a sequence of frames presented along the 0° axis for within-tolerance variables; the out-of-bounds variable, A3, appears as an offset - see the detail in part b of the figure. All sub-systems in B and C are within tolerance and will therefore appear in temporal succession along the 90° and the 270° axes, respectively. Nine of the 10 sub-systems of System C will appear in succession along the 180° axis, while C6, which is malfunctioning, will appear offset.

variables are within tolerance. Similarly, the second row of ten is presented sequentially along the 90° axis, the third row along the 180° axis, and the fourth row along the 270 axis. For out-of-bounds variables (like A3 and C6 in the example), the error appears as an easily detectable angular offset from the main axes. Assuming a refresh rate of 30 Hz (33.333 ms/frame) the complete set of 40 variables could be presented every 333 ms within one small display window. In effect, the 40 dials from Figure 10 would be replaced by a single small display. In an actual device, the offset of the out-of-bounds variables would be particularly noticeable because an illusory (stroboscopic) motion would be seen.

V. RECOMMENDATIONS

As noted in the General Discussion, the unexplained increase in the saccadic overhead with number of frames has important implications for the design of serial displays. It needs to be examined carefully. In addition, a full exploitation of the serial format's potential requires data on the space-savings aspect of the technology. Specifically, the following lines of further research would be useful:

1. Parametric studies of the effect of number of frames and task difficulty on saccadic overhead, using the methodology that was developed for Experiment 3 of the present report. These studies should be preceded by some exploration of the new methodology itsel

2. Parametric studies of the space-savings advantage of the serial format, using various versions of the serialized

check-reading display in Figure 11 as a vehicle. New methodologies will need to be developed for this problem, which involves an interesting mix of perceptual and cognitive factors. In addition to their inherent scientific interest, these experiments should yield information that is needed for the construction of a prototype display for support of the Air Force's technology transfer efforts.

REFERENCES

Matin, E. (1989). Breakdown of total information processing time into during-display and post-display components for serial and spatially distributed visual presentations. Final Report, AFOSR Faculty Fellowship.

Matin, E. & Boff, K. (1988). Information transfer rate with serial and simultaneous visual display formats. Human Factors, 30(2), 171-180.

Matin, E. & Boff, K. (1989). Sequential rapid communication visual displays. United States Patent, #4,845,645, issued on July 4, 1989.

Matin, E. & Boff, K. (1990). An adaptive (tracking) procedure for measuring visual search. Perceptual and Motor Skills, 70, 243-255.

Matin, E. & Boff, K. (1990). Human-machine interaction with serial visual displays. Society for Information Display, International Symposium, 21, 257-260.

Matin, E., Boff, K., & Donovan, R. (1987). Raising control/display efficiency with Rapid Communication Display Technology. Proceedings of the Human Factors Society, 31, 258-262.

Osgood, S., Boff, K., & Donovan, R. (1988). Rapid communication technology efficiency in a multi-task environment. Proceedings of the Human Factors Society, 32, 1395-1399.

1989 USAF-UES RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

A Blackboard Architecture for Landmark Identification on 3-Dimensional Surface Images of Human Subjects

Prepared by:	Randy Beth Pollack, Ph.D.
Academic Rank:	Assistant Professor
Department and	Computer Science
University:	Florida Institute of Technology
Research Location:	AAMRL/HEG Wright Patterson AFB Dayton, OH
USAF Researcher:	Kathleen Robinette

Date:	15 September 1990
Contract No:	F49620-88-C-0053/SB5881-0378

A Blackboard Architecture for Landmark Identification on 3-Dimensional Surface Images of Human Subjects

by

Randy Beth Pollack

ABSTRACT

An expert system was designed and implemented which used a blackboard architecture to generate and evaluate hypotheses regarding the positions of facial landmarks on 3-D digital images of human heads. The system was tested on a sample of 20 subjects and the results indicate that this approach was successful in automating the landmark identification task on a subset of landmarks. Further work is necessary to extend the process to include the remaining landmarks, to accommodate subjects in different poses, and to accommodate female subjects.

Acknowledgements

I would like to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsorship of this work. I would also like to thank Universal Energy Systems for their efficient and concerned administration of the program.

I would like to express my appreciation to the anthropologists and computer scientists in the Human Engineering Division of the AAMRL. Kathleen Robinette provided direction and advice and Joyce Robinson gave much needed assistance in the technical area. My graduate research assistant, Paresh Shah, must be credited with the coding; in addition, many important insights into the problem originated with him. I would like to thank all of these people for their help and support throughout this project.

I. INTRODUCTION

Anthropometric data have been used by the Air Force for many years to help in the design of clothing, equipment, cockpits, etc. The use of anthropometric statistics enables designers to provide a better-fitting product and ensure that the people who will be using the equipment are physically able to do so.

Traditionally, the tools of anthropometry have been physical measuring devices such as calipers and tape measures. Recently, the Human Engineering Division at Wright-Patterson Air Force Base has been exploring the use of 3-dimensional scanning technology to produce a digital representation of the surface of the human head. A vertical stripe of laser light is projected onto a stationary subject as the scanner rotates 360 degrees around him or her, taking video recordings which are converted into digital format. The digitized image is stored as a set of points representing the distance from the center of rotation to the surface of the subject at 256 points along each longitude. This technique has been used to collect data on both men and women to be used in the design of such equipment as night vision goggles for Air Force pilots.

Although the data collected in this way are much more complete than was possible using traditional methods, techniques have not yet been developed to easily analyze and use this type of data. Due to the large size and noisiness of the data set, and the difficulty of recognizing facial features which vary greatly in size and shape among individuals, the use of artificial intelligence methodologies seems appropriate to the problem.

My academic background incorporates two specialties that are directly relevant to the problem described above. My Ph.D. is in physical anthropology, which is the academic discipline that includes the study of anthropometry. More recently, I received a M.S. degree in computer science, with a specialty in artificial intelligence. I have been involved in AI research for the past several years. These factors led to my involvement in the UES/AFOSR 1989 Summer Faculty Research Program. My work during that time formed the basis for the research described in this report, which was carried out from January to September, 1990.

II. OBJECTIVES OF THE RESEARCH EFFORT

One major problem already encountered in the use of 3-D data

is the identification of facial landmarks on the image. The landmarks are used to define boundaries between different regions of the head, and to serve as endpoints for distance measurements. There are approximately 40 landmarks that must be recorded for each image and stored as a set of longitude/latitude coordinates.

Currently, about half the landmarks are being marked on the subjects prior to scanning using small felt patches which do not reflect the laser light. The remaining landmarks are those which can be visually identified on the image, such as the tip of the nose, or those which cannot be marked with patches, such as the pupil or the stomion (center of the mouth along the meeting line between the two lips). Someone must then record the longitude and latitude of each landmark in a separate file associated with the image file. This is done by using the cursor to locate each landmark on a graphic representation of the head and instructing the program to record the latitude and longitude of the point identified by the cursor.

A means of having the landmarks automatically identified by an AI program would save a great deal of time for both the technicians who currently identify the landmarks manually, and for the subjects being scanned. The goal would be to

eliminate the need for marking subjects with patches, or to reduce those to as few as possible, and to have a computer program identify and record the coordinates of each landmark. In addition to saving time, computerizing this process would ensure consistent landmark identification across subjects, which is especially difficult when more than one person does the marking.

The purpose of this project was to design and implement an expert system capable of performing the task of landmark identification. It was not expected that this system would successfully locate all landmarks with complete accuracy. Rather, the goal was to design a framework that would begin the process of landmark identification and would allow the later addition of different techniques and algorithms to complete the task. The work accomplished in the previous summer's research made it clear that a technique which might successfully locate a landmark on one subject would not be successful on another. Therefore, a major part of the work done in this project was designing a system that could integrate many techniques, evaluate the results obtained from each, and make an intelligent decision as to which results are most likely correct for a given subject. Given such a framework, it becomes relatively simple to add new techniques and expand the utility and reliability of the

system as they become available.

III. BLACKBOARD ARCHITECTURES

The framework chosen for the task described above was a knowledge-based system using a blackboard architecture. This type of system has been applied to symbol-processing problems that are in some ways analogous to landmark identification. One major advantage of the blackboard approach is the ease with which various types of data and analysis techniques may be integrated into the system.

Blackboard systems are characterized by more sophisticated control structures and data representation facilities than are found in more traditional expert systems. They frequently consider a problem from multiple levels of abstraction, with modules at one level evaluating hypotheses generated by modules at other levels and generating their own hypotheses in turn. Blackboard systems usually use 'opportunistic' reasoning, which gives them the ability to vary the problem solving algorithm according to the way in which the solution is developing. They are highly flexible systems and have been used to solve such problems as speech recognition (Erman, et.al. 1980), sonar signal analysis, (Nii, et.al. 1982), and analysis of aerial photographs

(Nagao, et.al. 1979). Good overviews of the blackboard architecture and applications may be found in Engelmore and Morgan (1988) and Jagannathan, Dodhiawala and Baum (1989).

IV. DESIGN AND IMPLEMENTATION

The language chosen for implementation of the blackboard system was C++. There are several reasons for that choice. First, much of the code already written for the analysis of this data has been written in C, and can therefore be incorporated into the system with little modification. It is also important that C++ has object-oriented facilities which make it easier to design a large system which is highly modular and yet respresents real-world objects in a more natural way than is possible with languages lacking object oriented facilities. The Zortech C++ compiler was used on an IBM-compatible computer with a 386 processor and 1 megabyte of memory. The major limitation experienced in this work is one of memory; any expansions to the system will require the addition of at least one more megabyte of memory.

The blackboard system designed for the landmark recognition task was simpler than many of those listed above. Its major components include knowledge sources, and the Blackboard itself.

Knowledge sources are individual techniques or rules used by the system for locating landmarks, integrating hypotheses posted by other knowledge sources, or taking care of other facets of data analysis. Each Knowledge Source (KS) has a triggering condition; when that condition becomes true, this indicates that the KS has something to contribute to the developing solution. KS's are completely independent of one another and respond only to their triggering conditions. The chief repository of procedural information in the system is the collection of KS's.

There are two general ways in which this triggering process can be handled. One is to test each KS's triggering condition each time a change is made to the blackboard. This becomes highly inefficient when the number of KS's becomes large. It is more efficient to use a simple data structure to keep track of the different types of changes to the blackboard that may occur, and associate with each type of change the specific KS's that it triggers. This is the approach used in the present system. A matrix is indexed along one dimension by the different landmarks being sought. Associated with each landmark is the set of KS's that are triggered when a hypothesis is posted for that landmark. Always included in this set is an integrating KS which

evaluates the newly posted hypothesis in the context of other hypotheses and related information. . The set may also include KS's that seek landmarks whose positions are constrained in some dimension by the landmark which has been posted.

The blackboard itself contains the following data structures and categories of information:

(1) Knowledge Source Activation Records (KSAR's): nodes which are created each time a KS's triggering condition is met. The information in each KSAR includes the number of the KS that has been triggered, the priority of that KS, and the iteration of the program's run in which the triggering took place.

(2) The Agenda: Since more than one KS may have knowledge which applies at any given time, an agenda must be maintained which consists of a list of all KSAR's which have been triggered but not yet chosen for execution. The KSAR with the highest priority at the beginning of each iteration is chosen from the agenda for execution and removed.

(3) The History List: When a KSAR is executed it is removed from the agenda and placed on the history list. At that time, other information is added to the record indicating

the environment in which it was executed. More specifically, each KS has a set of related landmarks whose coordinates are used in the analytical technique performed by that KS. The values of these landmarks are considered to be the context in which that KS executes. Saving the context in the KSAR and placing this information on the history list after KS execution allows for several important functions. First, a running record is kept of the system's performance for debugging, maintenance and improvement purposes. Second, the history list is checked before any KSAR chosen from the agenda is actually executed. If the most recent execution of that KS was performed in a context which is identical to the present context, it will not be executed again. This is necessary because multiple KSAR's for a particular KS may be placed on the agenda at once. For example, the posting of a hypothesis regarding the subnasale is a trigger for any KS that seeks the pronasale. Since there are at least two KS's with different techniques for locating the subnasale, each may execute and post a hypothesis. Each posting will trigger a KS dealing with the pronasale, so there will be two KSAR's on the agenda for that one KS. It would be inefficient and incorrect for both of these to result in KS executions if the two contexts associated with them are identical; the same code would simply execute twice, possibly posting two identical hypotheses. Checking

the history list before execution prevents this from happening.

Going one step further, the history list guards against cycles in the program. Cycles are possible because the system does not generally seek landmarks in a specific order. While the posting of a subnasale hypothesis triggers any KS seeking the pronasale, the opposite is also true. The posting of a pronasale hypothesis triggers KS's seeking the subnasale. This circularity is necessary because the system does not know which landmark will be located first, and the relationship between them can be used regardless of which value is discovered first. Checking the history list before KS execution prevents the cycles that would otherwise result from this circularity.

(4) The last major data structure on the blackboard is the set of hypotheses. This is implemented an array of pointers indexed by landmarks; each points to a list of hypotheses for that landmark. Each hypothesis has a confidence factor indicating the system's current belief in the validity of that hypothesis. The action of an integrating KS is to adjust the confidence levels of various hypotheses.

(5) Finally, a set of global variables is found on the blackboard which contains the current best hypothesis for

the coordinates of each landmark. These variables are set by the integrating KS's as hypotheses are posted and evaluated.

In addition to building a blackboard framework, the research done for this project included an investigation into several approaches for recognizing facial landmarks. This work focused on those landmarks located on or near the centerline of the face. Previous work indicated that these should be relatively easy to locate once that centerline has been established.

The first task, then, was to design knowledge sources capable of locating the centerline. Preliminary work had shown that in some subjects this is possible using a technique we have termed Min/Max features.

Traversing a single longitude, one can discern local maxima and minima in the radii which represent protrusions and indendations respectively. For this purpose, a local minimum is a point whose neighbors both have larger values; a local maximum is a point whose neighbors both have smaller values. The chin, for example, shows up as a local maximum at approximately the same latitude along a series of adjacent longitudes, and is therefore termed a Max Feature.

This is also the case for the nose, upper and lower lips, etc. Features that show up as minima include the mouth (which appears as an indentation between the two lips), the eyes, and the center of the ear.

In each of these cases, the local maximum or minimum occurs across a number of longitudes at approximately the same latitude. However, due to errors introduced by the imaging and digitization process, the maximum or minimum typically does not show up at all longitudes. An additional complication is that most features are not horizontal lines but have a curved or angled shape and show up at gradually changing latitudes. For example, following the line of the chin from the center to either side will take you to from lower to higher latitudes.

These problems were successfully solved for a number of Min/Max Features, and the use of these features enables the system to identify the centerline of the subject's face. Once a Mix/Max feature is located which is symmetric across the centerline, the center of the feature can be used to approximate the centerline longitude. The easiest Max Feature to locate is usually the chin because it is the sequence of maxima found at the lowest latitude. However, this is not always the case, depending on the neckline of the clothing worn by the subject, the prominence of an

Adam's apple, or other confusing elements. Searching for several Min/Max features and looking for agreement in their center points solves this problem and allows the system to identify the centerline with good reliability.

A first attempt is made to hypothesize a central longitude based on this technique. The orientation of the head may not be perfectly vertical, however, and other techniques are used by other knowledge sources to locate a centerpoint at higher latitudes (at the sellion, for example) and extrapolate the degree of tilt from the difference between that centerpoint and the one found at the promenton latitude.

Min/Max features were used by knowledge sources attempting to locate most of the landmarks that this work focused on. These include the menton, promenton, stomion, left and right chelion, subnasale, pronasale, sellion and glabella. However, the features discovered using this technique were not found on all subjects. For example, the sellion is located along the centerline between the nose and the forehead where the radius is at a minimum. In many people, however, this is not a local minimum, but a flat-bottomed valley. It frequently does not show up as a minimum using the Min/Max features technique described above. Other knowledge sources were written to use additional techniques to hypothesize landmark coordinates, and integrating knowledge sources

evaluated the results.

A technique that proved very useful involved the patterns of slopes across a given longitude or latitude. This method of approaching the data looks at a the points on a longitude (or latitude) as a series of slopes computed as the differences between adjacent radii. One can then identify points at which the slope changes, which may help to identify landmarks. The sellion is an example of a landmark that is more easily identified in this way compared to using Min/Max features. A pattern can be seen in which sellion is clearly discernable as the slope changes between the nose and the forehead. Changes in slope are useful in identifying a number of other landmarks as well, such as the promenton, subnasale, pronasale and left and right endocanthus.

The integrating knowledge sources deal with the hypotheses generated by each of the above techniques. In addition to these hypotheses, an integrating Knowledge Source may be aware of statistical relationships between landmarks as exhibited in the manually recorded values. For example, in the set of subjects used as a basis, the average distance between the stomion and the promenton latitudes is 23. If a hypothesis shows a distance of 40 between these two land-

marks, its confidence factor is decreased by the integrating Knowledge Source because it deviates too far from the mean. The use of statistical knowledge of this nature is useful in identifying hypotheses which are clearly off the mark. In the absence of other information, it is also occasionally useful to post a hypothesis based solely on a statistical mean.

VI. RESULTS

Tables 1 through 4 present the data on the eleven landmarks most reliably identified by this system. Table 1 shows the manually recorded values for 20 male subjects in the same pose. The only criterion used in choosing these subjects was that a complete set of manually recorded landmark coordinates had to be available for comparison purposes.

Table 2 shows the coordinates chosen by the program for the same subjects. Table 3 is more illuminating, showing the discrepancies between each manually recorded and computer generated coordinates. Finally, Table 4 summarizes these discrepancies, showing the mean, minimum and maximum discrepancy and standard deviation for each landmark.

It is clear from a study of these tables that the program has successfully located most of the landmark coordinates,

Subj	REndocan		Glabella		Sellion		Pronasale		Subnasale		Promenton		Menton		RChelion		Stomion		LChelion		LEndocan	
	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long
17	162	308	185	288	169	285	147	284	136	283	101	283	94	280	121	303	125	282	123	258	164	258
21	141	313	158	294	148	295	129	292	117	292	82	292	74	295	100	309	103	293	100	275	140	277
33	160	337	182	306	172	303	146	288	134	292	91	291	86	294	116	320	119	294	116	263	161	275
46	179	311	201	293	187	290	169	290	159	290	122	289	115	290	142	307	146	290	142	273	1	267
60	180	335	203	309	189	301	171	296	159	296	121	293	112	293	139	322	143	295	139	266	179	268
66	171	334	189	313	179	310	158	305	148	306	114	305	105	304	133	332	137	305	133	276	172	292
78	171	310	193	279	177	285	162	285	152	286	116	288	108	288	133	306	137	287	134	268	172	261
79	155	335	175	304	162	301	147	299	139	299	97	295	87	295	115	321	122	297	117	279	158	283
84	156	338	174	303	160	300	148	300	139	300	96	301	88	304	117	323	121	301	117	284	155	277
85	194	326	217	287	198	290	182	290	169	290	130	288	117	288	149	313	153	289	150	268	193	256
90	155	310	177	276	164	278	151	278	138	278	97	275	91	275	118	299	122	276	118	253	156	235
92	158	320	180	285	166	288	141	288	134	288	92	288	83	288	113	313	117	287	115	266	159	260
93	146	331	170	305	152	306	134	301	125	302	88	299	79	299	106	319	108	302	106	287	148	290
133	124	294	140	276	129	277	102	277	98	277	61	278	53	277	82	295	85	277	82	263	124	260
150	114	321	131	299	116	299	101	299	92	299	60	298	52	298	75	332	79	302	76	271	114	287
155	144	302	165	281	151	282	133	282	121	282	87	282	77	282	106	300	108	282	107	266	146	262
160	145	320	165	294	151	298	130	298	122	298	85	296	76	297	101	321	105	297	103	275	145	278
176	143	325	167	309	148	308	126	306	119	306	84	306	74	306	99	322	104	306	100	292	143	293
182	150	326	171	308	157	308	138	308	130	308	95	306	84	307	111	322	116	306	111	294	150	294
194	121	335	140	312	128	311	108	308	100	311	62	313	55	315	82	338	86	311	81	289	122	294

Table 1. Manually recorded landmark coordinates.

Subj	REndocan		Glabella		Sellion		Pronasale		Subnasale		Prosenton		Menton		RChelion		Stomion		LChelion		LEndocan	
	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long
17	162	312	183	287	169	285	147	282	141	282	105	282	95	282	122	303	126	282	122	261	162	258
21	141	313	162	295	148	295	127	293	118	293	83	293	73	293	99	313	103	293	99	274	141	277
33	166	331	184	305	173	301	146	296	136	296	94	296	84	296	117	321	121	296	117	268	166	277
46	181	308	201	291	188	291	168	290	161	290	124	290	114	290	144	309	148	290	144	275	181	276
60	181	334	200	307	188	302	171	295	161	295	123	295	113	295	139	319	143	295	139	273	181	266
66	171	339	193	315	178	311	159	306	149	306	114	306	104	306	132	330	136	306	132	279	171	283
78	171	305	193	279	178	283	160	289	152	289	115	289	105	289	133	311	137	289	133	270	171	261
79	158	330	179	310	165	305	146	298	137	298	97	298	87	298	116	318	120	298	116	279	158	280
84	151	335	172	302	158	302	147	302	136	302	97	302	87	302	115	324	119	302	115	281	151	265
85	192	319	216	291	199	291	182	289	168	289	128	289	118	289	150	309	154	289	150	267	192	263
90	159	302	180	276	166	276	149	277	136	277	100	277	90	277	117	301	121	277	117	255	159	247
92	158	316	179	287	165	287	142	287	132	287	94	287	84	287	113	310	117	287	113	265	158	261
93	146	331	168	305	153	305	133	303	124	303	89	303	79	303	104	322	108	303	104	282	146	283
133	121	293	146	279	128	279	105	278	99	278	63	278	53	278	80	298	84	278	80	259	121	264
150	108	328	130	301	115	301	99	299	93	299	62	299	52	299	76	321	80	299	76	276	108	277
155	145	298	164	282	152	282	131	282	123	282	85	282	75	282	103	306	107	282	103	261	145	261
160	142	322	165	299	149	299	130	298	120	298	86	298	76	298	103	318	107	298	103	277	142	274
176	140	323	162	308	147	308	126	306	120	306	85	306	75	306	104	327	108	306	104	289	140	293
182	151	328	167	309	158	309	138	308	129	308	92	308	82	308	112	324	116	308	112	293	151	293
194	122	338	140	313	129	313	108	312	100	312	67	312	57	312	81	335	85	312	81	289	122	288

Table 2. System generated landmark coordinates.

Subj	REndocan		Glabella		Sellion		Pronasale		Subnasale		Prosenton		Menton		RChelion		Stomion		LChelion		LEndocan	
	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long	Lat	Long
17	0	4	2	1	0	0	0	2	5	1	4	1	1	2	1	0	1	0	1	3	2	0
21	0	0	4	1	0	0	2	1	1	1	1	1	1	2	1	4	0	0	1	1	1	0
33	6	6	2	1	1	2	0	8	2	4	3	5	2	2	1	1	2	2	1	5	5	2
46	2	3	0	2	1	1	1	0	2	0	2	1	1	0	2	2	2	0	2	2	2	9
60	1	1	3	2	1	1	0	1	2	1	2	2	1	2	0	3	0	0	0	7	2	2
66	0	5	4	2	1	1	1	1	1	0	0	1	1	2	1	2	1	1	1	3	1	9
78	0	5	0	0	1	2	2	4	0	3	1	1	3	1	0	5	0	2	1	2	1	0
79	3	5	4	6	3	4	1	1	2	1	0	3	0	3	1	3	2	1	1	0	0	3
84	5	3	2	1	2	2	1	2	3	2	1	1	1	2	2	1	2	1	2	3	4	12
85	2	7	1	4	1	1	0	1	1	1	2	1	1	1	1	4	1	0	0	1	1	7
90	4	8	3	0	2	2	2	1	2	1	3	2	1	2	1	2	1	1	1	2	3	12
92	0	4	1	2	1	1	1	1	2	1	2	1	1	1	0	3	0	0	2	1	1	1
93	0	0	2	0	1	1	1	2	1	1	1	4	0	4	2	3	0	1	2	5	2	7
133	3	1	6	3	1	2	3	1	1	1	2	0	0	1	2	3	1	1	2	4	3	4
150	6	7	1	2	1	2	2	0	1	0	2	1	0	1	1	11	1	3	0	5	6	10
155	1	4	1	1	1	0	2	0	2	0	2	0	2	0	3	6	1	0	4	5	1	1
160	3	2	0	5	2	1	0	0	2	0	1	2	0	1	2	3	2	1	0	2	3	4
176	3	2	5	1	1	0	0	0	1	0	1	0	1	0	5	5	4	0	4	3	3	0
182	1	2	4	1	1	1	0	0	1	0	3	2	2	1	1	2	0	2	1	1	1	1
194	1	3	0	1	1	2	0	4	0	1	5	1	2	3	1	3	1	1	0	0	0	6

Table 3. Discrepancies between manually recorded and system generated landmark coordinates.

		Mean	Minimum	Maximum	Standard Deviation
REndocan	Lat	2.05	0.00	6.00	1.96
	Long	3.60	0.00	8.00	2.26
Glabella	Lat	2.25	0.00	6.00	1.75
	Long	1.80	0.00	6.00	1.56
Sellion	Lat	1.15	0.00	3.00	0.76
	Long	1.30	0.00	4.00	0.95
Pronasale	Lat	0.95	0.00	3.00	0.92
	Long	1.50	0.00	8.00	1.88
Subnasale	Lat	1.60	0.00	5.00	1.06
	Long	0.95	0.00	4.00	1.02
Promenton	Lat	1.90	0.00	5.00	1.22
	Long	1.50	0.00	5.00	1.24
Menton	Lat	1.05	0.00	3.00	0.80
	Long	1.55	0.00	4.00	1.02
RChelion	Lat	1.40	0.00	5.00	1.11
	Long	3.30	0.00	11.00	2.26
Stomion	Lat	1.10	0.00	4.00	0.99
	Long	0.85	0.00	3.00	0.85
LChelion	Lat	1.30	0.00	4.00	1.14
	Long	2.75	0.00	7.00	1.86
LEndocan	Lat	2.10	0.00	6.00	1.54
	Long	4.50	0.00	12.00	4.06

Table 4. A summary of discrepancies between manually recorded and system generated landmark coordinates.

if success is measured as nearness to the manually recorded points. A proposed acceptable level of difference between manually recorded and computer generated points is an average discrepancy of 3 or less. According to this measure, 19 of the 22 coordinates have been successfully located. If the acceptable discrepancy is reduced to 2 or less, then 15 of the 22 coordinates have been successfully located.

There are several landmarks that are located less successfully than others according to that definition; however, we would like to argue that in at least some cases the program may be superior in that a particular pattern has been identified across the 20 subjects and may therefore be more consistent than manual identification techniques.

One case in point is the glabella. This is a feature that is less marked than many others, and the program locates it using slopes. Although the mean discrepancy between recorded and generated latitudes is 2.25, it is possible in this case that the error is in the recorded rather than the generated values. This may be true for others, such as the right and left endocanthus longitudes, which show even more variability. Again, a pattern of slopes across the latitude was discovered which defines these points, and we would like to propose that it is a more consistent technique than

manual recording using the cursor.

On the other hand, this may not be true of the left and right chelion, which do not show up as consistently using either slopes or Min/Max features as do the other landmarks. These points are difficult to find using the knowledge sources currently working in the system, and it is likely that the manually recorded points may be superior. Clearly, this is an unresolved issue which must be addressed. How does one evaluate the performance of a system against a solution set that may not be entirely correct?

VII. RECOMMENDATIONS

This effort was judged to be successful in demonstrating that knowledge sources can be developed to generate reasonable hypotheses regarding the positions of a select group of landmarks on a sample of 20 subjects in one pose. Although these results are narrow in scope, they are sufficiently positive to justify further work on this project. Now that the blackboard framework has been developed, and integrating knowledge sources have been designed which are capable of dealing with possibly contradictory hypotheses generated by several different techniques, it is recommended that the program be tested further and expanded in several direc-

tions.

First, it is important to test the existing program on a larger number of subjects of the same sex and in the same pose to get a better evaluation of its reliability. The major difficulty in doing this during the present study was related to the transmission of voluminous data from Ohio to Florida, and the subsequent time-consuming task of transferring it from the network machine to floppy disk across phone lines.

Extensions to the program should proceed in at least two directions. The remaining landmarks which were not identifiable given the techniques used by the current knowledge sources should be sought using other techniques. In particular, learning techniques such as neural network algorithms may be better suited to some of them. This is the most difficult of the proposed extensions.

The second direction in which the work should be extended is to encompass subjects in different poses, and female as well as male subjects. These extensions should be more easily done, since they should only require modifications to existing techniques rather than the development of entirely new ones.

The key issue, however, is that the design of the blackboard system makes extensions of this nature a relatively straightforward task. The incorporation of addition Knowledge Sources is the area in which almost all the changes will take place, while the existing framework and control structure should not need alteration.

REFERENCES

Engelmore, R. and T. Morgan, Blackboard Systems. Addison-Wesley Publishing, 1988.

Erman, L.D., et.al., THe Hearsay-II Speech Understanding System: Integrating Knowledge to Resolve Uncertainty, ACM Computing Surveys 12:213-253, 1980.

Jagannathan, V., R. Dodhiawala and L. Baum, Blackboard Architectures and Applications. Academic Press, 1989.

Kjeldsen, R., A. Califano and R. Bolle, Evidence Integration for 3D Object Recordnition: A Connectionist Framework, Proceedings of the 5th Conference on AI Applications, pp.55-63, 1989.

Nagao, M., T. Matsuyama and H. Mori, Structural Analysis of Complex Aerial Photographs, Proceedings of the IJCAI pp.610-616, 1979.

Nii, H.P., et.al., Signal-to-Symbol Transformation: HASP/SIAP Case Study, AI Magazine 3:23-35, 1982.

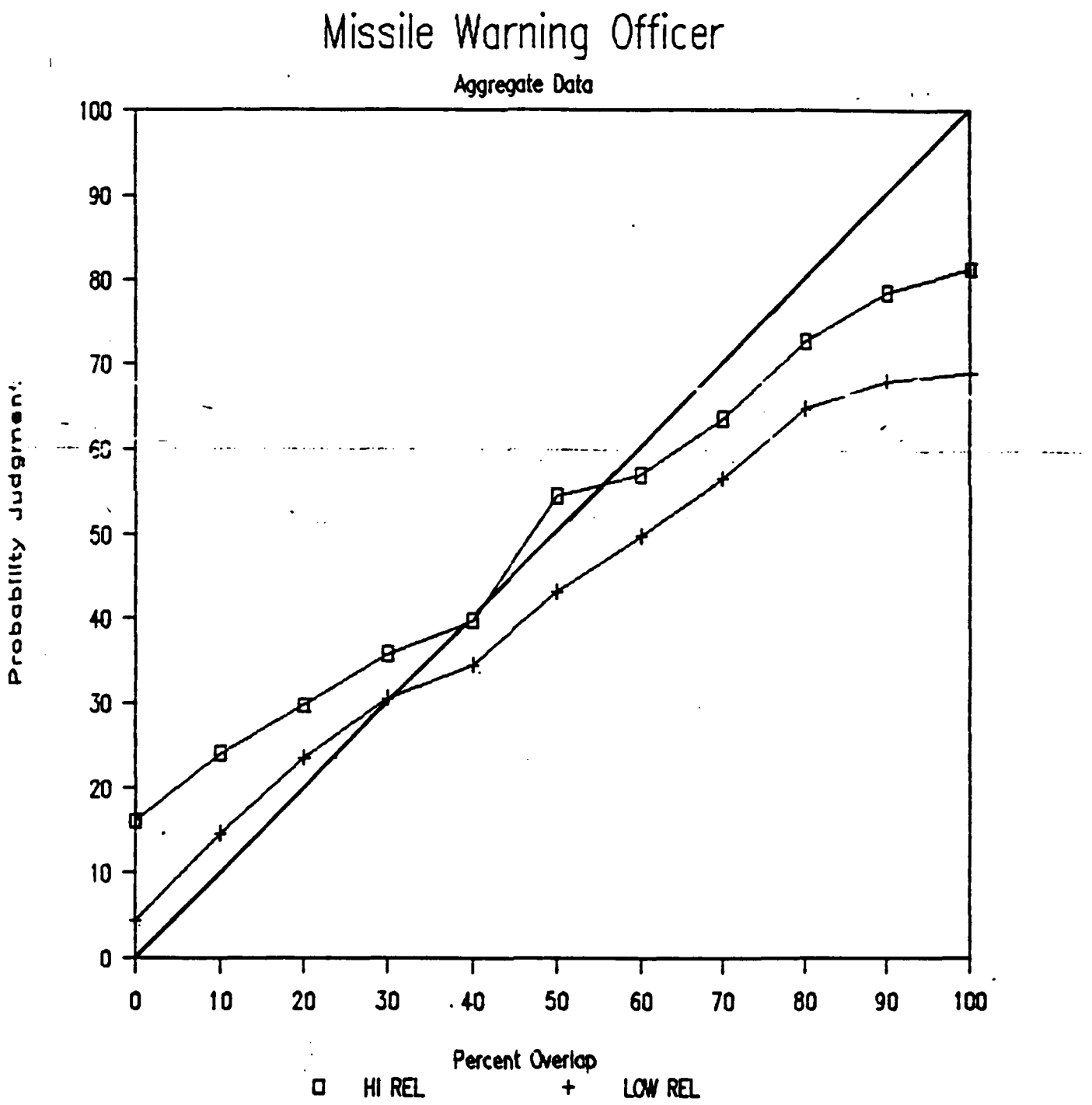
Effect of System Reliability on Probabilistic Inference

Donald U. Robertson
Indiana University of Pennsylvania

This research was supported by a grant from the Air Force Office of Scientific Research, RIP grant program.

The issue addressed in this study is how people use reliability information when they make probabilistic judgments. Previous research has suggested that people are sensitive to the reliability of information if its relationship to the judgment task is made sufficiently clear (Brehmer, 1970; Kahneman & Tversky, 1974; Kruglanski, Friedland, & Farkash, 1984). In fact, some researchers argue that not only do people attend to reliability, but will use it in a normatively correct fashion. That is, judgments based on low reliability information are more regressive than judgments based on high reliability information.

This finding is at variance with the results we obtained in studies conducted to explore the processes subjects use to make probability judgments in a missile warning officer (MWO) task (Robertson, Della-Rodolfa, & Forester, 1988). In that task, subjects are asked to judge the probability that a missile is on an attack trajectory. The judgment was based on information (a probability) generated from a "sensor system" that had high or low reliability. Figure 1 displays the aggregate results for the MWO task. The subjects estimated a higher probability of attack for high reliability information across the entire probability range. In a sense, subjects seemed to treat reliability as if it indicated that an attack was more likely, not that the information was more accurate. A brief digression about reliability will help to clarify why this pattern is not normatively correct.

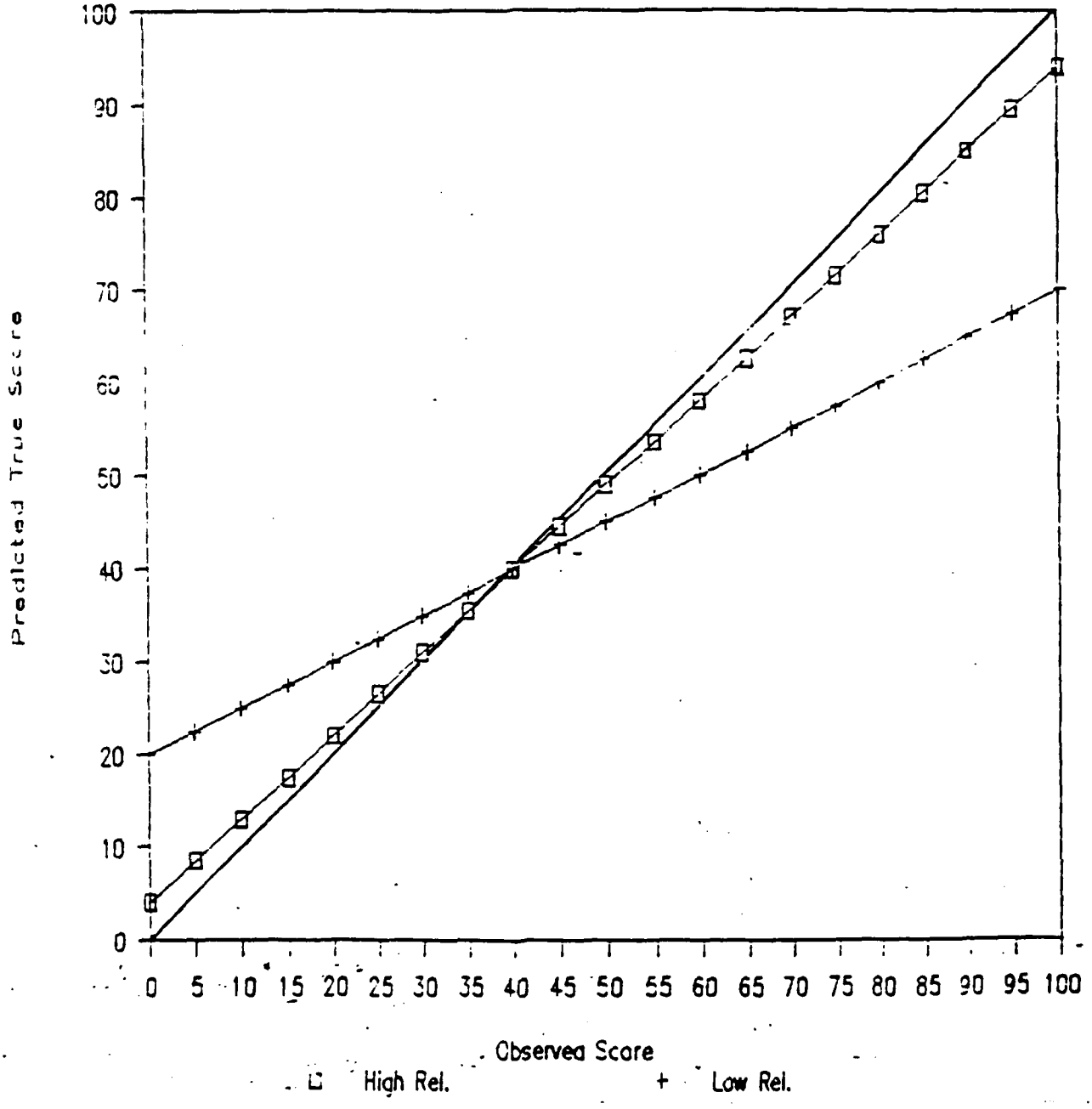

Missile Warning Officer
Aggregate Data
Probability Judgment
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
Percent Overlap
HI REL
LOW REL

Reliability Theory. The normative approach adopted here is the classical true score theory (Lord & Novick, 1968; Gulliksen, 1950). In classical true score theory, an observed measurement, X, is assumed to be the sum of two uncorrelated components, a true score (T) and an error score(E). Because T and E are uncorrelated, observed score variance is the sum of true and error score variances. Reliability is defined as the ratio of true to observed score variance, which is also the squared Pearson product moment correlation between the true and observed scores.

In the MWO study, as well as the other experiments reported in this paper, subjects were asked to judge the probability of an event based on a probability and the reliability of the "system" that generated that probability. This is analogous to being asked to predict a true score based on (a) an observed score (the sensor system output) and (b) observed score reliability (system reliability). A normatively correct response should be based on the regression of true scores on observed scores. The slope of the regression line is determined by the reliability of the measurement; the higher the reliability the greater the slope. Furthermore, the line will pass through the average true score, which is equal to the average observed score.

Two measures of the same construct that differ only in their reliabilities should pass through the same mean true score, but differ in their slopes. That is, they should be two intersecting

Regression of True on Observed Scores



straight lines. This is illustrated in Figure 2. Note that the estimated true scores for the low reliability measure exceed the estimated true scores for the high reliability measure when the observed score is below the mean. The opposite pattern occurs for observed scores above the mean.

To apply these concepts to the MWO task let us first consider the function for a perfectly reliable system. For a perfectly reliable system, the subjects judgment (S) would be identical to the system output (p). If the sensor system gave a probability of attack of 10% the subject would estimate a 10% probability of attack. The regression of S on p (analogous to the regression of T on X) would be the line S = p and would have a slope of one. As reliability decreases the slope of the line would decrease. This can be visualized as a rotation of the line S = p about a point that corresponds to the average true score. So long as the average true score is greater than zero, the line for a less than perfectly reliable system should intersect the line S = p.

An important question is what corresponds to the average true score in tasks like the MWO task. An answer to the question can be based on the notion that in the absence of any information, one should predict the mean if the goal is to minimize error. In the MWO task this would correspond to the base rate of attack. Previous research has demonstrated that the subjects were sensitive to different base rates, whether they

were provided in the form of specific probabilities or in the form of readiness state. So long as subjects attend to base rates, which are probabilities greater than zero and less than one, the following properties should hold for the empirical regression of S on p:

(1) The obtained functions for high and low reliability sensor systems should intersect.

(2) The point of intersection should be somewhere on the line defined by S = p.

(3) Subjects judgments for low reliability signals should be higher than judgments for high reliability signals if the signals are below the point of intersection.

(4) Subjects judgments for low reliability signals should be lower than judgments for high reliability signals if the signals are above the point of intersection.

Finally, two limiting conditions should be noted. If subjects adopt a base rate of zero, then all judgments should be less than the system output (i.e., $S < p$ for all p). If subjects adopt a base rate of one, then all judgments should be greater than the system output (i.e., $S > p$ for all p).

MWO Results. Now let us examine Figure 1 in terms of the four properties of normatively correct judgments. The two functions did not intersect, thus neither properties (1) or (2) were satisfied. Property (3) was not satisfied because the low reliability function was always lower than the high reliability

function. At high levels of p, the subjects' judgments appeared to be normative, but apparently this is an artifact of the heuristic that high reliability means higher probability. Finally, the pattern of results is not consistent with the subjects using a mean of zero, because all the data points are not below the line defined by S = p.

The four studies reported here were designed to test the generality of this misinterpretation of reliability hereafter referred to a the reliability bias. By the reliability bias, I mean the use of the rule that high reliability means higher probability as a substitute for the normatively correct, accuracy interpretation of reliability. Because the reliability bias has been demonstrated only with the MWO task, it is possible that it is the task itself that induces the bias and not a more general misunderstanding of the concept of reliability. The following series of four studies manipulated task components that have been found to affect human judgment in other decision making or judgment tasks.

EXPERIMENT ONE

One possibility is that the results obtained in the MWO task were in part a result of the fact that the subjects were judging the probability of a negative outcome. There is considerable research which indicates that probability judgments may change as a result of the valence of the outcome (e.g. Prospect Theory). In this Experiment 1, subjects were asked to judge the

probability that a missile would hit an enemy target. That is, the subject adopted the role of a missile launch offi[illegible] rather than a missile warning officer. The event to be judged, a missle hitting its intended target, was presumed tohave a more positive valence than the outcome in the MWO task.

Method

Subjects. Twenty-four students enrolled in introductory psychology classes received course credit for participating in the study. Each subject attended two one-hour sessions conducted approximately one week apart.

General Procedure. The general design of the study was similar to that of the MWO task. Subjects were presented with 20 warm-up trials, followed by one block of 42 trials in the first session. In the second session, subjects completed 5 warm-up trials and two blocks of 42 trials. Each block of 42 trials consisted of 21 trials for a low reliability signal and 21 trials for a high reliability signal. The 21 trials for low reliability corresponded to p values of 0 to 100% in increments of 5%. Trials within blocks were randomized for each subject.

To insure that subjects were attending to the information, they were required to keep a log which consisted of the trial number, the signal system probability, and the system reliability. Probabilities were classified as low, medium -, medium +, or high. System reliability was classified as high or low.

Missile Launch Officer Task. The task was identical to that used for the MWO studies except that the information was described as indicating the likelihood that a missile would hit its intended target. Subjects were seated in front of an IBM PC on which all stimuli were presented. The display consisted of three areas. The largest part of the display was devoted to a map of two fictitious countries, UZ and THOR. On each trial, the trajectory and potential impact region of a missile launch was represented as a section of a circle which partially overlapped with the intended target area. The size of the segment represented reliability information; the larger the segment the lower the reliability. The amount of overlap with the intended target country represented the probability of a hit.

A second area to the right of the map display gave quantitative information about the trial. In addition to the trial number, subjects were presented in the amount of overlap, p, and a number which indicated system reliability. The third area of the display, immediately below the map display, was used to prompt the subject for a response. Subjects entered their estimate of the probability of a hit for each trial, and the computer responded with a statement indicating what the subject had responded and a request for verification. After response verification, the information for the next trial was given.

Task Instructions. Subjects were given a typed copy of the following instructions which were read to him or her by the

experimenter. Phase II instructions were read immediately after completing the warm-up trials.

Missile Launch Officer Instructions

This is an experiment about how people make judgments when they are given information that is ambiguous or uncertain. The judgments you will be asked to make are whether or not a missile launched by the fictitious country UZ will hit the country THOR. Your judgment will be based on information provided by guidance systems which transmit information back to the launch officer, you. On the basis of this information you will be asked to judge the chances that the missile will hit THOR.

The uncertainty about your judgment is based on the fact that the guidance system is not completely reliable. The evil Thorians have jamming systems which will disturb the signals so we know only the general area in which the missile may land. Fortunately, we can measure this disturbance and the information you will be given will incorporate the potential unreliability of the guidance system information.

In order to perform the task you will need a thorough knowledge about the information you will be given. There are three major elements of the information: (a) geography, (b) guidance system reliability and missile trajectory, and (c) projected missile impact region or overlap.

GEOGRAPHY

In this simulation, you are a missile launch officer for the country UZ and will monitor all missile launches from several launch sites. The two countries, UZ and THOR, are displayed on the monitor and are separated by water. You are concerned only with whether a missile will hit the land area of THOR. That is, if the missile lands in the water it will have no effect on the wicked Thorians.

MISSILE TRAJECTORY AND GUIDANCE SYSTEM

The trajectory information you will be given is actually an integration of information from the missile guidance system and a satellite based heat sensor system. Because of the jamming of the guidance system signals, launches are also tracked by satellites which sense the amount of heat emitted from the missile. The problem with the satellite system is that other heat sources may look

much like a missile. As a result, we do not know the exact direction the missile is traveling.

These factors are taken into account by a computer based decision aid which represents trajectory information not as a single line (vector) but as a fan. The fan automatically takes into account all the factors that may produce error and the greater the amount of error the larger the fan is. Thus, two missiles launched toward the same target may give very different fan sizes depending upon the amount of error in the satellite and guidance systems.

PROJECTED MISSILE IMPACT REGION

Trajectory information then gives some idea of what direction the missile is traveling and how far it may go. These data along with the information about error result in a projected missile impact region. We are 99% sure that the missile will land somewhere within the impact region.

The amount of overlap between the projected missile impact region and THOR is one indication of how likely it is that the missile will hit THOR. You will be given the exact amount of overlap as part of the information you should use to make your judgment. The amount of overlap, however, does not depend on the size of the impact region. That is, highly reliable information can give the same amount of overlap as unreliable information. Also remember that the overlap is the proportion of the projected missile impact region that covers THOR, not the proportion of THOR covered by the impact region.

WARM-UP INSTRUCTIONS

The primary job of a missile launch officer is to monitor information, complete reports and make judgments. Each time there is a launch, you will complete a brief report or log and then estimate the probability of a hit.

The log functions as a back up for the display information. During the first phase you will record these pieces of information:

Event number
Reliability of guidance & satellite systems
Proportion of overlap for projected impact region

The reliability of the guidance and satellite systems will be coded as LOW or HIGH. The coding scheme is shown at the top of your log and indicates that 15 should be coded as

LOW and 30 as HIGH.

There are four classifications for the amount of overlap: LOW, MED-, MED+, and HIGH. Overlap percentages ranging from 0 to 25% are coded LOW, 26% to 50% MED-, 15% to 75% MED+, and 76% to 100% HIGH.

After you have completed the event log, you will enter what you think the probability of a hit is. Hit means simply that the missile impact will be somewhere on the land mass of THOR. The computer will prompt you with the estimate you entered. BE sure to check it before going on to the next event. Your estimate of the probability of a hit should be based on the information you are given on that trial only. Each event should be considered independently, i.e., what happened on the previous launch should not influence your judgment.

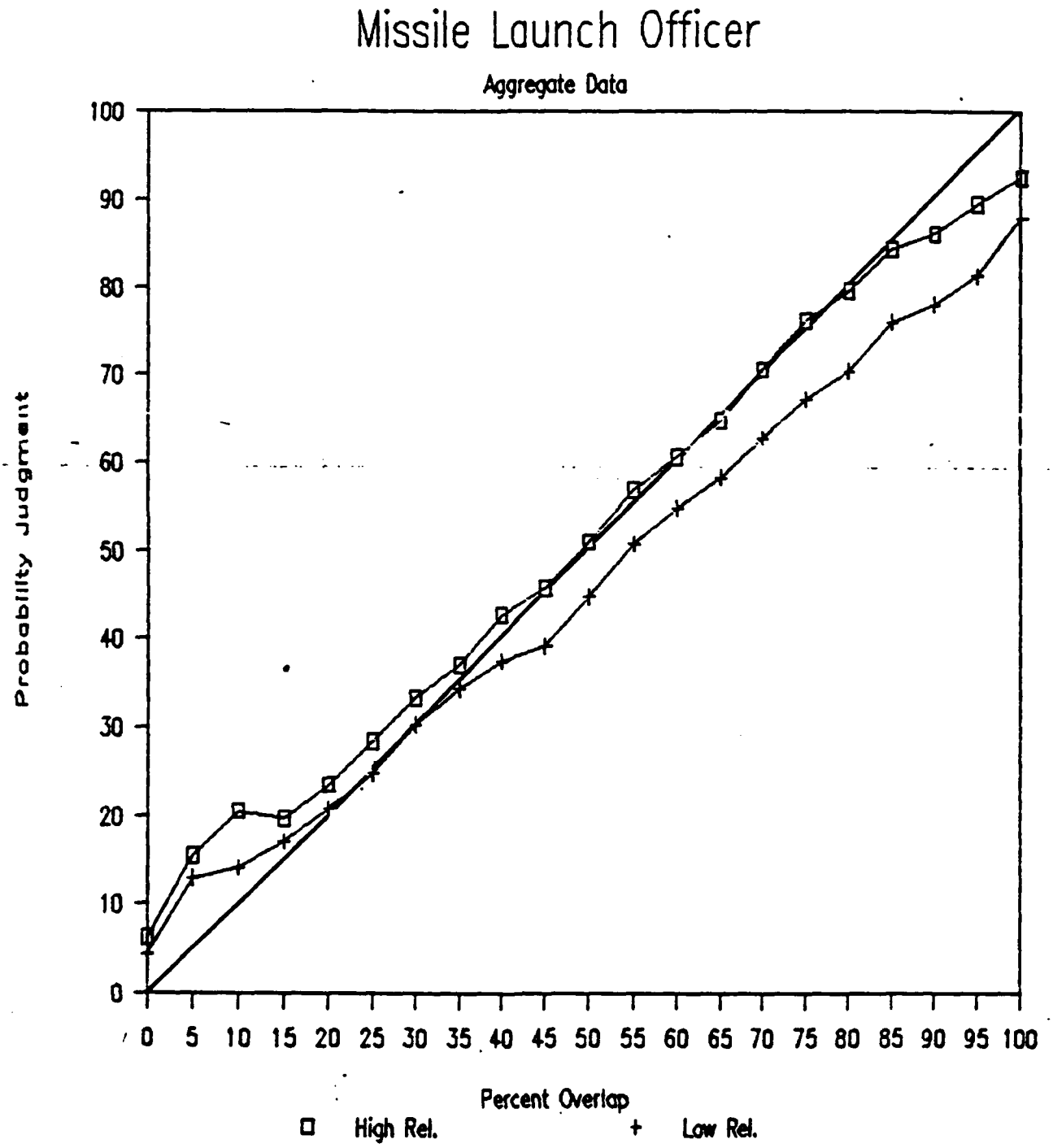
A final note. You should assume that on the average 50% of the missiles launched hit their intended target. This is based on historical records of other wars between UZ and THOR. The 50% average is based on what has happened in those previous wars.

PHASE II INSTRUCTIONS

Now that you have an idea of how this task will work, we would like you to make judgments for the following set of trials. Be sure to complete the log before you make your judgment and remember that on the average missiles will hit their intended target about 50% of the time.

Results

Figure 3 shows the aggregate results of this experiment. As can be seen, the average judgments for high reliability events were consistently higher than judgments for low reliability events. These data display a non-normative use of the reliability information which is consistent with the properties of the reliability bias. The two lines do not intersect; there is no point at which judgments of low reliability stimuli are higher than the corresponding judgments of stimuli produced by


Missile Launch Officer
Aggregate Data
Probability Judgment
100
90
80
70
60
50
40
30
20
10
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Percent Overlap
High Rel.
Low Rel.

the high reliability sensor system. The pattern is quite similar to that obtained in the MWO task.

The misinterpretation of the reliability information leads to non-normative judgments for signals with low probabilities. The approximate range of those low overlap percentages is between 0 and 30%. Subjects' judgments for overlap percentages in this range were used in a multivariate repeated measures analysis. The analysis was a 7 (probability: 0 to 30% in 5% increments) X 2 (reliability: high or low) X 3 (blocks of trials) completely within analysis. A multivariate analysis was conducted rather than a univariate repeated measures analysis because of the rather strong assumption of compound symmetry required in the univariate approach. The SPSSX MANOVA program was used for the analysis and the contrasts used to construct the repeated measures effects were the difference contrasts used in the REPEATED option.

The analysis revealed a significant effect of probability ($F(6,18) = 30.16$, $p < .001$) which indicates that subjects increased their probability judgments as the overlap percentage increased. The only other effect that approached significance was reliability ($F(1,23) = 3.30$, $p < .10$), which reflects the lower estimates for the low reliability judgments. If the subjects were obeying normative rules, the effect of reliability should have been significant, but the means should have been in the opposite direction. The absence of significant interactions

suggests that the effect of reliability does not vary across the range of probabilities in the analysis, nor did it change over the three blocks of trials.

EXPERIMENT TWO

The results of the first experiment suggest that subjects will continue to display the reliability bias when they are making judgments about the success of their efforts. However, in both the missile warning and missile launch officer scenarios, it is possible that inferences about the intentions of an enemy may have influenced the results. In the second experiment, the scenario was modified so that the event being judged was a natural event, the eruption of a volcano. In this case, inferences about the intentions of other people are not relevant.

Method

Subjects. Twenty-four students enrolled in introductory psychology classes received course credit for participating. As in the previous experiment, all subjects attended two individual one-hour sessions.

Design. The basic design for this study was identical to that used for the missile launch officer study. Session one was devoted to 20 warm-up trials and the first block of 42 events. Session two consisted of two blocks of 42 trials which were preceded by 5 warm-up trials. Subjects kept a log of the events which was similar to that used in experiment one.

Volcano scenario. The type of information provided in the

volcano task was essentially the same as that in the first experiment. Subjects were told that they were receiving information from heat sensors near the site of possible volcanic activity. A map was displayed on the computer and a volcanic "core" was identified. The basic notion was that if the sensor reading came from the volcanic core, then the volcano would erupt. The readings were displayed in the map as elliptical areas which partially overlapped the core. The greater the overlap the higher the chances were that the volcano would erupt, and the larger the size of the ellipse the less reliable the sensor reading was.

The information was also presented numerically in a portion of the screen next to the map display. Subjects were prompted to estimate the probability of eruption based on the sensor reading. The subject instructions were modelled after those used in the MLO task.

Results

Figure 4 shows the aggregate data for the volcano scenario. The functions are nearly identical to those obtained for the MLO task, although each function crosses the diagonal at a slightly lower point than in the MLO task. The two functions do not intersect and judgments for high reliability stimuli are always higher than those of low reliability stimuli.

A 7 X 2 X 3 repeated measures MANOVA was performed and the results indicated a significant effect of probability ($F(6,18) =$

30.30, $p < .001$) and a significant effect of reliability ($F(1,23) = 4.28$, $p < .05$). The reliability effect indicates that subjects made significantly lower judgments for the low reliability stimuli. No other effects were significant. Apparently, the reliability bias will occur in the context of natural events as well as in the context of missile launches.

EXPERIMENT THREE

The first two studies suggest that outcome valence and intentionality are not associated with the reliability bias. The third study was designed to focus on the dependent measure rather than the stimulus conditions. Previous research suggests that some judgment phenomena change as a function of whether one is making binary choices or continuous judgments. The third experiment was identical to the previous volcano experiment except that subjects were asked to judge whether they thought the volcano would or would not erupt, a yes/no judgment, rather than provide a continuous judgment of the probability of eruption.

Method

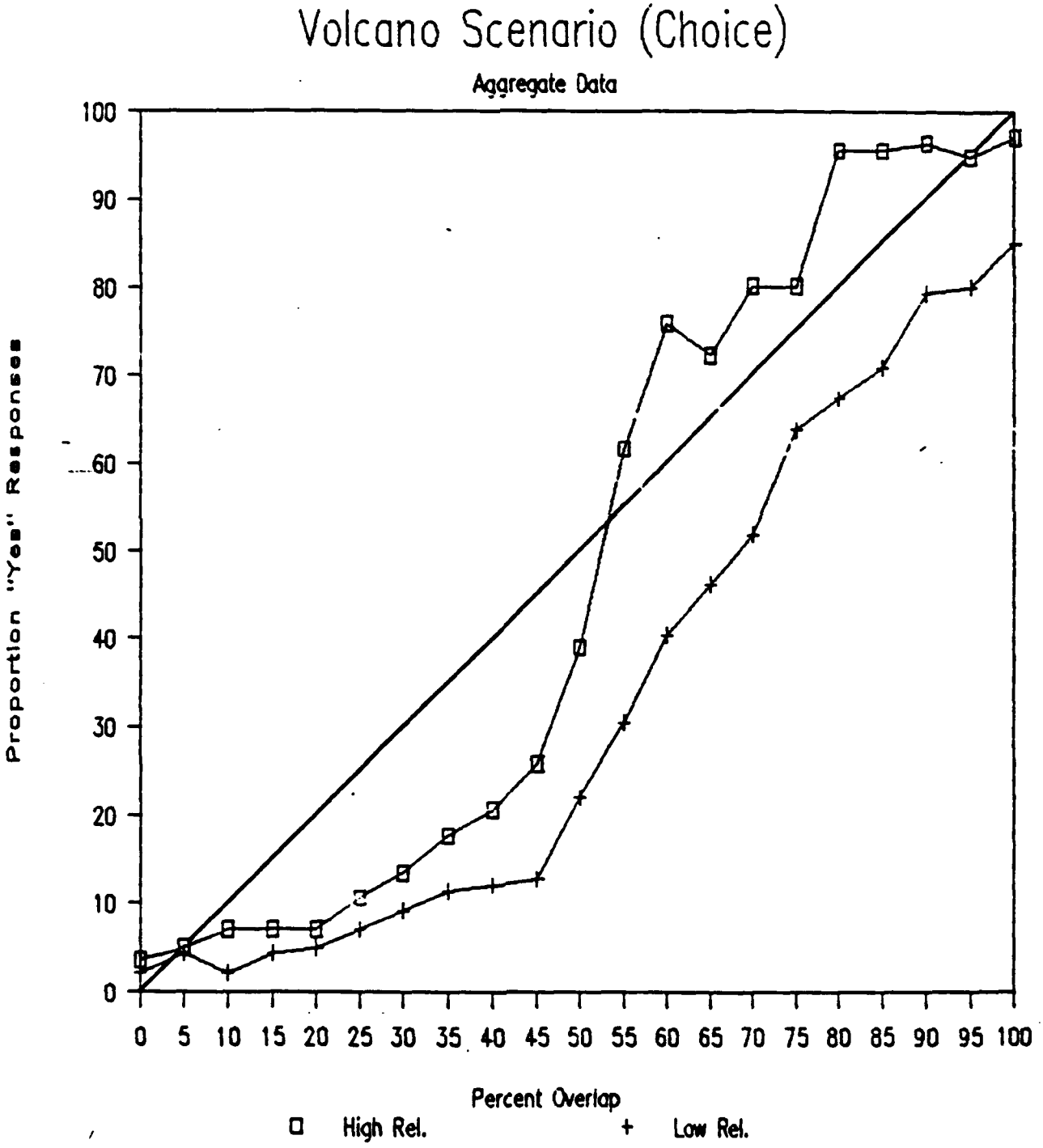
Subjects. Forty-seven undergraduate student were the subjects in this study; they received course credit for participating. All subjects were run at individual stations in a room which contained four microcomputers for two one-hour sessions.

Design and procedure. The design and procedure for this study was identical to that of Experiment Two except that

subjects responded by indicating yes or no in response to the question of whether they thought the volcano would erupt. Instructions were identical with the exception of the sentences describing the response format.

Results

In order to provide comparable data to the other studies, all responses were coded as zero or one for responses of "no" and "yes" respectively. Figure 5 shows the proportion of subjects who thought the volcano would erupt for each level of overlap. At each level of overlap, fewer subjects thought the volcano would erupt for low reliability signals than for high reliability signals. The pattern of results differs from the previous two experiments in that the function for the low reliability judgments is consistently lower than the overlap percentage. The high reliability function is lower than the overlap percentage up to about 55% overlap, and higher in the 55-100% range. The differences are probably a reflection of the choice response-format. That is, the proportion of subjects who say that the volcano will erupt is not strictly equivalent to a subjective probability estimate. However, if the choice response is conceptualized as an artificial dichotomy of an underlying continuous distribution similar to subjects probability, then there should be a region in which the proportion of "yes" responses is higher for the low reliability signals. This did not occur.


Volcano Scenario (Choice)
Aggregate Data
Proportion "Yes" Responses
100
90
80
70
60
50
40
30
20
10
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Percent Overlap
High Rel.
Low Rel.

The data were analyzed with a 7 x 2 x 3 repeated measures MANOVA. The results followed a pattern similar to those of the first two experiments. There was a significant effect of probability ($F(6,41) = 2.39$, $p < .05$) and a marginally significant effect of reliability ($F(1,46) = 3.50$, $p < .10$). None of the remaining effects approached significance.

EXPERIMENT FOUR

The results of the first three experiments are consistent in displaying the reliability effect across the scenario variations and response format. However, those three studies all displayed information about reliability in the same way. Reliability was indicated by the size of the sensor reading area and a numerical index. It could be argued that the effect is dependent on how reliability information is presented. Experiment 4 was designed to test whether a different method of representing reliability would eliminate the effect.

One way of presenting reliability information is by showing multiple sensor readings rather than a single value. Readings from a high reliability measuring instrument would be closer together, i.e., have less variance, than readings from a low reliability instrument. Furthermore, people may be more familiar with this way of representing measurement error. For example, the variability in body weight one can get by stepping on and off a scale is an indication of measurement error that many people are familiar with. Experiment 4 required subjects to make

probability judgments in the volcano scenario when reliability was represented by five "meter" readings as well as the size of sensor reading area and a numerical index.

Method

Subjects. Twenty-four undergraduate students received course credit for participating in the study. Subjects completed two one-hour sessions and were run individually in a room containing four microcomputer stations.

Design and procedure. The design and procedure was identical to Experiment 2 except that subjects were shown five overlap readings (meter readings) before they made their response. The five readings were obtained by a normal random number generator. The standard deviation of the low reliability readings was three times larger than that for high reliability readings that gave five values based on a mean overalp (the 21 values inthe 0-100% range employed in the previous experiments) and one of two standard deviations. If the random value was lower than zero, "No Reading" was displayed; if the value was higher than 100%, 100% was displayed. The subject controlled presentation of the five readings and after the fifth reading the average reading was displayed.

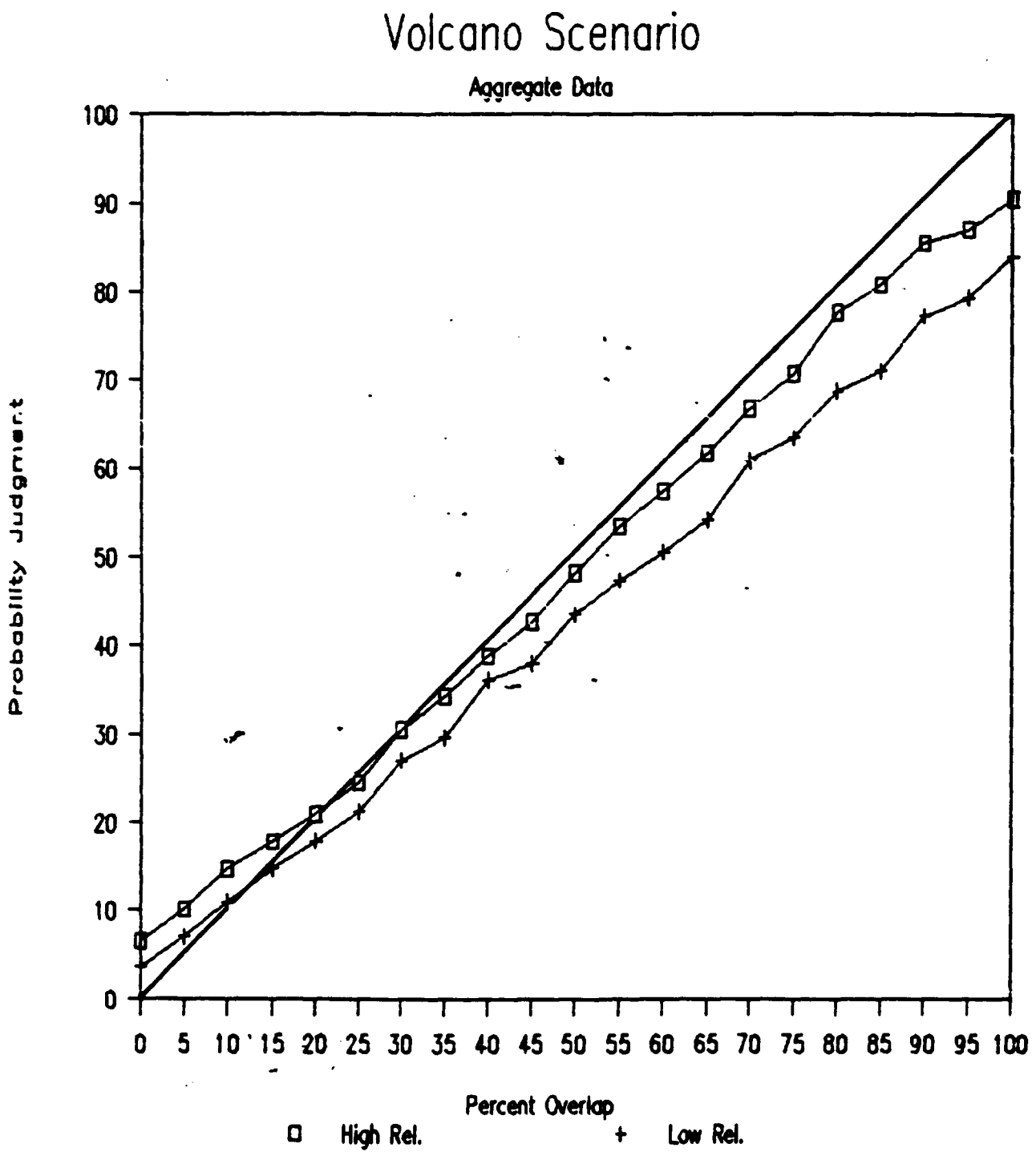
Each reading was shown on the screen in the area immediately below the core map. Readings were represented by a variable length rectangle above a scale that ranged from 0-100%. The exact reading was printed at the end of the rectangle. After the

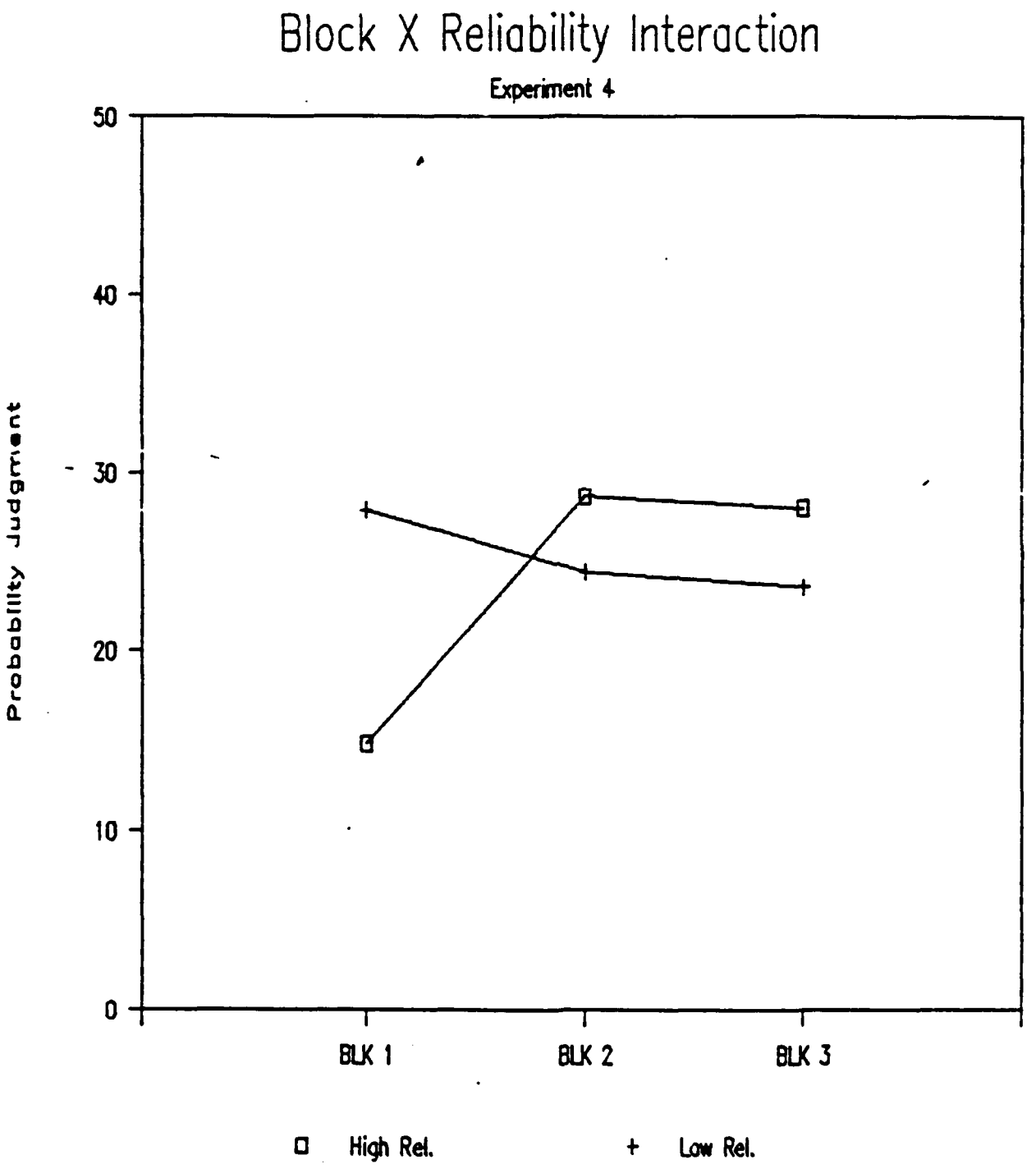
fifth reading, the meter display was replaced with the response prompts. Subjects completed the log using the average meter reading and the reliability information given in the information display to the right of the core map. Thus, the subject had available to him or her the same information as subjects in the previous three experiments; the information was supplemented by the meter display.

Results

Figure 6 shows the aggregate results for Experiment 4. This pattern differs somewhat from that obtained in Experiment 2, in which reliability was not represented by meter readings, in that the average judgments are higher. However, the aggregate functions for the low reliability judgments was consistently lower than that obtained for high reliability judgments.

The MANOVA results indicated significant main effects for probability ($F(6,18) = 38.72$, $p < .01$) and blocks ($F(2,22) = 4.07$, $p < .05$), but not for reliability ($F(1.23) = 1.42$, n.s.). In addition, there was a significant reliability by blocks interaction ($F(2,22) = 6.20$, $p < .01$). Figure 7 shows the interaction. The data points are the average judgments across the seven levels of probability (overlap percentages of 0-30%); if subjects' judgments were identical to the overlap percentage, the mean judgment would be 15%. The pattern of the interactions suggests that in the first block of trials, subjects were normatively correct because their judgments for the high


Volcano Scenario
Aggregate Data
Probability Judgment
100
90
80
70
60
50
40
30
20
10
0
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Percent Overlap
High Rel.
Low Rel.


Block X Reliability Interaction
Experiment 4
Probability Judgment
50
40
30
20
10
0
BLK 1
BLK 2
BLK 3
High Rel.
Low Rel.

reliability signals were closer to the overlap readings than judgments for low reliability readings. However, on blocks two and three, the pattern reversed with higher judgments for the high reliability signals. The addition of meter readings thus appears to have prevented the reliability bias from occurring in the first block of trials, but with "experience" the bias emerged on the last two blocks of trials.

GENERAL DISCUSSION

The results of these four experiments suggest that the reliability bias is a relatively robust effect that is not easily modified by task manipulations. Subjects appear to confuse high reliability with high probability regardless of whether the event has a positive or negative valence, whether event occurrence is a result of the actions of other people or nature, and whether the response is choice or judgment. The only factor that had any effect on the presence of the bias was presenting reliability information as a series of readings. However, the debiasing effect only occurred in the first of three blocks of trials.

Using reliability information correctly requires some counterintuitive judgments. Because judgments based on low reliability signals should be more regressive than judgments of high reliability signals, there are instances when high reliability should produce lower probability estimates. This occurs when the information, percentage overlap in these studies, is not s strong indicant of event occurrence. The judgment rule

is complicated, because if the information is highly indicative of event occurrence, high reliability should be associated with higher probability judgments. The complexity of mapping high (reliability) onto low (probability) may be central to the effect.

This linguistic explanation of the effect is consistent with the results of Experiment 4. The multiple meter readings during the initial stages of the experiment may have inhibited the linguistic association of high reliability with high probability, in part because much more information was being processed. As the link between the multiple readings and high or low reliability became stronger over time, in part because the subjects were coding the reliability on the log sheet, subjects may have tended to rely more exclusively on high/low reliability rather than the meter readings. Reliance on high/low reliability rather than meter readings then results in the reliability bias.

This account of the effect is consistent with a connectionist model in which the situation level connections between reliability information and the judgment become reinforced over time, while the connections between individual meter readings and the judgment are systematically weakened. This approach to studying the basis of the reliability bias is currently being studied.

References

Brehmenr, B. (1970). Inference behavior in a situation where the cues are not reliably perceived. Organizational Behavior and Human Performance, 5, 330-347.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80, 237-251.

Kruglanski, A. W., Friedland, N., & Farkash, E. (1984). Lay person's sensitivity to statistical information: The case of high perceived applicability. Journal of Personality and Social Psychology, 46, 503-518.

Lord, F. M., & Novick, M. R. (1968). Statiscial theories of mental test scores. Reading, Mass.: Addison-Wesley.

Robertson, D. U., Della-Rodolfa, T., and Forester, J. (1988). Probabilistic inference and ambiguity in a missile warning officer task. Paper presented at Eleventh Symposium on Psychology and the Department of Defense, Colorado Springs, Colorado.

Stable Grasping with the Utah/MIT Dexterous Robot Hand

FINAL REPORT TO:
Universal Energy Systems
4401 Dayton-Xenia Rd.
Dayton, Ohio 45432-1894

June 18, 1991

SUBMITTED BY:
Dr. Michael M. Stanišić
Assistant Professor of Mechanical Engineering
Aerospace and Mechanical Engineering Department
University of Notre Dame
Notre Dame, Indiana, 46556
(219) 239-7897

CONTRACT NO. F49620-88-C-0053/SB5881-0378

Abstract

This report contains a description of the accompanying algorithm for stable grasping with the Utah/MIT robot hand. All the variables used have been defined in the code and this report supplements the comments in the code to ensure proper understanding and usage of the variables and the setup of the algorithm. A detailed description of the structure of the input data file is also included as it is critical to the proper functioning of the algorithm.

1. Introduction

The code submitted with this report is for stable grasping with the Utah/MIT dexterous robot hand. The program titled "Fourfin.C" is for the case of three fingers and a thumb and the code titled "Threefin.C" is for the case of two fingers and a thumb. The file titled "data" contains the input data to both of the programs, since the case with two fingers and a thumb is just a modification of the original case with three fingers and a thumb. The structure of the input file will be discussed in reference to both codes later in this report. This autonomous control algorithm is based on Kerr and Roth's paper (Analysis of Multifingered Hands, IJRR, Vol. 4, No. 4 Winter 1986). It takes as its input the joint angles of each finger, the external forces acting on the object to be graspedand the coefficients of torsional (to be determined experimentally) and static friction. The algorithm then determines wether a stable grasp is possible or not and if it is, it outputs the joint torques necessary for a stable grasp. These joint torques are based on the determination of the magnitudes of the internal grasp forces, which are computed using the Simplex Method of linear programming. A step by step explanation of the implementation of this algorithm is presented in the following discussion and a flow chart has also been included in this report. The enclosed code has been extensively commented and all the important variables have been defined in those comment statements. It is therefore advised that reference should be made to the enclosed code for an explanation of terms used in the ensuing discussion.

2. Kinematic Analysis of Finger and Thumb

In this section the final transformation matrices that transform the coordinate system attached to the finger tip to the base coordinate system for both the finger and the thumb are presented, followed by the final position equations for the finger and thumb. For a more detailed kinematic analysis the reader is advised to refer to the paper titled "Kinematic Mappings from the ADL Handmaster to the Utah/MIT Dexterous Robot Hand".

To simplify writing the matrix, the following notations have been used,

$$\beta = \gamma + \psi_1 + 2\psi_2$$

$$\alpha = \gamma + \psi_1 + \psi_2$$

The transformation matrix for each finger is as follows,

$$(M_{03})_f = \begin{bmatrix} -c\beta & s\beta & 0 & l_3\, c\beta + r_1 + r_2\, c\gamma + l_1\, c(\gamma+\psi_1) + l_2\, c\alpha \\ -s\beta\, c\psi_0 & -c\beta\, c\psi_0 & -s\psi_0 & l_3\, s\beta\, c\psi_0 + (r_2\, s\gamma + l_1\, s(\gamma+\psi_1) + l_2\, s\alpha)\, c\psi_0 + h_f \\ -s\beta\, s\psi_0 & -c\beta\, s\psi_0 & c\psi_0 & l_3\, s\beta\, s\psi_0 + (r_2\, s\gamma + l_1\, s(\gamma+\psi_1) + l_2\, s\alpha)\, s\psi_0 + r_f \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The final position equations for the finger in the base coordinate system are,

$$(\bar{r_0})_f = \begin{bmatrix} r_1 + r_2\, c\gamma + l_1\, c(\gamma+\psi_1) + l_2\, c\alpha + l_3\, c\beta - x_c\, c\beta + y_c\, s\beta \\ (r_2\, s\gamma + l_1\, s(\gamma+\psi_1) + l_2\, s\alpha + l_3\, s\beta - x_c\, s\beta - y_c\, c\beta)\, c\psi_0 - z_c\, s\psi_0 + h_f \\ r_2\, s\gamma + l_1\, s(\gamma+\psi_1) + l_2\, s\alpha + l_3\, s\beta - y_c\, c\beta - x_c\, s\beta)\, s\psi_0 + z_c\, c\psi_0 + r_f \end{bmatrix}$$

The transformation matrix for the thumb is,

$$(M_{03})_t = \begin{bmatrix} -c(\psi_1 - 2\psi_2) & -s(\psi_1 - 2\psi_2) & 0 & l_3 c(\psi_1 - 2\psi_2) + l_1 c\psi_1 + l_2 c(\psi_1 - \psi_2) \\ -s(\psi_1 - 2\psi_2)c\psi_0 & c(\psi_1 - 2\psi_2)c\psi_0 & s\psi_0 & l_3 s(\psi_1 - 2\psi_2)c\psi_0 + (r_1 + l_1 s\psi_1 + l_2 s(\psi_1 - \psi_2))c\psi_0 \\ -s(\psi_1 - 2\psi_2)s\psi_0 & c(\psi_1 - 2\psi_2)s\psi_0 & -c\psi_0 & l_3 s(\psi_1 - 2\psi_2)s\psi_0 + (r_1 + l_1 s\psi_1 + l_2 s(\psi_1 - \psi_2))s\psi_0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The final position equations for the thumb in the base coordinate system are,

$$(\bar{r_0})_t = \begin{bmatrix} l_1\, c\psi_1 + l_2\, c(\psi_1 - \psi_2) + l_3\, c(\psi_1 - 2\psi_2) - x_c\, c(\psi_1 - 2\psi_2) - y_c\, s(\psi_1 - 2\psi_2) \\ (r_1 + l_1\, s\psi_1 + l_2\, s(\psi_1 - \psi_2) + l_3\, s(\psi_1 - 2\psi_2) + y_c\, c(\psi_1 - 2\psi_2) - x_c\, s(\psi_1 - 2\psi_2))\, c\psi_0 + z_c\, s\psi_0 \\ (r_1 + l_1\, s\psi_1 + l_2\, s(\psi_1 - \psi_2) + l_3\, s(\psi_1 - 2\psi_2) + y_c\, c(\psi_1 - 2\psi_2) - x_c\, s(\psi_1 - 2\psi_2))\, s\psi_0 - z_c\, c\psi_0 \end{bmatrix}$$

3. Geometry of Finger and Thumb Tip

The geometry of both the finger tips and the thumb is identical and is described by three geometric equations for the top and bottom of the fingers. The equation for the top of the finger is,

$$x - 4\,y^2 - 4\,z^2 = 0 \qquad y > 0 \qquad 0.0 < x < 0.78$$

The geometric description for the bottom of the finger tip is,

$$x - 2\,y^2 - 2\,z^2 = 0 \qquad y < 0 \qquad 0.0 < x < 0.30$$

$$z^2 + y^2 = (0.32)^2 \qquad y < 0 \qquad 0.30 < x < 0.78$$

It should be noted that discontinuities exist at the point where the geometric equations for the surface changes and contact cannot be specified for these points.

From the figure illustrating the finger geometry it is seen that the unit normal vector, $\bar{n}$, is in the C_{normal} direction, the vector, $\bar{o}$, is in the C_x direction and the vector, $\bar{a}$, is in the C_y direction. The

method of computing these unit vectors, that define the mapping between the contact forces and moments, C, and the external force and moment vector, F, is illustrated with the cas ein which contact is at the top of the finger tip, i.e.

$$x - 4\,y^2 - 4\,z^2 = 0 \qquad y > 0 \qquad 0.0 < x < 0.78$$

The contact points, which are inputs to the code are denoted by (x_c, y_c, z_c). Since the equation of the surface is given, the unit normal, $\bar{n}$, for the given contact points is found to be,

$$\bar{n} = \frac{1}{\sqrt{1 + 64\,{y_c}^2 + 64\,{z_c}^2}} \begin{bmatrix} 1 \\ -8\,y_c \\ -8\,z_c \end{bmatrix}$$

The quantity $\sqrt{1 + 64\,{y_c}^2 + 64\,{z_c}^2}$ will be referred to as $|\bar{N}|$. The vector, $\bar{o}$, is denoted in general as,

$$\bar{o} = \begin{bmatrix} o_x \\ o_y \\ o_z \end{bmatrix}$$

To calculate the three components of $\bar{o}$, there are three different conditions on the components of the unit vector, $\bar{n}$.

Case 1, $n_z \neq 0$ and $n_y \neq 0$:

In this case we take $o_z = 1$ *and* $o_y = 1$ arbitrarily. Knowing that the vectors $\bar{n}$ and $\bar{o}$ are orthogonal, the dot product o. . wo vectors has the relation,

$$\bar{n} \cdot \bar{o} = 0$$

From this dot product we can solve for o_x and thereby obtain the unit vector,

$$\bar{o} = \frac{1}{\sqrt{128\,{z_c}^2 + (1 - 8\,y_c)^2}} \begin{bmatrix} 8\,z_c \\ 8\,z_c \\ 1 - 8\,y_c \end{bmatrix}$$

where the quantity $\sqrt{128\,{z_c}^2 + (1 - 8\,y_c)^2}$ is referred to as 'mag'.

Crossing the the unit vector, $\bar{n}$ into the unit vector $\bar{o}$, we obtain the unit vector, $\bar{a}$ to be,

$$\bar{a} = \frac{1}{|\bar{N}|\,mag} \begin{bmatrix} -8\,y_c + 64\,{y_c}^2 + 64\,{z_c}^2 \\ -1 + 8\,y_c - 64\,{z_c}^2 \\ 8\,z_c + 64\,y_c\,z_c \end{bmatrix}$$

Case 2, $n_x \neq 0$ *and* $n_y \neq 0$:

In this case we arbitrarily take $o_x = 1$ *and* $o_z = 1$. Denoting the dot product as before,

$$\bar{n} \cdot \bar{o} = 0$$

After solving for o_y we obtain the unit vector, $\bar{o}$,

$$\bar{o} = \frac{1}{\sqrt{128\, y_c{}^2 + 1}} \begin{bmatrix} 8\, y_c \\ 1 \\ 8\, y_c \end{bmatrix}$$

where we will call $\sqrt{128\, y_c{}^2 + 1}$ as 'mag'. Crossing $\bar{n}$ into $\bar{o}$ we obtain,

$$\bar{a} = \frac{1}{|\bar{N}|\, mag} \begin{bmatrix} -64\, y_c{}^2 \\ -8\, y_c \\ 1 + 64\, y_c{}^2 \end{bmatrix} \quad -$$

Case 3, $n_x = 0$ *and* $n_y = 0$:

For this case we take $o_y = 1$ *and* $o_z = 1$ and from the dot product in a similar way as the previous two cases we obtain,

$$\bar{o} = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

where, as before $\sqrt{2}$ is referred to as 'mag'. Taking the cross product yields,

$$\bar{a} = \frac{1}{|\bar{N}|\, mag} \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$$

In a similar way as seen above the vectors $\bar{n}$,$\bar{o}$ *and* $\bar{a}$ are determined for contact at the bottom of the finger tips. The region of contact on each finger is determined through a number of 'if' conditions in the program.

4. Input File Structure

The input file, 'data', is the same for both the algorithms submitted and the variables read from it are structured as follows,

mu *mi*

*tmaxf*1[1] *tmaxf*1[2] *tmaxf*1[3]

*tminf*1[1] *tminf*1[2] *tminf*1[3]

*tmaxf*2[1] *tmaxf*2[2] *tmaxf*2[3]

*tminf*2[1] *tminf*2[2] *tminf*2[3]

*tmaxf*3[1] *tmaxf*3[2] *tmaxf*3[3]

*tminf*3[1] *tminf*3[2] *tminf*3[3]

tmaxtum[1] *tmaxtum*[2] *tmaxtum*[3]

tmintum[1] *tmintum*[2] *tmintum*[3]

force[1] *force*[2] *force*[3] *force*[4] *force*[5] *force*[6]

*cf*1[1] *cf*1[2] *cf*1[3]

*cf*2[1] *cf*2[2] *cf*2[3]

*cf*3[1] *cf*3[2] *cf*3[3]

ctum[1] *ctum*[2] *ctum*[3]

hf[1] *hf*[2] *hf*[3]

rf[1] *rf*[2] *rf*[3]

*psif*1[1] *psif*1[2] *psif*1[3]

*psif*2[1] *psif*2[2] *psif*2[3]

$psif3[1]$ $psif3[2]$ $psif3[3]$

$psitum[1]$ $psitum[2]$ $psitum[3]$

The maximum and minimum joint torques allowed by each joint of each finger and thumb are specified seperately in this file, for example tmaxfi[1] refers to the maximum torque allowable for the for the first joint of the first finger. The first joint refers to the joint closest to the palm of the hand and tmaxf1[3] refers to the last two joints of the first finger as we have constrained the last two joints to move through the same angle in the kinematic analysis of the hand. All the variables listed in the file structure have been defined in the accompanying code. The vectors containing the joint angles, for example $psif1[1] \ldots psif1[3]$ refer to the angles $\psi_0 \ldots \psi_2$ respectively.

For the case of two fingers and a thumb only the parts relevant to fingers one and three are used, the data for finger two is ignored but cannot be omitted from the data file as it is still read into the program but is not used. In this case the data to be modified for different grasps is,

$tmaxf1[i]$ $tminf1[i]$ $tmaxf3[i]$ $tminf3[i]$ $tmaxtum[i]$ $tmintum[i]$

$force[i]$ $cf1[i]$ $cf3[i]$ $ctum[i]$

$hf[1]$ $hf[2]$ $rf[1]$ $rf[2]$

$psif1[i]$ $psif3[i]$ $psitum[i]$

5. Description of Subroutines and Algorithm

The following discussion describes each subroutine used along with an explanation of the setup of the matrices that are used .

Subroutine 'data':

This subroutine reads in all the data from the input file and computes the direction cosines of the contact points on each finger in the base coordinate system, which is attached to the palm of the hand. After this, the matrix W which is in equation (1) of Kerr and Roth's paper is computed. This matrix

represents a mapping between the externally applied forces and moments on the object and the forces and moments acting on the finger tips of the hand. The set up of the matrix W is as follows,

$$W = \begin{bmatrix} o\bar{f}1 & a\bar{f}1 & n\bar{f}1 & 0 & \dots & o\bar{tum} & a\bar{tum} & n\bar{tum} & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots \\ r\bar{f}1 \times o\bar{f}1 & r\bar{f}1 \times a\bar{f}1 & r\bar{f}i \times n\bar{f}1 & n\bar{f}1 & \dots & r\bar{tum} \times o\bar{tum} & r\bar{tum} \times a\bar{tum} & r\bar{tum} \times n\bar{tum} & n\bar{tum} \end{bmatrix}$$

For the case of two fingers and a thumb the elements relating to finger two are omitted from the matrix W.

The Jacobian matrix for the hand is then assembled by arranging the jacobian matrices for each finger into a single matrix. The Jacobian for finger one, for example, is obtained by differentiating the position equations for the fingers and evaluating at the given input values. The set up of the Jacobian matrix for one finger is as follows, where J_1 denotes the Jacobian for finger 1,

$$J_1 = \begin{bmatrix} \frac{\partial r_x}{\partial \psi_0} & \frac{\partial r_x}{\partial \psi_1} & \frac{\partial r_x}{\partial \psi_2} \\ \frac{\partial r_y}{\partial \psi_0} & \frac{\partial r_y}{\partial \psi_1} & \frac{\partial r_y}{\partial \psi_2} \\ \frac{\partial r_z}{\partial \psi_0} & \frac{\partial r_z}{\partial \psi_1} & \frac{\partial r_z}{\partial \psi_2} \\ normf1[1] & -normf1[2]\ sin\psi_0 + normf1[3]\ cos\psi_0 & 2(-normf1[2]\ sin\psi_0 + normf1[3]\ cos\psi_0 \end{bmatrix}$$

The last row of the Jacobian comes from the torsion around the contact normal for a soft finger constraint. In a similar manner we write the Jacobian matrices for the other fingers and the thumb and assemble all of them into one matrix,

$$J = \begin{bmatrix} J_1 & 0 & 0 & 0 \\ 0 & J_2 & 0 & 0 \\ 0 & 0 & J_3 & 0 \\ 0 & 0 & 0 & J_4 \end{bmatrix}$$

Where $J_1 \dots J_4$ are the Jacobians for the fingers and the thumb respectively.

The constraint equations (5), (7), (8) and (9) from Kerr and Roth are then assembled into a matrix A and the constants on the right hand side of these equations are assembled into a matrix P.

Subroutine 'svdcmp':

This subroutine is used to compute the orthonormal basis for the null space of W. The number of linearly independent vectors that can be found for W, are given by $nc - nr$. The basis for the null space is found by,

$$W \cdot x = 0$$

where each column of the basis for the null space is occupied by a linearly independent combination of x. This is then normalized to find the orthonormal basis.

This subroutine is called by another subroutine, 'basis', which assembles the orthonormal basis vectors in a matrix called *space*. Since W is a non-square matrix in which the number of columns is greater than the number of rows, this matrix has to be augmented with extra rows of zero to make it square before it is passed into the singular value decomposition subroutine. On return this subroutine gives a vector, w, which contains the singular values and a matrix v. The columns of v whose same numbered elements w_j are zero form an orthonormal basis for the null space and are assembled into the matrix *space*. As an example, if the third element of w is zero then the third column of v would contain a vector that forms the basis for the null space.

Subroutine 'gaussj':

This subroutine computes the inverse of a square matrix by using Gauss-Jordan elimination. Since, in general, the matrix W that has to be inverted in this code is not square, we have to compute the right generalized inverse of this matrix, which is defined as,

$$W^{+} = W^{T}(WW^{T})^{-1}$$

Therefore, before calling 'gaussj' we have to find the product WW^T and then call the subroutine to compute the inverse of this product which gives a square matrix. The argument a in the subroutine is an $nr \times nr$ matrix containing the product WW^T and n is the dimension of this square matrix. The vector b just contains zeroes and is not of any concern in this code as this subroutine is not being used to find the solution vectors. On return the matrix a contains the inverse of WW^T.

Subroutine 'peninv':

This subroutine calls the subroutine 'gaussj' to compute the invers of WW^T and then assembles the right generalized inverse of the matrix W into the matrix *winv*. The matrix a contains the matrix that is to be inverted and its transpose is stored in a matrix called *wtr*.

Subroutine 'simplx':

This is a subroutine that employs the simplex method of linear programming to compute the magnitudes of the internal grasp forces within the constraints specified in Kerr and Roth's paper. This subroutine calls

three other subroutines internally, called 'simp1', 'simp2' and 'simp3'. What the simplex method essentially does is minimize a function of many variables by forming a polygon with the given constraints and maximizes the function within that polygon. There are three different kinds of constraints, m_1, m_2 and m_3. The m_1 constraints are of the form,

$$a_1 x_1 + b_1 x_2 + c_1 x_3 \leq d_1$$

m_2 of the constraints are of the form,

$$a_2 x_1 + b_2 x_2 + c_2 x_3 \geq d_2$$

m_3 of the constrints are of the form,

$$a_3 x_1 + b_3 x_2 + c_3 x_3 = d_3$$

The matrix a is the input tableau to the simplex method. Since equation (11) from Kerr and Roth's paper is what is used as the input tableau there are no m_3 constraints in this code. The constraint equation is of the form,

$$AN\lambda \geq P - A C_p \tag{1}$$

AN is a matrix of dimensions $dimen \times mak$ and $P - A C_p$ is a vector of length $dimen$. The magnitudes of the internal grasp forces are contained in the vector λ. For the set of equations in (1) the condition is that the right hand side should be greater than 0 when assembled in the simplex tableau. As an illustration of how the simplex tableau a is setup we take an example of some constraint functions and a function that is to be maximized. The following constraints are given,

$$x_1 - 2 x_2 + 3 x_3 - 4 x_4 \leq 1$$

$$2 x_1 + x_2 - 2 x_3 + x_4 \geq 3$$

$$x_1 - x_2 + 3 x_3 - x_4 \leq 4$$

the function to maximized is,

$$x_1 + x_2 + x_3 - 2 x_4$$

In this example $m_1 = 2, m_2 = 1$ *and* $m_3 = 0$. The simplex tableau is setup as follows,

	x_1	x_2	x_3	x_4
0	1	1	1	-2
1	-1	2	3	4
4	-1	1	-3	1
3	-2	-1	2	-1

The matrix a that is to be input, is contained within the double lines. The element a_{11} which is zero in the tableau on input will contain the maximum of the function that is to be maximized. The first row of the tableau contains the coefficients of the function that is to be maximized. The first column, apart from a_{11} which is zero, contains the constants on the right hand side of the constraint equations. Starting from the second column, the second, third and fourth rows contain the coefficients of the constraint equations after they have been moved to the right hand side. The order in which these constraints are arranged in the tableau is that the '$\leq$' constraints are placed first, followed by the '$\geq$' and then the '=' constraints. It is important that this order be adhered to strictly as this is the order in which the subroutine is set up to recognize these constraints.

Subroutine 'result':

This subroutine takes as its input the constraint equations (equations (7) - (9) in Kerr and Roth), the inverse of W, the orthonormal basis for the null space and the external forces and moments acting on the object to be grasped. It assembles the tableau to be input in the subroutine 'simplx' into the matrix a and calculates the magnitudes of the internal grasp forces by calling the subroutine 'simplx' in a loop to calculate each element of the vector λ. This subroutine also calculates the two vectors C_p and C_h of equation (2) from Kerr and Roth and adds them together to give the vector C, that contains the forces and moments acting on each finger tip. It is important to note that the matrix a that is the input to the subroutine 'simplx' has to be dimensioned with rows of $dimen + 2$ and columns of $mak + 1$, as internally in the subroutine, reference is made to the $n + 2$ rows and $m + 1$ columns of a. On output, this subroutine gives the vector C, that

the contact normal. The subroutine 'simplx' also returns a flag, *icase*, which is set to zero if a finite solution is found, otherwise it prints out an error message that says "No Solution Satisfies Given Constraints" and exits the program. This would indicate that a stable grasp is not possible for the given configuration of the hand or the maximum and minimum allowable joint torques for each finger are not sufficient.

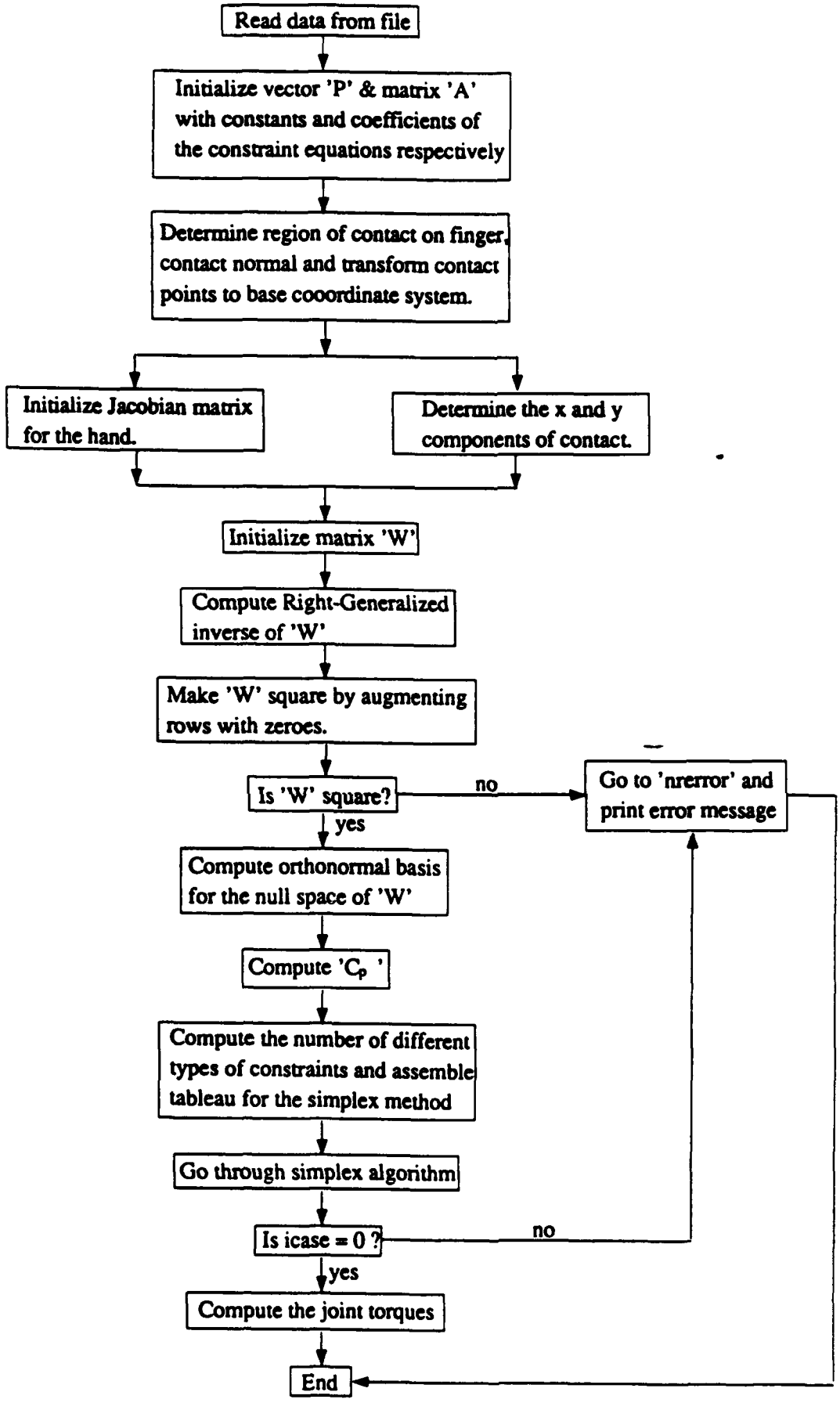
Main Program

This is where all the matrices passed in and out of the subroutines are dimensioned and the various subroutines are called to bring the control algorithm together. In all, four subroutines are called in the main program, namely, 'data', 'peninv', 'basis' and 'result'. The subroutine to multiply two matrices, 'mult', is called after 'result' and the joint torques required to achieve a stable grasp are assembled in the vector *torque* by multiplying the transpose of the Jacobian for the hand with the vector C.

6. Discussion

This algorithm was tested by giving a set of joint angles to configure the hand and as well as some numbers for maximum and minimum allowable joint torques and external forces and moments acting on a object to be grasped. The output was checked by taking a simple static force balance to see wether the forces and moments contained in the vector C were in static equilibrium with the external forces and moments applied to the object.

It is advised that the reader refer to the attached flow chart and the figures describing the kinematics of the finger and thumb to fully understand the algorithm.


Read data from file
Initialize vector 'P' & matrix 'A' with constants and coefficients of the constraint equations respectively
Determine region of contact on finger, contact normal and transform contact points to base cooordinate system.
Initialize Jacobian matrix for the hand.
Determine the x and y components of contact.
Initialize matrix 'W'
Compute Right-Generalized inverse of 'W'
Make 'W' square by augmenting rows with zeroes.
Is 'W' square?
no
Go to 'nrerror' and print error message
yes
Compute orthonormal basis for the null space of 'W'
Compute 'Cp '
Compute the number of different types of constraints and assemble tableau for the simplex method
Go through simplex algorithm
Is icase = 0 ?
no
yes
Compute the joint torques
End

```
/*************************************************/
/* THIS PROGRAM COMPUTES THE JOINT TORQUES
   REQUIRED FOR STABLY GRASPING AN OBJECT
   IN THE UTAH/MIT DEXTEROUS HAND. INPUT TO
   THE ALGORITHM IS THE EXTERNAL FORCES
   ACTING ON THE OBJECT AND THE JOINT ANGLES
   OBTAINED FROM THE INVERSE KINEMATICS OF
   THE HAND.                                   */
/*************************************************/

#include <math.h>
#include <stdio.h>
#include <alloc.h>

/*************************************************/
/* THESE ARE MACROS THAT DEFINE THE FUNCTIONS
   SWAP, fabs() AND FREEALL THAT ARE USED IN
   THIS PROGRAM BUT ARE NOT DEFINED IN ANY
   LIBRARY.                                    */
/*************************************************/
#define SWAP(a,b) {float temp=(a)\
;(a)=(b);(b)=temp;}
#define fabs(x) (((x)>0.0)?(x):-(x))
#define FREEALL free_ivector(13,1,m);\
free_ivector(12,1,m);\
free_ivector(11,1,n+1);

/*************************************************/
/* THE FOLLOWING STATEMENTS UNIVERSALLY
   DEFINE PARAMETERS AND SIZES OF MATRICES
   USED IN THIS PROGRAM.
   EPS      - DEFINES A COMPARISON PARAMETER
              THAT IS USED IN THE SUBROUTINE
              simplx (EPS CAN BE CHANGED TO
              SUITE THE MAGNITUDES OF THE
              NUMBERS IN THE DATA. HERE IT
              IS FOR NUMBERS AROUND 1.0).
   nr       - DEFINES THE NUMBER OF ROWS IN
              THE MATRIX 'W' THAT CORRESPONDS
              TO A MAPPING BETWEEN THE EXTERNAL
              FORCES ON THE HAND AND THE UNKNOWN
              CONTACT FORCES IN THE MATRIX 'C'.
   nc       - DEFINES THE NUMBER OF COLUMNS IN
              THE ABOVE MATRIX 'W'.
   normsiz  - DEFINES THE NUMBER OF COLUMNS IN
              THE ORTHONORMAL BASIS FOR THE NULL
              SPACE OF THE MATRIX 'W' (ALSO USED
              AS 'mak' IN THIS PROGRAM).
   jacsiz   - DEFINES THE NUMBER OF COLUMNS OF
              THE JACOBIAN MATRIX FOR THE HAND.
   dimen    - DEFINES THE NUMBER OF ROWS OF THE
              MATRICES 'A' AND 'P' THAT CONSIST
              OF ALL THE CONSTRAINT EQUATIONS AND
              THEIR CONSTANTS RESPECTIVELY.     */
/*************************************************/
#define EPS 1.0e-06
#define nr 6
#define nc 16
#define normsiz 10
#define jacsiz 12
#define dimen 52

/*************************************************/
/* THE FOLLOWING STATEMENTS DEFINE MACROS FOR
   FUNCTIONS THAT ARE USED IN THIS PROGRAM.
   PYTHAG    - THIS IS A FUNCTION THAT COMPUTES
```

```
                PYTHAGORA'S RULE WITHOUT
                DESTRUCTIVE OVERFLOW OR
                UNDERFLOW.
   MAX        - FINDS THE MAXIMUM OF TWO NUMBERS
   SIGN       - CHANGES THE SIGN OF A NUMBER
                IF THE SECOND ONE IS GREATER
                THAN OR EQUAL TO 0.0          */
/*********************************************/
static float at,bt,ct;
#define PYTHAG(a,b) ((at=fabs(a))> \
(bt=fabs(b)) ? \
(ct=bt/at,at*sqrt((double)(1.0+ct*ct))) \
: (bt ? (ct=at/bt,bt*sqrt((double)\
(1.0+ct*ct))):0.0))
static float maxarg1,maxarg2;
#define MAX(a,b) (maxarg1=(a),maxarg2=(b) \
,(maxarg1)>(maxarg2) ?\
(maxarg1):(maxarg2))
#define SIGN(a,b) ((b) >= 0.0 ? \
fabs(a) : -fabs(a))

/*********************************************/
/* THIS SUBROUTINE READS IN THE INPUT VALUES
   FROM A FILE AND INITIALIZES THE MATRICES
   THAT DEPEND ON THESE INPUT VALUES.
   w     - MATRIX THAT MAPS THE CONTACT FORCES
           OF THE FINGERS INTO THE EXTERNAL
           FORCES THAT ACT ON THE OBJECT TO BE
           GRASPED, BY A STATIC FORCE BALANCE.
   f     - VECTOR OF EXTERNAL FORCES ACTING ON
           THE OBJECT READ IN FROM THE DATA
           FILE.
   jt    - TRANSPOSE OF THE JACOBIAN MATRIX OF
           THE HAND.
   P     - VECTOR OF CONSTANTS CORRESPONDING
           TO THE CONSTRAINT EQUATIONS.
   A     - COEFFICIENTS OF ALL THE CONTACT
           FORCES IN THE CONSTRAINT EQUATIONS
           ARE CONTAINED IN THIS MATRIX.      */
/*********************************************/
data(w,f,jt,P,A)
float **w,**f,**jt,**P,**A;

/*********************************************/
/* fp IS A POINTER TO A FILE CONTAINING INPUT*/
/*********************************************/
FILE *fp;

/*********************************************/
/* THIS PART INITIALIZES ALL THE VARIBLES USED
   IN THIS SUBROUTINE.
   v          - A DUMMY VARIABLE USED IN READING
                IN THE DATA.
   force      - DUMMY MATRIX TO TAKE IN THE EXTERNAL
                FORCES ACTING ON THE OBJECT FROM
                THE DATA FILE.
   jacob      - JACOBIAN MATRIX FOR THE HAND.
   mu         - COEFFICIENT OF STATIC FRICTION
                BETWEEN THE OBJECT AND THE FINGERS.
   mt         - COEFFICIENT OF TORSIONAL FRICTION
                BETWEEN FINGERS AND OBJECT.
   cf1        - X,Y & Z COMPONENTS OF THE
                CONTACT POINTS ON FIRST FINGER
                GIVEN IN COORDINATE SYSTEM
                ATTACHED TO FINGERTIP.
   cf2        - CONTACT POINTS FOR SECOND FINGER
```

```
cf3       - CONTACT POINTS FOR THIRD FINGER.
ctum      - CONTACT POINTS FOR THE THUMB.
tmaxf1....tmaxtum - VECTORS CONTAINING THE
                    MAXIMUM JOINT TORQUES
                    THAT CAN BE AP  IED TO
                    EACH JOINT OF "HE FIRST,
                    SECOND & THIRD FINGER
                    AND THE THUMB.
tminf1....tmintum - VECTORS CONTAINING THE
                    MINIMUM TORQUES THAT CAN
                    BE APPLIED TO EACH JOINT
                    OF EACH FINGER AND THE
                    THUMB.
hf        - VECTOR CONTAINING THE OFFSET OF
            EACH FINGER IN THE Y DIRECTION
            OF A COORDINATE SYSTEM ATTACHED
            TO THE BASE OF THE THUMB IN THE
            PALM OF THE HAND.
rf        - VECTOR CONTAINING THE OFFSET IN
            Z DIRECTION OF THE ABOVE MENTIONED
            COORDINATE SYSTEM.
psif1....psitum - VECTORS CONTAINING THE
                  JOINT ANGLES OF EACH JOINT
                  OF EACH FINGER AND THUMB
                  COMPUTED THROUGH THE
                  INVERSE KINEMATICS OF THE
                  HAND AND READ IN FROM THE
                  DATA FILE.
gam       - PERMANENT OFFSET ANGLE OF FIRST
            DIGIT OF EACH FINGER FROM THE
            HORIZONTAL.
r1,r2,
l1,l2,l3 - CONSTANT LENGTHS OF EACH DIGIT
           IN FINGER AND THUMB (REFER TO
           FIGURE).
nf1....nftum - CONTACT NORMAL FOR FIRST
               SECOND & THIRD FINGER AND
               THUMB IN FINGER COORDINATE
               SYSTEM.
normf1....normtum - CONTACT NORMAL IN BASE
                    COORDINATE SYSTEM.
rf1....rtum - CONTACT POINT ON EACH FINGER
              AND THUMB EXPRESSED IN BASE
              COORDINATE SYSTEM.
Af1....Df3 - DUMMY VARIABLES CONTAINING
             PARTS OF TRANSFORMATION
             EQUATIONS FOR EACH FINGER
             FROM FINGERTIP COORDINATE
             TO BASE COORDINATE SYSTEM
             TO MAKE EQUATIONS SHORTER
             (SEE DERIVATIONS)
Atum....Ctum - DUMMY VARIABLES FOR ELEMENTS
               OF TRANSFORMATION EQUATION
               FOR THUMB (SEE DERIVATION)
nmagf1....nmagtum - MAGNITUDES OF THE
                    CONTACT NORMAL FOR
                    FINGER 1,FINGER 2,
                    FINGER 3 & THUMB
                    RESPECTIVELY.
omagf1....omagtum - MAGNITUDE OF THE 'X'
                    COORDINATE OF THE
                    CONTACT POINT IN THE
                    BASE COORDINATE SYSTEM.
of1....otum - 'X' COORDINATES OF THE CONTACT
              POINT IN THE BASE COORDINATE
              SYSTEM.
```

```
   af1....atum - 'Y' COORDINATES OF THE CONTACT
                 POINT IN THE BASE COORDINATE
                 SYSTEM.                          */
/************************************************/
int i,j;
float v,force[nr][1],jacob[nc][jacsiz];
float mu,mt,cf1[3],cf2[3],cf3[3],ctum[3];
float tmaxf1[3],tmaxf2[3],tmaxf3[3],tmaxtum[3];
float tminf1[3],tminf2[3],tminf3[3],tmintum[3];
float hf[3],rf[3];
float psif1[3],psif2[3],psif3[3],psitum[3];
float pi=4*atan(1.0);
float gam=0.0873;
float r1=0.50,r2=0.40,l1=0.78,l2=0.78,l3=0.78;
float nf1[3],nf2[3],nf3[3],ntum[3];
float normf1[3],normf2[3],normf3[3],normtum[3];
float rf1[3],rf2[3],rf3[3],rtum[3];
float Af1,Af2,Af3,Bf1,Bf2,Bf3,Cf1,Cf2,Cf3,Df1,Df2,Df3;
float Atum,Btum,Ctum;
float nmagf1,nmagf2,nmagf3,nmagtum;
float omagf1,omagf2,omagf3,omagtum;
float of1[3],of2[3],of3[3],otum[3];
float af1[3],af2[3],af3[3],atum[3];

/*****************************************************/
/* INITIALIZING ABOVE MENTIONED VARIABLES TO 0.0 */
/*****************************************************/
for(i=0;i<3;i++){
cf1[i]=0.0;
cf2[i]=0.0;
cf3[i]=0.0;
ctum[i]=0.0;
}

/************************************************/
/* THIS IS WHERE THE INPUT DATA IS READ IN
   FROM THE INPUT FILE 'data'                   */
/************************************************/
      fp=fopen("data","r");
      fscanf(fp,"%f%f",&mu,&mt);
      for(i=0;i<3;i++){
      fscanf(fp,"%f",&v);
      tmaxf1[i]=v;
      }
      for(i=0;i<3;i++){
      fscanf(fp,"%f",&v);
      tminf1[i]=v;
      }
      for(i=0;i<3;i++){
      fscanf(fp,"%f",&v);
      tmaxf2[i]=v;
      }
      for(i=0;i<3;i++){
      fscanf(fp,"%f",&v);
      tminf2[i]=v;
      }
      for(i=0;i<3;i++){
      fscanf(fp,"%f",&v);
      tmaxf3[i]=v;
      }
      for(i=0;i<3;i++){
      fscanf(fp,"%f",&v);
      tminf3[i]=v;
      }
      for(i=0;i<3;i++){
      fscanf(fp,"%f",&v);
```

```
    tmaxtum[i]=v;
    }
    for(i=0;i<3;i++){
    fscanf(fp,"%f",&v);
    tmintum[i]=v;
    }
    for(i=0;i<nr;i++){
    fscanf(fp,"%f",&v);
    force[i][0]=v;
    }
    for(i=0;i<3;i++){
    fscanf(fp,"%f",&v);
    cf1[i]=v;
    }
    for(i=0;i<3;i++){
    fscanf(fp,"%f",&v);
    cf2[i]=v;
    }
    for(i=0;i<3;i++){
    fscanf(fp,"%f",&v);
    cf3[i]=v;
    }
    for(i=0;i<3;i++){
    fscanf(fp,"%f",&v);
    ctum[i]=v;
    }
    for(i=0;i<3;i++){
    fscanf(fp,"%f",&v);
    hf[i]=v;
    }
    for(i=0;i<3;i++){
    fscanf(fp,"%f",&v);
    rf[i]=v;
    }
    for(i=0;i<3;i++){
    fscanf(fp,"%f",&v);
    psif1[i]=v*pi/180.0;
    }
    for(i=0;i<3;i++){
    fscanf(fp,"%f",&v);
    psif2[i]=v*pi/180.0;
    }
    for(i=0;i<3;i++){
    fscanf(fp,"%f",&v);
    psif3[i]=v*pi/180.0;
    }
    for(i=0;i<3;i++){
    fscanf(fp,"%f",&v);
    psitum[i]=v*pi/180.0;
    }
    fclose(fp);

/*************************************************/
/* VECTOR P IS INITIALIZED TO 0.0                */
/*************************************************/
for(i=1;i<=dimen;i++)P[i][1]=0.0;

/*************************************************/
/* THE MAXIMUM AND MINIMUM TORQUES FOR EACH
   JOINT ARE INITIALIZED TO THE VECTOR P      */
/*************************************************/
P[29][1]=tminf1[0];P[30][1]=tminf1[1];P[31][1]=tminf1[2];
P[32][1]=tminf2[0];P[33][1]=tminf2[1];P[34][1]=tminf2[2];
P[35][1]=tminf3[0];P[36][1]=tminf3[1];P[37][1]=tminf3[2];
P[38][1]=tmintum[0];P[39][1]=tmintum[1];P[40][1]=tmintum[2];
P[41][1]=-tmaxf1[0];P[42][1]=-tmaxf1[1];P[43][1]=-tmaxf1[2];
```

```
P[44][1]=-tmaxf2[0];P[45][1]=-tmaxf2[1];P[46][1]=-tmaxf2[2];
P[47][1]=-tmaxf3[0];P[48][1]=-tmaxf3[1];P[49][1]=-tmaxf3[2];
P[50][1]=-tmaxtum[0];P[51][1]=-tmaxtum[1];P[52][1]=-tmaxtum[2];

/*********************************************/
/* THE JACOBIAN MATRIX IS INITIALIZED TO 0.0 */
/*********************************************/
for(i=0;i<nc;i++){
for(j=0;j<jacsiz;j++){
jacob[i][j]=0.0;
}
}

/*********************************************/
/* COEFFICIENTS OF ALL THE CONSTRAINT EQUATIONS
   ARE PUT IN THE MATRIX A. THESE COEFFICIENTS
   ARE CONSTANTS.                            */
/*********************************************/
A[1][3]=1;A[2][7]=1;A[3][11]=1;A[4][15]=1;
A[8][1]=1;A[9][2]=1;A[10][4]=1;A[14][5]=1;
A[15][6]=1;A[16][8]=1;A[20][9]=1;A[21][10]=1;
A[22][12]=1;
A[26][13]=1;
A[27][14]=1;
A[28][16]=1.0;
A[5][1]=A[6][2]=A[7][4]=A[11][5]=A[12][6]=
A[13][8]=A[17][9]=A[18][10]=A[19][12]=A[23][13]=
A[24][14]=A[25][16]=-1.0;
A[5][3]=A[6][3]=A[8][3]=A[9][3]=A[11][7]=A[12][7]=
A[14][7]=A[15][7]=A[17][11]=A[18][11]=A[20][11]=
A[21][11]=A[23][15]=A[24][15]=A[26][15]=A[27][15]=mu;
A[7][3]=A[10][3]=A[13][7]=A[16][7]=A[19][11]=A[22][11]=
A[25][15]=A[28][15]=mt;

/*********************************************/
/* THESE IF CONDITIONS DETERMINE THE REGION
   OF THE CONTACT ON THE DIGIT OF EACH
   FINGER AND THUMB BASED ON THE X,Y & Z
   COMPONENTS OF THE CONTACT POINT READ IN
   FROM THE DATA FILE IN THE FINGERTIP
   COORDINATE SYSTEM. THE CONTACT NORMAL
   IN THE FINGERTIP COORDINATE SYSTEM IS ALSO
   DETERMINED IN THESE CONDITIONS.           */
/*********************************************/
if(cf1[1]>0.0 && cf1[0]>0.0 && cf1[0]<0.78){
nmagf1=sqrt((double)(1.0+64*cf1[1]*cf1[1]+64*cf1[2]*cf1[2]));
nf1[0]=1.0/nmagf1;nf1[1]=(-8*cf1[1])/nmagf1;nf1[2]=(-8*cf1[2])/nmagf1;
}
if(cf2[1]>0.0 && cf2[0]>0.0 && cf2[0]<0.78){
nmagf2=sqrt((double)(1.0+64*cf2[1]*cf2[1]+64*cf2[2]*cf2[2]));
nf2[0]=1.0/nmagf2;nf2[1]=(-8*cf2[1])/nmagf2;nf2[2]=(-8*cf2[2])/nmagf2;
}
if(cf3[1]>0.0 && cf3[0]>0.0 && cf3[0]<0.78){
nmagf3=sqrt((double)(1.0+64*cf3[1]*cf3[1]+64*cf3[2]*cf3[2]));
nf3[0]=1.0/nmagf3;nf3[1]=(-8*cf3[1])/nmagf3;nf3[2]=(-8*cf3[2])/nmagf3;
}
if(ctum[1]>0.0 && ctum[0]>0.0 && ctum[0]<0.78){
nmagtum=sqrt((double)(1.0+64*ctum[1]*ctum[1]+64*ctum[2]*ctum[2]));
ntum[0]=1.0/nmagtum;ntum[1]=(-8*ctum[1])/nmagtum;
ntum[2]=(-8*ctum[2])/nmagtum;
}
if(cf1[1]<0.0 && cf1[0]>0.0 && cf1[0]<0.3){
nmagf1=sqrt((double)(1.0+64*cf1[1]*cf1[1]+64*cf1[2]*cf1[2]));
nf1[0]=1.0/nmagf1;nf1[1]=(-8*cf1[1])/nmagf1;nf1[2]=(-8*cf1[2])/nmagf1;
}
if(cf2[1]<0.0 && cf2[0]>0.0 && cf2[0]<0.3){
```

```
nmagf2=sqrt((double)(1.0+64*cf2[1]*cf2[1]+64*cf2[2]*cf2[2]));
nf2[0]=1.0/nmagf2;nf2[1]=(-8*cf2[1])/nmagf2;nf2[2]=(-8*cf2[2])/nmagf2;
}
if(cf3[1]<0.0 && cf3[0]>0.0 && cf3[0]<0.3){
nmagf3=sqrt((double)(1.0+64*cf3[1]*cf3[1]+64*cf3[2]*cf3[2]));
nf3[0]=1.0/nmagf3;nf3[1]=(-8*cf3[1])/nmagf3;nf3[2]=(-8*cf3[2])/nmagf3;
}
if(ctum[1]<0.0 && ctum[0]>0.0 && ctum[0]<0.3){
nmagtum=sqrt((double)(1.0+64*ctum[1]*ctum[1]+64*ctum[2]*ctum[2]));
ntum[0]=1.0/nmagtum;ntum[1]=(-8*ctum[1])/nmagtum;
ntum[2]=(-8*ctum[2])/nmagtum;
}
if(cf1[1]<0.0 && cf1[0]>0.3 && cf1[0]<0.78){
nmagf1=sqrt((double)(1.0+64*cf1[1]*cf1[1]+64*cf1[2]*cf1[2]));
nf1[0]=1.0/nmagf1;nf1[1]=(-8*cf1[1])/nmagf1;nf1[2]=(-8*cf1[2])/nmagf1;
}
if(cf2[1]<0.0 && cf2[0]>0.3 && cf2[0]<0.78){
nmagf2=sqrt((double)(1.0+64*cf2[1]*cf2[1]+64*cf2[2]*cf2[2]));
nf2[0]=1.0/nmagf2;nf2[1]=(-8*cf2[1])/nmagf2;nf2[2]=(-8*cf2[2])/nmagf2;
}
if(cf3[1]<0.0 && cf3[0]>0.3 && cf3[0]<0.78){
nmagf3=sqrt((double)(1.0+64*cf3[1]*cf3[1]+64*cf3[2]*cf3[2]));
nf3[0]=1.0/nmagf3;nf3[1]=(-8*cf3[1])/nmagf3;nf3[2]=(-8*cf3[2])/nmagf3;
}
if(ctum[1]<0.0 && ctum[0]>0.3 && ctum[0]<0.78){
nmagtum=sqrt((double)(1.0+64*ctum[1]*ctum[1]+64*ctum[2]*ctum[2]));
ntum[0]=1.0/nmagtum;ntum[1]=(-8*ctum[1])/nmagtum;
ntum[2]=(-8*ctum[2])/nmagtum;
}

/***********************************************/
/* THESE ARE THE EQUATIONS THAT TRANSFORM THE
   INPUT CONTACT POINT FOR EACH FINGER AND
   THUMB FROM THE FINGERTIP COORDINATE SYSTEM
   TO THE BASE COORDINATE SYSTEM.              */
/***********************************************/
rf1[0]=r1+r2*cos((double)(gam))+l1*cos((double)(gam+psif1[1]))+
       l2*cos((double)(gam+psif1[1]+psif1[2]))+
       l3*cos((double)(gam+psif1[1]+2*psif1[2]))-
       cf1[0]*cos((double)(gam+psif1[1]+2*psif1[2]))+
       cf1[1]*sin((double)(gam+psif1[1]+2*psif1[2]));
rf1[1]=(r2*sin((double)(gam))+l1*sin((double)(gam+psif1[1]))+
       l2*sin((double)(gam+psif1[1]+psif1[2]))+
       l3*sin((double)(gam+psif1[1]+2*psif1[2]))-
       cf1[0]*sin((double)(gam+psif1[1]+2*psif1[2]))-
       cf1[1]*cos((double)(gam+psif1[1]+
       2*psif1[2])))*cos((double)(psif1[0]))-
       cf1[2]*sin((double)(psif1[0]))+hf[0];
rf1[2]=(r2*sin((double)(gam))+l1*sin((double)(gam+psif1[1]))+
       l2*sin((double)(gam+psif1[1]+psif1[2]))+
       l3*sin((double)(gam+psif1[1]+2*psif1[2]))-
       cf1[1]*cos((double)(gam+psif1[1]+2*psif1[2]))-
       cf1[0]*sin((double)(gam+psif1[1]+
       2*psif1[2])))*sin((double)(psif1[0]))+
       cf1[2]*cos((double)(psif1[0]))+rf[0];
rf2[0]=r1+r2*cos((double)(gam))+l1*cos((double)(gam+psif2[1]))+
       l2*cos((double)(gam+psif2[1]+psif2[2]))+
       l3*cos((double)(gam+psif2[1]+2*psif2[2]))-
       cf2[0]*cos((double)(gam+psif2[1]+2*psif2[2]))+
       cf2[1]*sin((double)(gam+psif2[1]+2*psif2[2]));
rf2[1]=(r2*sin((double)(gam))+l1*sin((double)(gam+psif2[1]))+
       l2*sin((double)(gam+psif2[1]+psif2[2]))+
       l3*sin((double)(gam+psif2[1]+2*psif2[2]))-
       cf2[0]*sin((double)(gam+psif2[1]+2*psif2[2]))-
       cf2[1]*cos((double)(gam+psif2[1]+
       2*psif2[2])))*cos((double)(psif2[0]))-
```

```
        cf2[2]*sin((double)(psif2[0]))+hf[1];
rf2[2]=(r2*sin((double)(gam))+l1*sin((double)(gam+psif2[1]))+
        l2*sin((double)(gam+psif2[1]+psif2[2]))+
        l3*sin((double)(gam+psif2[1]+2*psif2[2]))-
        cf2[1]*cos((double)(gam+psif2[1]+2*psif2[2]))-
        cf2[0]*sin((double)(gam+psif2[1]+
        2*psif2[2])))*sin((double)(psif2[0]))+
        cf2[2]*cos((double)(psif2[0]))+rf[1];
rf3[0]=r1+r2*cos((double)(gam))+l1*cos((double)(gam+psif3[1]))+
        l2*cos((double)(gam+psif3[1]+psif3[2]))+
        l3*cos((double)(gam+psif3[1]+2*psif3[2]))-
        cf3[0]*cos((double)(gam+psif3[1]+2*psif3[2]))+
        cf3[1]*sin((double)(gam+psif3[1]+2*psif3[2]));
rf3[1]=(r2*sin((double)(gam))+l1*sin((double)(gam+psif3[1]))+
        l2*sin((double)(gam+psif3[1]+psif3[2]))+
        l3*sin((double)(gam+psif3[1]+2*psif3[2]))-
        cf3[0]*sin((double)(gam+psif3[1]+2*psif3[2]))-
        cf3[1]*cos((double)(gam+psif3[1]+
        2*psif3[2])))*cos((double)(psif3[0]))-
        cf3[2]*sin((double)(psif3[0]))+hf[2];
rf3[2]=(r2*sin((double)(gam))+l1*sin((double)(gam+psif3[1]))+
        l2*sin((double)(gam+psif3[1]+psif3[2]))+
        l3*sin((double)(gam+psif3[1]+2*psif3[2]))-
        cf3[1]*cos((double)(gam+psif3[1]+2*psif3[2]))-
        cf3[0]*sin((double)(gam+psif3[1]+
        2*psif3[2])))*sin((double)(psif3[0]))+
        cf3[2]*cos((double)(psif3[0]))+rf[2];
rtum[0]=l1*cos((double)(psitum[1]))+l2*cos((double)(psitum[1]-psitum[2]))+
         l3*cos((double)(psitum[1]-2*psitum[2]))-
         ctum[0]*cos((double)(psitum[1]-2*psitum[2]))-
         ctum[1]*sin((double)(psitum[1]-2*psitum[2]));
rtum[1]=(r1+l1*sin((double)(psitum[1]))+
         l2*sin((double)(psitum[1]-psitum[2]))+
         l3*sin((double)(psitum[1]-2*psitum[2]))+
         ctum[1]*cos((double)(psitum[1]-2*psitum[2]))-
         ctum[0]*sin((double)(psitum[1]-
         2*psitum[2])))*cos((double)(psitum[0]))+
         ctum[2]*sin((double)(psitum[0]));
rtum[2]=(r1+l1*sin((double)(psitum[1]))+
         l2*sin((double)(psitum[1]-psitum[2]))+
         l3*sin((double)(psitum[1]-2*psitum[2]))+
         ctum[1]*cos((double)(psitum[1]-2*psitum[2]))-
         ctum[0]*sin((double)(psitum[1]-
         2*psitum[2])))*sin((double)(psitum[0]))-
         ctum[2]*cos((double)(psitum[0]));

/*************************************************/
/* THESE ARE THE DUMMY VARIABLES THAT SHORTEN
   TRANSFORMATION EQUATIONS FROM THE FINGERTIP
   COORDINATE SYSTEM TO THE BASE COORDINATE
   SYSTEM.                                       */
/*************************************************/
Af1=cos((double)(gam+psif1[1]+2*psif1[2]));
Bf1=sin((double)(gam+psif1[1]+2*psif1[2]));
Cf1=r2*sin((double)(gam))+l1*sin((double)(gam+psif1[1]))+
    l2*sin((double)(gam+psif1[1]+psif1[2]));
Df1=r1+r2*cos((double)(gam))+l1*cos((double)(gam+psif1[1]))+
    l2*cos((double)(gam+psif1[1]+psif1[2]));
Af2=cos((double)(gam+psif2[1]+2*psif2[2]));
Bf2=sin((double)(gam+psif2[1]+2*psif2[2]));
Cf2=r2*sin((double)(gam))+l1*sin((double)(gam+psif2[1]))+
    l2*sin((double)(gam+psif2[1]+psif2[2]));
Df2=r1+r2*cos((double)(gam))+l1*cos((double)(gam+psif2[1]))+
    l2*cos((double)(gam+psif2[1]+psif2[2]));
Af3=cos((double)(gam+psif3[1]+2*psif3[2]));
Bf3=sin((double)(gam+psif3[1]+2*psif3[2]));
```

```
Cf3=r2*sin((double)(gam))+l1*sin((double)(gam+psif3[1]))+
    l2*sin((double)(gam+psif3[1]+psif3[2]));
Df3=r1+r2*cos((double)(gam))+l1*cos((double)(gam+psif3[1]))+
    l2*cos((double)(gam+psif3[1]+psif3[2]));
Atum=cos((double)(psitum[1]-2*psitum[2]));
Btum=sin((double)(psitum[1]-2*psitum[2]));
Ctum=r1+l1*sin((double)(psitum[1]))+
     l2*sin((double)(psitum[1]-psitum[2]))+
     l3*sin((double)(psitum[1]-psitum[2]));

/*************************************************/
/* THE CONTACT NORMAL IS TRANSFORMED FROM
   THE FINGERTIP COORDINATE SYSTEM TO THE
   BASE COORDINATE SYSTEM.                       */
/*************************************************/
normf1[0]=-nf1[0]*Af1+nf1[1]*Bf1+Af1*l3+Df1;
normf1[1]=-nf1[0]*Bf1*cos((double)(psif1[0]))-
          nf1[1]*Af1*cos((double)(psif1[0]))+
          -nf1[2]*sin((double)(psif1[0]))+
          Bf1*l3*cos((double)(psif1[0]))+
          Cf1*cos((double)(psif1[0]))+hf[0];
normf1[2]=-nf1[0]*Bf1*sin((double)(psif1[0]))-
          nf1[1]*Af1*sin((double)(psif1[0]))+
          -nf1[2]*cos((double)(psif1[0]))+
          Bf1*l3*sin((double)(psif1[0]))+
          Cf1*sin((double)(psif1[0]))+rf[0];
normf2[0]=-nf2[0]*Af2+nf2[1]*Bf2+Af2*l3+Df2;
normf2[1]=-nf2[0]*Bf2*cos((double)(psif2[0]))-
          nf2[1]*Af2*cos((double)(psif2[0]))+
          -nf2[2]*sin((double)(psif2[0]))+
          Bf2*l3*cos((double)(psif2[0]))+
          Cf2*cos((double)(psif2[0]))+hf[1];
normf2[2]=-nf2[0]*Bf2*sin((double)(psif2[0]))-
          nf2[1]*Af2*sin((double)(psif2[0]))+
          -nf2[2]*cos((double)(psif2[0]))+
          Bf2*l3*sin((double)(psif2[0]))+
          Cf2*sin((double)(psif2[0]))+rf[1];
normf3[0]=-nf3[0]*Af3+nf3[1]*Bf3+Af3*l3+Df3;
normf3[1]=-nf3[0]*Bf3*cos((double)(psif3[0]))-
          nf3[1]*Af3*cos((double)(psif3[0]))+
          -nf3[2]*sin((double)(psif3[0]))+
          Bf3*l3*cos((double)(psif3[0]))+
          Cf3*cos((double)(psif3[0]))+hf[2];
normf3[2]=-nf3[0]*Bf3*sin((double)(psif3[0]))-
          nf3[1]*Af3*sin((double)(psif3[0]))+
          -nf3[2]*cos((double)(psif3[0]))+
          Bf3*l3*sin((double)(psif3[0]))+
          Cf3*sin((double)(psif3[0]))+rf[2];
normtum[0]=-Atum*ntum[0]-Btum*ntum[1]+l3*Atum+l1*cos((double)(psitum[1]))+
           l2*Atum;
normtum[1]=(-Btum*ntum[0]+Atum*ntum[1]+Ctum)*cos((double)(psitum[0]))+
           ntum[2]*sin((double)(psitum[0]));
normtum[2]=(-Btum*ntum[0]+Atum*ntum[1]+Ctum)*sin((double)(psitum[0]))-
           ntum[2]*cos((double)(psitum[0]));

/*************************************************/
/* EACH ELEMENT OF THE JACOBIAN MATRIX FOR THE
   HAND IS INITIALIZED.                          */
/*************************************************/
jacob[0][0]=0.0;
jacob[0][1]=-l1*sin((double)(gam+psif1[1]))-l2
            *sin((double)(gam+psif1[1]+psif1[2]))
            -l3*sin((double)(gam+psif1[1]+2*psif1[2]))+
            cf1[0]*sin((double)(gam+psif1[1]+2*psif1[2]))+cf1[1]*
            cos((double)(gam+psif1[1]+2*psif1[2]));
jacob[0][2]=-l2*sin((double)(gam+psif1[1]+psif1[2]))-
```

```
            2*l3*sin((double)(gam+psif1[1]+2*psif1[2]))+2*cf1[0]*
            sin((double)(gam+psif1[1]+2*psif1[2]))
            +2*cf1[1]*cos((double)(gam+psif1[1]+2*psif1[2]));
jacob[1][0]=-(r2*sin((double)(gam))+l1*sin((double)(gam+psif1[1]))+
            l2*sin((double)(gam+psif1[1]+psif1[2]))+l3*
            sin((double)(gam+psif1[1]+2*psif1[2]))-cf1[0]*
            sin((double)(gam+psif1[1]+2*psif1[2]))-cf1[1]*
            cos((double)(gam+psif1[1]+2*psif1[2])))
            *sin((double)(psif1[0]))-cf1[2]*cos((double)(psif1[0]));
jacob[1][1]=(l1*cos((double)(gam+psif1[1]))+l2*
            cos((double)(gam+psif1[1]+psif1[2]))+l3*
            cos((double)(gam+psif1[1]+2*psif1[2]))-
            cf1[0]*cos((double)(gam+psif1[1]+2*psif1[2]))+
            cf1[1]*sin((double)(gam+psif1[1]+2*psif1[2])))
            *cos((double)(psif1[0]));
jacob[1][2]=(l2*cos((double)(gam+psif1[1]+psif1[2]))+2*l3*
            cos((double)(gam+psif1[1]+2*psif1[2]))
            -2*cf1[0]*cos((double)(gam+psif1[1]+2*psif1[2]))+
            2*cf1[1]*sin((double)(gam+psif1[1]+2*psif1[2])))*
            cos((double)(psif1[0]));
jacob[2][0]=(r2*sin((double)(gam))+l1*sin((double)(gam+psif1[1]))+
            l2*sin((double)(gam+psif1[1]+psif1[2]))
            +l3*sin((double)(gam+psif1[1]+2*psif1[2]))-
            cf1[1]*cos((double)(gam+psif1[1]+2*psif1[2]))-
            cf1[0]*sin((double)(gam+psif1[1]+2*psif1[2])))
            *cos((double)(psif1[0]))-cf1[2]*sin((double)(psif1[0]));
jacob[2][1]=(l1*cos((double)(gam+psif1[1]))+l2*
            cos((double)(gam+psif1[1]+psif1[2]))+l3*
            cos((double)(gam+psif1[1]+2*psif1[2]))+cf1[1]*
            sin((double)(gam+psif1[1]+2*psif1[2]))-
            cf1[0]*cos((double)(gam+psif1[1]+2*psif1[2])))
            *sin((double)(psif1[0]));
jacob[2][2]=(l2*cos((double)(gam+psif1[1]+psif1[2]))+2*l3*
            cos((double)(gam+psif1[1]+2*psif1[2]))
            +2*cf1[1]*sin((double)(gam+psif1[1]+2*psif1[2]))))*
            sin((double)(psif1[0]));
jacob[3][0]=normf1[0];
jacob[3][1]=-normf1[1]*sin((double)(psif1[0]))+normf1[2]*
            cos((double)(psif1[0]));
jacob[3][2]=2.0*jacob[3][1];
jacob[4][3]=0.0;
jacob[4][4]=-l1*sin((double)(gam+psif2[1]))-l2
            *sin((double)(gam+psif2[1]+psif2[2]))
            -l3*sin((double)(gam+psif2[1]+2*psif2[2]))+
            cf2[0]*sin((double)(gam+psif2[1]+2*psif2[2]))+cf2[1]*
            cos((double)(gam+psif2[1]+2*psif2[2]));
jacob[4][5]=-l2*sin((double)(gam+psif2[1]+psif2[2]))-
            2*l3*sin((double)(gam+psif2[1]+2*psif2[2]))+2*cf2[0]*
            sin((double)(gam+psif2[1]+2*psif2[2]))
            +2*cf2[1]*cos((double)(gam+psif2[1]+2*psif2[2]));
jacob[5][3]=-(r2*sin((double)(gam))+l1*sin((double)(gam+psif2[1]))+
            l2*sin((double)(gam+psif2[1]+psif2[2]))+l3*
            sin((double)(gam+psif2[1]+2*psif2[2]))-cf2[0]*
            sin((double)(gam+psif2[1]+2*psif2[2]))-cf2[1]*
            cos((double)(gam+psif2[1]+2*psif2[2])))
            *sin((double)(psif2[0]))-cf2[2]*cos((double)(psif2[0]));
jacob[5][4]=(l1*cos((double)(gam+psif2[1]))+l2*
            cos((double)(gam+psif2[1]+psif2[2]))+l3*
            cos((double)(gam+psif2[1]+2*psif2[2]))-
            cf2[0]*cos((double)(gam+psif2[1]+2*psif2[2]))+
            cf2[1]*sin((double)(gam+psif2[1]+2*psif2[2])))
            *cos((double)(psif2[0]));
jacob[5][5]=(l2*cos((double)(gam+psif2[1]+psif2[2]))+2*l3*
            cos((double)(gam+psif2[1]+2*psif2[2]))
            -2*cf2[0]*cos((double)(gam+psif2[1]+2*psif2[2]))+
            2*cf2[1]*sin((double)(gam+psif2[1]+2*psif2[2])))*
```

```
   (REFER TO THE DERIVATIONS).
   a    - THIS IS THE SIMPLEX TABLEAU WHICH
          CONTAINS THE CONSTRAINTS FOR THE
          SYSTEM (REFER TO DERIVATIONS).
   m    - THE NUMBER OF ROWS IN 'a'.
   n    - THE NUMBER OF COLUMNS IN 'a'.
   ml   - THE NUMBER OF 'LESS THAN AND EQUAL TO'
          CONSTRAINTS.
   m2   - THE NUMBER OF 'GREATER THAN AND EQUAL
          TO' CONSTRAINTS.
   m3   - THE NUMBER OF 'EQUAL TO' CONSTRAINTS.
   icase - A FLAG THAT IS SET TO 0 IF A
           SOLUTION IS FOUND AND -1 IF NO
           SOLUTION SATISFIES THE GIVEN
           CONSTRAINTS.
   izrov & iposv - DUMMY ARRAYS THAT ARE ONLY
                   USED INTERNALLY.
/*************************************************/
simplx(a,m,n,ml,m2,m3,icase,izrov,iposv)
int m,n,ml,m2,m3,*icase,*izrov,*iposv;
float **a;
{
   int i,ip,ir,is,k,kh,kp,m12,nl1,nl2;
   int *l1,*l2,*l3,*ivector();
   float q1,bmax=0.0;
   if(m!=(ml+m2+m3))nrerror
   ("BAD INPUT CONSTRAINTS COUNT IN SIMPLX");
   l1=ivector(1,n+1);
   l2=ivector(1,m);
   l3=ivector(1,m);
   for(i=1;i<=n+1;i++){
   l1[i]=0.0;
   }
   for(i=1;i<=m;i++){
   l2[i]=0.0;
   l3[i]=0.0;
   }
   nl1=n;
   for(k=1;k<=n;k++){
   l1[k]=izrov[k]=k;
   }
   nl2=m;
   for(i=1;i<=m;i++){
      if(a[i+1][1]<0.0)nrerror
      ("Bad input tableau in SIMPLX");
      l2[i]=i;
      iposv[i]=n+i;
   }
   for(i=1;i<=m2;i++)l3[i]=1;
   ir=0;
   if(m2+m3){
     ir=1;
     for(k=1;k<=(n+1);k++){
        q1=0.0;
        for(i=m1+1;i<=m;i++)q1+=a[i+1][k];
        a[m+2][k]=-q1;
     }
     do {
        simp1(a,m+1,l1,nl1,0,&kp,&bmax);
        if(bmax<=EPS && a[m+2][1]<-EPS){
          *icase=-1;
          FREEALL return;
        } else if(bmax<=EPS && a[m+2][1]<=EPS){
            m12=m1+m2+1;
            if(m12<=m){
               for(ip=m12;ip<=m;ip++){
```

```
                if(iposv[ip] == (ip+n)){
                  simp1(a,ip,ll,nll,1,&kp,&bmax);
                  if(bmax>0.0)
                     goto one;
                }
              }
           }
           ir=0;
           --ml2;
           if(ml+1<=ml2)
              for(i=ml+1;i<=ml2;i++)
                 if(l3[i-ml] == 1)
                    for(k=1;k<=n+1;k++)
                       a[i+1][k]=-a[i+1][k];
           break;
        }
        simp2(a,n,l2,nl2,&ip,kp,&ql);
        if(ip == 0){
          *icase=-1;
          FREEALL return;
        }
one:    simp3(a,m+1,n,ip,kp);
        if(iposv[ip] >= (n+ml+m2+1)){
           for(k=1;k<=nll;k++)
              if(ll[k] == kp)break;
           --nll;
           for(is=k;is<=nll;is++)ll[is]=
                                   ll[is+1];
           a[m+2][kp+1] += 1.0;
           for(i=1;i<=m+2;i++)a[i][kp+1]=
                                -a[i][kp+1];
        } else {
           if(iposv[ip]>=(n+ml+1)){
             kh=iposv[ip]-ml-n;
             if(l3[kh]){
                l3[kh]=0;
                a[m+2][kp+1]+=1.0;
                for(i=1;i<=m+2;i++)
                   a[i][kp+1]=-a[i][kp+1];
             }
           }
        }
        is=izrov[kp];
        izrov[kp]=iposv[ip];
        iposv[ip]=is;
     } while(ir);
   }
   for(;;){
      simp1(a,0,ll,nll,0,&kp,&bmax);
      if(bmax <= 0.0){
        *icase=0;
        FREEALL return;
      }
      simp2(a,n,l2,nl2,&ip,kp &ql);
      if(ip ==0){
        *icase=1;
        FREEALL return;
      }
      simp3(a,m,n,ip,kp);
      is=izrov[kp];
      izrov[kp]=iposv[ip];
      iposv[ip]=is;
   }
}

/*************************************************/
```

```
/* THIS SUBROUTINE IS CALLED UP INTERNALLY IN
   THE SUBROUTINE simplx AND IS PART OF THE
   SIMPLEX METHOD.                                  */
/*************************************************/
simpl(a,mm,ll,nll,iabf,kp,bmax)
float **a,*bmax;
int mm,*ll,nll,iabf,*kp;
{
   int k;
   float test;
   *kp=ll[1];
   *bmax=a[mm+1][*kp+1];
   for(k=2;k<=nll;k++){
      if(iabf == 0)
        test=a[mm+1][ll[k]+1]-(*bmax);
      else
        test=fabs(a[mm+1][ll[k]+1])
                        -fabs(*bmax);
      if(test > 0.0){
        *bmax=a[mm+1][ll[k]+1];
        *kp=ll[k];
      }
   }
}

/*************************************************/
/* THIS SUBROUTINE IS CALLED UP BY THE
   SUBROUTINE simplx AND IS PART OF THE SIMPLEX
   METHOD.                                          */
/*************************************************/
simp2(a,n,l2,nl2,ip,kp,q1)
int n,*l2,nl2,*ip,kp;
float **a,*q1;
{
   int k,ii,i;
   float qp,q0,q;
   *ip=0;
   for(i=1;i<=nl2;i++){
      if(a[l2[i]+1][kp+1] < -EPS){
        *q1=-a[l2[i]+1][1]/a[l2[i]+1][kp+1];
        *ip=l2[i];
        for(i=i+1;i<=nl2;i++){
           ii=l2[i];
           if(a[ii+1][kp+1] < -EPS){
             q=-a[ii+1][1]/a[ii+1][kp+1];
             if(q < *q1){
               *ip=ii;
               *q1=q;
             } else if(q == *q1){
                for(k=1;k<=n;k++){
                   qp=-a[*ip+1][k+1]/
                       a[*ip+1][kp+1];
                   q0=-a[ii+1][k+1]/
                       a[ii+1][kp+1];
                   if(q0 != qp)break;
                }
                if(q0 < qp) *ip=ii;
             }
           }
         }
       }
    }
}

/*************************************************/
/* THIS ROUTINE IS CALLED UP BY THE SUBROUTINE
```

```
   simplx AND IS PART OF THE SIMPLEX METHOD. */
/*********************************************/
simp3(a,il,kl,ip,kp)
int il,kl,ip,kp;
float **a;
{
   int kk,ii;
   float piv;
   piv=1.0/a[ip+1][kp+1];
   for(ii=1;ii<=il+1;ii++)
      if(ii-1 != ip){
        a[ii][kp+1] *= piv;
        for(kk=1;kk<=kl+1;kk++)
           if(kk-1 != kp)
             a[ii][kk] -= a[ip+1][kk]*
                         a[ii][kp+1];
      }
   for(kk=1;kk<=kl+1;kk++)
      if(kk-1 != kp)a[ip+1][kk] *= -piv;
   a[ip+1][kp+1]=piv;
}

/*********************************************/
/* THIS SUBROUTINE ASSEMBLES THE MATRIX THAT
   IS PASSED INTO THE SUBROUTINE 'simplx'
   ASSEMBLES THE MAGNITUDES OF THE GRASP FORCES
   INTO ANOTHER MATRIX FROM WHICH THE JOINT
   TORQUES CAN BE COMPUTED.
   a       - MATRIX THAT CONTAINS THE SIMPLEX
             TABLEAU IN THE ORDER OF THE TYPE
             OF CONSTRAINTS.
   lam     - MATRIX THAT CONTAINS THE MAGNITUDES
             OF THE INTERNAL GRASP FORCES AS
             COMPUTED BY THE SIMPLEX ALGORITHM
   A       - MATRIX CONTAINING THE COEFFICIENTS
             OF THE CONSTRAINT EQUATIONS FOR THE
             HAND.
   winv    - RIGHT GENERALIZED INVERSE OF THE
             MATRIX 'w'.
   matll   - MATRIX CONTAINING THE PRODUCT OF
             MATRIX 'A' AND THE ORTHONORMAL
             BASIS FOR THE NULL SPACE OF 'w'
   con     - MATRIX CONTAINING THE CONSTANTS OF
             THE CONSTRAINTS.
   null    - MATRIX CONTAINING THE ORTHONORMAL
             BASIS FOR THE NULL SPACE OF THE
             MATRIX 'w'.
   C       - MATRIX CONTAINING THE FINGER CONTACT
             FORCES AND MOMENTS.
   wmat    - SQUARE MATRIX AUGMENTING THE MATRIX
             'w' BY ZEROES, THAT IS PASSED INTO
             THE SUBROUTINE 'svdcmp'.
   f       - MATRIX CONTAINING THE EXTERNAL FORCES
             ACTING ON THE OBJECT TO BE GRASPED.
   P       - MATRIX CONTAINING THE CONSTANTS OF
             THE FORCE AND MOMENT CONSTRAINTS.*/
/*********************************************/
result(a,lam,A,winv,matll,
       con,null,C,wmat,f,P)
float **a,**lam,**A,**winv,**matll,
      **con,**null,**C,**wmat,**f,**P;
{
int i,j,k,m1=0,m2=0,m3=0,icase=0,mark=nc-nr,
  *izrov,*iposv,imark=2,imark2=0,x=2;
float **jt,**eqn5,**eq78one,
  **eq78two,**meq78one,**meq78two;
```

```
float **matrix(),*vector(),**Cp,**Cpp,
      **dum,**Ch;
int *ivector();

izrov=ivector(1,100);
iposv=ivector(1,100);
for(i=1;i<=100;i++){
izrov[i]=0.0;
iposv[i]=0.0;
}
Cpp=matrix(1,nc,1,1);
Cp=matrix(1,nc,1,1);
Ch=matrix(1,nc,1,1);
for(i=1;i<=nc;i++){
Cpp[i][1]=0.0;
Cp[i][1]=0.0;
Ch[i][1]=0.0;
}
dum=matrix(1,dimen,1,1);
for(i=1;i<=dimen;i++){
dum[i][1]=0.0;
}
mult(A,null,dimen,nc,mark,mat11);
mult(winv,f,nc,nr,1,Cpp);
for(i=1;i<=nc;i++){
for(j=1;j<=1;j++){
Cp[i][j]=-Cpp[i][j];
if(fabs(Cp[i][j])==0.0)Cp[i][j]=0.0;
}
}
mult(A,Cp,dimen,nc,1,dum);
for(i=1;i<=dimen;i++)
for(j=1;j<=1;j++)con[i][j]=P[i][j]-dum[i][j];
for(k=1;k<=mark;k++){
for(i=1;i<=mark+1;i++)a[dimen+2][i]=0.0;
for(i=2;i<=dimen+1;i++)
if(con[i-1][1]<0.0){
m1=m1+1;
for(j=2;j<=mark+1;j++){
a[imark][1]=-con[i-1][1];
a[imark][j]=mat11[i-1][j-1];
}
imark=imark+1;
}
for(i=2;i<=dimen+1;i++)
if(con[i-1][1]>=0.0){
for(j=2;j<=mark+1;j++){
a[imark2+imark][1]=con[i-1][1];
a[imark2+imark][j]=-mat11[i-1][j-1];
}
imark2=imark2+1;
}
m2=dimen-m1;
a[1][x-1]=0.0;
a[1][x]=1.0;
for(i=1;i<=dimen+2;i++){
for(j=1;j<=mark+1;j++){
if(a[i][j]==-0.0){
a[i][j]=0.0;
}
}
}
simplx(a,dimen,mark,m1,m2,m3,&icase,
       izrov,iposv);
if(icase!=0)nrerror
("NO SOLUTION SATISFIES GIVEN CONSTRAINTS");
```

```
 [1]=a[1][1];
•1;i<=mark+1;i++)a[1][i]=0.0;

•2;imark2=0;
n2=0;m3=0;

ıull,lam,nc,mark,1,Ch);
•1;i<=nc;i++)C[i][1]=Cp[i][1]+Ch[i][1];



•**********************************/
[S IS THE MAIN PROGRAM IN WHICH ALL
: MATRICES PREVIOUSLY DEFINED ARE
[TIALIZED AND ALL THE SUBROUTINES
: CALLED. THE RESULTS OF THIS CODE
: PRINTED OUT AT THE END OF THIS
:TION.                              */
•**********************************/


j,n=nr,m=nc,mak=nc-nr,dim=4+6*4+2*jacsiz;
**winv,*sing,**null,**A,
**mat11,**con,**a,**lam,**C,
**wmat,**w,**f,**jt,**P,**torque;
atrix(1,m,1,n);
1;i<=m;i++){
1;j<=n;j++){
][j]=0.0;


ix(1,dimen,1,nc);
1;i<=dimen;i++){
1;j<=nc;j++){
]=0.0;


ix(1,nr,1,nc);
ix(1,nr,1,1);
rix(1,jacsiz,1,nc);
=matrix(1,jacsiz,1,1);
ix(1,dimen,1,1);
atrix(1,nc,1,nc);
1;i<=nc;i++){
1;j<=nc;j++){
][j]=0.0;


1;i<=nr;i++){
1;j<=nc;j++){
]=0.0;


1;i<=nr;i++)f[i][1]=0.0;
1;i<=jacsiz;i++){
1;j<=nc;j++){
j]=0.0;


1;i<=dimen;i++)P[i][1]=0.0;
,f,jt,P,A);
1;i<=nr;i++){
1;j<=nc;j++){
][j]=w[i][j];
```

```
   (REFER TO THE DERIVATIONS).
   a    - THIS IS THE SIMPLEX TABLEAU WHICH
          CONTAINS THE CONSTRAINTS FOR THE
          SYSTEM (REFER TO DERIVATIONS).
   m    - THE NUMBER OF ROWS IN 'a'.
   n    - THE NUMBER OF COLUMNS IN 'a'.
   ml   - THE NUMBER OF 'LESS THAN AND EQUAL TO'
          CONSTRAINTS.
   m2   - THE NUMBER OF 'GREATER THAN AND EQUAL
          TO' CONSTRAINTS.
   m3   - THE NUMBER OF 'EQUAL TO' CONSTRAINTS.
   icase - A FLAG THAT IS SET TO 0 IF A
           SOLUTION IS FOUND AND -1 IF NO
           SOLUTION SATISFIES THE GIVEN
           CONSTRAINTS.
   izrov & iposv - DUMMY ARRAYS THAT ARE ONLY
                   USED INTERNALLY.
/*************************************************/
simplx(a,m,n,ml,m2,m3,icase,izrov,iposv)
int m,n,ml,m2,m3,*icase,*izrov,*iposv;
float **a;
{
   int i,ip,ir,is,k,kh,kp,m12,nl1,nl2;
   int *l1,*l2,*l3,*ivector();
   float q1,bmax=0.0;
   if(m!=(ml+m2+m3))nrerror
   ("BAD INPUT CONSTRAINTS COUNT IN SIMPLX");
   l1=ivector(1,n+1);
   l2=ivector(1,m);
   l3=ivector(1,m);
   for(i=1;i<=n+1;i++){
   l1[i]=0.0;
   }
   for(i=1;i<=m;i++){
   l2[i]=0.0;
   l3[i]=0.0;
   }
   nl1=n;
   for(k=1;k<=n;k++){
   l1[k]=izrov[k]=k;
   }
   nl2=m;
   for(i=1;i<=m;i++){
      if(a[i+1][1]<0.0)nrerror
      ("Bad input tableau in SIMPLX");
      l2[i]=i;
      iposv[i]=n+i;
   }
   for(i=1;i<=m2;i++)l3[i]=1;
   ir=0;
   if(m2+m3){
     ir=1;
     for(k=1;k<=(n+1);k++){
        q1=0.0;
        for(i=m1+1;i<=m;i++)q1+=a[i+1][k];
        a[m+2][k]=-q1;
     }
     do {
        simp1(a,m+1,l1,nl1,0,&kp,&bmax);
        if(bmax<=EPS && a[m+2][1]<-EPS){
          *icase=-1;
          FREEALL return;
        } else if(bmax<=EPS && a[m+2][1]<=EPS){
           m12=m1+m2+1;
           if(m12<=m){
              for(ip=m12;ip<=m;ip++){
```

```
                if(iposv[ip] == (ip+n)){
                  simp1(a,ip,ll,nll,1,&kp,&bmax);
                  if(bmax>0.0)
                     goto one;
                }
              }
            }
            ir=0;
            --ml2;
            if(ml+1<=ml2)
               for(i=ml+1;i<=ml2;i++)
                  if(l3[i-ml] == 1)
                     for(k=1;k<=n+1;k++)
                        a[i+1][k]=-a[i+1][k];
            break;
         }
         simp2(a,n,l2,nl2,&ip,kp,&ql);
         if(ip == 0){
           *icase=-1;
           FREEALL return;
         }
one:     simp3(a,m+1,n,ip,kp);
         if(iposv[ip] >= (n+ml+m2+1)){
            for(k=1;k<=nll;k++)
               if(ll[k] == kp)break;
            --nll;
            for(is=k;is<=nll;is++)ll[is]=
                                    ll[is+1];
            a[m+2][kp+1] += 1.0;
            for(i=1;i<=m+2;i++)a[i][kp+1]=
                                 -a[i][kp+1];
         } else {
            if(iposv[ip]>=(n+ml+1)){
              kh=iposv[ip]-ml-n;
              if(l3[kh]){
                 l3[kh]=0;
                 a[m+2][kp+1]+=1.0;
                 for(i=1;i<=m+2;i++)
                    a[i][kp+1]=-a[i][kp+1];
              }
            }
         }
         is=izrov[kp];
         izrov[kp]=iposv[ip];
         iposv[ip]=is;
      } while(ir);
   }
   for(;;){
      simp1(a,0,ll,nll,0,&kp,&bmax);
      if(bmax <= 0.0){
        *icase=0;
        FREEALL return;
      }
      simp2(a,n,l2,nl2,&ip,kp,&ql);
      if(ip ==0){
        *icase=1;
        FREEALL return;
      }
      simp3(a,m,n,ip,kp);
      is=izrov[kp];
      izrov[kp]=iposv[ip];
      iposv[ip]=is;
   }
}

/*************************************************/
```

```
/* THIS SUBROUTINE IS CALLED UP INTERNALLY IN
   THE SUBROUTINE simplx AND IS PART OF THE
   SIMPLEX METHOD.                                  */
/*************************************************/
simpl(a,mm,ll,nll,iabf,kp,bmax)
float **a,*bmax;
int mm,*ll,nll,iabf,*kp;
{
   int k;
   float test;
   *kp=ll[1];
   *bmax=a[mm+1][*kp+1];
   for(k=2;k<=nll;k++){
      if(iabf == 0)
        test=a[mm+1][ll[k]+1]-(*bmax);
      else
        test=fabs(a[mm+1][ll[k]+1])
                        -fabs(*bmax);
      if(test > 0.0){
        *bmax=a[mm+1][ll[k]+1];
        *kp=ll[k];
      }
   }
}

/*************************************************/
/* THIS SUBROUTINE IS CALLED UP BY THE
   SUBROUTINE simplx AND IS PART OF THE SIMPLEX
   METHOD.                                          */
/*************************************************/
simp2(a,n,l2,nl2,ip,kp,ql)
int n,*l2,nl2,*ip,kp;
float **a,*ql;
{
   int k,ii,i;
   float qp,q0,q;
   *ip=0;
   for(i=1;i<=nl2;i++){
      if(a[l2[i]+1][kp+1] < -EPS){
        *ql=-a[l2[i]+1][1]/a[l2[i]+1][kp+1];
        *ip=l2[i];
        for(i=i+1;i<=nl2;i++){
           ii=l2[i];
           if(a[ii+1][kp+1] < -EPS){
             q=-a[ii+1][1]/a[ii+1][kp+1];
             if(q < *ql){
               *ip=ii;
               *ql=q;
             } else if(q == *ql){
               for(k=1;k<=n;k++){
                  qp=-a[*ip+1][k+1]/
                      a[*ip+1][kp+1];
                  q0=-a[ii+1][k+1]/
                      a[ii+1][kp+1];
                  if(q0 != qp)break;
               }
               if(q0 < qp) *ip=ii;
             }
           }
        }
      }
   }
}

/*************************************************/
/* THIS ROUTINE IS CALLED UP BY THE SUBROUTINE
```

```
   simplx AND IS PART OF THE SIMPLEX METHOD. */
/*********************************************/
simp3(a,il,kl,ip,kp)
int il,kl,ip,kp;
float **a;
{
   int kk,ii;
   float piv;
   piv=1.0/a[ip+1][kp+1];
   for(ii=1;ii<=il+1;ii++)
      if(ii-1 != ip){
        a[ii][kp+1] *= piv;
        for(kk=1;kk<=kl+1;kk++)
           if(kk-1 != kp)
             a[ii][kk] -= a[ip+1][kk]*
                          a[ii][kp+1];
      }
   for(kk=1;kk<=kl+1;kk++)
      if(kk-1 != kp)a[ip+1][kk] *= -piv;
   a[ip+1][kp+1]=piv;
}

/*********************************************/
/* THIS SUBROUTINE ASSEMBLES THE MATRIX THAT
   IS PASSED INTO THE SUBROUTINE 'simplx'
   ASSEMBLES THE MAGNITUDES OF THE GRASP FORCES
   INTO ANOTHER MATRIX FROM WHICH THE JOINT
   TORQUES CAN BE COMPUTED.
   a       - MATRIX THAT CONTAINS THE SIMPLEX
             TABLEAU IN THE ORDER OF THE TYPE
             OF CONSTRAINTS.
   lam     - MATRIX THAT CONTAINS THE MAGNITUDES
             OF THE INTERNAL GRASP FORCES AS
             COMPUTED BY THE SIMPLEX ALGORITHM
   A       - MATRIX CONTAINING THE COEFFICIENTS
             OF THE CONSTRAINT EQUATIONS FOR THE
             HAND.
   winv    - RIGHT GENERALIZED INVERSE OF THE
             MATRIX 'w'.
   mat11   - MATRIX CONTAINING THE PRODUCT OF
             MATRIX 'A' AND THE ORTHONORMAL
             BASIS FOR THE NULL SPACE OF 'w'
   con     - MATRIX CONTAINING THE CONSTANTS OF
             THE CONSTRAINTS.
   null    - MATRIX CONTAINING THE ORTHONORMAL
             BASIS FOR THE NULL SPACE OF THE
             MATRIX 'w'.
   C       - MATRIX CONTAINING THE FINGER CONTACT
             FORCES AND MOMENTS.
   wmat    - SQUARE MATRIX AUGMENTING THE MATRIX
             'w' BY ZEROES, THAT IS PASSED INTO
             THE SUBROUTINE 'svdcmp'.
   f       - MATRIX CONTAINING THE EXTERNAL FORCES
             ACTING ON THE OBJECT TO BE GRASPED.
   P       - MATRIX CONTAINING THE CONSTANTS OF
             THE FORCE AND MOMENT CONSTRAINTS.*/
/*********************************************/
result(a,lam,A,winv,mat11,
       con,null,C,wmat,f,P)
float **a,**lam,**A,**winv,**mat11,
      **con,**null,**C,**wmat,**f,**P;
{
int i,j,k,m1=0,m2=0,m3=0,icase=0,mark=nc-nr,
  *izrov,*iposv,imark=2,imark2=0,x=2;
float **jt,**eqn5,**eq78one,
  **eq78two,**meq78one,**meq78two;
```

```
float **matrix(),*vector(),**Cp,**Cpp,
      **dum,**Ch;
int *ivector();

izrov=ivector(1,100);
iposv=ivector(1,100);
for(i=1;i<=100;i++){
izrov[i]=0.0;
iposv[i]=0.0;
}
Cpp=matrix(1,nc,1,1);
Cp=matrix(1,nc,1,1);
Ch=matrix(1,nc,1,1);
for(i=1;i<=nc;i++){
Cpp[i][1]=0.0;
Cp[i][1]=0.0;
Ch[i][1]=0.0;
}
dum=matrix(1,dimen,1,1);
for(i=1;i<=dimen;i++){
dum[i][1]=0.0;
}
mult(A,null,dimen,nc,mark,mat11);
mult(winv,f,nc,nr,1,Cpp);
for(i=1;i<=nc;i++){
for(j=1;j<=1;j++){
Cp[i][j]=-Cpp[i][j];
if(fabs(Cp[i][j])==0.0)Cp[i][j]=0.0;
}
}
mult(A,Cp,dimen,nc,1,dum);
for(i=1;i<=dimen;i++)
for(j=1;j<=1;j++)con[i][j]=P[i][j]-dum[i][j];
for(k=1;k<=mark;k++){
for(i=1;i<=mark+1;i++)a[dimen+2][i]=0.0;
for(i=2;i<=dimen+1;i++)
if(con[i-1][1]<0.0){
ml=ml+1;
for(j=2;j<=mark+1;j++){
a[imark][1]=-con[i-1][1];
a[imark][j]=mat11[i-1][j-1];
}
imark=imark+1;
}
for(i=2;i<=dimen+1;i++)
if(con[i-1][1]>=0.0){
for(j=2;j<=mark+1;j++){
a[imark2+imark][1]=con[i-1][1];
a[imark2+imark][j]=-mat11[i-1][j-1];
}
imark2=imark2+1;
}
m2=dimen-ml;
a[1][x-1]=0.0;
a[1][x]=1.0;
for(i=1;i<=dimen+2;i++){
for(j=1;j<=mark+1;j++){
if(a[i][j]==-0.0){
a[i][j]=0.0;
}
}
}
simplx(a,dimen,mark,ml,m2,m3,&icase,
      izrov,iposv);
if(icase!=0)nrerror
("NO SOLUTION SATISFIES GIVEN CONSTRAINTS");
```

```
lam[k][1]=a[1][1];
for(i=1;i<=mark+1;i++)a[1][i]=0.0;
x=x+1;
imark=2;imark2=0;
m1=0;m2=0;m3=0;
}
mult(null,lam,nc,mark,1,Ch);
for(i=1;i<=nc;i++)C[i][1]=Cp[i][1]+Ch[i][1];
}


/*****************************************/
/* THIS IS THE MAIN PROGRAM IN WHICH ALL
   THE MATRICES PREVIOUSLY DEFINED ARE
   INITIALIZED AND ALL THE SUBROUTINES
   ARE CALLED. THE RESULTS OF THIS CODE
   ARE PRINTED OUT AT THE END OF THIS
   SECTION.                              */
/*****************************************/

main()
{
int i,j,n=nr,m=nc,mak=nc-nr,dim=4+6*4+2*jacsiz;
float **winv,*sing,**null,**A,
      **mat11,**con,**a,**lam,**C,
      **wmat,**w,**f,**jt,**P,**torque;
winv=matrix(1,m,1,n);
for(i=1;i<=m;i++){
for(j=1;j<=n;j++){
winv[i][j]=0.0;
}
}
A=matrix(1,dimen,1,nc);
for(i=1;i<=dimen;i++){
for(j=1;j<=nc;j++){
A[i][j]=0.0;
}
}
w=matrix(1,nr,1,nc);
f=matrix(1,nr,1,1);
jt=matrix(1,jacsiz,1,nc);
torque=matrix(1,jacsiz,1,1);
P=matrix(1,dimen,1,1);
wmat=matrix(1,nc,1,nc);
for(i=1;i<=nc;i++){
for(j=1;j<=nc;j++){
wmat[i][j]=0.0;
}
}
for(i=1;i<=nr;i++){
for(j=1;j<=nc;j++){
w[i][j]=0.0;
}
}
for(i=1;i<=nr;i++)f[i][1]=0.0;
for(i=1;i<=jacsiz;i++){
for(j=1;j<=nc;j++){
jt[i][j]=0.0;
}
}
for(i=1;i<=dimen;i++)P[i][1]=0.0;
data(w,f,jt,P,A);
for(i=1;i<=nr;i++){
for(j=1;j<=nc;j++){
wmat[i][j]=w[i][j];
}
```

```
     for (ll=1;ll<=n;ll++)
         if(ll != icol) {
            dum=a[ll][icol];
            a[ll][icol]=0.0;
            for (l=1;l<=n;l++) a[ll][l] -=
                               a[icol][l]*dum;
            for (l=1;l<=m;l++) b[ll][l] -=
                               b[icol][l]*dum;
         }
   }
   for (l=n;l>=1;l--) {
      if (indxr[l]!=indxc[l])
         for (k=1;k<=n;k++)
            SWAP(a[k][indxr[l]],a[k][indxc[l]]);
   }
   free_ivector(ipiv,1,n);
   free_ivector(indxr,1,n);
   free_ivector(indxc,1,n);
}

/************************************************/
/* THIS SUBROUTINE COMPUTES THE MOORE-PENROSE
   INVERSE FOR A RECTANGULAR MATRIX.
   winv   - THIS MATRIX IS THE OUTPUT FROM THIS
            SUBROUTINE AND CARRIES THE INVERSE
            OF THE RECTANGULAR MATRIX.
   m      - THE NUMBER OF COLUMNS IN THE INPUT
            MATRIX 'a'.
   n      - THE NUMBER OF ROWS IN THE INPUT
            MATRIX 'a'.
   a      - THE INPUT MATRIX WHOSE MOORE-PENROSE
            INVERSE IS TO BE TAKEN.
/************************************************/
peninv(winv,m,n,a)
int m,n;
float **winv,**a;
{
float **aa,**bb,**wtr,**matrix()
     ,*vector();
int i,j,dim=normsiz+6*jacsiz;
aa=matrix(1,n,1,n);
bb=matrix(1,n,1,1);
for(i=1;i<=n;i++){
for(j=1;j<=n;j++){
aa[i][j]=0.0;
bb[i][1]=0.0;
}
}
wtr=matrix(1,m,1,n);
for(i=1;i<=m;i++){
for(j=1;j<=n;j++){
wtr[i][j]=0.0;
}
}
for(i=1;i<=n;i++)
for(j=1;j<=m;j++)
wtr[j][i]=a[i][j];
mult(a,wtr,n,m,n,aa);
gaussj(aa,n,bb,1);
mult(wtr,aa,m,n,n,winv);
}

/************************************************/
/* THIS SUBROUTINE COMPUTES THE MAXIMUM ALLOWABLE
   MAGNITUDES OF THE INTERNAL GRASP FORCES
   USING THE SIMPLEX METHOD OF LINEAR PROGRAMMING
```

```
            cos((double)(psif2[0]));
jacob[6][3]=(r2*sin((double)(gam))+l1*sin((double)(gam+psif2[1]))+
            l2*sin((double)(gam+psif2[1]+psif2[2]))
            +l3*sin((double)(gam+psif2[1]+2*psif2[2]))-
            cf2[1]*cos((double)(gam+psif2[1]+2*psif2[2]))-
            cf2[0]*sin((double)(gam+psif2[1]+2*psif2[2])))
            *cos((double)(psif2[0]))-cf2[2]*sin((double)(psif2[0]));
jacob[6][4]=(l1*cos((double)(gam+psif2[1]))+l2*
            cos((double)(gam+psif2[1]+psif2[2]))+l3*
            cos((double)(gam+psif2[1]+2*psif2[2]))+cf2[1]*
            sin((double)(gam+psif2[1]+2*psif2[2]))-
            cf2[0]*cos((double)(gam+psif2[1]+2*psif2[2])))
            *sin((double)(psif2[0]));
jacob[6][5]=(l2*cos((double)(gam+psif2[1]+psif2[2]))+2*l3*
            cos((double)(gam+psif2[1]+2*psif2[2]))
            +2*cf2[1]*sin((double)(gam+psif2[1]+2*psif2[2])))*
            sin((double)(psif2[0]));
jacob[7][3]=normf2[0];
jacob[7][4]=-normf2[1]*sin((double)(psif2[0]))+normf2[2]*
            cos((double)(psif2[0]));
jacob[7][5]=2.0*jacob[7][4];
jacob[8][6]=0.0;
jacob[8][7]=-l1*sin((double)(gam+psif3[1]))-l2
            *sin((double)(gam+psif3[1]+psif3[2]))
            -l3*sin((double)(gam+psif3[1]+2*psif3[2]))+
            cf3[0]*sin((double)(gam+psif3[1]+2*psif3[2]))+cf3[1]*
            cos((double)(gam+psif3[1]+2*psif3[2]));
jacob[8][8]=-l2*sin((double)(gam+psif3[1]+psif3[2]))-
            2*l3*sin((double)(gam+psif3[1]+2*psif3[2]))+2*cf3[0]*
            sin((double)(gam+psif3[1]+2*psif3[2]))
            +2*cf3[1]*cos((double)(gam+psif3[1]+2*psif3[2]));
jacob[9][6]=-(r2*sin((double)(gam))+l1*sin((double)(gam+psif3[1]))+
            l2*sin((double)(gam+psif3[1]+psif3[2]))+l3*
            sin((double)(gam+psif3[1]+2*psif3[2]))-cf3[0]*
            sin((double)(gam+psif3[1]+2*psif3[2]))-cf3[1]*
            cos((double)(gam+psif3[1]+2*psif3[2])))
            *sin((double)(psif3[0]))-cf3[2]*cos((double)(psif3[0]));
jacob[9][7]=(l1*cos((double)(gam+psif3[1]))+l2*
            cos((double)(gam+psif3[1]+psif3[2]))+l3*
            cos((double)(gam+psif3[1]+2*psif3[2]))-
            cf3[0]*cos((double)(gam+psif3[1]+2*psif3[2]))+
            cf3[1]*sin((double)(gam+psif3[1]+2*psif3[2])))
            *cos((double)(psif3[0]));
jacob[9][8]=(l2*cos((double)(gam+psif3[1]+psif3[2]))+2*l3*
            cos((double)(gam+psif3[1]+2*psif3[2]))
            -2*cf3[0]*cos((double)(gam+psif3[1]+2*psif3[2]))+
            2*cf3[1]*sin((double)(gam+psif3[1]+2*psif3[2])))*
            cos((double)(psif3[0]));
jacob[10][6]=(r2*sin((double)(gam))+l1*sin((double)(gam+psif3[1]))+
            l2*sin((double)(gam+psif3[1]+psif3[2]))
            +l3*sin((double)(gam+psif3[1]+2*psif3[2]))-
            cf3[1]*cos((double)(gam+psif3[1]+2*psif3[2]))-
            cf3[0]*sin((double)(gam+psif3[1]+2*psif3[2])))
            *cos((double)(psif3[0]))-cf3[2]*sin((double)(psif3[0]));
jacob[10][7]=(l1*cos((double)(gam+psif3[1]))+l2*
            cos((double)(gam+psif3[1]+psif3[2]))+l3*
            cos((double)(gam+psif3[1]+2*psif3[2]))+cf3[1]*
            sin((double)(gam+psif3[1]+2*psif3[2]))-
            cf3[0]*cos((double)(gam+psif3[1]+2*psif3[2])))
            *sin((double)(psif3[0]));
jacob[10][8]=(l2*cos((double)(gam+psif3[1]+psif3[2]))+2*l3*
            cos((double)(gam+psif3[1]+2*psif3[2]))
            +2*cf3[1]*sin((double)(gam+psif3[1]+2*psif3[2])))*
            sin((double)(psif3[0]));
jacob[11][6]=normf3[0];
jacob[11][7]=-normf3[1]*sin((double)(psif3[0]))+normf3[2]*
```

```
            cos((double)(psif3[0]));
jacob[11][8]=2.0*jacob[11][7];
jacob[12][9]=0.0;
jacob[12][10]=-l1*sin((double)(psitum[1]))-
              l2*sin((double)(psitum[1]-psitum[2]))
              -l3*sin((double)(psitum[1]-2*psitum[2]))+ctum[0]*
              sin((double)(psitum[1]-2*psitum[2]))-ctum[1]*
              cos((double)(psitum[1]-psitum[2]));
jacob[12][11]=l2*sin((double)(psitum[1]-psitum[2]))+2*l3*
              sin((double)(psitum[1]-2*psitum[2]))-2*ctum[0]*
              sin((double)(psitum[1]-2*psitum[2]))+2*ctum[1]*
              cos((double)(psitum[1]-2*psitum[2]));
jacob[13][9]=-(r1+l1*sin((double)(psitum[1]))+l2*
             sin((double)(psitum[1]-psitum[2]))+l3*
             sin((double)(psitum[1]-2*psitum[2]))+ctum[1]*
             cos((double)(psitum[1]-2*psitum[2]))-ctum[0]*
             sin((double)(psitum[1]-2*psitum[2])))*sin((double)(psitum[0]))+
             ctum[2]*cos((double)(psitum[0]));
jacob[13][10]=(l1*cos((double)(psitum[1]))+l2*
              cos((double)(psitum[1]-psitum[2]))+l3*
              cos((double)(psitum[1]-2*psitum[2]))-ctum[1]*
              sin((double)(psitum[1]-2*psitum[2]))-ctum[0]*
              cos((double)(psitum[1]-2*psitum[2])))*cos((double)(psitum[0]));
jacob[13][11]=(-l2*cos((double)(psitum[1]-psitum[2]))-2*l3*
              cos((double)(psitum[1]-2*psitum[2]))+2*ctum[1]*
              sin((double)(psitum[1]-2*psitum[2]))+2*ctum[0]*
              cos((double)(psitum[1]-2*psitum[2])))*cos((double)(psitum[0]));
jacob[14][9]=(r1+l1*sin((double)(psitum[1]))+l2*
             sin((double)(psitum[1]-psitum[2]))+l3*
             sin((double)(psitum[1]-2*psitum[2]))+ctum[1]*
             cos((double)(psitum[1]-2*psitum[2]))-ctum[0]*
             sin((double)(psitum[1]-2*psitum[2])))*cos((double)(psitum[0]))+
             ctum[2]*sin((double)(psitum[0]));
jacob[14][10]=(l1*cos((double)(psitum[1]))+l2*
              cos((double)(psitum[1]-psitum[2]))+l3*
              cos((double)(psitum[1]-2*psitum[2]))-ctum[1]*
              sin((double)(psitum[1]-2*psitum[2]))-ctum[0]*
              cos((double)(psitum[1]-2*psitum[2])))*sin((double)(psitum[0]));
jacob[14][11]=(-l2*cos((double)(psitum[1]-psitum[2]))-2*l3*
              cos((double)(psitum[1]-2*psitum[2]))+2*ctum[1]*
              sin((double)(psitum[1]-2*psitum[2]))+2*ctum[0]*
              cos((double)(psitum[1]-2*psitum[2])))*sin((double)(psitum[0]));
jacob[15][9]=normtum[0];
jacob[15][10]=-normtum[1]*sin((double)(psitum[0]))+normtum[2]*
             cos((double)(psitum[0]));
jacob[15][11]=2.0*jacob[15][10];

/*************************************************/
/* THE FOLLOWING IF CONDITIONS DETERMINE THE
   THE 'X' AND 'Y' COMPONENTS OF THE CONTACT
   POINT ON THE FINGERTIP GIVEN THE CONTACT
   NORMAL, BY ASSUMING EITHER THE 'X' OR'Y'
   COMPONENTS TO BE 1.0 AND TAKING THE DOT
   PRODUCT TO BE 0.0. THESE ARE DETERMINED
   IN THE BASE COORDINATE SYSTEM.              */
/*************************************************/
if(normf1[2]!=0.0 && normf1[1]!=0.0){
omagf1=(sqrt((double)(2*normf1[2]*normf1[2]+(-normf1[0]-normf1[1])*
       (-normf1[0]-normf1[1]))))/normf1[2];
of1[0]=1/omagf1;of1[1]=1/omagf1;
of1[2]=-(normf1[0]+normf1[1])/(normf1[2]*omagf1);
}
if(normf1[2]==0.0 && normf1[1]!=0.0){
omagf1=(sqrt((double)(2*(normf1[1]*normf1[1])+
       normf1[0]*normf1[0]))))/normf1[1];
of1[0]=1/omagf1;of1[1]=-normf1[0]/(normf1[1]*omagf1);of1[2]=1/omagf1;
```

```
}
if(normf1[2]!=0.0 && normf1[1]==0.0){
omagf1=(sqrt((double)(2*(normf1[2]*normf1[2])+
       normf1[0]*normf1[0])))/normf1[2];
of1[0]=1/omagf1;of1[1]=1/omagf1;of1[2]=-normf1[0]/(normf1[2]*omagf1);
}
if(normf1[1]==0.0 && normf1[2]==0.0){
omagf1=sqrt((double)(2.0));
of1[0]=0.0;of1[1]=1/omagf1;of1[2]=1/omagf1;
}
af1[0]=normf1[1]*of1[2]-normf1[2]*of1[1];
af1[1]=-normf1[0]*of1[2]+normf1[2]*of1[0];
af1[2]=normf1[0]*of1[1]-normf1[1]*of1[0];
if(normf2[2]!=0.0 && normf2[1]!=0.0){
omagf2=(sqrt((double)(2*normf2[2]*normf2[2]+(-normf2[0]-normf2[1])*
       (-normf2[0]-normf2[1]))))
       /normf2[2];
of2[0]=1/omagf2;of2[1]=1/omagf2;
of2[2]=-(normf2[0]+normf2[1])/(normf2[2]*omagf2);
}
if(normf2[2]==0.0 && normf2[1]!=0.0){
omagf2=(sqrt((double)(2*(normf2[1]*normf2[1])+
       normf2[0]*normf2[0])))/normf2[1];
of2[0]=1/omagf2;of2[1]=-normf2[0]/(normf2[1]*omagf2);of2[2]=1/omagf2;
}
if(normf2[2]!=0.0 && normf2[1]==0.0){
omagf2=(sqrt((double)(2*(normf2[2]*normf2[2])+
       normf2[0]*normf2[0])))/normf2[2];
of2[0]=1/omagf2;of2[1]=1/omagf2;of2[2]=-normf2[0]/(normf2[2]*omagf2);
}
if(normf2[1]==0.0 && normf2[2]==0.0){
omagf2=sqrt((double)(2.0));
of2[0]=0.0;of2[1]=1/omagf2;of2[2]=1/omagf2;
}
af2[0]=normf2[1]*of2[2]-normf2[2]*of2[1];
af2[1]=-normf2[0]*of2[2]+normf2[2]*of2[0];
af2[2]=normf2[0]*of2[1]-normf2[1]*of2[0];
if(normf3[2]!=0.0 && normf3[1]!=0.0){
omagf3=(sqrt((double)(2*normf3[2]*normf3[2]+(-normf3[0]-normf3[1])*
       (-normf3[0]-normf3[1]))))/normf3[2];
of3[0]=1/omagf3;of3[1]=1/omagf3;
of3[2]=-(normf3[0]+normf3[1])/(normf3[2]*omagf3);
}
if(normf3[2]==0.0 && normf3[1]!=0.0){
omagf3=(sqrt((double)(2*(normf3[1]*normf3[1])+normf3[0]*
       normf3[0])))/normf3[1];
of3[0]=1/omagf3;of3[1]=-normf3[0]/(normf3[1]*omagf3);of3[2]=1/omagf3;
}
if(normf3[2]!=0.0 && normf3[1]==0.0){
omagf3=(sqrt((double)(2*(normf3[2]*normf3[2])+normf3[0]*
       normf3[0])))/normf3[2];
of3[0]=1/omagf3;of3[1]=1/omagf3;of3[2]=-normf3[0]/(normf3[2]*omagf3);
}
if(normf3[1]==0.0 && normf3[2]==0.0){
omagf3=sqrt((double)(2.0));
of3[0]=0.0;of3[1]=1/omagf3;of3[2]=1/omagf3;
}
af3[0]=normf3[1]*of3[2]-normf3[2]*of3[1];
af3[1]=-normf3[0]*of3[2]+normf3[2]*of3[0];
af3[2]=normf3[0]*of3[1]-normf3[1]*of3[0];
if(normtum[2]!=0.0 && normtum[1]!=0.0){
omagtum=(sqrt((double)(2*normtum[2]*normtum[2]+(-normtum[0]-normtum[1])*
        (-normtum[0]-normtum[1]))))/normtum[2];
otum[0]=1/omagtum;of3[1]=1/omagtum;
otum[2]=-(normtum[0]+normtum[1])/(normtum[2]*omagtum);
}
```

```
}
C=matrix(1,nc,1,1);
for(i=1;i<=nc;i++)C[i][1]=0.0;
sing=vector(1,nc);
for(i=1;i<=nc;i++)sing[i]=0.0;
peninv(winv,m,n,w);
null=matrix(1,nc,1,mak);
for(i=1;i<=nc;i++){
for(j=1;j<=mak;j++){
null[i][j]=0.0;
}
}
basis(sing,null,wmat);
mat11=matrix(1,dimen,1,mak);
con=matrix(1,dimen,1,1);
for(i=1;i<=dimen;i++){
for(j=1;j<=mak;j++){
mat11[i][j]=0.0;
con[i][1]=0.0;
}
}
a=matrix(1,dimen+2,1,mak+1);
for(i=1;i<=dimen+2;i++){
for(j=1;j<=mak+1;j++){
a[i][j]=0.0;
}
}
lam=matrix(1,mak,1,1);
for(i=1;i<=mak;i++)lam[i][1]=0.0;
result(a,lam,A,winv,mat11,con,null,C,w,f,P);
mult(jt,C,jacsiz,nc,1,torque);
/*****************************************/
for(i=1;i<=jacsiz;i++)printf("%5.3f\n",torque[i][1]);
}
```

```
if(normtum[2]==0.0 && normtum[1]!=0.0){
omagtum=(sqrt((double)(2*(normtum[1]*normtum[1])+normtum[0]*
        normtum[0])))/normtum[1];
otum[0]=1/omagtum;otum[1]=-normtum[0]/(normtum[1]*omagtum);otum[2]=1/omagtum;
}
if(normtum[2]!=0.0 && normtum[1]==0.0){
omagtum=(sqrt((double)(2*(normtum[2]*normtum[2])+normtum[0]*
        normtum[0])))/normtum[2];
otum[0]=1/omagtum;otum[1]=1/omagtum;otum[2]=-normtum[0]/(normtum[2]*omagtum);
}
if(normtum[1]==0.0 && normtum[2]==0.0){
omagtum=sqrt((double)(2.0));
otum[0]=0.0;otum[1]=1/omagtum;otum[2]=1/omagtum;
}
atum[0]=normtum[1]*otum[2]-normtum[2]*otum[1];
atum[1]=-normtum[0]*otum[2]+normtum[2]*otum[0];
atum[2]=normtum[0]*otum[1]-normtum[1]*otum[0];

/*************************************************/
/* THE ELEMENTS OF THE MATRIX 'w' ARE THE X,
   Y & Z COMPONENTS OF THE CONTACT POINTS AND
   THEIR CROSS PRODUCTS ARRANGED APPROPRIATELY
   IN THE PROPER COLUMNS TO GIVE A MAPPING
   BETWEEN THE EXTERNAL FORCES ACTING ON THE
   GRASPED OBJECT AND THE CONTACT FORCES IN THE
   FINGERS.                                      */
/*************************************************/
w[1][1]=of1[0];w[1][2]=af1[0];w[1][3]=normf1[0];w[1][4]=0;
w[2][1]=of1[1];w[2][2]=af1[1];w[2][3]=normf1[1];w[2][4]=0;
w[3][1]=of1[2];w[3][2]=af1[2];w[3][3]=normf1[2];w[3][4]=0;
w[4][1]=rf1[1]*of1[2]-rf1[2]*of1[1];
w[4][2]=rf1[1]*af1[2]-rf1[2]*af1[1];
w[4][3]=rf1[1]*normf1[2]-rf1[2]*normf1[1];
w[4][4]=normf1[0];
w[5][1]=rf1[2]*of1[0]-rf1[0]*of1[2];
w[5][2]=rf1[2]*af1[0]-rf1[0]*af1[2];
w[5][3]=rf1[2]*normf1[0]-rf1[0]*normf1[2];
w[5][4]=normf1[1];
w[6][1]=rf1[0]*of1[1]-rf1[1]*of1[0];
w[6][2]=rf1[0]*af1[1]-rf1[1]*af1[0];
w[6][3]=rf1[0]*normf1[1]-rf1[1]*normf1[0];
w[6][4]=normf1[2];
w[1][5]=of2[0];w[1][6]=af2[0];w[1][7]=normf2[0];w[1][8]=0;
w[2][5]=of2[1];w[2][6]=af2[1];w[2][7]=normf2[1];w[2][8]=0;
w[3][5]=of2[2];w[3][6]=af2[2];w[3][7]=normf2[2];w[3][8]=0;
w[4][5]=rf2[1]*of2[2]-rf2[2]*of2[1];
w[4][6]=rf2[1]*af2[2]-rf2[2]*af2[1];
w[4][7]=rf2[1]*normf2[2]-rf2[2]*normf2[1];
w[4][8]=normf2[0];
w[5][5]=rf2[2]*of2[0]-rf2[0]*of2[2];
w[5][6]=rf2[2]*af2[0]-rf2[0]*af2[2];
w[5][7]=rf2[2]*normf2[0]-rf2[0]*normf2[2];
w[5][8]=normf2[1];
w[6][5]=rf2[0]*of2[1]-rf2[1]*of2[0];
w[6][6]=rf2[0]*af2[1]-rf2[1]*af2[0];
w[6][7]=rf2[0]*normf2[1]-rf2[1]*normf2[0];
w[6][8]=normf2[2];
w[1][9]=of3[0];w[1][10]=af3[0];w[1][11]=normf3[0];w[1][12]=0;
w[2][9]=of3[1];w[2][10]=af3[1];w[2][11]=normf3[1];w[2][12]=0;
w[3][9]=of3[2];w[3][10]=af3[2];w[3][11]=normf3[2];w[3][12]=0;
w[4][9]=rf3[1]*of3[2]-rf3[2]*of3[1];
w[4][10]=rf3[1]*af3[2]-rf3[2]*af3[1];
w[4][11]=rf3[1]*normf3[2]-rf3[2]*normf3[1];
w[4][12]=normf3[0];
w[5][9]=rf3[2]*of3[0]-rf3[0]*of3[2];
w[5][10]=rf3[2]*af3[0]-rf3[0]*af3[2];
```

```
w[5][11]=rf3[2]*normf3[0]-rf3[0]*normf3[2];
w[5][12]=normf3[1];
w[6][9]=rf3[0]*of3[1]-rf3[1]*of3[0];
w[6][10]=rf3[0]*af3[1]-rf3[1]*af3[0];
w[6][11]=rf3[0]*normf3[1]-rf3[1]*normf3[0];
w[6][12]=normf3[2];
w[1][13]=otum[0];w[1][14]=atum[0];
w[1][15]=normtum[0];w[1][16]=0;
w[2][13]=otum[1];w[2][14]=atum[1];
w[2][15]=normtum[1];w[2][16]=0;
w[3][13]=otum[2];w[3][14]=atum[2];
w[3][15]=normtum[2];w[3][16]=0;
w[4][13]=rtum[1]*otum[2]-rtum[2]*otum[1];
w[4][14]=rtum[1]*atum[2]-rtum[2]*atum[1];
w[4][15]=rtum[1]*normtum[2]-rtum[2]*normtum[1];
w[4][16]=normtum[0];
w[5][13]=rtum[2]*otum[0]-rtum[0]*otum[2];
w[5][14]=rtum[2]*atum[0]-rtum[0]*atum[2];
w[5][15]=rtum[2]*normtum[0]-rtum[0]*normtum[2];
w[5][16]=normtum[1];
w[6][13]=rtum[0]*otum[1]-rtum[1]*otum[0];
w[6][14]=rtum[0]*atum[1]-rtum[1]*atum[0];
w[6][15]=rtum[0]*normtum[1]-rtum[1]*normtum[0];
w[6][16]=normtum[2];

/***********************************************/
/* THE ELEMENTS OF THE EXTERNAL FORCE VECTOR
   ARE READ INTO A VECTOR THAT IS DIMENSIONED
   FROM 1 RATHER THAN 0 AS IS THE CASE FOR
   ARRAYS IN THE 'C' PROGRAMMING LANGUAGE.
/***********************************************/
for(i=1;i<=nr;i++){
for(j=1;j<=1;j++){
f[i][j]=force[i-1][j-1];
}
}

/***********************************************/
/* THE TRANSPOSE OF THE JACOBIAN MATRIX IS
   TAKEN                                       */
/***********************************************/
for(i=1;i<=nc;i++){
for(j=1;j<=jacsiz;j++){
jt[j][i]=jacob[i-1][j-1];
}
}

/***********************************************/
/* THE MAXIMUM AND MINIMUM TORQUE CONSTRAINTS
   ARE INITIALIZED INTO THE MATRIX 'A'.     */
/***********************************************/
for(i=29;i<=40;i++){
for(j=1;j<=nc;j++){
A[i][j]=jt[i-28][j];
}
}
for(i=41;i<=52;i++){
for(j=1;j<=nc;j++){
A[i][j]=-jt[i-40][j];
if(A[i][j]==-0.0)A[i][j]=0.0;
}
}
}

/***********************************************/
/* WHEN CALLED, THIS SUBROUTINE PRINTS OUT
```

```
   A MESSAGE AND ENDS EXECUTION OF THE
   PROGRAM.                                    */
/*********************************************/
void nrerror(error_text)
char error_text[];
{
fprintf(stderr,"%s\n",error_text);
exit(1);
}

/*********************************************/
/* THIS SUBROUTINE DIMENSIONS A FLOAT
   VECTOR FROM 1 RATHER THAN 0 AS IS
   CONENTIONAL IN 'C' PROGRAMMING.             */
/*********************************************/
float *vector(nl,nh)
int nl,nh;
{
float *v;
v=(float *)malloc((unsigned)(nh-nl+1)*sizeof(float));
if(!v) nrerror("allocation failure in vector()");
return v-nl;
}

/*********************************************/
/* THIS SUBROUTINE FREES A VECTOR.             */
/*********************************************/
void free_vector(v,nl,nh)
float *v;
int nl,nh;
{
free((char*)(v+nl));
}

/*****************************************  **/
/* THIS SUBROUTINE COMPUTES THE ORTHONORMAL
   BASIS FOR THE NULL SPACE OF THE MATRIX
   'w' USING SINGULAR VALUE DECOMPOSITION.
   a    - THIS A SQUARE MATRIX OF DIMENSION
          nc * nc WITH THE FIRST 6 ROWS
          CONTAINING THE ELEMENTS OF THE
          MATRIX 'w' AND THE LAST nc-6 ROWS
          AUGMENTED BY ZEROES
   m    - NUMBER OF ROWS OF 'a'.
   n    - NUMBER OF COLUMNS OF 'a'.
   w    - DIAGONAL MATRIX OF SINGULAR VALUES
          OUTPUT AS A VECTOR.
   v    - AN n*n MATRIX WHOSE COLUMNS
          CORRESPONDING TO THE ZEROES IN
          VECTOR 'w' CONTAIN THE ORTHONORMAL
          BASIS FOR THE NULL SPACE.            */
/*********************************************/
svdcmp(a,m,n,w,v)
float **a,*w,**v;
int m,n;
{
int flag,i,its,j,jj,k,l,nm;
float c,f,h,s,x,y,z;
float anorm=0.0,g=0.0,scale=0.0;
float *rv1,**matrix(),*vector(),**w1,**c1,
      **c2,**vt;
if(m<n)nrerror("SVDCMP:You must augment \
'a' with extra zero rows");
rv1=vector(1,n);
for(i=1;i<=n;i++)rv1[i]=0.0;
w1=matrix(1,nc,1,nc);
```

```
c1=matrix(1,nc,1,nc);
vt=matrix(1,nc,1,nc);
c2=matrix(1,nc,1,nc);
for(i=1;i<=nc;i++){
for(j=1;j<=nc;j++){
w1[i][j]=0.0;
c1[i][j]=0.0;
vt[i][j]=0.0;
c2[i][j]=0.0;
}
}
for(i=1;i<=n;i++){
   l=i+1;
   rv1[i]=scale*g;
   g=s=scale=0.0;
   if(i<=m){
      for(k=i;k<=m;k++){
      scale += fabs(a[k][i]);
      }
      if(scale<0.0001)scale=0.0;
      if(scale!=0.0){
         for(k=i;k<=m;k++){
            a[k][i]/=scale;
            s += a[k][i]*a[k][i];
         }
         f=a[i][i];
         g=-SIGN(sqrt((double)s),f);
         h=f*g-s;
         a[i][i]=f-g;
         if(i!=n){
            for(j=l;j<=n;j++){
               for(s=0.0,k=i;k<=m;k++) s +=
                              a[k][i]*a[k][j];
               f=s/h;
               for(k=i;k<=m;k++) a[k][j] +=
                                 f*a[k][i];
            }
         }
         for(k=i;k<=m;k++)a[k][i]*=scale;
      }
   }
   w[i]=scale*g;
   g=s=scale=0.0;
   if(i<=m && i!=n) {
      for(k=l;k<=n;k++) scale += fabs(a[i][k]);
      if(scale<0.001)scale=0.0;
      if(scale!=0.0){
        for(k=l;k<=n;k++){
           a[i][k] /= scale;
           s += a[i][k]*a[i][k];
        }
        f=a[i][l];
        g=-SIGN(sqrt((double)s),f);
        h=f*g-s;
        a[i][l]=f-g;
        for(k=l;k<=n;k++) rv1[k]=a[i][k]/h;
        if(i != m){
          for(j=l;j<=m;j++){
             for(s=0.0,k=l;k<=n;k++) s +=
                            a[j][k]*a[i][k];
             for(k=l;k<=n;k++) a[j][k]+=
                               s*rv1[k];
          }
        }
        for(k=l;k<=n;k++) a[i][k] *= scale;
   }
```

```
    }
    anorm=MAX(anorm,(fabs(w[i])+fabs(rv1[i])));
  }
  for(i=n;i>=1;i--){
    if(i<n){
      if(g){
        for(j=l;j<=n;j++)
          v[j][i]=(a[i][j]/a[i][l])/g;
        for(j=l;j<=n;j++){
          for(s=0.0,k=l;k<=n;k++) s +=
                        a[i][k]*v[k][j];
          for(k=l;k<=n;k++) v[k][j] +=
                            s*v[k][i];
        }
      }
      for(j=l;j<=n;j++){
      v[i][j]=v[j][i]=0.0;
      }
    }
    v[i][i]=1.0;
    g=rv1[i];
    l=i;
  }
  for(i=n;i>=1;i--){
    l=i+1;
    g=w[i];
    if(i<n)
       for(j=l;j<=n;j++) a[i][j]=0.0;
    if(fabs(g)>0.0001){
      g=1.0/g;
      if(i != n){
        for(j=l;j<=n;j++){
          for(s=0.0,k=l;k<=m;k++) s +=
                        a[k][i]*a[k][j];
          f=(s/a[i][i])*g;
          for(k=i;k<=m;k++)a[k][j] +=
                            f*a[k][i];
        }
      }
      for(j=i;j<=m;j++) a[j][i] *= g;
    } else {
        for(j=i;j<=m;j++) a[j][i]=0.0;
    }
    ++a[i][i];
  }
  for(k=n;k>=1;k--){
    for(its=1;its<=30;its++){
        flag=1;
        for(l=k;l>=1;l--){
          nm=l-1;
          if(fabs(rv1[l])+anorm==anorm){
            flag=0;
            break;
          }
          if(fabs(w[nm])+anorm==anorm)break;
        }
        if(flag){
          c=0.0;
          s=1.0;
          for(i=l;i<=k;i++){
            f=s*rv1[i];
            if(fabs(f)+anorm!=anorm){
              g=w[i];
              h=PYTHAG(f,g);
              w[i]=h;
              h=1.0/h;
```

```
                c=g*h;
                s=(-f*h);
                for(j=1;j<=m;j++){
                   y=a[j][nm];
                   z=a[j][i];
                   a[j][nm]=y*c+z*s;
                   a[j][i]=z*c-y*s;
                }
             }
          }
      }
z=w[k];
if(l==k){
   if(z<0.0){
      w[k] = -z;
      for(j=1;j<=n;j++) v[j][k]=(-v[j][k]);
   }
   break;
}
if(its==30) nrerror("No converg. in 30 \
                     SVDCMP iter.");
x=w[l];
nm=k-1;
y=w[nm];
g=rv1[nm];
h=rv1[k];
f=((y-z)*(y+z)+(g-h)*(g+h))/(2.0*h*y);
g=PYTHAG(f,1.0);
f=((x-z)*(x+z)+h*((y/(f+SIGN(g,f)))-h))/x;
c=s=1.0;
for(j=l;j<=nm;j++){
   i=j+1;
   g=rv1[i];
   y=w[i];
   h=s*g;
   g=c*g;
   z=PYTHAG(f,h);
   rv1[j]=z;
   c=f/z;
   s=h/z;
   f=x*c+g*s;
   g=g*c-x*s;
   h=y*s;
   y=y*c;
   for(jj=1;jj<=n;jj++){
      x=v[jj][j];
      z=v[jj][i];
      v[jj][j]=x*c+z*s;
      v[jj][i]=z*c-x*s;
   }
   z=PYTHAG(f,h);
   w[j]=z;
   if(z){
     z=1.0/z;
     c=f*z;
     s=h*z;
   }
   f=(c*g)+(s*y);
   x=(c*y)-(s*g);
   for(jj=1;jj<=m;jj++){
      y=a[jj][j];
      z=a[jj][i];
      a[jj][j]=y*c+z*s;
      a[jj][i]=z*c-y*s;
   }
}
```

```
   rvl[l]=0.0;
   rvl[k]=f;
   w[k]=x;
}
}
free_vector(rvl,1,n);
}

/*************************************************/
/* THIS SUBROUTINE CALLS THE SUBROUTINE svdcmp
   AND PUTS THE COLUMNS OF THE MATRIX 'v' THAT
   CORRESPOND TO THE ORTHONORMAL BASIS FOR THE
   NULL SPACE OF THE MATRIX 'w' INTO THE MATRIX
   'space'.
   w      - NOT TO BE CONFUSED WITH THE MATRIX
            'w' CONTAINING THE STATIC FORCE
            BALANCE, THIS IS THE VECTOR THAT
            CONTAINS THE SINGULAR VALUES.
   space  - THIS MATRIX CONTAINS THE ORTHONORMAL
            BASIS FOR THE NULL SPACE OF THE
            MATRIX 'w' THAT CONTAINS THE FORCE
            BALANCE.
   wmat   - SQUARE MATRIX THAT IS AUGMENTED BY
            ZEROES TO BE PASSED TO THE SUBROUTINE
            svdcmp.                              */
/*************************************************/
basis(w,space,wmat)
float *w,**space,**wmat;
{
int m=nr,n=nc,i,j,b=1,dim=
                4+6*4+2*jacsiz;
float **matrix(),*vector(),**v;
float **f,**jt,**p;
v=matrix(1,nc,1,nc);
if(m<n)m=n;
svdcmp(wmat,m,n,w,v);
for(i=1;i<=nc;i++)
if(w[i]<=0.0001){
for(j=1;j<=nc;j++){
if(fabs(v[j][i])<=0.0001)v[j][i]=0.0;
space[j][b]=v[j][i];
}
b=b+1;
}
}

/*************************************************/
/* SUBROUTINE TO MULTIPLY TWO MATRICES.
   ml   - FIRST MATRIX.
   m2   - SECOND MATRIX.
   sr1  - NUMBER OF ROWS OF FIRST MATRIX.
   sc   - NUMBER OF COLUMNS OF FIRST AND
          SECOND MATRICES.
   sr2  - NUMBER OF ROWS OF SECOND MATRIX.
   c    - MATRIX CONTAINING THE RESULT OF THE
          MULTIPLICATION.
/*************************************************/
mult(ml,m2,sr1,sc,sr2,c)
int sr1,sc,sr2;
float **c,**ml,**m2;
{
int i,j,k;
float sum=0.0,mul;
for(i=1;i<=sr2;i++){
for(j=1;j<=sr1;j++){
for(k=1;k<=sc;k++){
```

```
mul=ml[j][k]*m2[k][i];
sum=sum+mul;
}
c[j][i]=sum;
sum=0.0;
}
}
}

/*************************************************/
/* THIS FUNCTION DIMENSIONS THE ROWS AND
   COLUMNS OF A FLOAT MATRIX TO START FROM 1
   INSTEAD OF 0.
   nrl   - LOWER DIMENSION OF ROWS, IN THIS
           CASE IT IS ALWAYS 1.
   nr2   - UPPER DIMENSION OF THE ROWS, I.E.
           THE NUMBER OF ROWS OF THE MATRIX.
   ncl   - LOWER DIMENSION OF COLUMNS, IN
           THIS CASE IT IS ALWAYS 1.
   nc2   - UPPER DIMENSION OF THE COLUMNS,I.E.
           THE NUMBER OF COLUMNS IN THE
           MATRIX.                              */
/*************************************************/
float **matrix(nrl,nr2,ncl,nc2)
int nrl,nr2,ncl,nc2;
{
int i;
float **m;
m=(float **) malloc((unsigned)
   (nr2-nrl+1)*sizeof(float*));
if (!m) nrerror("allocation failure \
in matrix()");
m -= nrl;
for(i=nrl;i<=nr2;i++){
m[i]=(float *) malloc((unsigned)
     (nc2-ncl+1)*sizeof(float*));
if (!m[i]) nrerror("allocation failure\
 in 2 matrix()");
m[i] -= ncl;
}
return m;
}

/*************************************************/
/* THIS FUNCTION DIMENSIONS THE ROWS OF AN
   INTEGER VECTOR TO START FROM 1.
   m    - THE LOWER DIMENSION ON THE NUMBER OF
          ROWS, IN THIS CASE IT IS 1.
   n    - THE UPPER LIMIT ON THE NUMBER OF ROWS
          I.E. THE MAXIMUM NUMBER OF ROWS
          ALLOWABLE.                            */
/*************************************************/
int *ivector(m,n)
{
int *v;
v=(int *)malloc((unsigned) (n-m+1)
                    *sizeof(int));
if (!v) nrerror("allocation failure\
 in ivector()");
return v-1;
}

/*************************************************/
/* THIS FUNCTION FREES AN INTEGER VECTOR.
   ipiv    - INTEGER NUMBER.
   m       - LOWER DIMENSION ON THE NUMBER
```

```
          OF ROWS, IN THIS CASE IT IS 1
   n      - THE MAXIMUM NUMBER OF ROWS IN THE
          VECTOR.                              */
/*********************************************/
void free_ivector(ipiv,m,n)
int *ipiv;
{
free((char*) (ipiv+1));
}

/*********************************************/
/* THIS SUBROUTINE COMPUTES THE INVERSE OF A
   SQUARE MATRIX USING GAUSS-JORDAN ELIMINATION
   a    - THE SQUARE MATRIX WHOSE INVERSE IS
          TO BE COMPUTED. ON OUTPUT THE ELEMENTS
          OF THIS MATRIX CONTAIN THE MATRIX
          INVERSE.
   n    - THE NUMBER OF ROWS AND COLUMNS IN THE
          MATRIX 'a'.
   b    - DUMMY VECTOR CONTAINING ALL ZEROES.
   m    - NUMBER OF ROWS IN 'b'.              */
/*********************************************/
 gaussj(a,n,b,m)
   float **a,**b;
   int n,m;
   {
   int *indxc,*indxr,*ipiv;
   int i,icol,irow,j,k,l,ll,*ivector();
   float big,dum,pivinv,**matrix();
   indxc=ivector(1,n);
   indxr=ivector(1,n);
   for(i=1;i<=n;i++){
   indxc[i]=0;
   indxr[i]=0;
   }
   ipiv=ivector(1,n);
   for (j=1;j<=n;j++) ipiv[j]=0;
   for (i=1;i<=n;i++) {
     big=0.0;
     for (j=1;j<=n;j++)
       if(ipiv[j] !=1)
          for(k=1;k<=n;k++) {
            if(ipiv[k] == 0) {
              if(fabs(a[j][k])>=big) {
                big=fabs(a[j][k]);
                irow=j;
                icol=k;
              }
            } else if (ipiv[k]>1)nrerror
             ("GAUSSJ: Singular Matrix-1");
          }
     ++(ipiv[icol]);
     if(irow!=icol) {
        for (l=1;l<=n;l++)
        SWAP(a[irow][l],a[icol][l])
        for (l=1;l<=m;l++)
        SWAP(b[irow][l],b[icol][l])
     }
     indxr[i]=irow;
     indxc[i]=icol;
     if (a[icol][icol]==0.0) nrerror
        ("GAUSSJ: Singular Matrix-2");
     pivinv=1.0/a[icol][icol];
     a[icol][icol]=1.0;
     for (l=1;l<=n;l++) a[icol][l] *=pivinv;
     for (l=1;l<=m;l++) b[icol][l] *=pivinv;
```

```
}
C=matrix(1,nc,1,1);
for(i=1;i<=nc;i++)C[i][1]=0.0;
sing=vector(1,nc);
for(i=1;i<=nc;i++)sing[i]=0.0;
peninv(winv,m,n,w);
null=matrix(1,nc,1,mak);
for(i=1;i<=nc;i++){
for(j=1;j<=mak;j++){
null[i][j]=0.0;
}
}
basis(sing,null,wmat);
mat11=matrix(1,dimen,1,mak);
con=matrix(1,dimen,1,1);
for(i=1;i<=dimen;i++){
for(j=1;j<=mak;j++){
mat11[i][j]=0.0;
con[i][1]=0.0;
}
}
a=matrix(1,dimen+2,1,mak+1);
for(i=1;i<=dimen+2;i++){
for(j=1;j<=mak+1;j++){
a[i][j]=0.0;
}
}
lam=matrix(1,mak,1,1);
for(i=1;i<=mak;i++)lam[i][1]=0.0;
result(a,lam,A,winv,mat11,con,null,C,w,f,P);
mult(jt,C,jacsiz,nc,1,torque);
/*****************************************/
for(i=1;i<=jacsiz;i++)printf("%5.3f\n",torque[i][1]);
}
```

1989 USAF-UES RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

UNIVERSAL ENERGY SYSTEMS, Inc.

FINAL REPORT

ARTICULATED TOTAL BODY (ATB) "VIEW" PROGRAM

Prepared by : Chi-Ming Tang, Ph.D.

Academic Rank : Associate Professor

Department and University : Department of Mathematics
State University of New York, College at Geneseo

Research Location : AAMRL/BBM
WPAFB, OH. 45433-6573

USAF Researcher : Ints Kaleps, Ph.D.

Date : April 30, 1991

Contract No. : F49620-88-C-0053/SB5881-0378

Table of Contents

Abstract

The Articulated Total Body (ATB) model is currently being used by the AFAAMRL to study the human body biomechanics in various dynamic environments. The VIEW program provides a graphical representation of the simulation output from the ATB model. Over the course of several years, a number of changes and additions have been made to the ATB program, but the corresponding modifications needed to show all of the ATB elements had not been made to the VIEW program until 1989.

During the summer of 1989, the VIEW program was upgraded under the USAF Summer Faculty Research Program (SFRP), conducted by Universal Energy System, Inc.. This modification provides an option to draw a quality picture that includes harness belts with hidden line capability (Ref. 8). The goal of this project, which is a continuation of the SFRP work, is to modify the VIEW program so that it has the capability

(1) to show "additional ellipsoids" as defined by the ATB program, and

(2) to provide an option of hyperellipsoids.

Acknowledgements

I wish to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsorship of this research. Universal Energy Systems must be mentioned for their concern and help to me in all administrative and directional aspects of this research project.

My experience was rewarding and enriching because of many different influences. I would like to express my deepest appreciation to Louise A. Obergefell, Annette Rizer and Dr. Ints Kaleps for their valuable knowledge, suggestions, and support during the development of this research project. Finally, I would like to take this opportunity to thank the State University of New York at Geneseo for supporting this research, especially, many thanks to the professors B. W. Ristow and D. J. Harke.

I. Introduction

The Articulated Total Body (ATB) model is currently being used by the Armstrong Aerospace Medical Research Laboratory (AAMRL) to study the human body biomechanics in various dynamic environments. The ATB model is a rigid body dynamics computer simulation program which outputs the linear and angular position of each object at user defined time intervals. The output data is used by the VIEW graphics program to plot the projected image of the objects on the viewing plane. Over the course of several years, a number of changes and additions have been made to the ATB program, but the VIEW program was not upgraded to fit the corresponding modifications so that it could show all of the ATB elements.

The original VIEW program development was started in 1981 by Leetch and Bowman and ended in 1983 (Ref. 3 and Ref. 4). The graphics output of this version was developed to suit the early ATB model version. That is, the graphics output only depicts the ellipsoids (body segments) and polygons (contact planes). Also, the program can only display one ellipsoid for each body segment.

The concept of the harness-belt was introduced and added to the ATB model in 1975. A new Harness-belt algorithm was added to the ATB-II model in 1980. However, a harness-belt option was not included in the program until 1989. A modification that provides an option to draw a quality picture including harness belts with hidden line capability was developed by Tang, Rizer, and Kaleps in 1989 (Ref. 8).

The concept of the additional ellipsoid for each segment was intrduced and added to the ATB model in the early years. In addition, hyperellipsoid

option was added to the ATB-IV model in 1988. Therefore, the new feature developed for this research program shall have a capability to depict the additional ellipsoid(s) and hyperellipsoid(s) for each simulation of the ATB model.

In compliance with these specific needs, a number of modifications and six additional subroutines have been made to the VIEW program. The details of this implementation will be discussed in section III and an example will be given in section IV.

II. Objectives and Approach

There are three objectives in this research project, the additional ellipsoid, the tangent point and the hyperellipsoid.

The latest ATB model allows additional ellipsoids for each segment instead of only allowing one ellipsoid per each segment. Modification of the VIEW program to show additional ellipsoids as defined by the ATB program are essential and important for the simulation. Figure 1 below is a good example to show the drawback of the old VIEW program. The number 1 segment contains two ellipsoids, number 1 and number 23. However, the original VIEW program does not have the capability to draw the second ellipsoid on the same segment, therefore, the part of the harness belt is misplotted on the top portion of the number 1 ellipsoid.

The revised VIEW program has the capability to draw each ellipsoid no matter where it is in the segement. Figure 2 below shows the improvement of the new version of the VIEW program. The additional ellipsoid number 23 is displayed instead of ellipsoid number 2. The misplotted portion of the harness belt is now displayed clearly.

In accordance with the additional ellipsoid task, the all DO loop in the VIEW program will be from 1 to 40 instead from 1 to NSEG (the number of segments) everywhere.

In addition, a new record appending to the record 6.0 was added to the input control data file. This new record contains the numbers that represent ellipsoids/hyperellipsoids to be removed from VIEW program output. The details will be explained in section III(A) below on this report.

Most the time, when the harness belts are to be depicted on the

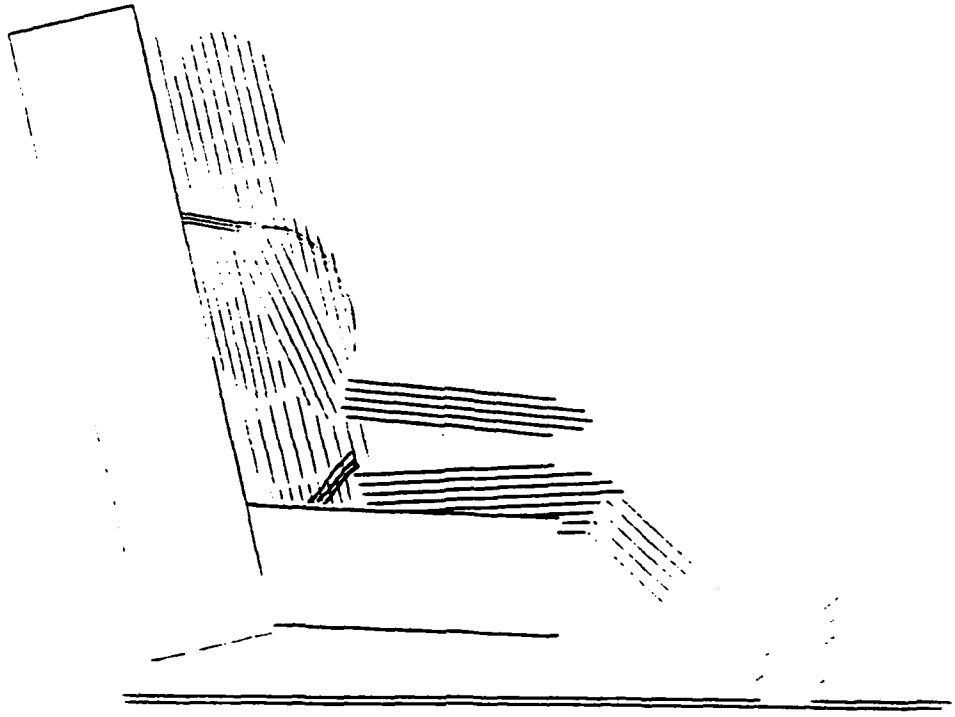


Figure 1

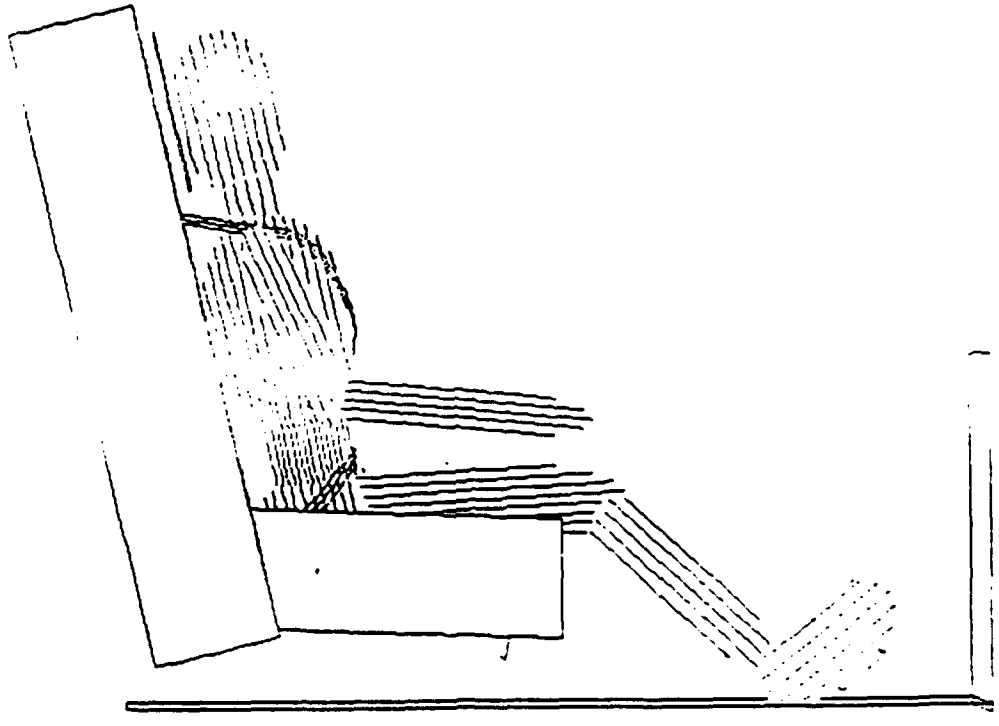


Figure 2

graphical output from the VIEW program, there is a symptom exhibited. That is, if the line segment between the anchor point and the first reference point penetrates the reference ellipsoid, we shall find the supplement point on the ellipsoid such that the line segment between the anchor point and the supplement point will be tangent to the ellipsoid. Otherwise, a partial part of the belt will not display. The reason is that these points of this part are not the surface point of the ellipsoid. When applying the hidden algorithm, these points will be hidden and so they will not be plotted.

Let the supplement point described above refer to a tangent point. The cause of this symptom was initially thought due to the lack of the tangent point. Once the additional ellipsoid version was implemented, we found that the major reason was not the tangent point. Because some points of the harness belt are attached on the additional ellipsoid. If this additional ellipsoid will not be displayed, certainly, the portion of the belt attached on this additional ellipsoid could not be explicitly depicted as well as other parts of the belt. Therefore the algorithm of the proposed tangent point is no longer needed in this research. Besides, the tangent point was already discussed in the ATB model simulation.

Figures 3 and 4 below show the improvement of the new version of the VIEW program again. The additional ellipsoid number 22 is displayed. Thus, the parts of the belts on the left upper corner are displayed very clearly.

In order to improve the modeling of corners and other geometries the option to use hyperellipsoids as contact surfaces rather than standard ellipsoids was added to the ATB model. This option was originally developed for General Motors Corporation and has been incorporated into the ATB model with their permission.

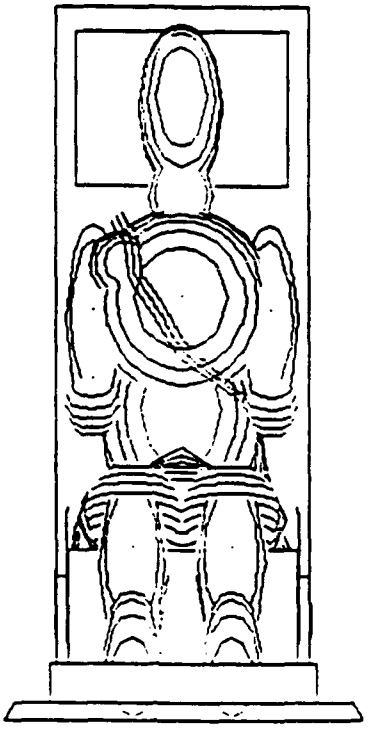


Figure 3

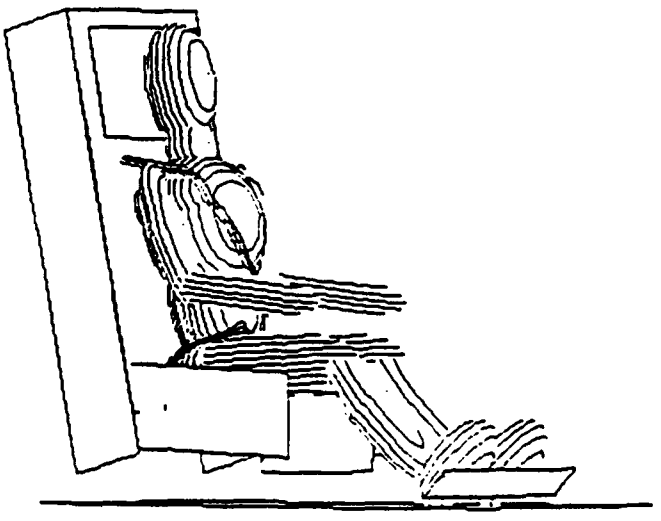


Figure 4

A hyperellipsoid is defined as the surface generated by

$$(x/a)^m + (y/b)^m + (z/c)^m = 1$$

where

a, b, c are the semi-axes lengths, and

m is the power of the hyperellipsoid, an even integer.

If m = 2, the surface is an ellipsoid. For larger values of m the figure "squares off" at the corners. As m approaches infinity the figure approaches a rectangular parallelpiped with the same dimensions as the hyperellipsoid. This makes the hyperellipsoid very useful for describing contact surfaces.

A number of modifications have been be made to fit the needs of an option of hyperellipsoid. Such as:

To reformat of BD array from the ATB output,

To create a new integer array IHYPEP to store the information about the ellipsoid/hyperellipsoid in the sunroutine INPUT,

To upgrade the algorithm to find three radical vectors of the hyperellipsoid point to a surface point that forms the contour of the projected shadow in the subroutine PRJELR,

To upgrade the algorithm to calculate the vector M for the hidden line problem between ellipsoids/hyperellipsoids in the subroutines XYZ, YZ, and Z.

The details of these modifications will be explained later in the next section III below on this report.

III. Modifications and New Subroutines

To fit the specific needs of the additional ellipsoid and hyperellipsoid versions, a number of modifications and new subroutines have been made to the VIEW program. In order to distinguish from the original program code, these changes are labeled by the names "ADDELP" and "HYPELP" in columns 75-80 on each program code.

(A) Modifications:

(1) Upgraded Common Blocks

There are several variables added into the common blocks, ELLIPSE and REMOVE. In accordance with the hyperellipsoid version, the variables NSEGP, ISEGEP, and IHYPEP are appended to the common block ELLIPSE. The variables NEREM and IREMVE are inserted in the common block REMOVE. The definitions of those variables are described below:

NSEGP	Integer scalar, represents the total number of segments to be displayed.
ISEGEP(2,40)	Integer array, represents the pointer to the number of segment for each ellipsoid/ hyperellipsoid in the first subscript and the corresponding identification number of the ellipsoid/hyperellipsoid are stored in the

second subscript.

IHYPEP(6,40) Integer array, contains the information of each ellipsoid or hyperellipsoid which are input from ATB input file. The definition of each subscript is described below:

Subscript	Definition
1	Power of x
2	Power of y
3	Power of z
4	Flag of hyperellipsoid Is "1" if hyperellipsoid, "0" otherwise.
5	Min power of x, y, and z
6.	Max power of x, y, and z

NEREM Integer scalar, represents the total number of ellipsoids/hyperellipsoids to be removed from plot output. Read from input control file in subroutine INPUT.

IREMVE(40) Integer array, contains number that represents ellipsoids/hyperellipsoids to be removed from VIEW program output. Read from input control file in subroutine INPUT.

(2) Extended Dimension Declarations

The dimensions of some variable arrays on the common blocks

have been extended to fit the needs of the "additional ellipsoid" version. The following is a list of these changes in alphabetical order.

Variable Name	Original Dimension Declaration	New Dimension Declaration
CONVEC	(2,4,110)	(2,4,120)
D	(3,3,110)	(3,3,120)
ICOLOR	111	121
IE	(110,110)	(120,120)
NIE	110	120
NPPP	110	120
NSTEPS	90	100
POS	(2,110)	(2,120)
SEGLP	(3,110)	(3,120)
SEGLPO	(3,110)	(3,120)
SIGN	110	120

Also, the dimension of the variable array D1 in the subroutine POLYD was extended from 990 to 1080, that is, D1(1080).

(3) Input Data File on Logical Unit Number 1:

This file is a data file generated by the ATB model simulation program. In order to input the information regarding the additional ellipsoids and hyperellipsoids, the subroutine named UNIT1 of the ATB program must be changed so that it can fulfill the corresponding need in the VIEW program. For example, the variable LBD is contained in the subroutine INPUT. The definition of the variable is described as follows:

LBD(40) Integer array, contains the idenfication number of segment for each ellipsoid or hyperellipsoid. Read from the ATB input file in the subroutine INPUT. The information is then restored into the variable IHYPEP. Note that the variable LBD is a local vaiable in the subroutine INPUT. But the variable IHYPEP ia a global variable.

Thus, the subroutine UNIT1 must provide the values of the array LBD so that we can define the values of the array IHYPEP in the subroutine INPUT.

In order to accomplish the mission of inputing the information of the variable LBD, the input values of the variable D, the directional cosine matrix for each segment, have to replaced. The new variable DIP is defined as the directional cosine matrix from the inertial coordinate system to the principal coordinate system. In other words, the output from the subroutine UNIT1 of the ATB program will be the value of the variable DIP instead of the value of the variable D as before.

(4) Input Data file on Logical Unit Number 3:

This file is so called the input control file read from FORTRAN unit number 3. Two records sets, RECORD 6.0 and RECORD 6.1 have been changed. One new variable NEREM is inserted in the RECORD 6.0. The RECORD 6.1 is renamed as RECORD 6.2. Therefore, a new record set, RECORD 6.1 is included in the input control file of the

new VIEW program. These records are described below:

(a) RECORD 6.0

Variable List : NFAST, NEREM, NPREM, NISG

Format : (4I2)

Discussion : This card contains the total number of segments, ellipsoids/hyperellipsoids, and polygons to remove from plotting and the substitute coordinate system for the MPL array. The detailed description can be referred to in Reference 4.

Variable	Description
NFAST	NFAST represents the total number of segments from the ATB program output which are not to be plotted. The detailed description can be referred to in Reference 4.
NEREM	NEREM represents the total number of ATB model simulation ellipsoids/ hyperellipsoids to be removed from plotting consideration. This variable is used in combination with record 6.1, IREMOV array.
NPREM	NPREM represents the total number of ATB model simulation polygons to be removed

from plotting consideration. The detailed description can be referred to in Reference 4.

NISG — NISG is a variable containing the number of the coordinate system that will be substituted for MPL array values which is zero. The detailed description can be referred to in Reference 4.

(b) RECORD 6.1

Variable List : IREMVE (1-NEREM)

Format : 4012

Discussion : This record contains the ellipsoid/ hyperellipsoid numbers from the ATB model simulation to be removed from consideration for the output graphics file.

Variable	Description
IREMVE(40)	IREMVE is an array containing NEREM numbers identifying the numbers of the ellipsoids/hyperellipsoids to be removed from the output graphics file.

(c) RECORD 6.2

Variable List : IREMOV (1-NPREM)

Format : 3012

Discussion : This record is exactly same as old record 6.1 before. Only rename the record

number. The detailed description about the variable IREMOV can be refered to in Reference 4.

(5) Subroutine INPUT

The subroutine INPUT has been modified to fit the needs for the new requirement which is in regards to the additional ellipsoid. As mentioned in the section (3) above, the input value of the directional cosine matrix is, DIP, from the inertial system to the principal system. Thus, the calculation of the directional cosine matrices, D's, from the insertial system to each local ellpsoid system is contained in the subroutine INPUT.

Let the variable PPE be the directional cosine matrix from the principal system to the local ellipsoid system. Then, the product of two matrix multiplication [DIP] [PPE] will produce the matrix D, namely,

$$D = [DIP]\,[PPE].$$

The above calculation is implemented in the sunroutine INPUT The matrices PPE are defined as follows:

For the ellipsoid case,

$$PPE(I,J,*) = BD(M,*)$$

where $M = 3 * (J + 4) + I$, $I = 1, 2, 3$ and $J = 1, 2, 3$.

For the hyperellipsoid case,

$$PPE(I,J,*) = BD(M,*),$$

where M = 3 * (J + 1) + I + 1, I = 1, 2, 3 and J = 1, 2, 3.

Otherwise,

PPE = I, the 3x3 identity matrix.

Besides, the variable A array has been modified as follows:

For the ellipsoid case,

it is the ellipsoid matrix defined as same as before, namely,

$$A = 1/(BD)^2.$$

For the hyperellipsoid case,

$$A = 1/BD$$

Note that there is no square sign on the variable BD.

(6) Subroutines XYZ and YZ

The hidden line problem between two ellipsoids was discussed on the Appendix A of the Reference 4. The algorithm is to solve the equation

$$(\mathbf{S} - \mathbf{M})^T [\mathbf{A}] (\mathbf{S} - \mathbf{M}) = 1$$

or equivalent to

$$(1/a^2)(S_1 - M_1)^2 + (1/b^2)(S_2 - M_2)^2 + (1/c^2)(S_3 - M_3)^2 = 1 \qquad (1)$$

for the vector $\mathbf{M} = (M_1, M_2, M_3)^T$.

But, for the hyperellipsoid case, the above equation becomes

$$(1/a^m)(S_1 - M_1)^m + (1/b^m)(S_2 - M_2)^m + (1/c^m)(S_3 - M_3)^m = 1 \qquad (2)$$

The Newton-Raphson method is to be applied to solve the

above equation (2). However, we need a good initial estimate for the equation (2). The starting point of the iterations is obtained by solving the equation (1), that is, the output from subroutine XYZ and YZ depending the situation we consider. The subroutine XYZNEW and YZNEWT is to perform these numerical calculations. The detail is given below in the section (B) New Subroutines.

In order to pass the flag of the hyperellipsoid to the subroutines XYZNEW and YZNEWT, the subroutine XYZ and YZ have been changed their input parameter list. Two parameters IHFLAG and MPOWER are appended in the input parameter list. The definitions of IHFLAG and MPOWER are decribed below:

IHFLAG	Integer scalar, flag of hyperellipsoid. IS "1" if hyperellipsoid, "0" otherwise.
MPOWER	Integer scalar, uniform power of x, y, and z.

(7) Subroutine Z

Referring to the Appendix A in Reference 4, the purpose of the subroutine Z is to solve the following equation for M_3:

$$A(S_1)^2 + B(S_2)^2 + C(S_3 - M_3)^2 = 1 \qquad (3)$$

where $A = 1/a^2$, $C = 1/b^2$, and $C = 1/c^2$. But, for the hyperellipsoid case the above equation becomes

$$A(S_1)^m + B(S_2)^m + C(S_3 - M_3)^m = 1 \qquad (4)$$

where the coefficients A, B, and C are defined as follows:

$$A = 1/a^m, \quad B = 1/b^m, \quad \text{and} \quad C = 1/c^m,$$

where m is the maximum power of x, y, and z terms in the hyperellipsoid equation.

Therefore, the subroutine Z has been upgraded to fit this need, namely, it will solve the equation (3) or (4) depending on the case of ellipsoid or hyperellipsoid. If the case of the hyperellipsoid occurs, the solution will be

$$M_3 = S_3 \pm [(1 - AS_1^m - BS_2^m)/C]^{1/m}.$$

(8) Subroutine PRJELR

The function of PRJELR is to circumscribe a projected shadow of an ellipsoid with a rectangle. The resulting rectangle is used as a polygon by the overlap routines to determine what objects overlap after they are projected. Referring to pages 183-185 on the Appendix B of Reference 4, these three radical vector r's satify the equation

$$\mathbf{S}\mathbf{S}^T \mathbf{A}\cdot\mathbf{r} = -1$$

or equivalent to, after the algebraic manuipulation,

$$\beta_1 r_{x|y} + \beta_2 r_z = -1 \qquad (5)$$

Since $\mathbf{r}^T \mathbf{A}\cdot\mathbf{r} = 1$, again, after the algebraic manuipulation, we have

$$\gamma_1 r_{x|y}^2 + \gamma_2 r_{x|y} r_z + \gamma_3 r_z^2 = 1 \qquad (6)$$

Three radical vectors r's can be obtained by solving the pair of nonlinear equations (5) and (6). The above equations are good for

the ellipsoid case. For the hyperellipsoid case, these equations become

$$\beta_1 r_{x|y}^{m-1} + \beta_2 r_z^{m-1} = -1 \qquad (7)$$

$$\gamma_1 r_{x|y}^{m} + \gamma_2 r_{x|y}^{m/2} r_z^{m/2} + \gamma_3 r_z^{m} = 1 \qquad (8)$$

where m is the uniform power of x, y, and z.

The subroutine SOLVH is to solve the system of nonlinear equations (7) and (8) by using the Newton's method. Similary, same as subroutines XYZ's, we use the solutions from the system of equations (5) and (6) as the initial estimates of the iterations, that is, the output from the subroutine SOLVR.

(B) New Subroutines

To fit the specific needs of the hyperellipsoid, there are three new subroutines being attached to the original VIEW program. Their names are called SOLVH, XYZNEW, and YZNEWT. The following is the description of the purpose of each subroutine.

(1) Subroutine SOLVH

This subroutine is called by the subroutine PRJELR to find three radical vestors r's by solving the system of nonlinear equations (7) and (8) defined above using the Newton's method. The starting points are obtained from the output of the subroutine SOLVR.

(2) Subroutine XYZNEW

This subroutine is called by the subroutine XYZ to find the value M_1. Namely, the goal of the subroutine XYZNEW is to solve the following equation:

$$A(S_1 - M_1)^m + B(S_2 - \alpha M_1)^m + C(S_3 - \beta M_1)^m = 1 \qquad (9)$$

for M_1. The coefficients A, B, and C are defined as follows:

$A = 1/a^m$, $B = 1/b^m$, and $C = 1/c^m$,

where m is the maximum power of x, y, and z terms in the hyperellipsoid equation. The constants S_1, S_2, S_3, α, and β are same as in the subroutine XYZ. The detailed description of the above equation (9) can be referred to the page 175 in Reference 4. The numerical method used here to solve the equation (9) is so called the Newton-Raphson method. The starting point of the iteration can be obtained from the solution of the equation:

$$A(S_1 - M_1)^2 + B(S_2 - \alpha M_1)^2 + C(S_3 - \beta M_1)^2 = 1$$

where $A = 1/a^2$, $C = 1/b^2$, and $C = 1/c^2$. In other words, we first solve for M_1 as same as the ellipsoid case in the subroutine XYZ. Then use this output M_1 as the initial estimate to find the better approximate solution M_1 of the equation (9).

(3) Subroutine YZNEWT

This subroutine is called by the subroutine YZ. The basic

principle and approach in this routine are same as the subroutine XYZNEW. That is, the purpose of the subroutine YZNEWT is to solve the following equation, using the Newton-Raphson method,

$$A(S_1)^m + B(S_2 - M_2)^m + C(S_3 - \alpha M_2)^m = 1 \qquad (10)$$

for M_2. The coefficients A, B, and C are defined as follows:

$$A = 1/a^m, \quad B = 1/b^m, \quad \text{and} \quad C = 1/c^m,$$

where m is the maximum power of x, y, and z terms in the hyperellipsoid equation. The constants S_1, S_2, S_3, and α are same as in the subroutine XYZ. The detailed description of the above equation (10) can be referred to the page 178 in Reference 4. The initial estimate for M_2, is also obtained from the following equation

$$A(S_1)^2 + B(S_2 - M_2)^2 + C(S_3 - \alpha M_2)^2 = 1$$

,where $A = 1/a^2$, $C = 1/b^2$, and $C = 1/c^2$, from the subroutine YZ.

IV. Sample Input and Output

This section provides an example of the visual data display of an ATB model simulation. The output of the simulation from ATB program consists of two additional ellipsoids. These two additional ellipsoids are number 22 and number 23 which are attached to the ellipsoids number 3 and number 2, respectively. Its input control file and output graphics file plot show as follows.

(A) Sample Input Control File

The following table is an example of the input control file. This file consists of the information on the following records:

Record Number	Variable List
1.0	DEVFLG
2.0	NFRME
3.0	ID(1-10)
4.0	STIME, DTIME, ETIME
5.0	IDEBUG(1-4)
6.0	NFAST, NEREM, NPREM, NISG
6.1	IREMVE(1-NEREM)
6.2	IREMOV(1-NPREM)
7.0	NSP
8.0	ICOLOR(1-40)
9.0	ICOLOR(41-100)
10.0	ICOLOR(101)
10.1	ICOLOR(102-121)
11.0	NSTEPS(1-40)
12.0	NSTEPS(41-66)
12.2	NSTEP, IHARN(1-2), WFAC
13.0	INT, SFACTR
14.0	OFSETX,OFSETY
15.0	VP(1-3), RA(1-3), IVP, ICODE
16.0	XMIN,XMAX

For further information on these variables, see Reference 4. The corresponding data information is given below Table 1:

Table 1 Sample Input Control File

```
12345678901234567890123456789012345678901234567890123456789012345678901234567890

1
 0
(blank line)
0.0000  0.0100  0.0000
1111
 024 116
2161718192021242526272829303132333435363738394O
 7
 0
        2       2       2       2       2       2       2       2
        2       2       2       2       2       2       2       2
        2       2       2       2       2       2       2       2
        2       2       2       2       2       2       1       1
        1       1       3       1       1       3       3       1
        1       3       3       1       1       1       1       1
        1       1       1       1       1       1       1       1
        1       1       1       1       1       1       1       1
        1       1       1       1       1       1       1       1
        1       1       1       1       1       1       1       1
        1       1       1       1       1       1       1       1
        1       1       1       1       1       1       1       1
        1       1       1       1
        4
        2       2       2       2       2       2       2       2
        2       2       2       2       2       2       2       2
        2       2       2       2
 8 5 8 3 5 5 3 3 5 3 3 3 3 3 3 3 3 3 3 3 8 8 8 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
20202020202020202020202020202020202020202020202020202020
20 0 0      0.50
  4        110.0
      2.0       3.0
   1000.0    1000.0    -20.0      10.0       0.0       0.0       16        2
     0.00      9.90
```

(B) Sample Output Graphics File Plot

The graphics outputs of the Table 1 are shown on Figures 5

and 6 at the viewpoints which are located at (1000.0, 1000.0, -20.0), and (750.0, 250.0, -20.0), respectively, in the IVP segment number coorinate system. These two pictures display the hyperellipsoid number 24 without hidden algorithm.

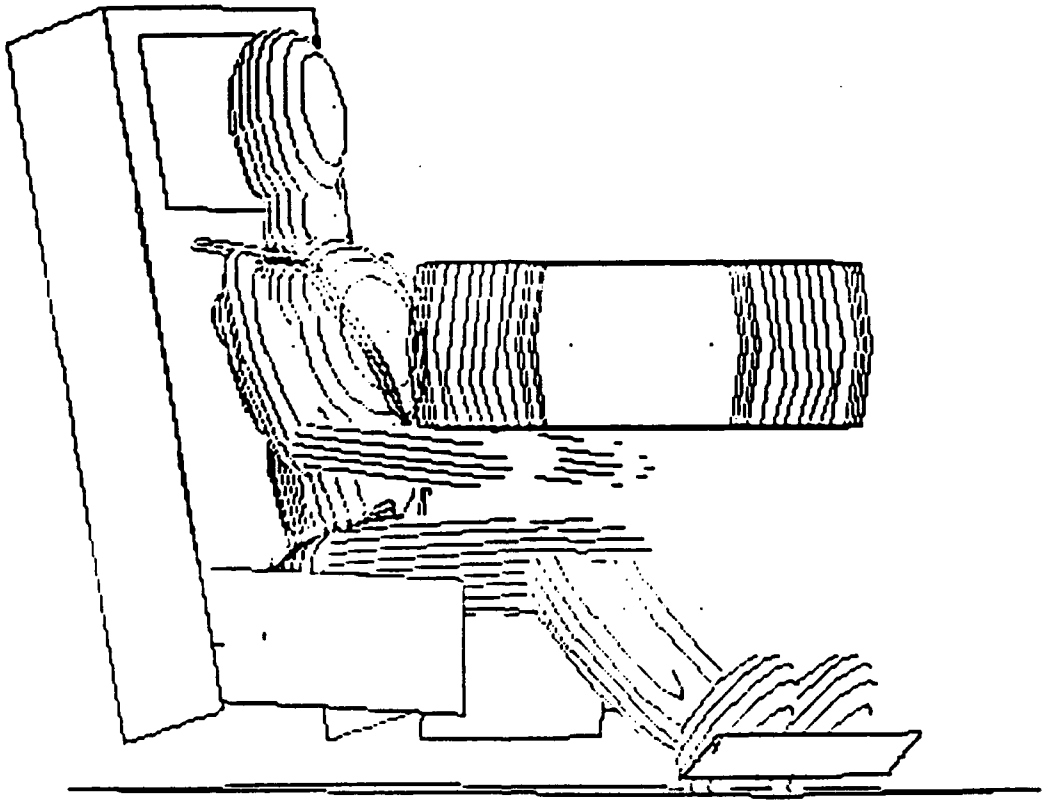


Figure 5

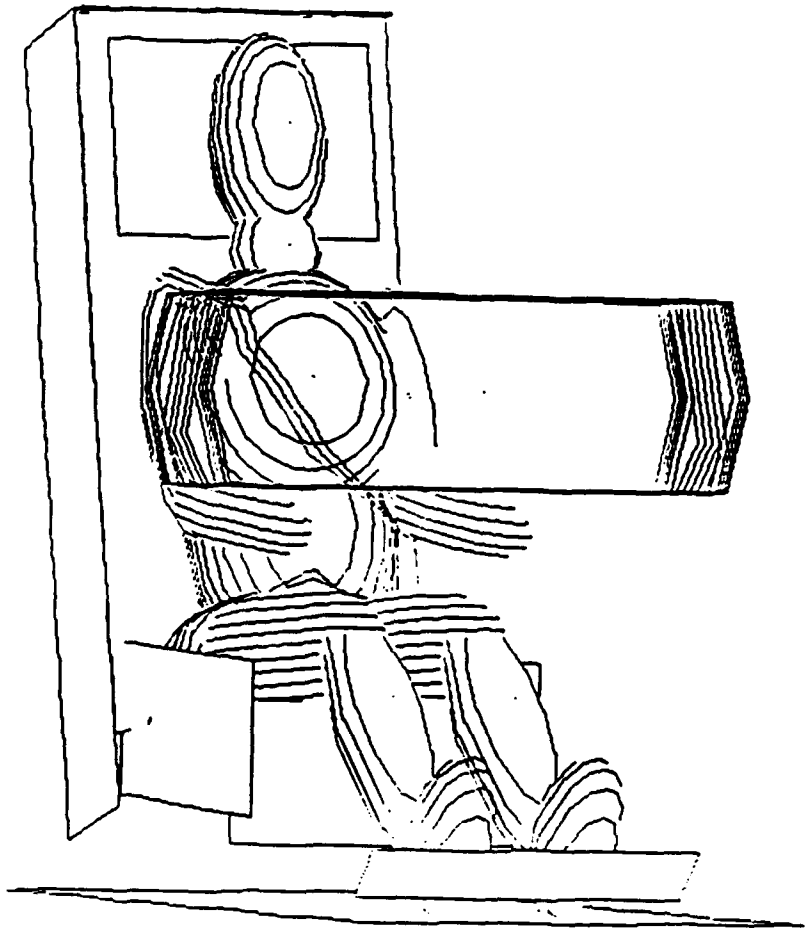


Figure 6

V. Recommendations

The revised VIEW program with additional ellipsoid option generates a nice visual data display for the users of the ATB model program. However, the program with the hyperellipsoid option has not yet reached its final development. Somehow, the hidden algorithm does not work properly. If we reduce the powers of x, y, and z terms in the equation of the hyperellipsoid to 2, that is, change back to the ellipsoid, the hidden algorithm works very well. Figure 7 below shows the evidence. This figure is the same picture as figure 6. The only difference between these two pictures is that the hyperellipsoid is replaced by the ellipsoid. This indicates that the subroutines XYZNEW and YZNEWT do not work out to the potential that they should. The Newton-Raphson method is good and fast, but only near the root. In using Newton-Raphson method, a starting point near the root must be specified. Therefore, the main strategy is to search for a good starting point. In future research, a new strategy would be to consider relocating a good starting point so that the vector **M** can be solved exactly or as accurately as possible.

In the figure 6 we can see a drawback of the picture in which the contour lines of the hyperellipsoid do not show a desirable picture. There is a large gap between the center and end parts. In addition, these are not symmetric on the two end portions. In either end part, the middle of the contour lines does not break at the center. The reason is the nature of an implicit equation can not produce the symmetric broken points. One possible solution is to take smaller increments for x and y, but that produces little improvement in covering the large gap and the broken points. Another

possible solution is to take an unequally spaced increments for x and y, still, this produces little improvement. According to the suggestion from Franklin and Barr (Ref. 2), the explicit equation of the hyperellipsoid is expressed z as a function of x and y in the b nomial expansion. Again, the output does not show the desired picture we expect.

In summary, this research has not yet reached the final development in the VIEW program. We need more time to fulfill its complete potiential. Looking for a good starting point for the equations (9) and (10) in the section III above is the primary strategy. Certainly, we should spend more time to work on the suggestion from Franklin and Barr (Refs. 1 and 2),. Another possible idea is to treat the hyperellipsoid as a ploygon instead of a square ellipsoid.

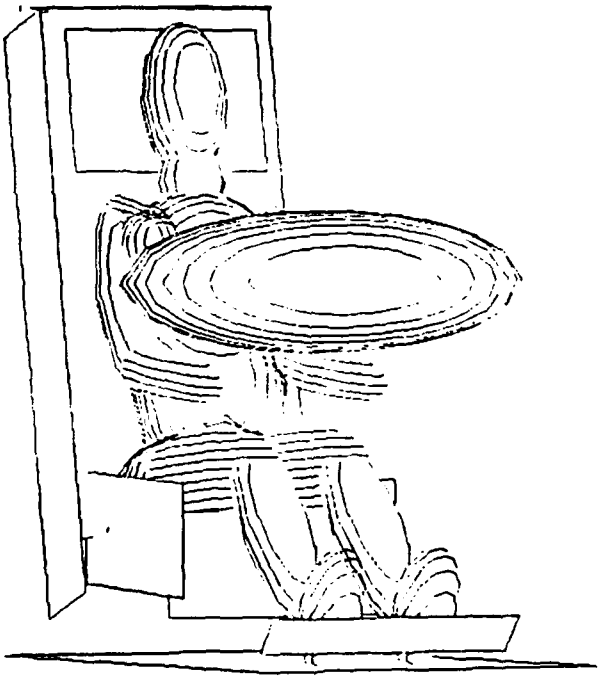


Figure 7

VI. Refferences

1. Barr, A. H., "Superquadrics and Angle-Preserving Transformations", IEEE Computer Graphics and Applications, Vol. 1, No. 1, Jan. 1981.
2. Franklin, W. R. and Barr, A. H., "Faster Calculation of Superquadric Shapes", IEEE, Computer Graphics and Applications, Vol. 1, NO. 3, July 1981.
3. Leetch, B. D. and Bowman, W. L., "Articulated Total Body (ATB) 'VIEW' Program, Software Report, Part I, Programmer's Guide", AFAMRL-TR-81 -111, Volume 1, June 1983.
4. Leetch, B. D. and Bowman, W. L., "Articulated Total Body (ATB) 'VIEW' Program, Software Report, Part II, User's Guide", AFAMRL-TR-81-111, Volume II, June 1983.
5. Obergefell, L. A., Fleck, J. T., Kaleps, I., and Gardner, T. R., "Articulated Total Body Model Enhancements" Volume 1: Modifications, AAMRL-TR-88-009, January 1988.
6. Obergefell, L. A., Gardner, T. R., Kaleps, I., and Fleck, J. T., "Articulated Total Body Model Enhancements" Volume 2: User's Guide, AAMRL-TR-88-043, January 1988.
7. Obergefell, L. A., Kaleps, I., Gardner, T. R., and Fleck, J. T., "Articulated Total Body Model Enhancements" Volume 3: Programmer's Guide, AAMRL-TR-88-007, January 1988.
8. Tang, C. M., Rizer, A. L., and Kaleps, I., "Articulated Total Body (ATB) View Program with Harness-Belt Implementation", AAMRL, September, 1989.

1989 USAF-UES FACULTY RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

Prepared by:	Ebo Tei, Ph.D.
Academic Rank:	Professor
Department and	Psychology
University:	University of Arkansas Pine Bluff
Research Location:	University of Arkansas Pine Bluff Pine Bluff, Arkansas 71601
USAF Researcher:	Rik Warren, Ph.D.
Date:	24 Jun 91
Contract No:	F49620-88-C-0053/SB5881-0378

EXPLORATIONS INTO THE VISUAL PERCEPTUAL FACTORS OPERATING IN HIGH-SPEED LOW-ALTITUDE TURNS

by

Ebo Tei, Ph.D.

ABSTRACT

Ultimately, to develop empirically based training programs, decrease loss of life due to crashes, and increase tactical capability, there is the need to understand and establish the visual, cognitive, and physical limits imposed on the pilot as he operates in the low altitude environment. The general goal was to initiate a series of exploratory research that will eventually lead to a research program geared toward isolating the visual perceptual and cognitive factors that contribute to a disproportionate number of accidents in maneuvers involving turns in low altitude flights. The objectives of this research initiation program were achieved.

ACKNOWLEDGEMENTS

I wish to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsoring this research initiation program. Appreciation is also extended to Universal Energy Systems for their professionalism in administering the program.

I wish to personally thank Dr. Rik Warren of the Armstrong Aerospace Medical Research Laboratory for his invaluable advise and help in making the project possible. Sincere appreciation is also extended to Robert Todd and Jeff Maresh of Engineering Solutions for their patience and professionalism in dealing with my equipment problems and developing the software for the project. Special thanks go to Ms. Tammy Sellers for running the study and developing the training protocols. And finally special appreciation is extended to the University of Arkansas at Pine Bluff for providing the facilities and release time that made it possible for me to carry out the project.

I. INTRODUCTION

For a variety of reasons pilots may be forced to operate in the low altitude environment. This may come about as a result of the weather, the mission requirements, and threat to mission success. Whatever the reason for operating in the low altitude arena, the pilot's task is a formidable one. Not only does he have to maneuver his aircraft throughout its full operating envelope with 100 per cent safety, but also he has to have the ability to monitor himself while managing and controlling all the mission tasks without ever compromising terrain clearance. The task is made much more difficult by the fact that the pilot may be required to operate from 50 to 100 feet of the ground at 500 knots while maneuvering in excess of 5 g's to accomplish a variety of tasks both inside and outside the cockpit.

It is not surprising then that a large percentage of operations factor accidents occur in the low altitude environment (Miller, 1983). However, what is surprising is that a disproportionately high percentage of aircrew fatalities resulting from the low altitude accidents were attributable to vertical maneuvering and turning.

There are three categories of maneuvers in the low altitude environment: straight and level, vertical maneuvering, and turning. The analysis of accidents covering only day VFR accidents indicates that even though pilots are exposed to vertical maneuvering and turning only 8 percent of the time,

these two maneuvers account for most of the ground impacts (in excess of 85 percent). Even more important is the fact that exposure to turning is only about 5 percent and yet it accounts for over 50 percent of the accidents.

The question is why do most of the accidents occur during the turns? The experience of the pilot does not seem to be an important variable. Several explanations including those dealing with physics, perception, and cognition have been advanced (eg. Kellog, 1982; Haber, 1984, 1987; McNaughton, 1981, 1984; Roscoe, 1980, Wiener, 1977; 162d Tactical Fighter Group, 1986). What is needed is a systematic exploration of the possible causes through basic research.

A great deal of research effort has been concentrated on the visual perceptual aspects of low altitude flights (eg. Rinalducci, 1981; Rinalducci, et al., 1984, 1985; Warren, 1982,1988; Warren and Mcmillan, 1984; Wolpert et al., 1989; Flack, et al., 1986; Hettinger & Owen, 1985; Levinson, et al., 1982; Owen and his colleagues at the Ohio State Aviation Psychology Lab, 1981-1989; Kellogg & Miller, 1984; De Maio, et al., 1983, 1985; McCormick, 1985). Unfortunately, in reviewing the literature, one finds a common thread in all the research -- the experimental tasks involve investigation of perceptual factors during straight and level maneuvering.

It is true that the straight and level maneuver accounts for about 90 percent of the exposure time in low altitude flight. However, in terms of ground impacts, it accounts for

only 9 percent of the accidents. While a great deal of the effort involving straight and level maneuvering has been carried out in an attempt to design better simulators for low altitude training, it is still no substitute for understanding the perceptual and cognitive constraints that are brought to bear on the pilot during turns and vertical maneuvers.

The Low Altitude Training program developed by the 162D Tactical Fighter Group took into account the perceptual and cognitive variables that are important in turn maneuvers. The most relevant among these were time control (mission cross-check time) and constant self-monitoring to keep track of terrain clearance and mission tasks. Future enhancements in the training program can be achieved by providing empirical support for some of the major perceptual assumptions underlying the program. Furthermore, only by understanding the perceptual and cognitive dynamics of the turning maneuver can there be some chance of reducing the disproportionately high rates of accidents occurring in turns. It was to address some of these concerns that the current exploratory project was proposed.

II. OBJECTIVES

The current project was intended to be the beginning of a series of pilot experiments with the goal of systematically exploring and gaining insights into the visual perceptual and

cognitive limitations imposed on pilots as they are forced to operate in the low altitude environment. It was to serve as a basis for future expanded research effort that will be specifically geared toward: a) identifying the sources of visual perceptual factors needed to operate safely in high-speed low altitude flights, especially during maneuvers involving turns, b) gaining insights into why the accident rate is so high in the turns during low altitude maneuvers, c) isolating the psychological factors that contribute to a disproportionate number of accidents in maneuvers involving turns, d) identifying the metacognitive skills and strategies necessary to monitor performance safely during turns.

Since the available literature provided little guidance in this area of research and the performance period of the project itself was very short, the objectives of the current project were very modest:

1. Train subjects to fly a circular ground track over a flat ground while maintaining a designated altitude. The question was how readily could subjects learn to do the task given the constraints of the experimental conditions.
2. Develop training protocols for future experiments based on the performance of subjects in the current project. How do you train subjects to fly a circular path when their only window to the world is the view from the cockpit and they can

only see a very small portion of the circular track at any one time?

3. Develop and refine the various displays that will serve as a foundation for later research. This also implied a refinement of the computer software for generating the displays.
4. Determine the limits of the various experimental parameters used in developing the various conditions in the experiment. For example, what is the maximum speed that can be used for the subjects to perform the task within acceptable limits? That is, at what speed does performance break down to the extent that subjects can no longer perform their assigned task?

III. METHODS

1. Logic and Rationale

The basic rationale of current and future research was to first train subjects to fly a circular ground track while maintaining a designated altitude. After subjects were performing the tracking task sufficiently well, the next step was to both perceptually and cognitively load the tracking task with other demands. The purpose would be to explore how well subjects perform the task under additional demands. In other words, how readily do subjects learn the task and how

well do they adapt to new cognitive and perceptual demands without affecting their performance on the primary ground tracking task?

2. Apparatus

The flight simulation program was implemented on a 386-based computer system. The system performance requirements included the specification that the processor exceeded 8800 Dhrystones for CPU performance and 2500k Whetstones for math processing performance. The system was compatible with Metrabyte™ I/O boards including PIO-1224-bit parallel digital I/O interface and DAS-88 channel high speed A/D converter. There was also compatibility with Microsoft™ MS-DOS™, IBM™ PC-DOS™, and X-TAR FA-2000-64KNM geometry accelerator and PG2000-4096-768 graphics board. The display screen provided for horizontal scan rates between 16 and 35KHZ, and capable of displaying VGA images.

Other system components included Tower Case and 275 W power supply, 1.2 Mb Teac™ floppy drive, Seagate™ 128 Mb - 14ms (ESDI) formatted hard drive, WD 1007 (ESDI) disk controller, Orchid Pro™ designer card (512 kb, 16 bit), AT I/O card, Mtek™ 101-key keyboard, Mitsubishi™ HL6905 20" monitor and base, Everex™ 60 Mb tape backup, 80387-33 Math Coprocessor, AMI 33 Mhz 386 4 Mb - 64 cache, and joystick with an analog interface.

3. Flight Simulation and Display Characteristics

3.1 Software

The realtime simulation FLIGHT Software was developed from scratch for the project. The development was a major part of this project and actually went through three major revisions. The program allows the user to design, run, and analyze curvilinear flight experiments. It simulates a generic aircraft with equations of motion that properly simulate a coordinated level turn of an aircraft. The radius of turn and turn rate for a given velocity and roll angle are correctly related. The program incorporates equations of physics for a coordinated turn which help the experimenter set the parameters for the experimental conditions desired. The equations describe the relationship between the speed, turning radius, and the roll angle of all aircraft making a coordinated level turn. The size, weight, or thrust of the aircraft is not a factor in a coordinated level turn, thus reducing the problem to simple physics. (See Appendix B for documentation on coordinated turn equations).

Another feature of the FLIGHT Software was that it included plotting capabilities to aid the experimenter in visualizing the experimental results. At the close of a trial, two plots automatically appear. The lower half of the display shows a plot of the altitude (AGL) profile as a function of time. In the upper right window, a bird's eye view of the flight path is superimposed over the ideal circular path. This gives

feedback as to how faithfully the circular path was tracked by the subject. The plots always appear during the review mode of the program. However, as part of the experiment configuration, the software allows the experimenter to determine if the plots will appear between trials or not.

The FLIGHT Software has several features designed to make its use as convenient as possible. All the major components are menu driven. (See Appendix A for a full documentation of the FLIGHT Software Instructions). The major menu items are:

1. **Condition File Edit** -- This allows the specification of the ground track texture(s) used in the experiment.
2. **Order File Edit** -- This specifies the starting conditions for each trial.
3. **Experiment Configuration Edit** -- This allows a number of general experiment parameters to be specified.

[Refer to Section 3.3 (Developing the Displays) for details on these three menu items].

4. **Run Experiment** -- This permits the running of a Flight experiment that has been defined by the Condition, Order, and Configuration files above. It also allows information about the subject to be specified.
5. **Review Experiment** -- This allows a trial-by-trial replay of the experiment including initial

conditions of each trial and the subject's performance for those trials.

6. **Analyze Experiment** -- This allows the conversion of .bin files into .dat files which can then be used with many statistical analysis programs.
7. **Exit Program** -- The option to terminate the Flight program.

3.2 Flight Simulation Displays

The display consisted of a VGA-resolution horizon and circular ground track. When tracking properly and directly over the track, subjects would see the line representing the horizon and a small portion of a circular track disappearing off to the lower right of the visual display screen. The direction of flight was clockwise. The viewport from the cockpit of an aircraft following a circular path in the real world would preclude seeing the whole track at any one time. So the display was actually simulating the real world. The higher the velocity of the aircraft, the smaller the segment of the ground track that would be seen if the subject was faithfully following the circular track. This feature of the display presented a lot of tracking problems to the subjects in the initial phases of the experiment. However, the training and the plots of subject's performance that appeared at the end of trials helped the subjects gain insight into the problem.

The display itself was dynamic in the sense that active control of the flight stick by the subject introduced corresponding dynamic changes in the display. This was in accordance with the active psychophysics proposed by Warren and McMillan (1984). If subjects descended below ground level, there was an inversion such that the track was seen suddenly above the horizon line. This was an indication that the subject had crashed. For the few subjects who had this experience, the appearance of the display was explained and they were able to recover from this on subsequent trials.

3.3 Developing the Displays

Developing the displays and determining the experimental parameters and configurations constituted a major part of the project. The displays determined the various flight environments used. The development involved several phases. First a Condition File was developed. The specification of parameters in this file created the various characteristics of the ground track used on each trial. The items specified included the radius of the track, the number of lanes within the track, the width of each lane, the density of the texture across the ground track, and the number of segments ahead of the current position to be drawn by the program to give the subject a complete window to the world. (Refer to Flight Software Instructions in Appendix A).

The Condition File was then used to develop the Order File.

The Order File was simply a list of the starting conditions for each trial. This included the ground track to be used for the trial, the subject's initial position (in feet) on the X-coordinate (where +X corresponds to North and -X to the South), the subject's initial position (in feet) on the Y-coordinate (where +Y corresponds to East and -Y to the West), the subject's initial altitude (in feet), the initial roll of the aircraft (in degrees), the initial pitch (in degrees), the initial heading (in degrees), and the initial velocity (in knots).

The next step involved the development of the Experiment Configuration File which allowed the experimenter to specify a number of experiment parameters. This file specified how each session of the experiment was to be run. The menu for this file allowed the user to specify which of up to three separate experimental phases were to be used in the experiment. Within each phase, the user could specify whether that phase was enabled or disabled, the Condition File name, and the Order File name. The other menu item in this file was the Run Configuration which allowed the experimenter to specify the testing and display parameters. The parameters included the trial length, sample rate, and debug which toggled the debug mode on or off to indicate roll, pitch, and heading during the trials. The other parameters were viewport distance (the distance in feet from the subject's eyes to the center of the video display), viewport width (the width in

feet of the video display), and viewport width (the height of the video display in feet). The viewport settings were critical in ensuring that the graphics generated were appropriate for the display size and the viewer's distance from the display.

The Experiment Configuration also allowed the user to set the type of control for altitude only, roll only, or both. This made it possible to limit the control dynamics to only one dimension for certain experiments. Another feature was the capability to set the vertical rate and the roll rate factors from the experiment menu. The user could set these rates as desired for a given experiment.

4. Procedure

4.1 Phase 1

Ten subjects (5 males and 5 females) enrolled in a General Psychology class participated in the study to earn extra credits. They were informed that they were participating in a flight experiment, and that their major task would be to follow a circular track while maintaining a given altitude. They were given a description of how the track would look like. As soon as the trial begun, they were to "find" the track, establish contact with it, try to maintain a consistent flight path, and also hold the same altitude.

The basics of aircraft movement were carefully explained:

banking left or right by moving the flight stick in the corresponding direction, and changing the pitch of the aircraft by pulling back or pushing forward to gain or lose altitude. The experimenter then went through one trial to demonstrate how the task was to be performed, and the experiment begun. The first session was to be a practice session lasting about an hour.

4.1.1 Design

Three main factors were varied. There were three different velocities set at 100, 300, or 500 knots. There were three flight environments specified by three texture densities of one-, two-, or three-lane ground tracks. Finally, initial roll was set at three levels of 30, 45, or 60 degrees. Radius of the ground track was fixed at a 1000 feet, and altitude was set at 100 feet. Both initial heading and pitch were set at zero degrees.

This was a within-subject design. There were to be four sessions with the first one serving as a training session. There were to be 54 trials, with each lasting 60 seconds. The Order File for the training session was established by randomly selecting two trials from each of the experimental order files. The other three sessions were to be run in blocks of 54 trials for each of the velocities at 100, 300, and 500 knots. The order of presentation was to be counterbalanced. Thus for each block of trials, each subject would have supplied 6 data points per cell.

4.1.2 Results

The initial performance of the subjects in the first few trials led immediately to changes in the focus of the experiment. The task was impossible to perform by the subjects. It had not been anticipated that the two higher velocities of 300 and 500 knots constrained at a radius of 1000 feet would require the subjects to be flying at bank angles generally characteristic of acrobatic air displays. It became a challenge to see whether the subjects could be trained to fly the track. They either could not find the track at all, or when they did, it just flew by them. For the subjects, the control on the flight stick was very erratic. The focus of the experiment quickly changed from one of gathering data to one of developing training protocols to teach the subjects how to perform the task.

4.2 Training Protocols

The training was carried out by the research assistant who was a licensed pilot. A "hands-on" approach to the training was adopted whereby the experimenter actually demonstrated most of the possible flight situations to the subjects. Using Fitts and Posner's (1967) three main phases of learning psychomotor skills, the early or cognitive phase of the training involved identifying the critical and separate parts of the tracking task. The skill involved in each part was demonstrated by the trainer. The subjects were then given the

opportunity to practice each part.

The level of difficulty of the task made the training more interactive than had originally been anticipated. Subjects were given demonstrations of flight path corrections and explanations of the relationship between movements of the flight stick and corresponding changes in the display. For example, the change in orientation of the horizon line as a function of pitch and bank angle of the aircraft was explained. The important visual cues necessary to assess the status and condition of the aircraft were pointed out. Subjects were then shown how to use these visual cues to initiate action, guide their performance, and when an action should be ended. This understanding of the relationship between flight stick movement, aircraft movement, and changes in the visual display was very crucial in the training.

Another major facet of the training was teaching subjects how to "find" or "return" to the track once they had "lost" it. Losing the track meant that it was no longer visible on the screen, and the subject saw only the horizon line in a particular orientation depending on the bank angle of the aircraft. Subjects would lose the track when they did not bank sharply enough or correct their flight path quickly enough. To "return" to the track, they needed to wait for the track to become visible again. When it was visible again, it was generally far away from the subject's current position and seen to be lying toward the middle of the screen. The task

of the subject then was to level the aircraft, fly directly toward the track, and then depending on the position of the aircraft either bank right or left to reestablish contact.

What made this skill difficult to learn was that when the track became visible again, it flashed by on the screen so quickly that subjects had great difficulty "catching" it or keeping it on the screen. The problem was solved by teaching subjects to remove small amounts of their right bank angle until the track came across the screen more slowly. This required gaining an insight into how much bank angle was "enough" for a particular airspeed.

Subjects received a great deal of feedback. The plotting capabilities of the software which showed the performance of the subject at the end of each trial came in very handy. This was supplemented by heavy doses of verbal guidance and actual demonstrations on the part of the trainer.

4.2.1 Training Outcomes

When each trial began, the initial position of the aircraft was directly over the track. However, depending on the initial bank angle in the Experiment Configuration set for that particular trial, subjects had to quickly make adjustments in the bank angle to correct the flight path; otherwise they immediately lost the track. Subjects were not informed of the different initial bank angles. However, they quickly learned that the airspeed of the aircraft determined how quickly they should correct their bank angle to maintain

contact with the track. The greater part of the training consisted in learning to control the flight stick accurately and quickly enough to maintain contact with the track without overcompensating or undercompensating.

Different problems were encountered with the different velocities used in this phase of the study. Surprisingly, the 500 knot velocity seemed to be the easiest for subjects to maintain some "reasonable" contact with the track. At this speed, subjects were much less likely to overbank to the right. The reason immediately became clear when those trials were examined. The maximum bank angle in the Experiment Configuration was set at 90 degrees. At the speed of 500 knots and a track radius of 1000 feet, the necessary bank angle to maintain a flight path was 80 to 85 degrees. Subjects learned to go to a 90-degree right bank, establish contact with the inside of the track, and make tight turns on the inside without having to make further flight path corrections. Subjects were actually "forced" to stay on the inside because if they tracked at the correct bank angle, the track was hardly visible on the screen. At this velocity, it was much easier to lose the track.

At a velocity of 100 knots, the biggest problem was overcorrection. The display changed at a much slower rate than at 500 knots, and a bank angle of only 41 degrees was necessary to maintain contact with the track. Initially, subjects had a tendency to bank right to an angle of 90

degrees. When they did this they immediately lost the track. The track would flash by from right to left so quickly that subjects were unable to reestablish contact until the bank angle had been reduced.

Trials involving a velocity of 300 knots seemed to be the most difficult. Subjects would overbank and "fly by" the track, fail to change their flight path quickly enough, and thus "lose" the track altogether. Once lost, it was also much more difficult to reestablish contact with the track. Subjects were more likely to "lose" the track in this condition than the other two velocities.

Finally, there were more problems encountered with a one-lane ground track than with the other two. It was much more difficult to track and easier to "lose".

4.3 Phase 2

The training protocols developed were used in the second phase of the project. Five male subjects participated. There were several changes made in the experimental conditions and configurations.

In the first phase, a maximum of 1000 feet of track radius had been built into the software inadvertently. This limitation had placed major constraints on the bank angles that would be required to track the ground path since velocity, roll, and radius of the track were linked factors (see Warren and Owen, 1982). This led to revisions in the software to remove this limitation.

4.3.1 Design

There were minor revisions made in the original design. Again there were three velocities set at more realistic values of 100, 150, and 200 knots. To avoid the problems associated with extreme bank angles encountered in Phase 1, the maximum bank angle that had to be maintained in order for a ground path to be tracked correctly was limited to 60 degrees. Three bank angles were selected -- 30, 45, and 60 degrees. Once velocity and bank angle were specified, this determined the radius of the ground track to be used (see Table 1). Initial roll was also set at three levels -- 0, 20, and 40 degrees. Altitude, pitch, and heading were the same as in Phase 1. Ground texture was restricted to a ground path made up of three lanes.

TABLE 1

	Roll	Radius
100 knots:	30	1533
	45	866
	60	511
150 knots:	30	3450
	45	1993
	60	1150
200 knots:	30	6134
	45	3542
	60	2045

4.3.2 Results

After receiving one training session along the lines developed in the first phase of the project, all subjects were able to learn to track the ground path with no difficulty. This was an indication that the training protocol had been very successful. Changing certain aspects of the original experimental configuration (eg. reducing aircraft velocity) minimized the frustrations experienced by subjects in the first phase of the study.

IV. SUMMARY CONCLUSIONS

All the objectives of the project were achieved and some important lessons were learned. Perhaps the most important of these objectives was the development of a software package that will serve as the core program around which future curvilinear experiments will be designed. The important lesson was that the establishment of training protocols that will lead to skilled performance is a critical first step in carrying out curvilinear flight experiments. From this experiment, a detailed and systematic analysis of the skills for which training is to be provided, and a systematic definition of the knowledge and insight necessary for the performance of the skills have been developed. This will aid future studies.

V. RECOMMENDATIONS

This was the beginning of what is hoped to be a series of studies in curvilinear experiments to gain insight into why there are disproportionately more accidents involving turn maneuvers. The major limitation of this initiation project was the sophistication of the computer system used. Graphics were generated using VGA resolution. The display tended to have some flicker even though the flicker rate was not serious enough to be intolerable or impair performance. It was obvious however, that any attempt to generate a display any more complex than what was used in this study would have posed some major difficulties. .e future direction in this program is to explore more complicated terrain profiles with different texture densities. A more sophisticated and powerful dedicated graphics board like an XTAR board will be needed.

REFERENCES

Academic Text: Low Altitude Training (1986). 162D Tactical Fighter Group. Tucson, Arizona.

DeMaio, J., Rinalducci E. J., Brooks R. and Brunderman (1983). Visual cuing effectiveness: comparison of perception and flying performance. Proceedings of the Third Symposium on Aviation Psychology. Columbus, Ohio: The Ohio State University, Aviation Psychology Laboratory, Department of Aviation, 499 - 503.

Fitts. P. M. & Posner, M. I. (1967). Human Performance. Belmont, Calif: Wadsworth.

Flack, J. M., Riccio, G. E., McMillan, G. R. and Warren, R. (1986). Psychophysical methods for equating performance between alternative motion simulators. Ergonomics, 29, 1421 - 38.

Haber, R. N. (1984). Perceptual factors in low altitude flight. Proceedings of the 9th Symposium on Psychology in the Department of Defense. USAF Academy, Colorado Springs, Colorado.

Haber, R. N. (1987). Why low-flying fighter planes crash: Perceptual and attentional factors in collisions with the ground. Human Factors, 29(5), 519 -532.

Hettinger, L. J. and Owen, D. H. (1985). Increasing sensitivity to optical information specifying loss in altitude. Proceedings of the Third Symposium on Aviation Psychology, Columbus, Ohio: The Ohio State University,

Aviation Psychology Laboratory, Department of Aviation, 483 - 90.

Kellogg, R. S. (1982). Visual/perceptual aspects of an F-4G accident. In proceedings of the 26th Annual Meeting of the Human Factors Society, Seattle, WA: 316 - 318.

Kellogg, R. S. and Miller, M. (1984). Visual perceptual aspects of low level high-speed flight and flight simulation. Proceedings of the IMAGE III conference, Phoenix, AZ.

Levinson, W. H., Zacharias, G. L., and Sinacori, J. B. (1982). Design of an experiment to study the pilot's use of visual and motion cues in a height regulation task. Cambridge, MA: Bolt, Beranek and Newman; Technical Report 5028.

McCormick, D. (1985). Simulated terrain following flight: Visual and radar terrain correlation requirements. Proceedings of the Third Symposium on Aviation Psychology. Columbus, Ohio: The Ohio State University, Aviation Psychology Laboratory, Department of Aviation, 499 - 503.

McNaughton, G. B. (1981). Notes on human factors mishaps. Life Sciences - United States Air Force Safety Journal, 1 - 9.

McNaughton, G. B. (1984). Collisions with the ground -- what we learn from mishap analyses. Panel session: Perceptual factors in Low Altitude Flight. Proceedings of the 9th Symposium on Psychology in the Department of Defense. USAF Academy, Colorado Springs, Colorado.

Miller, M. (1983). Low altitude training: How long can you go? Tucson: 162D Fighter Weapons School, Arizona Air National Guard, Tucson, Arizona.

Owen, D. H., Warren, R., Jensen, R. S., Mangold, S. J., and Hettinger, L. J. (1981). Optical information for detecting loss in one's own forward speed. Acta Psychological, 48, 203 - 213.

Owen, D. H. (Ed.) (1982). Optical flow and texture variables useful in simulating self motion. (AFOSR Project 762550/7-13531 Interim Technical Report. Grant No. AFOSR - 81 - 0078). Columbus, Ohio: The Ohio State University, Department of Psychology, Aviation Psychology Laboratory.

Owen, D. H., Wolpert, L. and Warren, R. (1984). Effects of optical flow acceleration, edge acceleration, and viewing time on the perception of egospeed acceleration. In D. H. Owen (Ed.), optical flow and texture variables useful in detecting decelerating and accelerating self motion (AFHRL-TP-84-4, AD-A148718). Williams AFB, AZ: Operations training Division, Air Force Human Resources Laboratory.

Owen, D. H. (Ed.) (1985). Optical and event-duration variables affecting self motion perception. (Interim Tech. Rep. for contract No. F33615-83-K-0038). Columbus, OH: The Ohio State University, Department of Psychology, Aviation Psychology Laboratory.

Owen, D. H. and Warren, R. (1987). Perception and control of self motion; implications for visual simulation of vehicu-

lar locomotion. In L. S. Mark, J. S. Worm, and R. L. Huston (Eds.), Ergonomics and human factors: Recent research. New York: Springer - Verlag, 40 - 70.

Rinalducci, E. J. (1981). Visual cues in the simulation of low level flight. USAF - SCEE Summer Faculty Research Program, AFHRL/OF, Williams AFB, Arizona.

Rinalducci, E. J., Patterson, M. J., and DeMaio, J. (1984). Static vs. dynamic presentation of visual cues in simulated low level flight. Proceedings of the 9th Symposium on Psychology in the Department of Defense. USAF Academy, Colorado Springs, Colorado.

Rinalducci, E. J., Patterson, M. J., Forren, M. and Andes, Jr. R. (1985). Altitude estimation of pilot and non-pilot observers using real-world scenes. Proceedings of the Third Symposium on Aviation Psychology. Columbus, Ohio: The Ohio State University, Aviation Psychology Laboratory, Department of Aviation, 491 - 498.

Roscoe, S. (1980). Aviation Psychology, Iowa City, IA: University of Iowa Press.

Warren, R. (1982). Optical transformation during movement: review of the optical concomitants of egomotion. AFOSR-TR-82-1028, Final Tech. Report Grant No. AFOSR-81-0108. Bolling Air Force Base, DC: Air Force Office of Scientific Research (NTIS No. AD-A122 275).

Warren, R. (1988). Visual perception in high-speed low-altitude flight. Aviation, Space, and Environmental

Medicine, November, A116-A124.

Warren, R. and McMillan, G. R. (1984). Event perception needs an active psychophysics: An altitude control example. Proceedings of the IMAGE III Conference, Phoenix, AZ.

Warren, R. and Owen, D. H. (1982). Functional optical invariants: A new methodology for aviation research. Aviation, Space, and Environmental Medicine, 53(10), 977 - 983.

Weiner, E. L. (1977). Controlled flight into terrain accidents. System-induced errors. Human Factors, 19, 171 - 181.

Wolpert, L., Owen, D. H., and Warren R. (1983). Eye-height-scaled versus ground-texture-unit-scaled metrics for the detection of loss in altitude. In Proceedings of the Second Symposium on Aviation Psychology. Columbus, Ohio: The Ohio State University Aviation Psychology Laboratory, 513 - 521.

Wolpert, L., Reardon, K. A. and Warren, R. (1989). The effect of changes in edge and flow rates on altitude control. Proceedings of the Fifth Symposium on Aviation Psychology. Columbus, Ohio: The Ohio State University, Aviation Psychology Laboratory, Department of Aviation.

A P P E N D I X A

Engineering Solutions Inc.

FLIGHT Software Instructions

Introduction

Enclosed is the FLIGHT Software diskette. The FLIGHT program allows the user to design, run, and analyze curvilinear flight experiments.

Running FLIGHT

Before you use this software, you must copy it onto the hard disk on your system. Create a directory named "Flight" on drive C: (type "md Flight" then hit the RETURN key, hereafter abbreviated <CR>). Next, log onto the Flight directory (type "cd \Flight"<CR>). Now you are ready to copy the contents of the FLIGHT diskette onto your hard drive.

While in the Flight directory, insert the FLIGHT diskette into drive A of your computer and type "copy *.*"<CR>. After the files are copied, you are ready to run the program. To run the program, type:

FLIGHT <CR>

At this time an introductory screen will be displayed followed automatically by a list of available experiment configuration files. In addition to the parent directory ("..\"), there will be only one file available ("test.cfg"). To load the "test.cfg" file into the FLIGHT program, use the arrow keys to highlight the file in the list and press <CR>. Now you will be presented with the Main Menu. If you wish to enter a different experiment configuration file, you may enter the file name by typing it and pressing <CR>.

FLIGHT features

The FLIGHT Software has several features designed to make its use as convenient as possible. All menu selections are accomplished with single key presses. Each option is selected either by pressing a number which corresponds to that option, or by pressing the highlighted letter found within the desired option.

Another convenient feature of the FLIGHT program is the Escape key (<ESC>). Any time you press <ESC>, you will immediately terminate any option you have selected and be returned to the previous menu.

The use of default values/selections whenever possible also helps make the FLIGHT program easy to use. The default value can be selected simply by pressing <CR>. If you do not wish to use the default, just type in the desired selection and press <CR>.

The Main Menu

The following is a description of the Main Menu options and their submenus:

<1>Condition File Edit

This option allows the user to view and edit the Condition Files used in the experiment to specify the ground track texture(s) used in the experiment. These files contain a table which specify the parameters necessary to create the ground texture of your experiment. Any time you make changes to this file, be sure to save your changes before exiting.

The items you must specify to create your ground track are:

- Number — A numeric label for each line in the table. When you create/edit the Order File, you will refer to the various lines in the Condition File by this number.
- Radius — The distance from (0, 0, 0) of the three-dimensional coordinate system (i.e., the center of the world) to the midline of the ground track.
- #Wide — The number of concentric circles which specify n - 1 "lanes" within the ground track. (e.g., when #Wide = 3, the ground track will consist of two concentric lanes.)
- Width — When #Wide $\geq$ 2, this option specifies the width of the concentric lanes which comprise the ground track.
- #Around — The innermost circle of the ground track is constructed by connecting points with straight lines. The greater the number of points, the smoother the circle, the fewer the number of points, the less round the "circle" appears. #Around specifies the number of points used to create each circle. Additionally, at each point specified by #Around, a line crosses the entire width of the ground track which divides the ground track into segments; this provides texture across the ground track.
- #Draw — This specifies the number of segments ahead of the current position to be drawn by the program. By specifying #Draw to be at least 1/4 the value of #Around, subjects will be presented with a complete window to the world, while easing the amount of work required of the computer.
- Type — This option is not yet operational.

<2>Order File Edit

The Order File is simply a list of the starting conditions for each trial. Again, any time you make changes to an Order File, be sure to save them before you exit.

Order File Edit (cont'd)

The initial conditions you specify are:

- Cnd — This option specifies which Condition Number (i.e., ground track) to use for the trial.
- X — This specifies the subject's initial position on the X-coordinate (in feet) where X corresponds to North and -X corresponds to South.
- Y — This specifies the subject's initial position on the Y-coordinate (in feet) where Y corresponds to East and -Y corresponds to West.
- Alt — The subject's initial altitude (in feet). (In the X, Y, Z coordinate system, Altitude corresponds to -Z.)
- Roll — The subject's initial roll (in degrees).
- Pitch — The subject's initial pitch (in degrees).
- Hdg — The subject's initial heading (in degrees).
- Vel — The subject's initial velocity (in MPH).

<3>Experiment Configuration Edit

The Experiment Configuration File allows the user to specify a number of general experiment parameters. Before you are allowed access to the Experiment Configuration File, you will be prompted for a password. At the prompt respond: FLIGHT <CR>, and you should gain access to the Experiment Configuration Menu.

The Experiment Configuration Menu contains the following items:

<1>First Phase General Configuration
<2>Second Phase General Configuration
<3>Third Phase General Configuration

These three selections allow the user to specify which of up to three separate phases are to be used in the experiment. Within each of these choices, you must specify the following information:

- Phase On/Off — "Off" indicates that the phase is disabled, "On" indicates that the phase is enabled.
- Condition File Name — Enter the prefix of the .cnd file you wish to use in the experiment.
- Order File Name — Enter the prefix of the .ord file you wish to use in the experiment.
- Instance File Name — This option is not operational at this time.

Experiment Configuration Menu (cont'd)

<4>Run Configuration

This selection allows the experimenter to specify testing and display parameters. These include:

• Trial Length	The duration of each trial in the experiment (in seconds).
• Sample Rate	You may choose from six different sampling rates which will then be used for every trial in the experiment.
• Store Rate	This option is not operational at this time.
• Debug On/Off	Toggles the debug mode on and off. While debug is on, X, Y Z, roll, pitch, and heading are displayed during the trials.
• Run XTAR	Specify whether you are using XTAR graphics boards on this system.
• VGA Graphics	If you are not using XTAR graphics boards, you must answer yes that your system has VGA graphics capabilities. If not, the program will not run.
• Viewport Distance	Specify the distance from the subject's eye to the center of the video display (in feet).
• Viewport Width	Specify the width of the video display (in feet).
• Viewport Height	Specify the height of the video display (in feet).
	The three Viewport settings are critical to ensure that the graphics generated by FLIGHT are appropriate for the display size and the viewer's distance from the display.
<5>Instance File	This option is not operational at this time.
<6>Password Change	This option allows the user to change the password.
<7>Exit Configuration Menu	This option returns the user to the Main Menu.

<4>Run Experiment

This option permits the user to run a FLIGHT experiment which has been defined by the Condition, Order, and Configuration files discussed previously. Upon selecting this option, you will be asked whether you wish to edit the Configuration File. After a negative response or following your changes, you will be prompted for subject information. Type in the requested information followed by <CR> for each entry. If you wish to make use of a default "subject" press <Esc> and a generic set of subject data will be entered. This is convenient when verifying experimental designs.

After you have entered the subject information, you will be ready to begin the experiment. Press any key to begin the next trial. At this point, the initial trial conditions will be displayed. Press any key again to begin the actual trial. If you wish to terminate a trial and then proceed with the next one, press "Q" during a trial. If you wish to terminate the trial and return to the Main Menu (thereby ending the testing session), press <Esc> anytime during the testing session.

<5>Review Experiment

This option allows the user to replay a subject's performance. After selecting this option, you will be prompted for the prefix of the .bin file to review. If you do not know the file name, press <F10> for a listing of all available .bin files. The review option provides a trial-by-trial replay of the experiment including the initial conditions of each trial and the subject's performance for those trials you request.

<6>Analyze Experiment

This option allows the user to convert the .bin files into .dat files. The data (.dat) files can then be used with many statistical analysis programs. The .dat files contain 11 columns, they are:

• Time	The time (in seconds) of the data point.
• X, Y, & Z	The subject's position on the three-coordinate system (in feet).
• Roll, Pitch, and Heading	The subject's roll, pitch, and heading (in degrees).
• Velocity	The subject's velocity (in MPH).
• Aileron and Elevator	The subject's control input for the three controls on a scale from -100 (minimum) to +100 (maximum) for each.
• Throttle	The subject's control input for the throttle on a scale from 0 (minimum) to +100 (maximum).

<7>Exit Program

This option terminates the FLIGHT program and returns the user to the DOS prompt. The current Configuration File is automatically saved before you program ends.

Input Controls

The aileron (right/left stick motion) is a roll-rate controller and roll position determines turn rate. The elevator (forward/backward stick motion) is a rate-of-change in altitude controller. At this time, there is no controller for forward speed.

If you have any further questions pertaining to this software, contact us at: **Engineering Solutions, Inc.; 5688 Duquesne Place; Columbus, OH 43235; (614) 459-0344.**

A P P E N D I X B

Flight Version 1.2a *26 February 1991*

Installation Instructions

Copy the new flight.exe" to the working directory on the computer. The software displays the new version number 1.2a at the top of the screen. It should be verified that the new software is being used for all subsequent work. Any experiment configurations that were created will work with the new program. In addition, any data collected with the old program can be reviewed or analyzed.

Changes

1. Modified the equations of motion to properly simulate a coordinated level turn of an aircraft. The radius of turn and turn rate for a given velocity and roll angle are now correctly related.

2. Added a flag in the experiment configuration to determine the type of control. The type of control can be set for altitude only, roll only, or both. This allows the experimenter to limit the control dynamics to only one dimension for certain experiments.

3. Corrected the display height and width parameters in the configuration table to accept the correct range of numbers. The input range was incorrectly limited from 0.0 to 1.0.

4. A problem was found where memory was not correctly deallocated and files were not being closed properly while the software was running. This could lead to memory allocation errors while running the software and errors loading font and other files when several trials were run or reviewed. These problems should not occur any longer.

5. Added the capability to set the vertical rate and the roll rate factors from the experiment configuration menu. This allows the experimenter to set these rates as desired for a given experiment.

6. Added plotting capabilities to aid the experimenter in visualizing the experimental results. At the close of a trial two plots automatically appear. The lower half of the display shows a plot of the altitude profile as a function of time. In the upper right window, a birds eye view of the flight path is superimposed over the ideal circular path. The plots will always appear during review mode. There is a flag in the configuration table that allows the experimenter to determine if the plots will appear between trials or not.

7. Modified the velocity units to be knots instead of ft/sec. this affects the displayed velocity in the order file and on the report window before each trial.

Coordinated Turn Equations

26 February 1991

Introduction

The following equations describe the relationship between the speed, turning radius, and the roll angle of all aircraft making a coordinated level turn. The size, weight, or thrust of the aircraft is not a factor in a coordinated level turn thus reducing the problem to simple physics.

Definition of Variables and Constants

g	32.174 feet/sec^2 acceleration due to gravity
V	Velocity of aircraft in feet/sec
Vn	Velocity of Aircraft in Knots
R	Radius of flight path in feet
ψ	Heading of aircraft in degrees
$\Delta\psi$	Turn rate of aircraft in radians/sec
T	Turn time of the aircraft to complete 360 degrees (2π radians) in seconds
ϕ	Roll of aircraft in degrees
G	G load factor in G's

Radius

The radius of the flight path as a function of speed and roll position is as follows. Note that the velocity must be specified in feet per second.

$$R = \frac{V^2}{g \tan(\phi)} \quad (1)$$

If the velocity is specified in knots then the equation requires a conversion constant as follows.

$$R = \frac{(1.688\ Vn)^2}{g \tan(\phi)} \quad (2)$$

The radius may also be specified as a simple function of velocity and turn rate.

$$R = \frac{V}{\Delta\psi} \quad (3)$$

Specified in terms of velocity in knots and turn time in seconds the equation becomes.

$$R = 1.688\ Vn\ \frac{T}{2\pi} \tag{4}$$

Roll Angle

The next set of equations define the roll position of the aircraft. The first equation defines roll as a function of velocity, and turn rate.

$$\phi = \tan^{-1}\left(\frac{\Delta\psi\ V}{g}\right) \tag{5}$$

The same equation specified by velocity in knots and turn time in seconds is as follows.

$$\phi = \tan^{-1}\left(\frac{1.688\ Vn\ 2\pi}{g\ T}\right) \tag{6}$$

Roll angle can also be specified in terms of velocity in knots and radius.

$$\phi = \tan^{-1}\left(\frac{(1.688\ Vn)^2}{g\ R}\right) \tag{7}$$

Turn Time

The turn time equation is as follows.

$$T = \frac{2\pi\ R}{1.688\ Vn} \tag{8}$$

G Load Factor

The equation used to calculate the G load factor is as follows.

$$G = \frac{1}{\cos(\phi)} \tag{9}$$

Examples

If the turn time is known to be 120 seconds (2 minute turn) and the velocity is 100 knots the roll, radius, and G load factor can be calculated using equations 4, 6, and 9 respectively as shown below.

$$R = 1.688(100)\frac{(120)}{2\pi} = 3224 ft \tag{10}$$

$$\phi = \tan^{-1}\left(\frac{1.688(100)2\pi}{g\ (120)}\right) = 15.36 degrees \tag{11}$$

$$G = \frac{1}{\cos(15.36)} = 1.04G \tag{12}$$

If the velocity is known to be 100 knots and the roll angle is 45, the radius, G load factor, and turn time can be calculated using equations 2, 9, and 8 respectively.

$$R = \frac{(1.688(100))^2}{g\tan(45)} = 885.6\ ft \tag{13}$$

$$G = \frac{1}{\cos(45)} = 1.41\ G \tag{14}$$

$$T = \frac{2\pi(885.6)}{1.688(100)} = 33\ \text{sec} \tag{15}$$

Report # 110
760-0MG-071
Prof. Yin-min Wei
Report Not Publishable

1991 USAF-UES RESEARCH INCENTIVE PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

Final Report

Computer-based training for complex, dynamic tasks.

Prepared by:	Kevin B. Bennett, Ph. D.
Academic Rank:	Assistant Professor
Department and	Psychology Department
University:	Wright State University
Research Location:	AL/HRGA
	WPAFB, OH 45433-6503
USAF Researcher:	Michael Young
Date:	31 May 1991
Contract No:	F49620-88-C-0053

Computer-based training for complex, dynamic tasks.

by

Kevin B. Bennett

ABSTRACT

For a variety of reasons the Air Force is interested in improving the effectiveness of performance in complex, dynamic tasks representative of those in the command and control environment. This final report describes four studies that were conducted to investigate issues in the design of graphic displays for computer-based training in these domains. In display design techniques of direct perception and direct manipulation can be used to provide alternative conceptual perspectives of a target world that will support an individual in learning to understand, or in learning to control a complex system. The experiments investigated issues in the development of two types of graphic displays: animated functional mimics and configural displays. Experiments 1 and 2 investigated the role of emergent features and graphical elements in the design of configural displays. It was found that a direct mapping between the emergent features produced by the configural display and the semantics of the domain assisted individuals in tasks that required the integration of information, relative to a non-configural display (bar-graphs). In addition, color-coding the lower-level graphical elements may have helped to off-set some of the costs that are typically associated with the configural display of information. Experiments 3 and 4 investigated fundamental issues in the design of animated mimic displays, focusing on the effectiveness of chromatic and luminance contrast for the provision of animation. It was found that observers were extremely sensitive to luminance contrast, and that chromatic contrast was less effective.

Acknowledgments

I would like to thank those individuals and organizations who have contributed their time, effort, and resources to making the RIP appointment successful and mutually beneficial. First, I would like to thank the Air Force Systems Command, the Air Force Office of Scientific Research, Universal Energy Systems, and the AL/HRGA (especially Bert Cream and Larry Reed), for making the program possible. I would also like to thank Bob Pressel who assisted in development and Al Nagy, David Woods, John Flach and Mike Young for discussions of various aspects of the research reported. The part-task simulation used in Experiments 1 and 2 was originally developed at the Westinghouse Electric Corporation, primarily in the department of Human Sciences Research with the assistance of numerous individuals and organizations.

I. INTRODUCTION:

Effective training is a major concern to the Air Force. Air Force personnel are required to operate and maintain extremely complex equipment, and the consequences are potentially catastrophic: poorly maintained or operated equipment and a lack of adequate preparation for critical tasks have the potential for loss of human life, the resources invested in training these personnel, and expensive equipment. Advances in computer science and artificial intelligence are providing powerful new tools that expand the potential to provide computer-based training in complex problem solving domains. These technological advances include low-cost memory and processing power, graphics processors, and specialized software such as rapid-prototyping, simulation, expert system shells, and object-oriented languages.

An alternative to the use of this computational power for the development of expert systems that explicitly tutor an individual is the development of an effective learning environment that facilitates the acquisition of knowledge: the environment module (Burton, 1988). Graphic representations of domains can facilitate learning by providing both direct manipulation (to facilitate control) and direct perception (to facilitate situation assessment or diagnosis). Through direct perception, graphic displays can facilitate learning by assisting an individual in finding and integrating relevant data, by "envisioning" information (that is, to make the abstract concrete), and by restructuring of an individual's view of the problem (the provision of alternative conceptual perspectives). The term "representational aiding" (Kerlik, 1991; Woods and Roth, 1988; Woods, 1991; Zachary, 1986) has been used to refer to the role that graphic displays may assume in supporting an individual to complete tasks in complex, dynamic domains.

For the most part, technical issues associated with the implementation of graphic displays for representational aiding have been resolved. The problem that faces designers is the development of design approaches that allow this potential to be used effectively. A number of researchers have been developing these approaches (Bennett, in press; Bennett,

Toms, and Woods, in preparation; Flach and Vicente, in press; Rasmussen, 1986; Vicente and Rasmussen, 1990; Woods, 1991; Woods and Roth, 1988). At a high level of abstraction, these design approaches are characterized by the consideration of both the specific perceptual and cognitive capabilities of the human and the specific characteristics of the domain. In particular, the goal is to map the domain semantics (the critical variables, the relationships between these variables, and the relevant goals and constraints) into the static appearance and dynamic behavior of the graphic displays in such a fashion that the extraction or decoding of this information is easily achieved.

From the perspective of the cognitive system triad (Woods and Roth, 1988) the quality of performance in complex, dynamic domains is the result of three interactive and mutually constraining components: the cognitive demands produced by the domain of interest, the cognitive agent(s) that meet those demands, and the representation of the domain through which the agent experiences and interacts with the domain. The characteristics of each component, and the interactions between them, determine the ease or difficulty of problem solving. Typically, research on graphic displays has emphasized only two of these components: the representation (the form of the graphic display), and the information processing characteristics of the individual using that representation. For example, Cleveland and his colleagues (Cleveland, 1985) investigated the ability of individuals to discriminate identical information mapped into different low-level graphical forms (e.g., area of a circle vs. length of a line). One very useful result of this work is the provision of a hierarchical listing of the effectiveness of these graphic forms for conveying information to human problem solvers. When viewed from the perspective of the cognitive system triad, however, these investigations are incomplete: understanding the general utility of various graphic forms and the general processing characteristics of the individuals who will use them is necessary, but not sufficient. These two factors also interact with, and are constrained by, the characteristics of the domain.

A critical aspect of this approach is the realization that to accomplish tasks in complex domains the operator must understand and consider the system from a number of different conceptual perspectives. Rasmussen (1986) has developed a conceptual

framework, the "abstraction hierarchy", that can be used as a guide to developing graphic representations that provide these conceptual perspectives (Rasmussen, 1986; Flach and Vicente, in press). The levels of abstraction range from the physical form of a system (e.g., what are the s_ .em components?, what do they look like?, where are they located?) to the higher-level purposes it serves (e.g., what is the system's purpose?, what constraints does the system operate under to fulfill this purpose?).

Configural displays (sometimes referred to as object displays) have the potential to be particularly effective at representing information at higher levels in the abstraction hierarchy. High-level, abstract concepts are determined by the relationships between lower-level variables. Mapping information about these lower-level variables into the graphical elements of an object results in "emergent features." Emergent features are highly salient visual properties that result from the interaction of lower-level graphical elements which are not present in digital or separate presentation of the variables. Because of this property configural displays are useful in representing higher-level relationships between variables, in providing information about the status of higher-level goals, and are typically independent of the actual physical implementation.

Research on configural displays has yielded mixed results. In some experiments the display of information in a configural format has improved integration performance (Barnett and Wickens, 1988; Carswell and Wickens, 1987; Goldsmith and Schvaneveldt, 1984; Wickens and Andre, 1990) while in other experiments it has not (Sanderson, Flach, Buttigieg, and Casey, 1989; Coury, Boulette, and Smith, 1989). In contrast to the findings for integration tasks, the literature has revealed a rather consistent pattern of results indicating that in tasks where performance depends upon the consideration of an individual variable (focused attention) or when irrelevant variables need to be ignored (divided attention) the display of data in a configural format hinders performance (Carswell and Wickens, 198 Casey and Wickens, 1986; Coury et al., 1986; Goettl, Kramer and Wickens, 1986; Wickens and Andre, 1988).

In addition to information about high-level abstract relationships and system goals

individuals will also require explicit information at the level of physical implementation in complex, dynamic domains: what are the alternative system resources that are available to avoid or recover from violation of system goals? For example, imagine that the level of a storage tank is low. What alternative resources are physically connected to the tank? How might these alternative resources be redirected to increase the level of the tank? Information at this level of abstraction corresponds to the lower levels of Rasmussen's hierarchy. In particular, the level of physical function "represents the physical (i.e., the mechanical, electrical, or chemical) processes of the system or its parts" (Rasmussen, 1986, p. 16). One type of display that presents information at this level of abstraction is the "mimic" or "pictorial" display.

In process control domains static labels and markers are added to control panels to assist the operators in understanding the complex physical connections that exist in the system. In aviation pictorial displays have been used to convey the status of fuel, hydraulic, or electrical systems (Hawkins, Reising, and Gilmore, 1983; Stokes, Wickens, and Kite, 1990). Thus, animated mimic displays provide representations of the important components, systems, or subsystems and the flow of information or resources between them. Very little, if any, research has directly addressed fundamental issues in the design of animated mimic displays. What method should be used to provide the animation? How should information concerning the rates of flow that exist in the domain be mapped into the perceptual characteristics of the graphic representation?

II. OBJECTIVES OF THE RESEARCH EFFORT:

Previous research has revealed contradictory results concerning the effectiveness of configural displays in conveying two types of information 1) high-level, abstract information, and 2) low-level information concerning the values of individual variables. In Experiments 1 and 2 these issues are reexamined using a laboratory analog of a complex, dynamic task that is representative of a real world domain. Performance for a configural

display with color-coded graphical elements was compared to performance for a bar graph display in two empirical investigations. Based on the results, principles to achieve these dual design goals were developed and are described.

When completing tasks in complex, dynamic domains an individual must consider information from a variety of conceptual perspectives. High-level information concerning the relationship between variables can provide an indication of progress towards system goals. However, lower-level information concerning alternative system resources and the physical connections between these resources must also be made available if an individual is to avoid or recover from trouble. In many domains this information is provided through static mimic displays. With advancing computational power it is now possible to graphically illustrate the important components, systems, or subsystems and the flow of information or resources between them. Animated mimic displays have the capability to facilitate diagnostic performance in real-time as well as improve the effectiveness of computerized training. Despite this potential, very little empirical research has addressed issues in implementation. Experiments 3 and 4 investigate fundamental issues in providing animation through the use of apparent motion. One general method for providing animation and two encoding alternatives (chromaticity and luminance) are examined. Design recommendations for animated mimic displays were developed and are discussed.

III. DEVELOPING GRAPHICAL DISPLAYS FOR THE ENVIRONMENT MODULE OF AN INTELLIGENT TUTORING SYSTEM

III A. Configural displays: Emergent features and graphical elements

One important line of research contrasts "separate" and "configural" (or pattern) displays as alternative representations for domain semantics. A configural display represents higher order properties of the domain through the relationships among the lower order data that define the property. Instead of directly noting the value of the higher order

property, it is represented as an emergent property of the structure and behavior of the lower order data. The emergent property is based on a relationship between low-level graphical elements or the configuration of these elements -- in other words, there is an emergent pattern or configural property of the display. The pattern or configural relationships can be structural or dynamic, i.e., the behavior or movement of the graphical elements relative to others. Because the relationships that define the emergent property sometimes refer to how graphical elements fit together to create higher order visual objects (e.g., four lines making a rectangle or a geometric pattern), they are sometimes referred to as object displays.

This type of display can be contrasted to separate displays in which individual state variables have their own unique graphic form (e.g., two bar graphs, rather than a rectangle). This holds the interaction between the graphical elements to a minimum, and therefore reduces the range and quality of emergent features that are produced. However, as Sanderson, Flach, Buttigieg, and Casey (1989) have illustrated, under certain conditions bar graphs can produce emergent features that facilitate performance relative to configural displays. Actually the configural-separable dichotomy is more appropriately characterized as a continuum with geometric pattern displays at one end, digital displays at the other end, and bar graphs falling somewhere in the middle of the continuum.

To complete tasks in complex, dynamic domains individuals must consider information regarding both higher-level properties (those defined by the relationships between one or more variables) and lower-level data (the individual variables themselves). The empirical research on configural displays has revealed mixed results with respect to these dual requirements. First, there is inconclusive evidence indicating that configural displays can improve performance at tasks that require consideration of the relationship between several variables (high-level information, sometimes referred to as an "integration" task). In contrast to these findings there is a relatively consistent pattern of results indicating that presenting information in a configural display results in performance decrements when the task requires the consideration of individual variables (sometimes referred to as a "focused" task). The present paper provides design principles for configural displays that can help achieve these dual design goals. Empirical evidence that supports

the utility of these principles was obtained using a laboratory version of a real world complex, dynamic control task. A set of state variables, which included a new type of predictive information, were displayed to system controllers in two forms: 1) a bar graph display that presented each state variable as a separate graphic form; and 2) a configural display that integrated the state variables into a single graphic form with emergent properties that provide information about the relationships among the state variables.

Research on configural displays

One of the characteristics of complex, dynamic domains (e.g., command and control, process control) is that individuals must consider information from a variety of sources to reach decisions with regard to issues that arise in the domain. In some cases the interface has been designed with a "one-sensor, one-display" design philosophy which forces the operator to collect, mentally maintain, and mentally integrate information. This places severe cognitive demands on the limited-capacity cognitive resources of attention and working memory, and increases the potential for error (Goodstein, 1981). One option open to system designers is to provide configural displays that collect and integrate information for the operator. For example, a configural display that is currently in use in some nuclear power plants is the polar graphic display: a polygon with eight spokes (Woods, Wise, and Hanes, 1981). The spokes of the polygon are dynamically scaled so that a regular polygon always represents normal conditions, while distortions in the polygon represent a developing abnormality. This display integrates information from over one hundred individual sensor values in order to represent process health. The display exploits the excellent pattern-recognition capabilities of the human information processing system by literally transforming decision-making from a predominately cognitive activity to a predominately perceptual activity.

However, despite the intuitive appeal, the literature has revealed inconclusive evidence regarding the effectiveness of configural displays in tasks that require the

integration of information from a number of variables. In some experiments the display of information in a configural format has improved integration performance (Barnett and Wickens, 1988; Carswell and Wickens, 1987; Goldsmith and Schvaneveldt, 1984; Wickens and Andre, 1990) while in other experiments it has not (Sanderson, Flach, Buttigieg, and Casey, 1989; Coury, Boulette, and Smith, 1989). One interpretation of these seemingly contradictory results is that it is not the presentation of domain information in a configural format, per se, that determines the effectiveness of a display. Rather, the crucial determinant of success is how well the critical data relationships have been mapped into the appearance and dynamic behavior of the configural display.

For graphic displays that map higher-level information the concept of "emergent features" (Pomerantz, 1986) is crucial for effective design. Emergent features refers to the fact that the lines, contours, and shapes that are produced by the individual parts of a graphic display can "configure" to produce higher-level, global perceptual features that are independent of any single part. For example, mapping two variables into a single graphic form of a rectangle (one variable in the x axis and one variable in the y axis) produces a rectangle that has the emergent feature of area. Sanderson et al. (1989) illustrated that the success of a graphic display in representing higher-level properties depends upon the mapping between the emergent features produced by a display and the demands of a task, rather than upon objectness per se. In this study the emergent features associated with separate displays (the inferred linearity between the heights of bars in a bar chart) were more effective in representing the critical domain semantics than those of a configural display (the angles of a triangle).

Thus, one critical issue for configural display design is how to combine low-level graphical elements so that they configure to produce higher-level emergent features in a manner that reflects the critical domain semantics (in particular, the relationships between variables). Pomerantz (1986), working with simple visual stimuli and tasks, states that "The emergent features associated with grouping and configuration result from idiosyncratic and unpredictable interactions of parts, and as a result they can produce unusual effects that can easily be misinterpreted" (p.13). In complex, dynamic domains the

problem is exacerbated by the additional requirement that the emergent features must reflect the complicated domain semantics. The following section provides one example of the mapping of domain properties into the dynamic behavior of a configural display for process control.

Discovering the semantics of a complex domain

The particular task that serves as a vehicle for these studies of configural and separate display formats is a laboratory version of an actual process control task. The real world analog for the laboratory task is the control of water level in a boiler during the start-up of a power plant. This is an actual task where human operators control a non-minimum phase dynamic system. The laboratory task is a second order simulation of these dynamics that was designed to capture the critical demands that the actual task places on human performance (Roth and Woods, 1988). A similar laboratory process system with simpler dynamics has been used in studies of process control skill, cf., Crossman and Cooke (1974), Moray, Lootsteen and Pajak (1986) and Moray (1987). In the real world version of the task, energy from a source (in this particular case, a nuclear reactor) is used to convert water to steam in multiple boilers (or steam generators -- SG). When the supply of steam is sufficient, it is used to load a turbine to generate electricity. During the start-up of this process, the task of the feedwater operator (one of several human operators and automatic control loops who control the process) is to control feedwater flow in order to maintain the water level in the steam generators between high and low set-point boundaries. Exceeding these limits results in a plant shutdown, which has high economic penalties, and the start-up must begin anew.

There are several types of task demands that make performance difficult. First, two goals interact in feedwater/level control -- (a) generate electricity (from the point of view of steam generator level control, generate sufficient steam to meet the electricity goal) and (b) maintain steam generator (SG) level within limits. The processes that effect these goals

are the energy inflow/outflow through the SG (energy inflow versus the energy leaving in the form of steam to either the turbine or some other location) and the water (mass) inflow/ outflow through the SG (mass inflow in the form of feedwater versus mass outflow in the form of steam) process.

Goal competition can arise because changes in each process affects both goals. Thus, some changes that will help to satisfy one of the goals, at the same time, can degrade the status of the other goal. Changes in steam and feed flow affect level by changing the mass balance. Changes in these processes also affect the energy balance (the temperature of the feedwater relative to the water in the steam generator and the rate of steam flow is proportional to the rate of energy outflow). Although changes in the energy balance do not affect the amount of water mass in the steam generator, they do affect the energy content of the water and therefore the measured level. The level control goal is specified in level units, not mass (one can think of the measured level as a measure of the volume occupied by the mass of water present). As a result, changes in energy affect the level goal, but at a different time constant than the mass effects. Strong pressure to complete the start-up rapidly intensifies competition between goals. Rapid maneuvers produce large disturbances in SG level which must be compensated for and introduce multiple forces acting on level simultaneously which complicates situation assessment (especially where is level going to be in the future given past actions and influences).

The demand characteristics of the task can also be described in terms of the dynamic system properties of feedwater/boiler systems. Most feedwater/boiler systems are characterized by non-minimum phase dynamics. This means that there are time delays so that there is a gap between the time an event occurs or an action is taken, and the time its effect on variables like indicated SG level can be observed. Another kind of non-minimum phase behavior which this system exhibits has been referred to as shrink/swell effects by operators. In shrink/swell, the initial effect of a change in a state variable on SG level is the opposite of its long term effect. For example, when the flow of relatively cold feedwater is increased, the fact that the energy of the feedwater is less than the energy level in the SG causes the net energy in the SG to decrease. The energy drop decreases indicated

level at one time constant, while the net increase in mass inflow increases indicated level but at another longer time constant. The result is that an increase in feedwater flow results in indicated level first decreasing (shrink) due to the energy effect and then increasing due to the mass effect. Because of time delays and shrink/swell effects, only weak evidence about the critical state variables is available to controllers. In addition, accurate measures of steam flow and feed flow do not exist at low power.

The complex dynamics place a premium on a controller's ability to anticipate the effects of changes in plant state or control actions on SG level, and make compensatory responses before the ultimate effect of the event on SG level is seen. In order to do this the controller needs to know what energy and mass factors have been introduced into the system. However, the dynamics and the poor state information make this assessment difficult (cf., Roth and Woods, 1988). For example, is a change in indicated SG level a longer term change in mass or a transitory energy effect?

Representational aiding in the feedwater/boiler control task

The cognitive analysis of this task (Roth and Woods, 1988) showed that performance could be enhanced if information were provided which helped a controller better anticipate level behavior. Effective assistance should help the problem solver build a better situation assessment of where level is, what factors are influencing level (when are changes due to shrink and swell and when are they due to changes in the mass balance), and where level will go given these influences and possible interventions. To enhance operator anticipation, a new form of predictive information was developed for non-minimum phase dynamic systems (cf., Woods and Roth, 1988). Performance at the feedwater/level task fundamentally revolves around separating the relative contributions of two independent functional processes on indicated level (water mass effects from the effects of energy changes or shrink/swell) or providing cues that would assist operational personnel to do this.

Decision support was developed to assist the operators in separating the contributions. A measure of the mass contribution to indicated SG level (calibrated in terms of indicated level units because the shutdown limits are expressed in terms of indicated level) was developed and will be referred to as compensated steam generator level (CSGL). CSGL is an estimate of current SG level that is not confounded by any shrink/swell effects or, alternatively, is an estimate of SG mass transformed into level units. Compensated level provides information about the future course of indicated level. It can be thought of as an indication of where measured SG level will be in the future once shrink and swell effects dissipate. Note that compensated level is calculated for each time sample of data; it is not a literal projection of level behavior into the future (although it can be used to derive such a projection). It computes what indicated level would be, if the energy state of the SG was nominal. Initial studies have shown that compensated state variables aid human control performance in non-minimum phase dynamic systems (Bennett, Woods, Roth, and Haley, 1986).

To support human control performance, then, we need to design a visual representation or display (Woods, 1991) which portrays state values in a form so that the operator can better anticipate and control the feedwater/boiler process during the start-up (i.e., avoid unnecessary plant shutdowns). The information to-be-communicated in this representation revolves around showing the operator the relationship between state variables related to the goal of the task -- keep level within limits. Data on the state variables associated with the energy and mass balance processes that influence level behavior -- feedwater temperature, feedwater flow, steam flow, energy inflow -- must be integrated and inter-related with the goal data. The operator must integrate this diverse set of data in order to assess where level is, what factors are influencing level (when are changes due to shrink and swell and when are they due to changes in the mass balance), and where level will go given these influences and possible interventions.

We used this dynamic control task as a vehicle to investigate issues about how separate and integral or configural displays of data support human performance. This task is appropriate to this end on several grounds., First, task performance requires

consideration of several interacting state variables. Second, it is a dynamic task where we could produce a tractable laboratory version which possesses the same critical demand characteristics as the real world analog. Third, in contrast to past studies which have examined separate and configural display issues in laboratory tasks that have no relation to actual settings, actual practitioners and actual displays, the data obtained in this study directly relates to a real world situation and the displays investigated relate more closely to displays that real practitioners might actually use.

Mapping the domain semantics into computer-based configural displays

Two displays, a configural and a separate[1] display, were developed for the manual control of feedwater task (see Figures 1, 2, 3, and 4). The separate display mapped the four task variables (ISGL, CSGL, SF, and FF) into separate, color-coded bars with a common baseline. Figures 1 and 2 illustrate two snapshots of the separate display. As discussed in the previous section, two critical relationships in the domain semantics are mass balance (differences between steam and feed flow) and energy balance (difference between ISGL and CSGL when mass are equal). The emergent features corresponding to these critical data relationships in the separate display are the inferred linear relationship between the heights of adjacent bars. For example, the relationship between steam and feed flow is represented by an imaginary ascending or descending line between the bars representing these two variables. In Figure 1 there are large mass and energy imbalances, as indicated by the steep inferred linear relationships between the appropriate bargraphs. A large swell effect is present, and ISGL is likely to cross the lower set-point boundary soon. In Figure 2 mass and energy are both balanced and the plant is relatively stable.

1. As previously mentioned, bar graph displays do produce emergent features. These labels have been chosen to be consistent with previous research.

FIGURE 1. The separate bar graph display (large mass and energy imbalances).

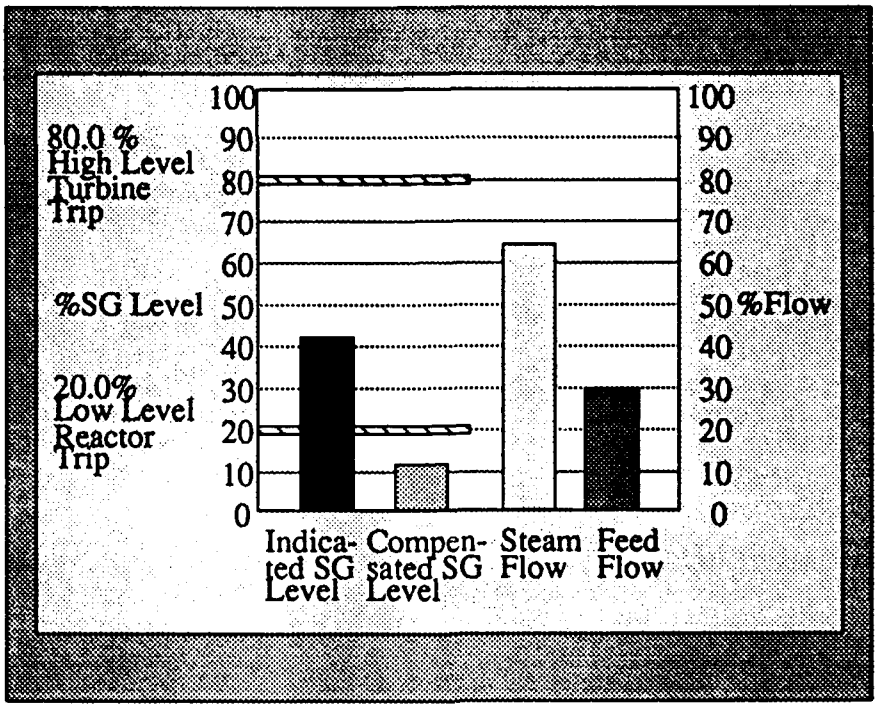


FIGURE 2. The separate bar graph display (small mass and energy imbalances).

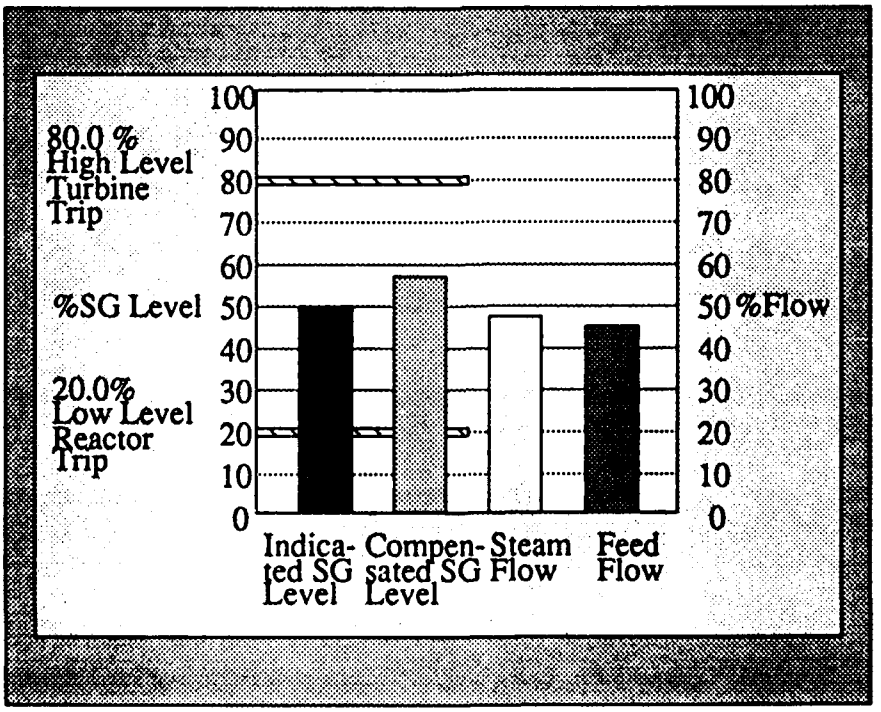


FIGURE 3. The configural bar graph display (large mass and energy imbalances).

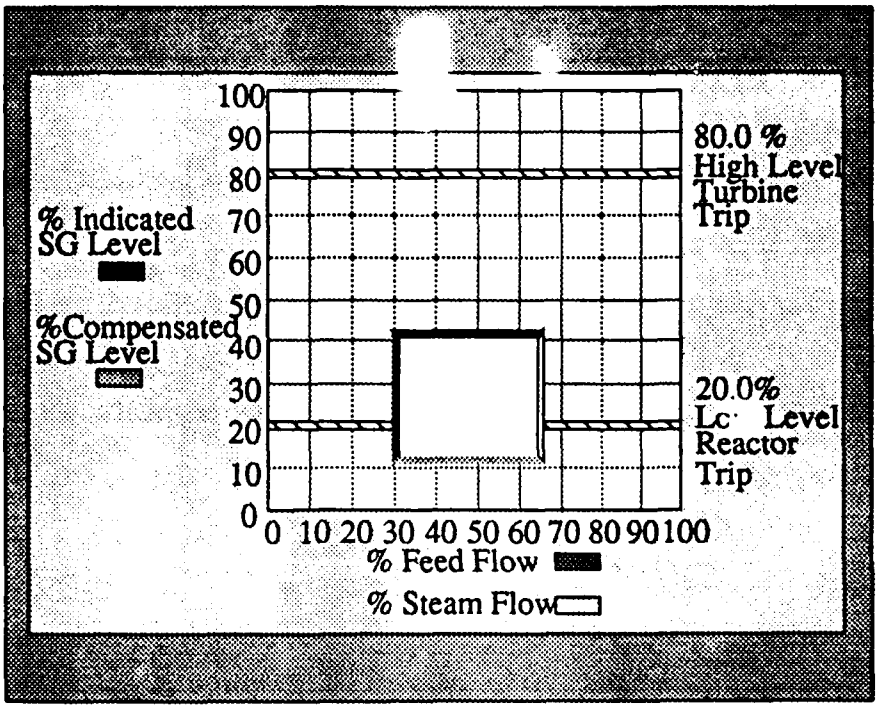
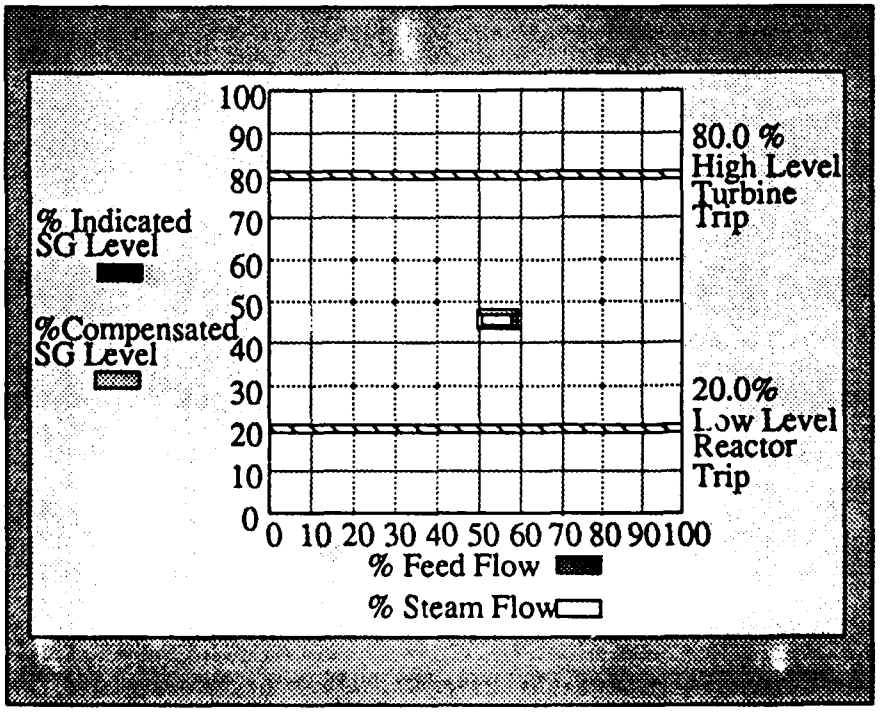


FIGURE 4. The configural bar graph display (small mass and energy imbalances).



The configural display mapped the four variables into a single geometric object: a rectangle. The values for indicated and compensated SGL are plotted in the horizontal axis and values for steam flow and feedwater flow are plotted in the vertical axis. Where the values for the four variables intersect in the display grid a rectangular shape is drawn and the sides of the rectangle are color-coded to reflect the contribution of an individual variable. Figure 3 illustrates a snapshot of the configural display when there are relatively large mass and energy balances (the same values are used for Figures 1 and 3). Figure 4 illustrates how the configural display would appear with a relatively stable and safe system state (the same values were used in Figures 2 and 4).

Mapping these four variables into a single geometric object results in a number of emergent perceptual features. For example, the critical relationships between SF/FF and ISGL/CSGL are directly mapped into the height and width of the rectangle, respectively. Additional emergent features including the area and shape of the rectangle, the location of the rectangle in the display grid, and the direction and rate of movement within this grid. For example, the area of the rectangle is an emergent feature which roughly corresponds to plant stability.

The configural display provides a more effective mapping of the domain semantics than the separate display. To successfully complete the manual control of feedwater task the user must consider several higher-level properties (mass balance, energy balance) to assess system state and to determine the correct control input. Part of the inherent difficulty of the task arises because of the tight inter-coupling between subsystems. Changes in a primary variable (e.g., feedwater flow or steam flow) will produce changes in both the mass balance and the energy balance, which in turn, effect the critical performance variable, ISGL. The configural display provides a representation that highlights this tight coupling by providing a number of emergent features, including the height, width, and area of the rectangle, and its location, direction of movement, and rate of movement within the display grid. The only emergent features that the separate display provides is the inferred linear relationship between the bar graphs. Thus, the configural display represents the high-level information about the relationships between variables through emergent perceptual

features that can be directly perceived and decoded; the separate display maps this information into emergent features that are less direct and must be inferred. Experiment 1 compares performance at the MCF task with the configural and the separate displays. It is predicted that the configural display will improve performance at the manual control of feedwater task.

Experiment 1

Method

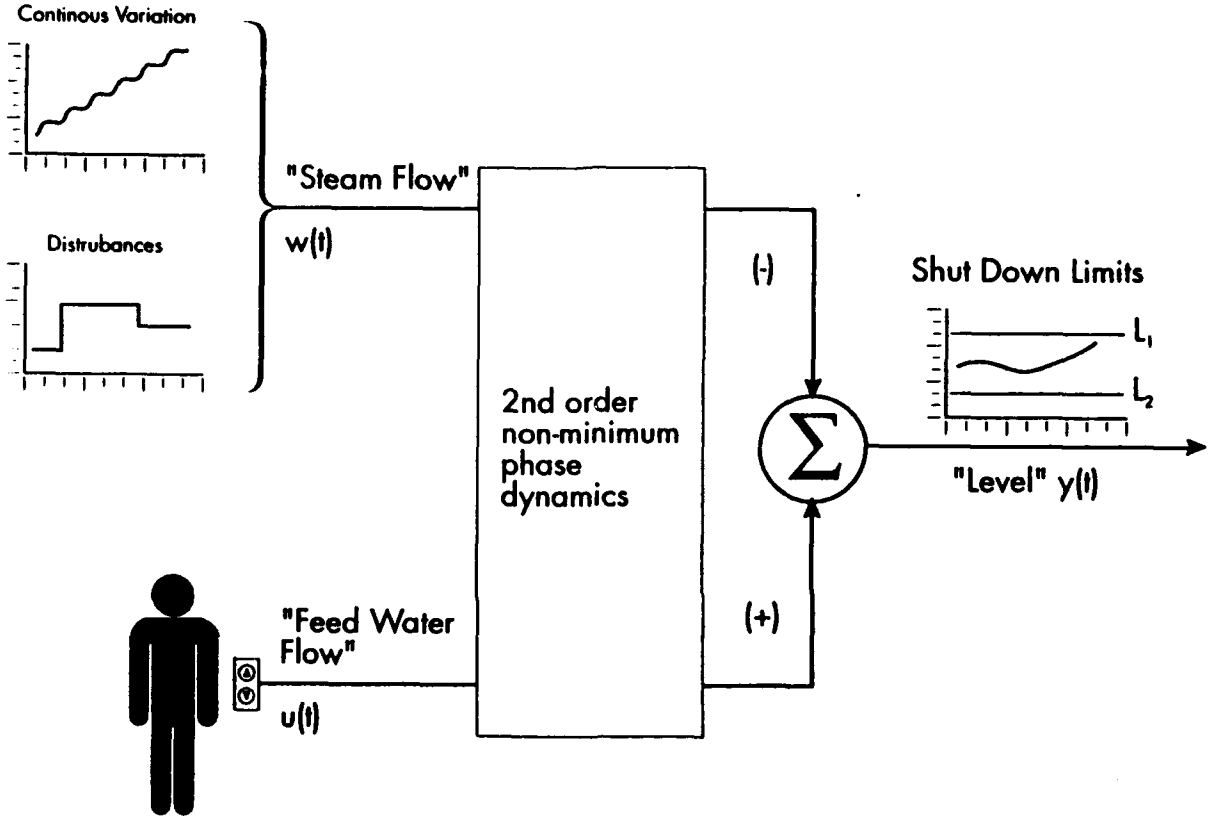
Subjects. Twenty students (10 male and 10 female) participated in the experiment and were paid $5.00 an hour. The subjects ages ranged from 19 to 35 years of age and all subjects had normal or normal-corrected vision with no color-blindness deficiencies. The subjects were randomly assigned to one of the two display conditions.

Apparatus. All experimental events were controlled by a general purpose laboratory computer (Sun Microsystem 4-110 Workstation). Subjects were seated in an enclosed experimental room. A 16" color video monitor with a resolution of 1152 by 900 pixels was used to present experimental prompts and a standard keyboard was used to record user responses.

Simulation model. The subject's task is to control a second-order non-minimum phase dynamical system (see Figure 5).

The simulated system possesses the same basic dynamical characteristics as a single nuclear power plant steam generator, e.g., time delays, shrink/swell behavior. These differential equations incorporate the influence of a number of factors on the indicated steam generator level, including steam flow, feedwater flow, rate of power, and temperature of the feedwater. Another set of differential equations reconstructs an estimate of plant states (using the "observer" method) which is used to generate compensated SGL.

FIGURE 5. The MCF simulation and control task.



System dynamics. The reactor power level and the feedwater temperature remained fixed throughout the experiment. Programmed changes to steam flow were introduced to produce the primary challenges that the subjects had to respond to. There were two types

of changes: 1) continuous variation, and 2) disturbances (see Figure 5). The continuous variations were constant changes to steam flow resulting from the combination of three sine waves and a ramp. The net result was to produce a trial in which the steam flow rate was either oscillating, oscillating with a gradual rise, or oscillating with a gradual fall (a rising ramp is illustrated in Figure 5). The second type of changes to steam flow were a result of random changes, or disturbances, to steam flow. The number and size of these disturbances varied as a function of the elapsed time of an experimental trial. During the first 30 seconds of an experimental trial one randomly-timed disturbance was introduced. For each thirty second increment in time that followed, the number of disturbances was increased by 2. The direction of each disturbance was either positive or negative, and was also randomly determined. The size of the disturbance was randomly selected within a time-dependent range: disturbances within the first 30 second time interval ranged between 0 and 1% and for each additional 30 seconds the range was increased by .5%. The effect of the programmed changes to steam flow was to produce an experimental trial that constantly changed in a relatively unpredictable fashion, and that became progressively more difficult as time-on-task increased.

Stimuli. Two different graphic displays were used to present the ISGL, CSGL, SF, and FF (see Figures 1, 2, 3, and 4). The background matte for each graph was 11.5 cm high and 14 cm wide, was colored a light gray, and the lines and letters were colored black. Both displays were updated with information from the simulation model every 2 seconds and random noise ranging from -2% to +2% was added to the value of each variable displayed (these changes were not permanently added to the mathematical model, only displayed to the subject). For the separate display each variable was presented as an individual bar with a common baseline in the x-axis. Each bar was color-coded: blue for ISGL, white for SF, green for FF, and yellow for CSGL. Each bar was .6 cm wide with a maximum height of 5.5 cm and equally spaced in the x-axis. Assuming a subject seating distance of 50 cm, the bar chart display subtended a visual angle of 4.9 degrees horizontally and a maximum of 6.3 degrees vertically. The y-axis was labelled using a scale of 0 to 100% and black horizontal grid lines extended the length of the x-axis and were placed at 10% intervals.

The trip set points were placed at 20% and 80% as horizontal red lines that began on the left side of the display extended half the length of the *x*-axis.

Whenever possible, the configural display used the same sizing, scaling, and coloring conventions as the separate display. In the configural display the *x*-axis was also scaled and labelled and the trip set points were extended to cover the length of the *x*-axis. The configural display maps the four critical variables into a single geometric object: a rectangle. The difference between indicated and compensated SGL was mapped on the *y*-axis and the difference between steam flow and feedwater flow is mapped on the *x*-axis. Where the values for the four variables intersect in the display grid a black rectangular shape was drawn and the sides of the rectangle were color-coded (using the same colors as the separate display) to reflect the contribution of an individual variable. The maximum height of the rectangular object was 4.5 cm while the maximum width was 5.5 cm. Assuming a subject seating distance of 50 cm, the configural display subtended a maximum visual angle of 6.3 degrees horizontally and a maximum of 5.1 degrees vertically.

Procedure. The experiment was conducted during a one-week period with one experimental session per day (lasting one hour), for a total of five sessions. The subjects were individually tested in an enclosed room. During the first experimental session the subjects were provided with both a written and a verbal explanation of the task and a verbal description of their respective display. The experimenter remained in the room during this session to answer any general questions about the task, but did not provide any information regarding particular control strategies. During each experimental session the subject completed at least 12 trials, a factorial combination of 3 steam flow ramp types (rising, null, falling) and 4 indicated steam generator level starting positions (35, 45, 55, and 65%). The order of these trials was randomly determined. To ensure equal training time each subject was required to complete additional trials until an hour had expired (a randomly determined subset of the original 12 trials). Since the dependent measure was time-on-task this manipulation provides an additional control for experience. The additional trials were not considered in the data analyses.

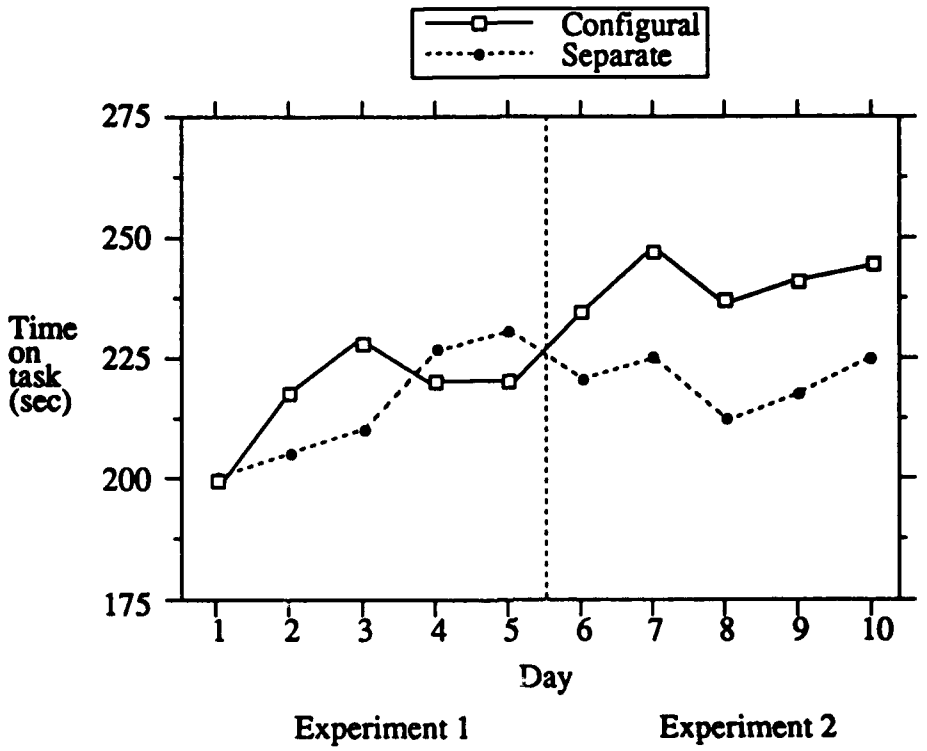
During each experimental trial the rate of feedwater flow was under the control of the subject, whose task was to adjust this rate (by pressing one of two keys on the keyboard: an up arrow to increase feedwater flow, a down arrow to decrease feedwater flow) in an attempt to maintain the indicated steam generator level between the upper and lower trip set points (see Figure 5). The subject initiated a trial by pressing a designated key. Each trial could last up to 5 minutes and ended either when 5 minutes had elapsed or when the indicated steam generator level surpassed one of the two trip set points. The subject was provided with feedback concerning time-on-task after each trial.

In summary, the experimental design contained 4 independent variables: display (separate vs. configural, a between-subjects variable), day (1 through 5, a within-subjects variable), starting position of ISGL (35, 45, 55, and 65%, a within-subjects variable) and steam flow ramp type (rising, null, falling, a within-subjects variable). The primary dependent variable was time-on-task, which was recorded at 1/100 second accuracy.

Results

A 2 x 5 x 4 x 3 mixed ANOVA was performed on the regular TOT scores. The main effects of day $F(4,72) = 2.64$, $p < 0.04$, starting position $F(3,54) = 12.74$, $p < .0001$, and ramp $F(2,36) = 6.60$, $p < .004$, were significant, as well as the interaction between ramp and starting position $F(6,108) = 19.70$, $p < .0001$. The main effects can be summarized by stating that performance improved with experience at the task, the rising and falling ramps were more difficult than the null ramp, and that performance was degraded when the ISGL starting position was closer to the trip set points. The interaction effect indicates that performance was particularly poor when 1) the ISGL starting position was low and the steam generator ramp was rising or 2) when the ISGL starting position was high and the ramp was falling. All other effects, including the predicted main effect of display, $F(1,18) = .03$, $p < .85$, were not significant. For comparison purposes the means for the display by task by day interaction effect are illustrated in Figure 6.

FIGURE 6. Means for time-on-task in Experiments 1 and 2.



Discussion

Experiment 1 compared performance in a complex dynamic task when information was presented with a separate and a configural graphic display format. Contrary to experimental predictions, the configural display did not facilitate performance at the manual control of feedwater task. One possible explanation is that performance of the task requires not only the integration of information, but a high degree of focused attention on individual variables. Coury et al. (1989) found that when performance of an integration task required attention to be focused on individual variables (due to a high degree of uncertainty about system state), the benefits of configural displays disappeared. In the MCF task the indicated steam generator level is clearly the most critical performance variable, since its value determines whether or not the plant trips, and this may have

required an inordinate amount of focused attention on individual variables. A second possible explanation is that the emergent features produced by the separate display (the inferred linearity between the heights of bars) are equally as effective in facilitating performance at the task as the emergent features produced by the configural display (height, width, area, shape, location, movement of the rectangle).

An alternative explanation is that the predicted performance differences do exist, but that the time-on-task variable was not sensitive enough to capture them. Previous research on configural and separate displays has typically analyzed performance on simple, usually static, experimental tasks such as sorting (Coury et al., 1989) or detection (Barnett and Wickens, 1988; Sanderson et al., 1989) where the dependent measures are reaction time and accuracy. In contrast, the MCF task is a complex, dynamic task where performance is inherently variable. For example, a subject who changes an upward trend in ISGL and narrowly avoids the upper trip set point could add, regardless of input, an additional one to two minutes of time-on-task because of the system dynamics. Also, because it is a complex task operators can adopt a variety of high-level strategies that vary in effectiveness (Roth and Woods, 1988). The measure of time-on-task has a high degree of face validity (it is the only performance criterion that really counts in the real world), yet it may not be sufficiently fine-grained for display evaluation. The difficulty of assessing performance differences in MCF task is a reflection of those that are encountered in the real world: the more complex and dynamic that a task is, the more inherent variability is increased, experimental control is sacrificed and the ability to evaluate display design is reduced. Yet it is from the study of these complex and dynamic settings that we stand to profit the most.

One possible methodological alternative is the "memory probe" technique, where a subject performing one task is occasionally (and therefore somewhat unexpectedly) required to perform a different task. The probe technique has been used by Wickens and his colleagues (Andre and Wickens, 1988; Wickens and Andre, 1988, 1990; Wickens, Kramer, Barnett, Carswell, Fracker, Goettl, and Harwood, 1985 - Experiments 1 and 4) to assess performance on tasks that require selective or focused attention to individual

variables. In a typical scenario the subject was required to perform an integration task on the majority of experimental trials, but on a small percentage of trials (e.g., 20%) was asked to perform a focused task instead (e.g., remember the value of an individual variable). The second experiment explored the possibility that the memory probe technique could be used to provide sensitive behavioral measures for display evaluation in complex settings. To be useful the technique would have to be extended so that measures of performance for integration tasks, as well as focused tasks might be obtained. In Experiment 2 the subject's primary task was to perform the MCF task. However, the primary task was occasionally interrupted by the memory probe task: the screen was blanked and subjects were asked to recall information about the state of the system.

Increasing the perceptual salience of low-level data

The second issue to be investigated in Experiment 2 concerns the relative accessibility of low-level data in configural displays. In contrast to the findings for integration tasks, the literature has revealed a rather consistent pattern of results indicating that in tasks where performance depends upon the consideration of an individual variable (focused attention) or when irrelevant variables need to be ignored (divided attention) the display of data in a configural format hinders performance (Carswell and Wickens, 1987; Casey and Wickens, 1986; Coury et al., 1986; Goettl, Kramer and Wickens, 1986; Wickens and Andre, 1988). Since information about individual variables is often needed to complete domain tasks (e.g., the critical performance variable ISGL) this cost needs to be eliminated or minimized.

One explanation of the cost of configural displays for low-level data is that perception of the individual elements that configure to form the overall object are secondary to perception of the object itself. That is, as a function of being part of an object the individual parts are somehow "lost" and less accessible. There are several lines of research that do not support this interpretation. The "object file" theory of attention and

object perception (Kahneman and Treisman, 1984) is one theoretical perspective. In this theory the first stage of object perception is the automatic and parallel extraction of information from a small number of separate perceptual analyzers such as color, shape, and motion (Treisman and Gelade, 1980). This is followed by the recombination of this spatially-located information into objects. This information (the features that pertain to an object) are stored in a temporary object file that is dynamically updated (Kahneman and Henik, 1977; Kahneman and Treisman, 1984). In the object file theory of attention competition for attentional resources occurs between objects (they must be processed serially) but not within objects (once attention has been allocated to an object it is easy to focus on the features that define an object, an effect similar to semantic priming). Thus, from the perspective of the object file theory information about the graphical elements of an object display should be processed automatically and in parallel.

The second line of evidence evolves from Pomerantz's theory of form perception (Pomerantz, 1986). The notion of individual elements configuring to produce higher-level emergent features (e.g., closure, symmetry) and the role for high-level properties has already been discussed. With respect to the graphical elements (the low-level data) Pomerantz argues that there is no evidence of "perceptual glue" that binds individual elements into a whole, therefore making the individual parts less accessible. He states that ..."the groupings of parts into wholes does not make parts imperceptible or inaccessible; the evidence for a perceptual glue binding parts together is thin. Rather, when parts configure into wholes, new emergent features arise that are available alongside the parts but that are more salient perceptually and will therefore be attended to instead of parts when it is advantageous to do so" (Pomerantz, 1986, p. 28). From this perspective, the low-level data that configures to produce emergent features is still available, but less distinctive.

Pomerantz's theory suggests that one way to improve the availability of low-level data is to make the graphical elements more salient perceptually. This might be accomplished through a variety of methods (for example, maintaining or emphasizing the scale, separating the contributing elements spatially, color-coding the contributing elements). At the same time, it is important not to destroy the emergent features that are

relevant to higher-level issues. In this case, we chose to use color-coding to accomplish these dual goals. Color-coding the representations of the individual parameters (the low-level graphical elements) makes the contributions of individual variables more salient perceptually, and therefore more accessible. In addition, it has been shown that color and shape do not configure to produce emergent features (Carswell, 1988; Treisman and Gelade, 1980). Thus, color-coding the graphical elements is unlikely to destroy the emergent features and, therefore, is unlikely to disrupt performance on integration tasks.

Color-coding with spatially separated graphical elements

Wickens and Andre (1990) investigated these issues with a configural display that combined three variables. The graphical elements were presented as connected lines with each successive element alternating in the x and y dimensions of the display frame. These lines were anchored in the lower left-hand portion of the display frame. The low-level elements configured to produce an imaginary rectangle that was anchored in the upper-left corner of the display frame and spatially separated from the graphical elements that contributed to its form. In comparing a monochrome configural display and a version of the same display in which the elements were color-coded Wickens and Andre (1990) found that color-coding produced a speed-accuracy trade-off: color-coding the graphical elements facilitated accuracy at the focused task, and did not disrupt accuracy at the integration task. However, the improvement in accuracy performance was associated with a decrement in latency performance: the chromatic version of the configural display increased the latencies associated with both types of tasks. Direct comparisons between performance between the monochrome separate display and the chromatic configural display were not made, perhaps because of the confounds between chromaticity and objectness.

If our goal is to design configural displays that facilitate the extraction of information regarding both high-level properties and low-level data these findings are

encouraging, but not definitive. In their configural display the emergent features resulted from an imagined rectangle, and the graphical elements that determined its size and shape were spatially distinct from the rectangle itself. An important question to ask is what might be expected if the graphical elements were incorporated as part of the object? On the one hand, color-coding the graphical elements of a clearly defined object might enhance the within-object processing (consistent with the object file theory), and eliminate the speed-accuracy trade-off that Wickens and Andre observed. On the other hand, spatially separating the graphical elements may have contributed to the improved ability to extract low-level information. Viewed from Pomerantz's theory of form perception, spatially separating the graphical elements increases their perceptual salience and therefore the availability of low-level information. The observed facilitation for the processing of low-level data may have been due to the combined effect of both color-coding and spatial separation, rather than color-coding alone. This issue was investigated in Experiment 2.

In Experiment 2 performance with the separate and configural displays was compared for two types of memory probe tasks. In the focused memory probe task individuals were asked to recall the value of one of the four variables relevant to the MCF task (SF, FF, ISGL, or CSGL). Differences in performance at this task will indicate whether or not color-coding the individual contributions will enhance the perceptual salience of the low-level data, and therefore offset the performance costs usually associated with configural displays. In the integration memory probe task the subjects were probed on critical differences between variables. The choice of which differences to probe for was determined by the analyses of the domain semantics: the mass and energy balances play a critical role in the evaluation of current system state and future control inputs. Thus, the integrated memory probes tested for memory of differences between SF/FF (mass balance) and for differences between ISGL/CSGL (energy balance). The predictions for the integration probe task remain the same as in Experiment 1: performance should be facilitated with the configural display since the number and quality of the emergent features that it produces is superior to those produced by the separate display. It is possible that the effectiveness of the emergent features produced by the two displays might vary as the

relationships between the critical variables change. To check for this possibility, an additional variable, plant "state" was included. The memory probes were administered in one of two situations: when the difference between SF and FF or ISGL and CSGL was 1) small (less than 5%) or 2) large (greater than 15%). The dependent measure of time-on-task was also recorded, as in Experiment 1.

Experiment 2

Method

The subjects, apparatus, simulation model, system dynamics, and stimuli were identical to Experiment 1.

Procedure. The procedure was essentially the same as Experiment 1, with the only changes being those necessary to implement the memory probes. All subjects were tested on the five consecutive days immediately following Experiment 1. On the first day the subjects were informed of the memory probes and were instructed to respond as quickly and as accurately as possible. The subjects performed the MCF task, as in Experiment 1, and were interrupted to complete memory probes. During a memory probe the simulation was stopped, the screen was blanked, a probe was presented, and the subjects responded by entering a numeric value via the keyboard. The simulation was then restarted in the previously existing system state, including the appropriate adjustments for simulation time constants. Response time was measured from the time that the screen was blanked until the first digit of the subject's response, and was measured with 1/100 second accuracy. Accuracy (error magnitude) was measured by computing the absolute value of the difference between the subject's estimate and the actual value (as it appeared on the subject's screen, that is, the value of simulation variable with the random noise for that particular screen update). Feedback on accuracy was provided before the simulation was restarted.

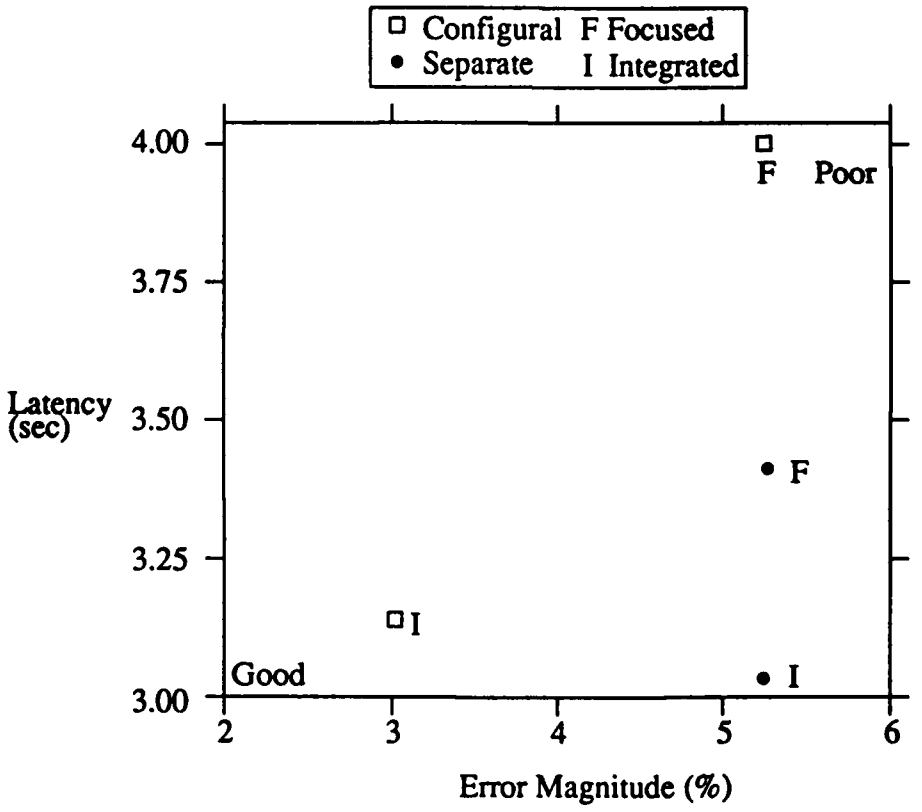
There were two different types of memory probe tasks: a focused task or an integration task. In the focused task subjects were asked to estimate either 1) the rate of steam flow, 2) the rate of feedwater flow, 3) the indicated steam generator level, or 4) the compensated steam generator level. In the integration task the subjects were asked to estimate either 1) the difference between steam and feedwater flow, or 2) the difference between indicated and compensated steam generator level. Memory probes could also occur during two different system states: 1) when the differences between SF/FF or ISGL/ CSGL were large (greater than 15%) and 2) when these differences were small (less than 5%). Approximately 128 probes were obtained in an experimental session (an algorithm was developed to ensure that approximately equal numbers and distributions of probes were obtained in an experimental session). The 64 probes for the focused task consisted of eight probes for each of four categories (SF, FF, ISGL, and CSGL) in both states (large and small differences). For the integration task the 64 probes consisted of 16 probes for each of two categories (differences between SF and FF, and differences between ISGL and CSGL) in both states (large and small differences). In summary, the experimental design contained 4 independent variables: display (separate vs. configural, a between-subjects variable), day (1 through 5, a within-subjects variable), task (focused vs. integration, a within-subjects variable) and state (large vs. small differences, a within-subjects variable).

Results

Accuracy. All memory probe scores in an experimental session were averaged across probe categories (Task x State) for a total of 4 scores for each individual. A 2 x 5 x 2 x 2 mixed ANOVA was performed on these data. The main effect of display was significant, $F(1,18) = 4.40$, $p < .05$, indicating that performance with the configural display (average error magnitude = 4.13 %) was more accurate than performance with the separate display (5.21 %). The main effect for task was significant, $F(1,18) = 18.75$, $p < .0007$, indicating that performance in the integration task (4.12 %) was significantly better than

performance in the focused task (5.22 %). A significant display by task interaction, $F(1,18) = 16.41$, $p < .002$, indicated that the configural display facilitated performance in the integration task (see Figure 7). F-tests for simple effects indicated that the differences between displays was not significant for the focused task, $F(1,18) = .01$, but highly significant for the integration task, $F(1,18) = 34.33$, $p < .00007$. As Figure 7 illustrates, the configural display facilitated the accuracy of performance for the integration memory task.

FIGURE 7. Latency and accuracy means for the display by task interaction.



The main effect for state, $F(1,18) = 105.65$, $p < .000002$, indicated that performance was significantly better when the differences between flows or levels were small (3.62 %) than when they were large (5.71 %). A significant state by task interaction, $F(1,18) = 27.85$, $p < .0002$, indicated that the accuracy of performance on the two types of task was influenced by the size of the differences between variables. F-tests for simple effects

indicated that performance in the focused task was significantly better for small differences than for large differences, $F(1,18) = 10.95$, $p < .005$, and that these differences were in the same direction and more pronounced in the integration task, $F(1,18) = 116.70$, $p < .000001$. All other effects were not significant.

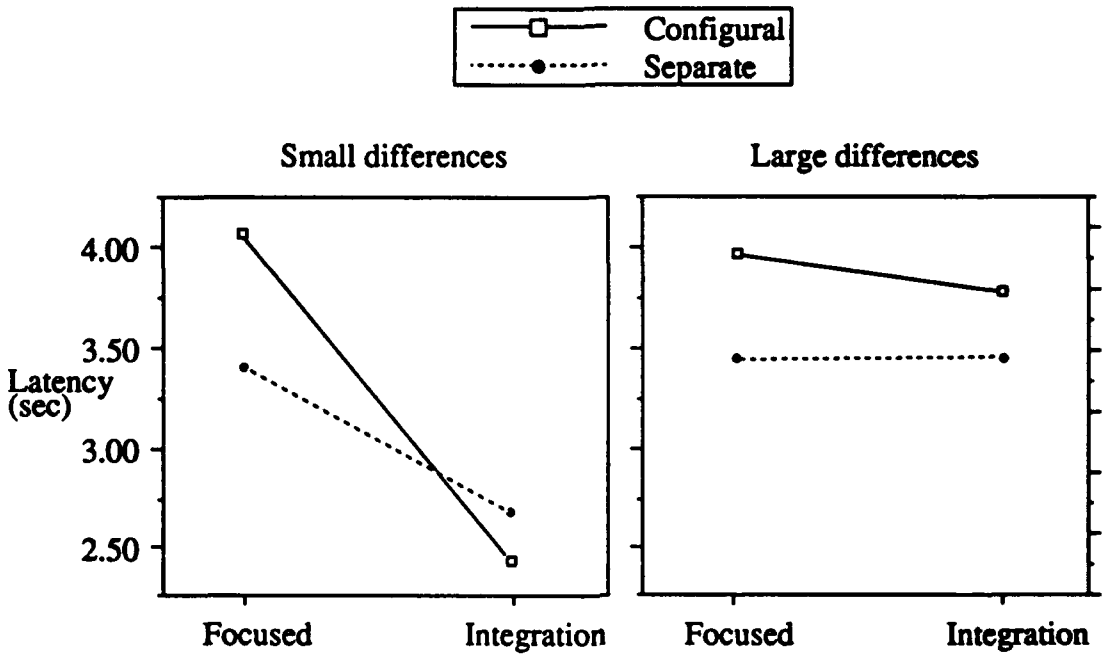
Latency. All memory probe scores for an experimental session were averaged across probe categories for a total of 4 scores (Task x State) for each individual. A 2 x 5 x 2 x 2 mixed ANOVA was performed on these data. The display by task interaction was significant $F(1,18) = 13.25$, $p < .003$ (see Figure 7). F-tests for simple effects indicated that performance at the focused task was significantly slower with the configural display (4.03) than with the separable display (3.43), $F(1,18) = 30.21$, $p < .0002$, and that there were no significant differences between displays for the integration task, $F(1,18) = .13$. The display by task by state interaction effect was significant, $F(1,18) = 4.51$, $p < .05$ indicating that the interaction between display and task was dependent upon the size of the differences (see Figure 8).

F-tests for simple effects indicated that the interaction between display and task was significant when the differences between variables was small, $F(1,18) = 63.47$, $p < .000008$, and marginally significant when the differences between variables was large, $F(1,18) = 4.24$, $p < .052$. Further F-tests revealed that for small differences between variables the configural display facilitated performance at the integration task, $F(1,18) = 7.76$, $p < .02$, while the separable display facilitated performance at the focused task, $F(1,18) = 71.94$, $p < .000005$.

The four-way interaction, display by task by state by day interaction, was significant, $F(4,72) = 2.54$, $p < .05$, indicating that the nature of the interaction between display and task was not only dependent upon the size of the differences between variables, but also changed over time (see Figure 9). Separate F-tests for the overall simple effects (display, task, and display by task) were computed in each of the 10 day by state conditions and the results of these comparisons are presented in Table 1. Additional F-tests were computed to compare the means for the two displays and the two tasks at each of these 10

conditions. The results of these comparisons are presented in Table 2.

FIGURE 8. Latency means for the display by task by state interaction.



The main effect of day was significant, $F(4,72) = 42.55$, $p < .000001$, indicating that performance improved across experimental sessions (4.22, 3.54, 3.32, 3.00, and 2.95 sec). The display by day interaction effect was significant, $F(4,72) = 3.32$, $p < .02$ (see Figure 10). F-tests for simple effects indicated that although performance with the configural display was significantly worse than performance with the separate display during the first, $F(1,72) = 24.85$, $p < .00005$, and the second, $F(1,72) = 5.75$, $p < .02$, experimental sessions the third, $F(1,72) = 2.69$, fourth, $F(1,72) = .40$, and fifth, $F(1,72) = .1951$, experimental sessions were not significantly different.

FIGURE 9. Latency means for the display by task by state by day interaction effect.

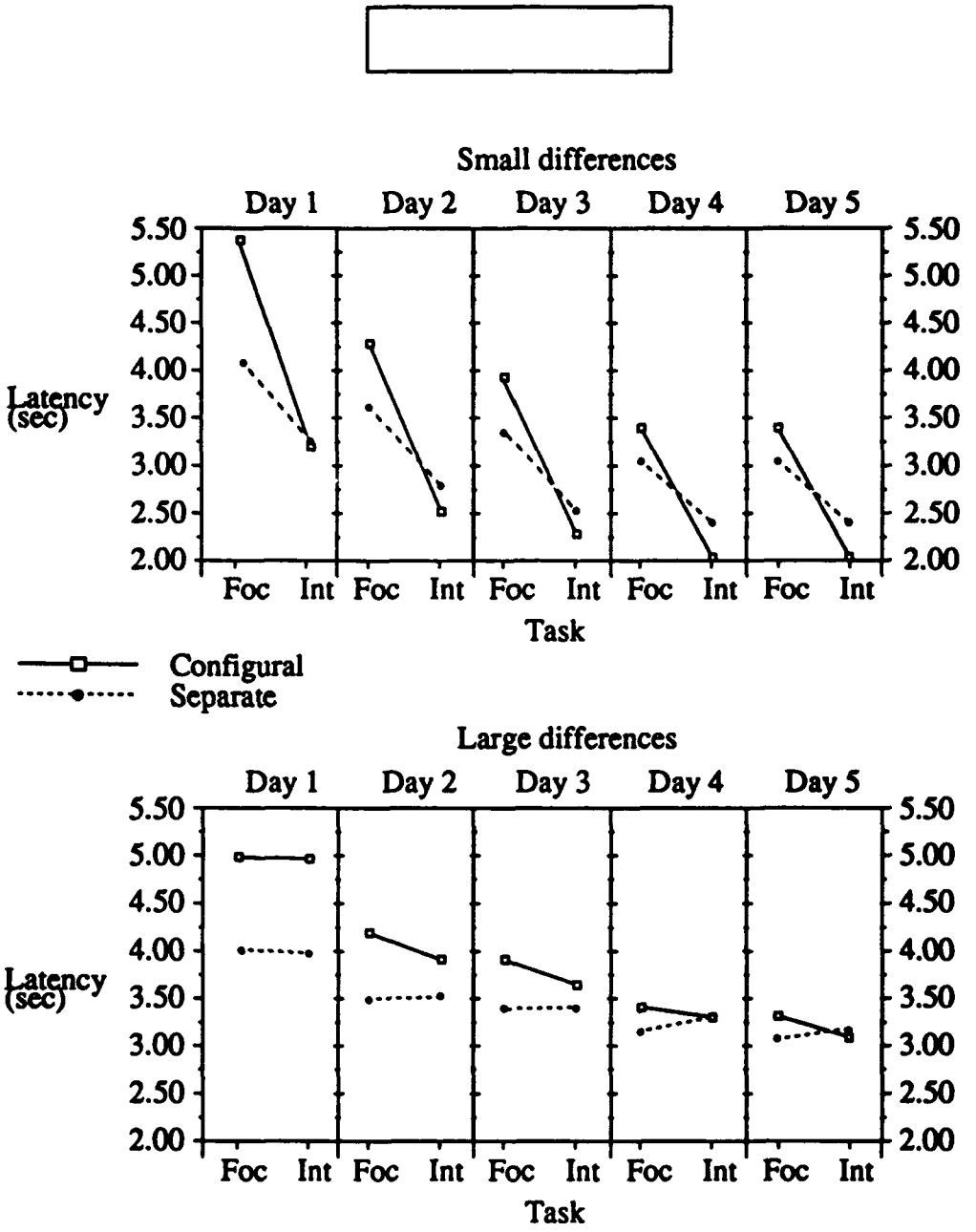


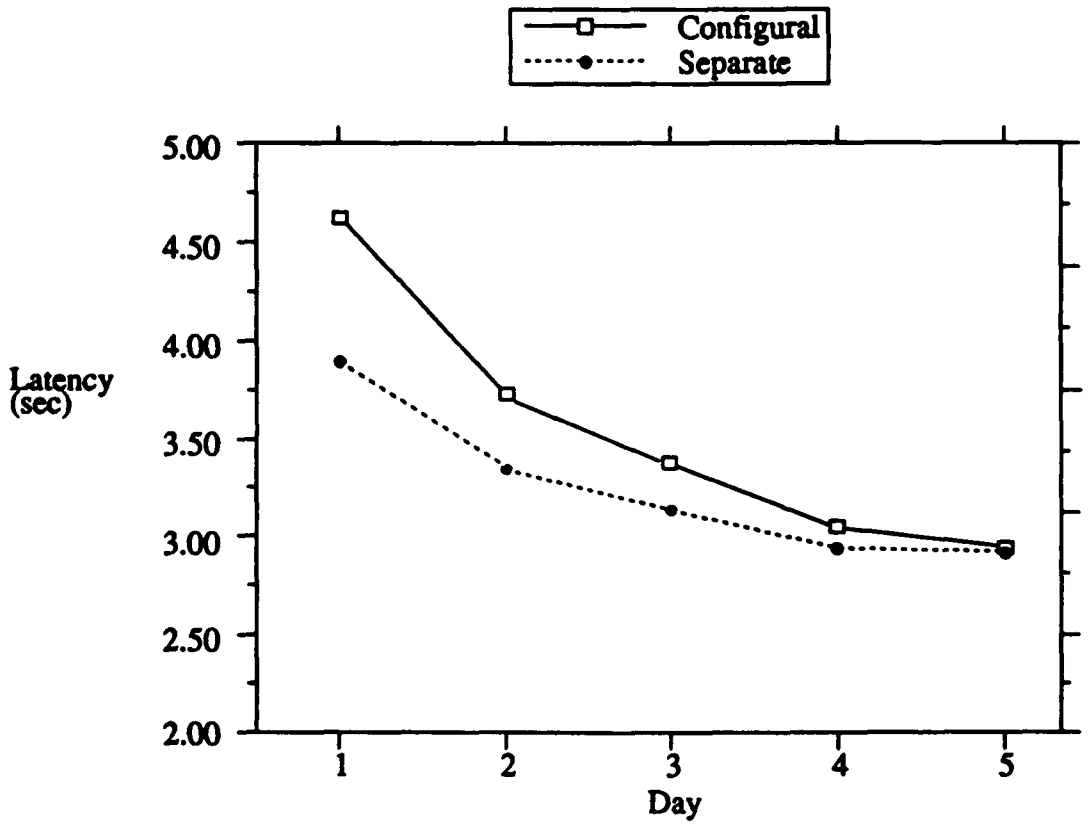
TABLE 1. Simple effects for display by task by state by day latency interaction effect.

		Display		Task		Display x Task	
State	Day	F	p	F	p	F	p
Small	1	41.87	<.000003	234.19	<.000001	47.35	<.000002
Small	2	4.88	<.03	176.90	<.000001	22.81	<.00007
Small	3	3.81	<.06	164.82	<.000001	19.50	<.0002
Small	4	.17	<.70	103.39	<.000001	13.67	<.0008
Small	5	.26	<.63	105.49	<.000001	10.80	<.002
Large	1	92.22	<.000001	1.39	<.25	.32	<.59
Large	2	31.33	<.00002	1.65	<.21	2.87	<.10
Large	3	12.19	<.002	1.05	<.31	3.81	<.06
Large	4	2.70	<.11	.09	<.78	1.52	<.22
Large	5	.95	<.68	.66	<.59	2.22	<.14

TABLE 2. Simple effects (latency) for configural vs. separate displays.

		Day 1		Day 2		Day 3		Day 4		Day 5	
State	Task	F	p <	F	p <	F	p <	F	p <	F	p<
Small	Int	.01	.907	3.30	.070	3.04	.082	5.40	.022	3.85	.051
	Foc	89.14	.0000	24.39	.0001	20.27	.0002	8.44	.006	7.22	.009
Large	Int	51.69	.0000	7.62	.007	1.19	.281	.08	.782	.13	.732
	Foc	40.84	.0000	26.59	.0001	14.82	.0005	4.14	.044	3.03	.083

FIGURE 10. Latency means for display by day interaction effect.



There was also a significant main effect of state, $F(1,18) = 60.46$, $p < .000009$, indicating that performance was significantly faster when the differences between variables were small (3.15 sec), as opposed to large (3.66 sec). The main effect of task was significant, $F(1,18) = 69.31$, $p < .000006$, indicating that performance was significantly faster in the integration task (3.09 sec) than in the focused task (3.73 sec). The day by task interaction was significant, $F(4,72) = 3.89$, $p < .007$, indicating that performance in the focused task was initially worse than performance in the integration task, but improved at a faster rate. The state by task interaction effect was significant, $F(1,18) = 50.72$, $p <$ .00002. F-tests for simple effects indicated that performance for the focused task was not significantly different when large or small differences existed between variables, $F(1,18) =$

.29, but that performance for the integration task was facilitated when small differences existed, $F(1,18) = 91.74$, $p < .000002$. All other effects were not significant.

Time-on-task. As in Experiment 1, time-on-task was recorded for each experimental trial. The experimental design contained 4 independent variables: display (separate vs. configural, a between-subjects variable), day (1 through 5, a within-subjects variable), starting position of ISGL (35, 45, 55, and 65%, a within-subjects variable) and steam flow ramp type (rising, null, and falling, a within-subjects variable). A 2 x 5 x 4 x 3 mixed ANOVA was performed on the regular TOT scores. The main effects of starting position $F(3,54) = 7.63$, $p < .0002$, and ramp $F(2,36) = 34.20$, $p < .0001$, were significant, as well as the interactions between ramp and starting position $F(6,108) = 16.66$, $p < .0001$ and between day by starting position $F(12,216) = 2.11$, $p < .02$. The main effects can be summarized by stating that the rising and falling ramps were more difficult than the null ramp, and that performance was degraded when the ISGL starting position was closer to the trip set points. The ramp by starting position interaction effect indicates that performance was particularly poor when 1) the ISGL starting position was low and the steam generator ramp was rising or 2) when the ISGL starting position was high and the ramp was falling. The day by starting position interaction effect indicates that with experience on the task performance differences due to ISGL starting position disappeared. All other effects, including the predicted main effect of display, $F(1,18) = 1.33$, $p < 0.26$, were not significant.

Discussion

Previous research has yielded mixed results concerning the effectiveness of configural displays in improving performance at tasks that require the integration of information. The integration probe task in Experiment 2 measured memory for the difference between two sets of variables: 1) steam flow and feedwater flow and 2) indicated and compensated steam generator level. These mass and energy balances are an important

part of the MCF task: they must be assessed to determine the current state of the system, the future state of the ISGL, and therefore, the appropriate control input (Roth and Woods, 1988). The results of the integration probe task revealed that accuracy performance was significantly better with the configural display than with the separate display (see Figure 7). This advantage was not obtained at a cost in latency. In fact, under some experimental conditions integration probe performance was also significantly faster. When the differences between SF/FF or ISGL/CSGL were small, responses to the integration probe were significantly faster with the configural display (see Figure 8) and this advantage remained relatively stable with additional experience at the task (see Figure 9 and Table 2). When the differences between SF/FF or ISGL/CSGL were large, the latency of responses to the integration probe were initially slower with the configural display, and these differences were quite pronounced. However, with additional experience at the task this cost disappeared (see Figure 9 and Table 1).

Thus, the emergent features that were provided by the configural display were superior to those provided by the separate display. With the latter display the balance information was mapped into the emergent features of the inferred linear relationship between the heights of bars in a bar chart. With the configural display the critical relationships were mapped directly into the height and the width of the rectangle. This direct physical representation made the user's task of extracting the relevant information from the configural display easier than extracting it from the separate display.

An interesting, and somewhat surprising result, is that the separate display supported performance more effectively in the integration probe task than in the focused task. When the differences between the variables to be integrated were small, the latency of responses for the separate display were significantly faster for the integration probe task (see Figures 8 and 9 and Table 1). For the accuracy of responses there was no difference in performance (see Figure 7). Although the separate display was less effective than the configural display, both displays produced emergent features that supported integration task performance, relative to focused task performance. These results compliment the results of Sanderson, et al.(1989), and emphasize the need to re-conceptualize the separate-

configural dichotomy as a continuum. Most graphic displays that provide information about more than one variable are likely to produce emergent features, although the number and quality will vary with the graphic form. Configural displays (especially object displays) are likely to produce a larger number, and more salient emergent features than a bar graph. Similarly, a bar graph is likely to produce a larger number, and more salient emergent features than a digital display of information.

Color-coding of graphical elements to improve salience of low-level data

Previous research has yielded a fairly consistent pattern of results indicating that performance is degraded when the experimental task requires focused or divided attention and the information is presented in a configural format (Carswell and Wickens, 1987; Casey and Wickens, 1986; Coury et. al., 1986; Goettl, Kramer and Wickens, 1986; Wickens and Andre, 1988). The present experiment tested the hypothesis that color-coding the graphical elements of a configural display could offset this cost. In the focused probe task, where individuals were tested for their memory of individual variables, there was no significant difference in the accuracy of responses between the two graphic displays (see Figure 7). However, this level of accuracy appears to have been achieved at a cost for the configural display, as indicated by the significantly higher latency scores (see Figures 7 and 8).

This processing cost was dependent upon both system state and experience at the task. There was a general trend for the cost of the configural presentation of low-level data to be quite large initially and to dissipate with additional experience at the task. When there were large differences between the variables to be integrated the cost was highly significant initially ($p < .000004$), but became progressively smaller with experience (see Figure 9 and Table 1). By the end of the experiment this cost was only marginally significant ($p < .083$). The same general trend was evident when the difference between variables to be integrated were small. There was a highly significant initial cost in latency for the configural display

($p < .00000007$), and, as in the previous case, this cost became progressively smaller during the course of the experiment. However, this cost was still significant at the end of Day 5 ($p < .009$).

Wickens and Andre (1990) investigated closely-related issues using different displays and methodological procedures. Two of the displays in their study were a monochrome bar graph and a color-coded configural display, and they gathered performance measures for both a focused and an integrated memory probe task. Performance at the focused task was less accurate and required longer response times for the configural display than for the bar graph display (only statements concerning the direction of these performance differences can be made, since direct statistical comparisons were not computed). It is important to note that since the bar graph was monochrome and the configural display was color-coded, performance should have been biased in favor of the configural display.

Considered together, these results do not provide strong evidence that color-coding graphical elements can alleviate the costs that are associated with the configural display of low-level data. On the other hand, there is not sufficient evidence to rule out the possibility. In their study Wickens and Andre (1990) also compared monochromatic and chromatic versions of a configural display and found that color-coding the graphical elements resulted in a significant increase in response accuracy at the cost of a marginally significant increase in response latency. In fact, the results of the present study could be interpreted in a more positive fashion. First, there was no cost with respect to the accuracy of responses. Second, although there were large initial costs for response latency, these costs tended to dissipate with experience at the task and were only marginally significant in one of two conditions by the end of the experiment.

The utility of the memory probe methodology

A final note on the results of Experiment 2 concerns the utility of the probe

methodology. For the manual control of feedwater task the dependent measure that has the most face validity is time-on-task: in the real world the only criterion that matters is whether or not the setpoint boundaries are crossed. Unfortunately, the results of Experiment 2 indicate that the TOT measure was not effective for evaluating display design. The results show apparently clear and consistent differences indicating that performance was better for the configural display than for the separate display (an average difference of nearly 22 seconds, see Figure 6), yet these results were not statistically significant. These evaluation problems are a reflection of the difficulties that are encountered when attempting to evaluate display designs in complex, dynamic domains: performance is inherently variable due to the nature of the tasks and the wide range of strategies that are available for their completion. In contrast to the TOT methodology, the memory probe technique provides a very fine-grained analysis of performance. An additional benefit, as the results of Experiment 2 indicate, is that the methodology can be extended to evaluate both the information processing benefits and costs associated with configural displays. Thus, the memory probe technique can be used in the evaluation of configural display design in complex, dynamic settings. The utility of the technique seems to be limited only by the determination of appropriate integration probes. As in the case of configural display design, the appropriate question to ask will be determined by the cognitive task analysis and an analysis of the domain semantics.

General Discussion

The graphic display of information is a form of decision support that provides the very real potential to improve the overall performance of human-machine systems in complex, dynamic task domains. This vast potential arises, in part, because of continuing advances in hardware and software technology. Yet, a better understanding of how these technological capabilities can be used effectively is needed. Previous research has discounted the role that the semantics of the domain must assume in the design of effective

configural displays for complex, dynamic domains. A perspective is needed that considers the cognitive triad: the cognitive demands of the domain, the cognitive resources of the agents meeting these demands, and the interface (graphic displays) as a representation of the domain that supports demand/resource mismatches (Woods and Roth, 1988). A similar approach from a different theoretical orientation is the ecological approach to interface design (Flach and Vicente, in press; Vicente and Rasmussen, 1990).

To accomplish tasks in complex domains the operator must consider the system from different levels of abstraction, and alternate between these levels (Rasmussen, 1986). Thus, information about high-level properties that are defined by the relationships between variables needs to be available, as well as the low-level information about individual variables that contribute to them. One of the implications is that there is a need for either multiple graphic displays that portray separate levels of abstraction, or individual displays that provide access to information from multiple levels. For the design of individual configural displays the critical issue is to provide a representation that allows the easy extraction of information with respect to both high-level properties and low-level data. Principles for achieving these dual goals will be discussed in greater detail.

Designing configural displays for the extraction of high-level properties

The results of the present experiment add to accumulating evidence that the capability of a display to improve performance at integration tasks depends upon how well the semantics of a domain have been mapped into the form and dynamic behavior of a display. In particular, the results indicate that if the low-level graphical elements configure to produce emergent features that correspond to the critical relationships between variables in the domain, then performance at integration tasks will be improved. There are two separate activities that must be completed to achieve this goal, and failure at either will reduce the effectiveness of the resulting display. First, the semantics of the domain must be determined. Second, a display must be designed that produces emergent features that

directly reflect the domain semantics. Each of these activities will be discussed in greater detail.

Discovering the domain semantics. In complex dynamic domains discovering the domain semantics is not an easy task. Researchers in cognitive engineering (Hollnagel and Woods, 1983; Norman, 1986; Rasmussen, 1986; Woods, 1991; Woods and Hollnagel, 1987) have developed disciplined, top-down approaches to identify the semantics of a domain. The "goal/means" hierarchy (Rasmussen, 1986) is a description of domain semantics organized in five separate levels of abstraction, ranging from the physical form of a system (e.g., what are the system components?, what do they look like?, where are they located?) to the higher-level purposes it serves (e.g., what is the system's purpose?, what constraints does the system operate under to fulfill this purpose?). It describes the higher-level properties of interest, the lower-level data that is relevant to those properties, the relationships between these lower-level data, and the relevant goals and constraints. The semantics of the domain must then be mapped into the static form and dynamic behavior of a display.

Mapping domain semantics into emergent features. Designing a configural display to facilitate performance with respect to higher-level issues involves more than just integrating variables into an object. The key to designing a successful configural display is to integrate the variables into a graphic form that directly reflects the critical data relationships that exist in the domain. As the work of Pomerantz (1986) and unsuccessful configural displays (compare the results of Sanderson et al., 1989, and Carswell and Wickens, 1987) this is not an easy task. As an example of some of the problems that might be encountered consider the "face" (Chernoff, 1973) graphic form. Due to its very nature the mapping from data to facial features is complex and arbitrary. Kleiner and Hartigan (1981) provide an example where the same set of data has been mapped into three different versions of a facial display. Each of the three versions substantially changes the resulting perceptual cues. A second problem is that the individual features of the face are not equal in perceptual salience (Brown, 1985; MacGregor and Slovic, 1986; Naveh-Benjamin and Pachella, 1982), which causes the information presented on some features to dominate the

information presented on other, less salient features. Finally, there are difficulties with maintaining scale (this is discussed in further detail in the ensuing section on low-level data.

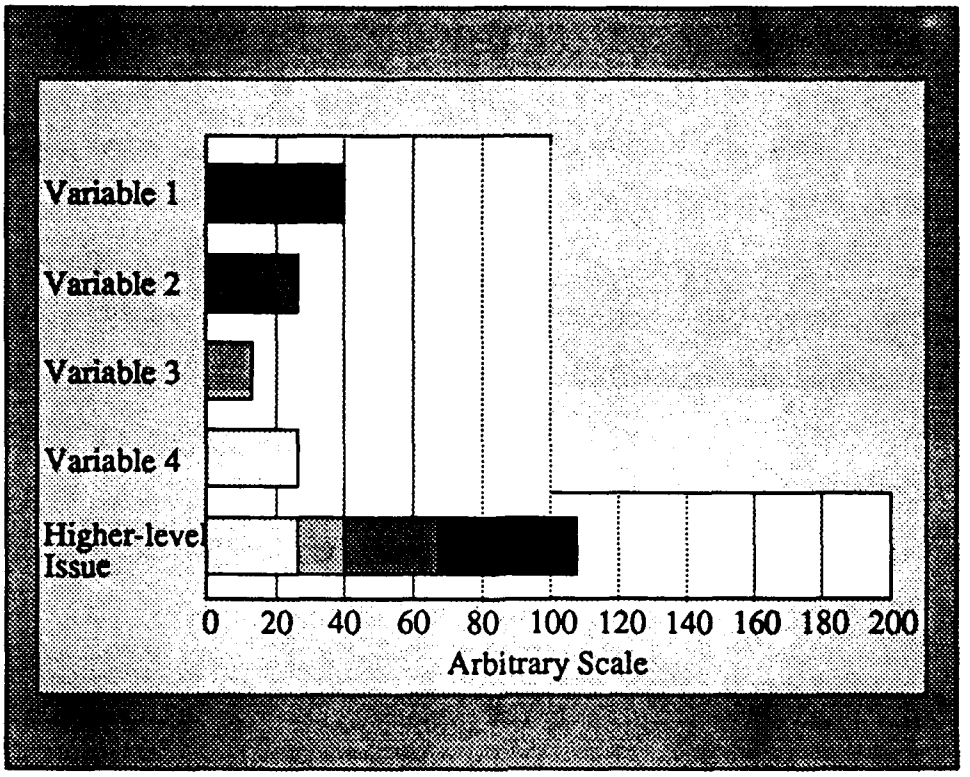
Some of the problems associated with the face display can be eliminated outright by choosing a geometric object as the graphical form for a display. However, many issues still arise. Which of the myriad of geometric objects is most appropriate for a particular application? Which variables need to be included in the graphic form? How should the individual elements be assigned to the dimensions of the object (e.g., does the relationship between two variables suggest that they should be juxtaposed, or that they should be adjacent)? Should all variables be converted to a common scale? How can the decision-making context (relevant goals and constraints) be represented? To further elaborate these issues, the design rationale for the configural display investigated in the present experiment will be described, and contrasted to a design for different domain semantics.

The goals/means analysis for the MCF task revealed that the mass and energy balances are the critical information for performing the task. The mass balance is determined by the difference between the steam and feed flow, while the energy balance is determined by the difference between indicated and compensated steam generator level. This suggests that these two sets of variables should be juxtaposed in the graphic form. In addition, the four variables are functionally dependent, and highly inter-coupled, which suggests that a single object should be used (this point is discussed in the ensuing). Since there are four variables to be portrayed an appropriate graphic form is a rectangle. Thus, a single rectangular object mapping the difference between the two related sets of variables was chosen. The differences between level variables (steam and feed flow) were mapped into the y dimension to provide a conceptual similarity to the level of fluid inside a vessel. Since all four variables could be measured in the same metric (0-100%) no changes in scaling were necessary. Finally, this object was placed in the context of the critical performance boundaries: the upper and lower set point boundaries. This mapping of domain variables resulted in emergent features that had direct meaning for the MCF task, as described previously.

Now consider how this design might have changed with different domain semantics. Imagine that there are four individual variables that must be considered to reach a decision about a higher-level issue, but that these four variables are not functionally related as in the MCF task, and instead have an additive relationship. Combining these four variables in the rectangular configural display would result in arbitrary emergent features that convey little or no information with respect to the task. For example, the emergent feature of area would not map directly onto the task demands. With large values for all four variables (with the additive relationship producing a large total value) the display would be a small rectangle located in the upper right portion of the display grid. With large differences between variables mapped in the same dimension (average total value) the display would be a large rectangle covering most of the display grid. Thus, the emergent features would not reflect or emphasize the relationships between variables. In fact, these emergent features would produce very salient perceptual cues that would have to be ignored.

The display in Figure 11 illustrates a configural display that more appropriately reflects the domain semantics. It is an example of what Woods and Roth (1988) have referred to as a "contribution graphic." The value of each individual variable is represented as a separate bar graph in the upper portion of the display. The contributions from each of the four variables have been integrated and presented in a single bar graph at the bottom of the display. The individual contributions have also been maintained in the integrated bar graph. This graphic format suppresses many of the emergent features (e.g., height, width, area, movement) that are irrelevant because the individual variables are not functionally related. It provides a direct graphic representation of the additive relationship between the low-level data and the high-level issue. It also provides a direct graphic representation explaining how the individual elements that contribute to the high-level issue.

FIGURE 11. A contribution configural display for alternative domain semantics.



Designing configural displays for the extraction of low-level data

Although research on configural displays has reported a fairly consistent pattern of results indicating that information about lower-level data (individual variables) is less accessible in configural displays than in separate displays, basic theories of attention and form perception suggest that this design goal can be achieved. Pomerantz's theory of form perception seems particularly appropriate. From this perspective, the low-level graphical elements of a form configure to produce higher-level emergent features that are very salient perceptually. However, Pomerantz (1986) emphasizes that these low-level elements are not "lost in" or "glued to" the higher-level form. The graphical elements that produce these emergent features are still available and can be attended to, but are simply less salient perceptually.

From this perspective improving an individual's ability to extract low-level information from a display is dependent upon increasing the perceptual salience of these graphical elements. Thus, the performance deficits that have been reported for configural displays may be due to the fact that the low-level graphical elements are not sufficiently salient. In fact, this is one potential explanation for the equivocal results of the present experiment. The edges of the rectangle in the configural display were only two pixels wide, and therefore not particularly salient (contrast this to the large color-coded areas of the bar graphs). In addition, the edges of the rectangle would become smaller as the differences between variables decreased, further lowering their perceptual salience. This potential explanation is supported to some degree by the differences in performance observed between the small and large state variables (see Figure 9 and Table 1).

Since there are limits to how wide the edges of the rectangle can be, another coding convention is required. One design improvement would be to extend the edges of the rectangle until they reach the boundaries of the coordinate frame. This would increase the salience of the color-coding and maintain this salience regardless of system state. In addition to color-coding there are at least two methods that might be used to make low-level graphical elements more salient: emphasizing scale and spatial separation.

Maintaining and emphasizing scale. The factors that influence an individual's ability to extract information from static graphic displays have been discussed at length (Cleveland, 1985; Tufte, 1983). The design goal is to provide a context that increases the salience of the data itself (relative to non-data elements of the graph), the relationship between a datum and other data, or the relation between a datum and the potential values that it might assume. Both Cleveland (1985) and Tufte (1983) provide recommendations for the design of frames, labels, coordinate grids and other graph elements that can be used to achieve these goals. These recommendations can be used to improve the salience of the low-level graphical elements representing low-level data.

Once again the face display will be used as an example of how not to design configural displays. When lower-level information is mapped into the low-level graphical

elements of a face, all traces of its contribution disappears: data that was originally ordinal is mapped into nominal facial features. This results primarily from the fact that no scale is available to provide a context to evaluate individual values. Rather than removing scale, the contributions of low-level data should be emphasized. For example, extending the edges of the rectangle would not only increase the perceptual salience due to color-coding, it would also emphasize scale. Thus, maintaining and emphasizing scale is one method to improve the ability to extract low-level data in a configural display.

Spatial separation. Another method to maintain the contribution of graphical elements is spatial separation. It is convenient to think of spatial separation as a design-space continuum, and the configural displays discussed previously provide examples. The contribution graphic that was described for the alternative domain semantics in the previous section is at the high end of the spectrum. In that display each graphical element is mapped into a spatially separated bar graph, while the contributions are mapped into an integrated bar graph. As previously described, spatially separating the elements both suppresses irrelevant emergent features and emphasizes the contributions of lower-level data. An intermediate level of spatial separation is used in the configural display described by Wickens and Andre (1988; 1990). In that display the contributions of low-level data were independent of the high-level object that they ultimately defined. The configural display in the present experiment is located at the low end of the spatial separation continuum: the graphical elements directly define the object.

From the perspective of Pomerantz's theory of form perception spatially separating the graphical elements is a logical way to improve their perceptual salience, and, therefore, the ability of an individual to extract low-level data. However, the research conducted on this issue has revealed mixed results, and indicates that the role of spatial separation for graphic displays in general may be a complicated one. Cleveland (1985, p. 252) describes research in which increasing the spatial distance in an integration task (comparing a test object to a standard object) increased the error of accuracy performance "by amounts that were nontrivial." On the other hand, the findings of Wickens and Andre (1988; 1990) indicate that the role of spatial separation is mediated by several factors. An interesting

empirical investigation might be devised to clarify this issue by comparing performance on focused tasks for configural displays at low, intermediate, and high levels of spatial separation.

III B. PROVIDING ANIMATION IN ANIMATED FUNCTIONAL MIMICS

Introduction: Animated mimic displays

Animated mimic displays have the potential to improve the effectiveness of both training and real-time performance. A graphic depiction of physical components of the system and the flow of information or resources between them (or, alternatively, a lack of expected flow) will contribute to the development of appropriate mental models, or understandings of a complex domain (Hollan, Hutchins, and Weitzman, 1987; 1984). This type of display provides a dynamic "explanation" of the myriad of events that occur simultaneously in complex, dynamic domains. In addition to the benefits for training, the animated mimic display can play an important role in real-time control of systems. These displays can facilitate an individual's ability to assess current system state, and to determine the causal factors that underlie the current system state. This information will also define the range of alternative control input that can be used to avoid or recover from violation of system goals. An animated mimic display also provides immediate feedback concerning the effectiveness of control input as the effects propagate through the causal network.

Animation through apparent motion

Despite the potential benefits to real-time and training performance, very little empirical research has addressed fundamental issues in the implementation of animated mimic displays. One critical issue concerns how the animation will be provided. One

approach is to provide arrows that illustrate the rate and direction of flow. The direction and displacement, and perhaps the physical size, of the arrows could provide visual cues illustrating the direction and rate of flow. An alternative approach is to place "graphical elements" inside the physical connections and to periodically shift or rotate their perceptual characteristics to illustrate the rate and direction of flow. This produces apparent motion inside the physical connections that is related to the Phi phenomenon and produces a perceptual effect that is similar to that produced by some electric signs and marquees. The second method of providing animation has several advantages, including computational efficiency and the existence of a large body of empirical research that can be drawn upon to guide design. It appears that Hollan et al.(1987; 1984) used this approach in the development of the STEAMER instructional system, and it is also the approach that was taken in the development of an animated mimic display for the manual control of feedwater task (Bennett, in press).

From the perspective of display design outlined previously, the effectiveness of an animated mimic display will be determined by the extent to which information is encoded or mapped into a graphic representation that results in psychological impressions of flow that directly correspond to the physical rates of flow existing in the domain. There are many factors that might influence the quality of this mapping, including spatial frequency (the number and size of the graphical elements), temporal frequency (the speed at which the perceptual characteristics of the graphical elements are changed), and orientation of the physical connections. However, a fundamental question concerns the perceptual characteristics of the graphical elements themselves. For the rotation of adjacent graphical elements to produce effective apparent motion they must be distinct perceptually. Should the graphical elements differ in shape, luminance, or chromaticity? What are the luminance or chromaticity contrasts, or the differences in shape that are required? The present study focuses on the role of luminance and chromaticity in the design of animated mimic displays. The literature on the perception of motion suggests that these two factors may play a critical role in determining the effectiveness of these displays.

Perception of motion

There is evidence for the existence of two distinct processes in the perception of motion: "short-range" and "long-range" processes. One example of long-range processes is the "phi" phenomenon observed by early gestalt psychologists. The phi phenomenon refers to the fact that two separate and stationary lights that are alternatively turned on and off in succession will be perceived as a single light moving back and forth in space. There are limits on the spatial distance between lights and the length of alternation time that determine whether or not apparent motion will occur. Common examples of apparent motion are the global percepts of motion that exist in some signs and marquees and airport runways that result from the discrete on-off cycling of individual lights. The effect is believed to be the result of high-level neural mechanisms.

Braddick (1974) demonstrated that there are also lower-level neural mechanisms short-range processes) in the perception of motion. He found that when a body of random dots are uniformly displaced a small distance (less than approximately 15 arc minutes) while the surrounding dots are displaced randomly, a clearly-defined figure will emerge. Since the conditions of displacement, as well as other conditions (Anstis, 1980; Braddick, 1980; Ivry and Cohen, 1990), are outside the spatial and temporal limits for long-range processes, these results suggest the existence of additional, low-level neural mechanisms. It is these perceptual processes that are believed to be responsible for the perception of real, continuous motion.

It is clear that luminance is a critical stimulus dimension for the ..ception of motion. For example, Kelly (1979b) obtained contrast-sensitivity functions for moving sine-wave gratings while controlling involuntary eye movements. The results indicate that under certain spatial and temporal frequencies the amount of luminance contrast required to perceive moving gratings is less than 1%. In fact, in a previous study Kelly found that adding motion to sine-wave gratings actually reduced the luminance contrast

requirements by a factor of 2 (Kelly, 1979a).

The role of chromaticity in the perception of motion is more ambiguous, and initial studies seemed to indicate that chromatic input could be used for the perception of motion in long-range, but not short-range processes. Ramachandran and Gregory (1978) replicated Braddick's experiment using random dots that varied in chromaticity but not luminance. Under these conditions the figure did not emerge from the random dots. Cavanagh, Tyler, and Favreau (1984) reported that when sine-wave gratings were compared at equiluminance there was a significant slowing of the perceived motion. Other studies drawing similar conclusions (Carney, Shadlen, and Switkes, 1987; Lindsey and Teller, 1990; Livingstone and Hubel, 1987), would suggest that the processes underlying the perception of motion are relatively insensitive to differences in chromaticity. On the other hand, recent studies have indicated that chromatic input can be used for the perception of motion, even for short-range processes (Gorea and Papathomas, 1989; Simpson, 1990). Perhaps Lindsey and Teller (1990, p.1752) summarize the ambivalent role of chromatic input for the perception of motion by stating that "... the impairment of motion perception at isoluminance is not all-or-none, but varies considerably with the specific motion perception task and with variation of stimulus parameters."

A series of experimental studies were conducted to investigate the effectiveness of chromaticity and luminance contrast for encoding rate of flow information in animated mimic displays. One goal of these studies was to determine an estimate of the amount of luminance contrast that is required to produce accurate perception of motion under conditions that more closely approximate the use of animated mimic displays in applied settings. Existing estimates of luminance contrast (e.g., Kelly, 1979b) were obtained under much more controlled conditions, including pure sine-wave gratings and stabilized retinal images. A second goal was to determine the role of chromatic contrast in the design of animated mimic displays. It is not sufficient to show that chromatic input can be used, it must also be shown that chromatic input can be used effectively : an appreciable slowing of the perceived rates of flow is unacceptable. In addition, an easy design error

would be to assume that chromatic differences that are clearly discriminable when the display is static will be equally discriminable when animated.

Two experiments are reported in which a standard psychophysical procedure was employed. Subjects adjusted the rate of apparent motion in a comparison bar to match the rate of motion in a standard bar. The perceptual characteristics of adjacent graphical elements in the bar (squares) were systematically altered by a factorial combination of differences for both luminance and chromaticity. Measures of both accuracy and latency were obtained.

Experiment 3

Method

Subjects. Four observers (1 male and 3 female) participated in the experiment and were paid $5.00 an hour. The observers ages ranged from 15 to 33 years of age and all observers had normal or normal-corrected vision with no color-blindness deficiencies. All four observers had participated in three previous experiments using similar procedures but different stimuli.

Apparatus. All experimental events were controlled by a general purpose laboratory computer (Sun Microsystem 4-110 Workstation). A 16" color video monitor (SONY) with a resolution of 1152 by 900 pixels was used to present the stimuli and experimental prompts. This monitor had a refresh rate of 66 Hz, non-interlacing.

Stimuli. All chromaticity and luminance measurements were made with a Minolta Chroma Meter (model CS101). The photometer measured chromaticity in x and y coordinates (CIE). All measurements were translated into the CIELUV u' and v' chromaticity coordinates using the formulas $u' = 4x / -2x + 12y + 3$ and $v' = 9y / -2x + 12y + 3$, as described in Merrifield and Silverstein (1986). Two horizontal bars were presented

on a medium grey background (u' = .2126, v' = .4652, cd/m^2= 3.64). Each bar was 7.6 cm wide and .38 cm high and separated vertically by a distance of 4.8 cm. The two bars were centered in the horizontal and vertical dimensions of the screen. Assuming an observer seating distance of 50 cm, each bar subtended a visual angle of 8.64 degrees horizontally and 24.4 arc min vertically, while the distance between the bars subtended a visual angle of 5.48 degrees. Each bar was filled with 20 squares that were .38 cm high and .38 cm wide (subtending a visual angle of 24.4 arc min both horizontally and vertically).

Each bar had three chromaticity/luminance values (CL) that repeated every third square (see Figure 12). During an individual trial the squares inside both the standard and the comparison bars always had the exact same chromaticity coordinates and luminance contrast. However, across trials the chromaticity coordinates and luminance contrast of the three repeating squares were systematically varied. For changes in chromaticity this variation consisted of four ranges of differences along a line in the color space that connected the red and green primary colors. These levels ranged from no differences between squares to chromaticity differences that comprised approximately one-sixth of the full range between red and green (see Figure 12). For Chromatic Contrast 0 the u' and v' chromaticity coordinates of the three repeating squares were all .2097 - .5454. For Chromatic Contrast 1 these values were .2097 and .5454, .2163 and .5447, and .2229 and .5440; for Chromatic Contrast 2 they were .2097 and .5454, .2229 and .5440, and .2360 and .5425; for Chromatic Contrast 3 they were .2097 and .5454, .2294 and .5433, and .2492 and .5411.

Four sizes of luminance contrast were factorially combined with the four chromatic contrasts. The four luminance contrasts were determined by successively decrementing cd/m^2 values of 0, 1.5, 3.0, or 4.5 from the luminance of the initial repeating square (15 cd/m^2). Thus, for Luminance Contrast 0 the three repeating squares had luminance values of 15 cd/m^2; for Luminance Contrast 1 these squares had values of 15, 13.5, and 12 cd/m^2; for Luminance Contrast 2 they had values of 15, 12, and 9 cd/m^2; for Luminance Contrast 3 they had values of 15, 10.5, and 6 cd/m^2. The Michelson contrast

was used to measure luminance contrast: C = (Lmax - Lmin) / 2 ($\bar{L}$), where Lmax is the maximum luminance, Lmin is the minimum luminance, and $\bar{L}$ is the mean luminance. Thus, the targeted luminance contrasts were C = 0%, 11.1%, 25%, and 42.86%. The measured luminance contrasts (averaged across chromatic contrast) were C = 0.49%, 11.82%, 25.98%, and 43.30%. The C values for the individual CL differences are listed in Table 3.

The four chromatic contrasts and the four luminance contrasts were factorially combined for a total of 16 CL differences between squares. Three separate sets of stimuli were developed for use in individual experimental sessions by rotating the luminance contrast within levels of chromatic contrast. For example, with a Luminance Contrast of 3 the luminance values of the three repeating squares of a bar would be 15, 10.5, and 6 cd/ m^2 for Set 1, 10.5, 6, and 15 cd/m^2 for Set 2, and 6, 15, and 10.5 cd/m^2 for Set 3.

The CIELUV color difference equations were applied to measured chromaticity and luminance values to estimate the relative perceptual salience of the 16 CL differences. A variation of the procedure described in Merrifield and Silverstein (1986, pp. 29 - 37) was used. The perceptual difference between two stimuli with small angular subtense is described by the equation $\Delta E^*_{SF} = [(k_L \Delta L^*)^2 + (k_u \Delta U^*)^2 + (k_v \Delta V^*)^2]^{1/2}$ where

$L^* = (116 (Y/Yn)^{1/3}) - 16;$
$U^* = 13 (L^* (u' - u'n));$
$V^* = 13 (L^* (v' - v'n));$
$u' = 4x / ((-2x) + (12y) + 3);$
$v' = 9y / ((-2x) + (12 y) + 3);$
$k_L = 0.2310$ (light/dark small-field correction factor);
$k_u = 0.0912$ (red/green small-field correction factor);
$k_v = 0.8150$ (violet/green-yellow small-field correction factor).

and

u'n = 0.1978 (1976 UCS u' coordinate of neutral chromatic point D65);
v'n = 0.4684 (1976 UCS v' coordinate of neutral chromatic point D65);
Yn = 72.9 (maximum display luminance in cd/m^2 measured by photometer);
Y = luminance in cd/m^2 measured by photometer
x = CIE x coordinate measured by photometer
y = CIE y coordinate measured by photometer

FIGURE 12. The bars and color space for Experiment 3.

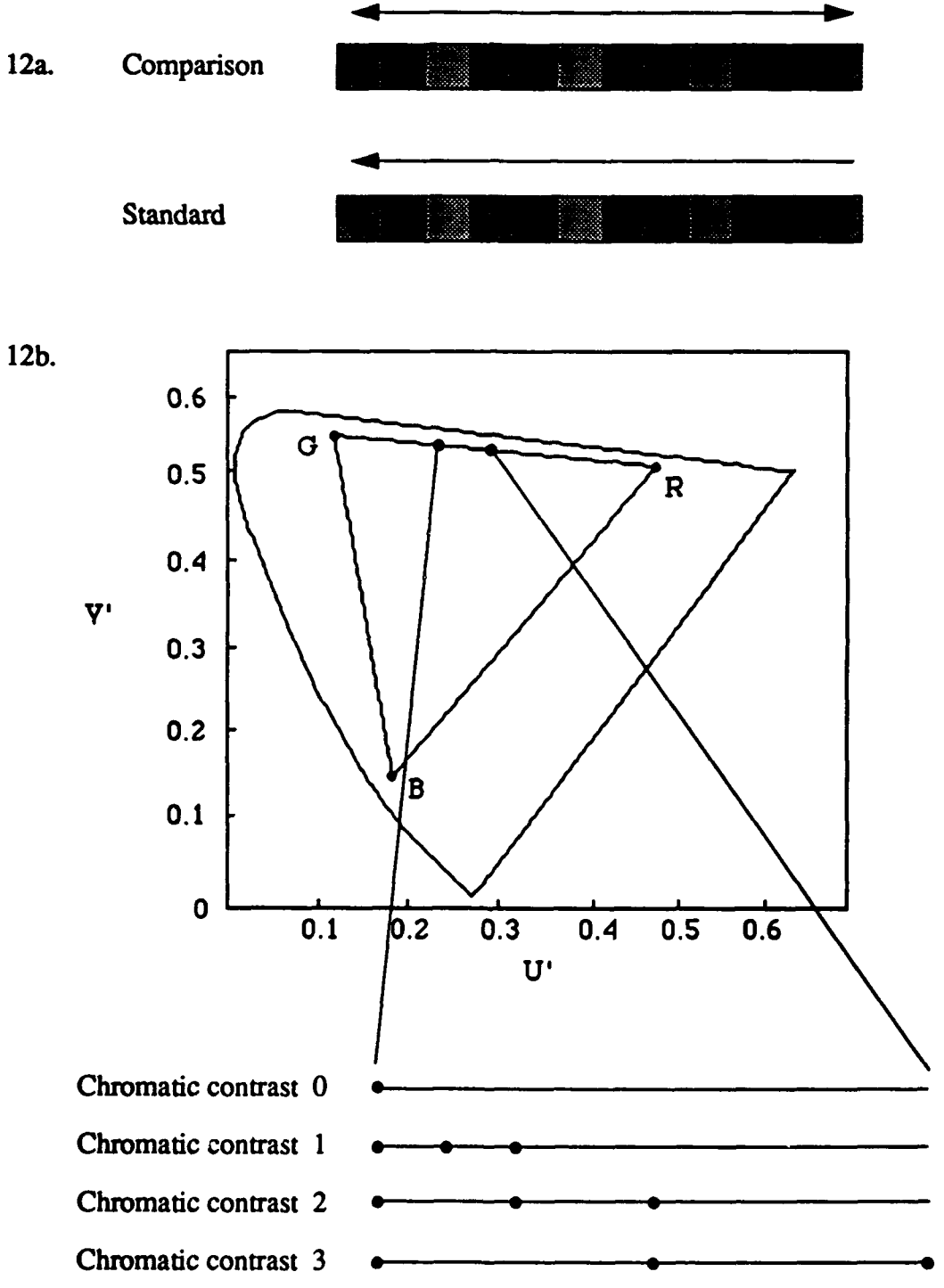


TABLE 3. Average luminance contrast in Experiment 3.

		Chromatic contrast			
		0	1	2	3
	0	0.00%	0.58%	0.46%	0.92%
Luminance contrast	1	11.95%	11.80%	11.92%	11.60%
	2	26.39%	25.77%	25.63%	26.12%
	3	43.43%	43.29%	43.28%	43.21%

TABLE 4. Average perceptual differences (ΔE^*_{SF}) in Experiment 3.

		Chromaticity contrast			
		0	1	2	3
	0	.06	1.36	2.92	4.40
Luminance contrast	1	2.86	3.21	4.08	5.20
	2	6.13	6.16	6.63	7.42
	3	9.90	9.97	10.28	10.74

Average ΔE^* values were calculated for the 16 C/L differences using these equations. A ΔE^* value was calculated for each contrast of the three repeating squares of a bar (square 1 vs. square 2, square 1 vs. square 3, and square 2 vs. square 3). These values were averaged for an estimate of that particular CL difference, and then averaged across Sets 1, 2, and 3 for an experiment-wide estimate of the CL difference. The resulting values are shown in Table 4.

Procedure. The observers were seated in an enclosed room with flat-black walls, and during an experimental session all ambient lighting was removed. The experiment was conducted during a three-day period with one experimental session per day (lasting approximately 40 minutes). During a previous experiment the observers were provided with both a written and a verbal explanation of the task, including instructions to respond as quickly and accurately as possible. The observers were informed that the stimuli would be slightly different in the present experiment, but that all other procedures would remain the same.

The lower bar was the "standard" bar. The CL values of the squares in this bar were shifted at discrete intervals from the right to the left (i.e., at each interval the chromaticity values of each square would be changed to that of the square on its right). The arrow in Figure 12 above the standard bar is used to indicate that boxes can only go from right to left. This shift was accomplished efficiently by a single call to rotate the computer's color table. The perceptual attributes of the squares in the standard bar were shifted at one of two rates (15 or 30 times per second), and the rate remained constant throughout an experimental trial.

The upper bar was the "comparison" bar and its rate was controlled by the observer. During a trial the perceptual attributes of the comparison bar were changing in one of two directions (to the right or to the left, as indicated by the arrows in Figure 12). The direction of flow in the comparison bar remained the same during an experimental trial and could not be reversed by the observer (the observer could stop the apparent motion by continually reducing its rate, but could not reverse the direction for a particular trial).

The observer's task was to adjust the rate of motion in the comparison bar to match the rate of motion in the standard bar. Using an optical mouse, observers pointed and clicked at separate boxes in the upper left portion of the screen to increase or decrease the rate of flow in the comparison bar and to start an experimental trial. A standard psychophysical procedure was used. The first input by the observer increased or

decreased the rate of the comparison bar by 5 updates per second. From that point the size of the rate change depended upon both the current rate change and the direction of previous observer input. An observer input in the opposite direction from the previous input (a reversal) decreased the rate change by half the current amount. An observer input in the same direction as the previous input resulted in a change in rate equal to the size of the current rate change, unless the two previous observer inputs were in the same direction. In this case the size of the rate change was doubled. After the eighth reversal the trial was automatically ended. Although there was a lower limit on rate change (the rate change could not be less than 0), there was no upper limit.

Measures of both accuracy and latency were obtained for each experimental trial. An error magnitude score was obtained by averaging the rates of the comparison bar before and after the eighth reversal, subtracting this value from the rate of the standard bar, and taking the absolute value of the difference. Observers were provided with feedback for error magnitude. Measures of latency (accurate to 1/100 of a second) were also recorded, although no feedback was provided.

To summarize, in each of 3 experimental sessions (Days 1 through 3, within subjects factor) an observer completed 64 trials: a factorial combination of the 16 chromaticity/luminance contrasts (four chromatic contrasts factorially combined with four luminance contrasts, within subjects factors), the two directions of the comparison bar (left or right, within-subjects factor), and the two rates for the standard bar (15 or 30 changes per second, within-subjects factor). The order of these trials was randomly determined for each experimental session, and the order of sets was counter-balanced across observers.

Results

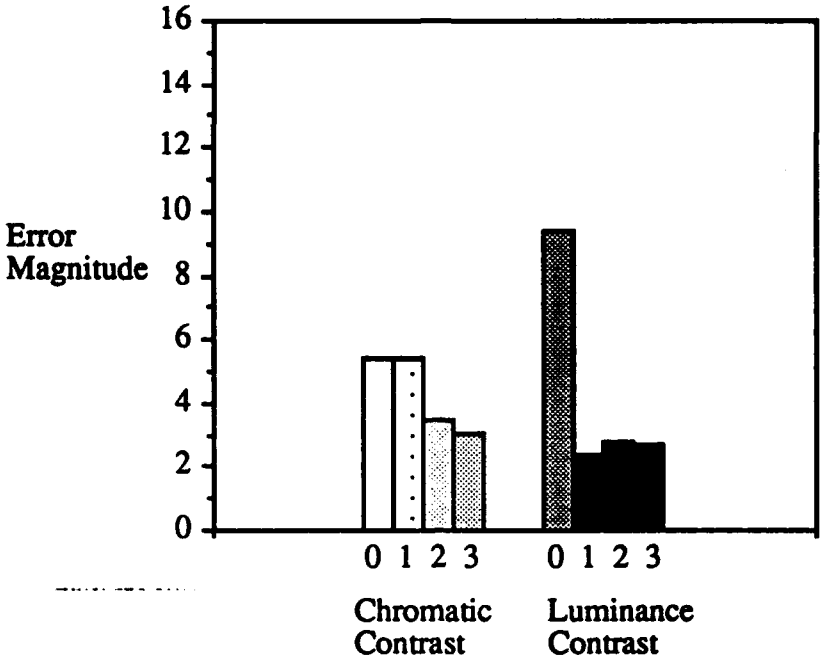
Accuracy. The accuracy scores contained a number of outlier scores, some of which were quite large (e.g., the largest error magnitude score was 4,529,852,585). The

most likely cause of these outliers is the lack of an upper boundary on the number of updates per second for the comparison bar. When observers increased the change in rate to a level that was greater than the refresh rate of the monitor apparent forward motion, apparent backward motion, or a lack of apparent motion could result. The origin of this perceptual effect is described in greater detail in the discussion section. All scores in which the final rate of the comparison bar was greater than the refresh rate of the screen (66 Hz) were dropped from the analyses. Of the 768 total scores, 13 scores (1.69%) were removed using this criterion. Pilot studies had shown that the direction of the bars had no impact on matching performance and the resulting scores were averaged across the two directions of the comparison bar for a total of 32 scores per observer per experimental session.

A 3 x 4 x 4 x 2 repeated-measures ANOVA was performed on these scores. The assumption of non-correlation between repeated measures were checked by calculating the Greenhouse-Geiser estimate of epsilon. For effects where this assumption was violated the appropriate adjustments to degrees of freedom were made and are reflected in the probability levels that are reported (for ANOVA's and post-hoc comparisons). The main effects of chromaticity, $F(3,9) = 10.94$, $p < .02$, and luminance, $F(3,9) = 26.33$, $p < .02$, and the interaction effect between chromaticity and luminance $F(9,27) = 9.01$, $p < .02$ were significant. All other effects were not significant.

The means for the main effects of both chromaticity and luminance are shown in Figure 13. Pair-wise comparisons for luminance revealed that performance with Luminance Contrast 0 was significantly less accurate than Luminance Contrast 1, $F(1,9) = 56.92$, $p < .004$, Luminance Contrast 2, $F(1,9) = 49.66$, $p < .005$, and Luminance Contrast 3, $F(1,9) = 51.00$, $p < .005$. All other comparisons were not significant.

FIGURE 13. Accuracy means for chromaticity and luminance (Exp. 3).

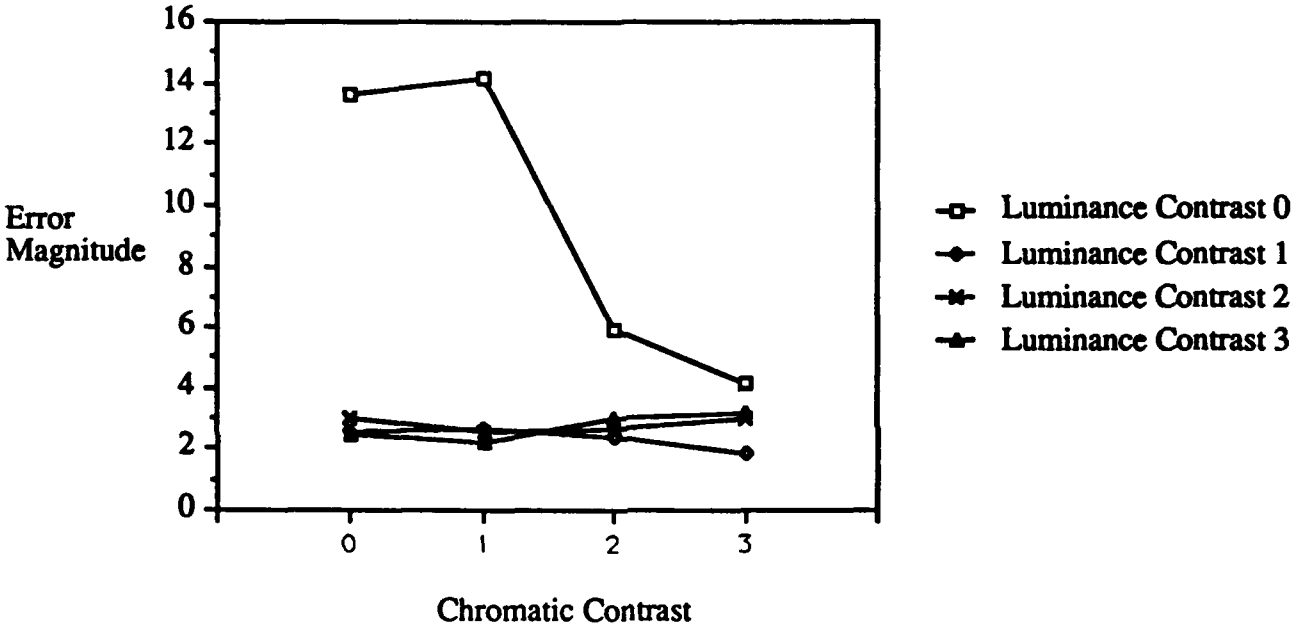


Pair-wise comparisons for chromatic contrast indicated that the mean error magnitude for Chromatic Contrast 0 was significantly less accurate than the means for Chromatic Contrast 2, $F(1,9) = 13.10$, $p < .02$, and Chromatic Contrast 3, $F(1,9) = 19.73$, $p < .007$. Chromatic Contrast 1 was significantly less accurate than Chromatic Contrast 2, $F(1,9) = 12.79$, $p < .02$, and Chromatic Contrast 3, $F(1,9) = 19.34$, $p < .008$. All other comparisons failed to reach significance.

The significant interaction indicates that the quality of matching performance was dependent upon both chromatic and luminance contrasts. The means for this effect are illustrated in Figure 14. F-tests for simple interaction effects between chromaticity and luminance were conducted by calculating the interaction for all individual pairs of luminance contrasts (e.g., Luminance Contrast 0 vs. 1). These tests revealed that all simple interactions between luminance and chromaticity were significant when Luminance Contrast 0 was included in the contrast (0 vs. 1, $F(3,27) = 15.32$, $p < .008$; 0 vs. 2, $F(3,27) = 17.69$, $p < .006$; 0 vs. 3, $F(3,27) = 20.37$, $p < .005$) while all simple interactions not

involving Luminance Contrast 0 were not significant (1 vs. 3, $F(3,27) < 1.0$; 1 vs. 2, $F(3,27) < 1.0$; 2 vs. 3, $F(3,27) < 1.0$). These results indicate that the overall interaction effect between chromaticity and luminance was due to the effects of chromatic contrasts at Luminance Contrast 0. When no luminance contrast was available observers were able to improve matching performance as the chromatic contrasts became larger.

FIGURE 14. Accuracy means for the chromaticity by luminance interaction (Exp. 3).



Latency. The 13 latency scores associated with comparison bar updates that were faster than the refresh rate of the screen were not considered in the analysis; the remaining latency scores were averaged across the two directions of the comparison bar. A 3 x 4 x 4 x 2 repeated-measures ANOVA was performed on these scores. The main effect of chromaticity, $F(3,9) = 6.22$, $p < .0618$ was marginally significant. All other effects were not significant.

Discussion

The results suggest that the psychological processes responsible for the perception of motion are sensitive to both luminance and chromatic input, at least under the stimulus conditions employed in the present experiment. The accuracy of matching performance was improved when non-zero luminance contrasts existed in the repeating squares (see Figure 13). Luminance contrasts of approximately 12%, 26%, and 43% improved matching performance relative to luminance contrasts of less than 1%. In addition, there was no incremental improvement in performance associated with relative increases in luminance contrast (as indicated by the lack of significant performance differences between the higher levels of luminance contrast). These results suggest that only a sufficient level of luminance contrast is necessary for performance of the matching task. The results are consistent with a large body of literature indicating that luminance contrast plays a central role in the perception of motion.

The results also indicate that observers were able to use chromatic information to perform the matching task. Matching performance improved with larger chromatic differences (Chromatic Differences 2 and 3) relative to smaller chromatic differences (Chromatic Differences 0 and 1), as indicated in Figure 13. Previous research investigating the perception of motion has shown that the effectiveness of chromatic information depends upon the nature of the perceptual process that is being investigated as well as the task requirements (Cavanagh et al., 1984; Gorea and Papathomas, 1989; Ramachandran and Gregory, 1978; Simpson, 1990). In particular, the higher-level, long-range processes responsible for the perception of motion have been shown to be more sensitive to chromatic input than lower-level short-range processes, where investigations have produced mixed results. Since the size of the stimuli in the present experiments (24 min arc) was outside the spatial limits of short-range processes (15 min arc, Braddick, 1974) it is likely that long-range processes were being investigated.

The significant interaction effect between chromaticity and luminance provides additional insight concerning the roles of chromaticity and luminance contrast in performance of the rate-matching task. The results indicate that with any luminance contrast above approximately 0% the observers were able to perform the matching task

accurately (see Figure 14). When there was no luminance contrast between adjacent squares observers could successfully perform the task using chromatic differences, provided that this difference was sufficiently large. These results indicate that, at least with the C/L contrast levels employed in Experiment 3, observers were relying primarily upon luminance contrast to perform the matching task, but could use chromatic contrast when insufficient luminance contrast was present.

Although overall accuracy performance on the matching task was quite good (an average error magnitude of 4.33) on a low percentage of trials (1.69%) rather large errors in accuracy were present (the largest outlier was 4,529,852,585). The most likely explanation of these outlier scores is related to the physical constraints associated with the monitor and aspects of the methodology that was employed. The physical displacement of the squares was always in one direction, and normally resulted in apparent motion that was consistent with the physical displacement. This is illustrated in Figure 15a. Both the physical displacement of the squares (indicated by the stippled arrow) and the subjective impression of apparent motion (indicated by the solid arrow) is from left to right. However, when the observer increased the update rate of the comparison bar to a level that was higher than the refresh rate of the monitor (66 Hz) apparent forward motion, apparent backward motion, or a lack of apparent motion could result. The direction and rate of apparent motion depended upon the number of physical displacements that occurred between screen updates. In Figure 15b two displacements to the right have occurred between successive screen updates (stippled arrow) and the end result is apparent motion from the right to the left (solid arrow). In Figure 15c three displacements have occurred between screen updates, resulting in the absence of apparent motion; in Figure 15d four displacements have occurred, resulting in apparent forward motion. These perceptual effects are the most likely explanation of the outliers in the data.

The roles of chromaticity and luminance in the perception of motion were investigated in greater detail in Experiment 4. The results of Experiment 3 indicate that although observers were able to use chromatic information to match the motion of squares in the pipes, they appeared to do so only when luminance contrast was not available. One

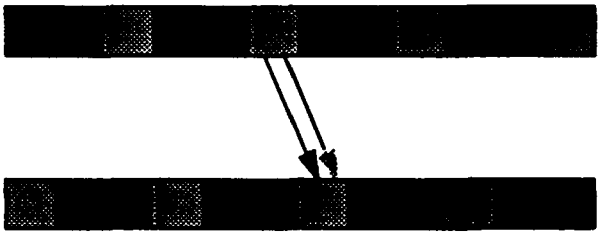
possible interpretation of these results is that luminance contrast is more critical for the perception of motion than chromatic information. However, an alternative explanation is that the observers were simply using the information that was most salient perceptually. As Table 3 reveals, the range of luminance contrasts that were chosen were more discriminable than the range of chromatic contrasts. With smaller differences in luminance observers might be more able, or more willing, to use chromatic information. To explore this possibility the range of luminance contrasts was reduced in Experiment 4: the largest luminance contrast (9.79%) was smaller than the smallest luminance contrast in Experiment 3 (11.82%). The range of chromatic contrasts was slightly enlarged, but remained at approximately the same levels as in Experiment 3. This resulted in a reversal of the relative discriminability of luminance and chromaticity. All other aspects of the stimuli and method remained the same as in Experiment 3.

Experiment 4

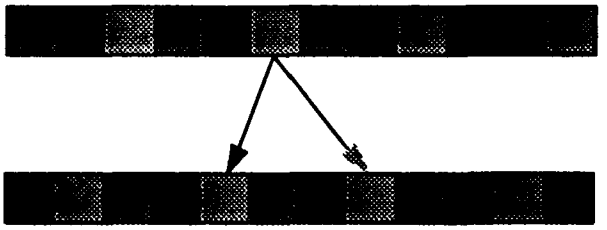
Method

Subjects. Seven observers (5 male and 2 female) participated in the experiment and were paid $5.00 an hour. None of the observers participated in Experiment 3, but all had participated in a previous experiment (5 sessions of approximately 40 minutes). The observers had been given previous instructions regarding the nature of the task, including instructions to respond as quickly and accurately as possible. The observers were informed that only the difference in the present experiment was the nature of the perceptual differences between squares in the bars. The observers ages ranged from 20 to 25 years of age and all observers had normal or normal-corrected vision with no color-blindness deficiencies.

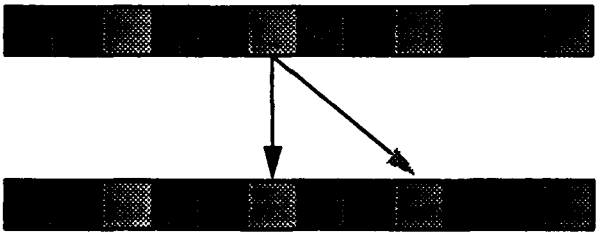
FIGURE 15. Discrepancy between apparent motion (solid arrows) and physical displacement (stippled arrows).



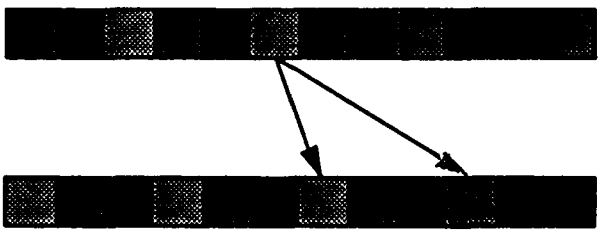
3a. Forward apparent motion



3b. Backward apparent motion



3c. No apparent motion



3d. Forward apparent motion

Apparatus. The apparatus was identical to the previous experiments.

Stimuli. All aspects of the stimuli remained the same as in Experiment 3 with the exception of differences in chromaticity and luminance values. The primary difference was in the magnitude of the luminance contrasts that were employed. The four luminance contrasts consisted of successive decrements of 0, .45, .90, or 1.35 cd/m^2 to the initial repeating square which had a baseline luminance of 15 cd/m^2. Thus, for Luminance Contrast 0 the three repeating squares had luminance values of 15 cd/m^2; for Luminance Contrast 1 these squares had values of 15, 14.55, and 14.10 cd/m^2; for Luminance Contrast 2 they had values of 15, 14.10, and 13.20 cd/m^2; for Luminance Contrast 3 they had values of 15, 13.65, and 12.3 cd/m^2. The corresponding target luminance contrast values were C = 0%, 3.1%, 6.4%, and 9.9%. The measured C values were 0.39%, 2.80%, 6.21%, and 9.79%; the C values for the 16 individual CL combinations are listed in Table 5.

TABLE 5. Average luminance contrast in Experiment 4.

		Chromatic contrast			
		0	1	2	3
Luminance contrast	0	0.11%	0.22%	0.56%	0.67%
	1	2.77%	3.01%	2.77%	2.66%
	2	6.19%	6.21%	6.23%	6.22%
	3	9.34%	9.76%	10.17%	9.88%

Four chromatic contrasts were employed. Although the maximum difference in chromaticity was slightly expanded relative to Experiment 3, the maximum range of chromatic contrasts covered approximately one sixth of the range between the red and green primaries. The primary difference with respect to Experiment 3 is that five sets of

stimuli were developed for Experiment 4 that covered different portions of the line connecting the red-green primary's. The u' and v' chromaticity coordinates of the chromatic contrasts were determined in the same manner as in Experiment 3. The left-most endpoints for Sets 1, 2, 3, 4, and 5 were u'=.1176 and v'= .5554, u'=.1638 and v'=.5506, u'=.2100 and v'=.5458, u'=.2562, and .5410, and u'=.3024, v'=.5362, respectively. The chromaticity coordinates for each chromatic contrast was obtained by adding constant values to these initial endpoints. For Chromatic Contrast 0 the constants for both u' and v' were 0; for Chromatic Contrast 1 the constants for u' and v' were +.0077 and -.0008; for Chromatic Contrast 2 the constants for u' and v' were +.0154 and -.0016; for Chromatic Contrast 3 the constants for u' and v' were +.0231 and -.0024. For example, in Set 1 the three repeating chromaticity values for Chromatic Contrast 3 were u' = .1176 and v' = .5554 (initial endpoint), u' = .1407 (.1176 + .0231) and v' = .5530 (.5554 + -.0024), and u' = .1638 (.1407 + .0231) and v' = .5506 (.5530 + -.0024).

The four chromatic contrasts and the four luminance contrasts were factorially combined for a total of 16 CL differences between squares. The CIELUV color difference equations were used to determine the average ΔE^*_{SF} for these CL differences using the same method described in Experiment 3, and the results are provided in Table 6.

TABLE 6. Average perceptual differences (ΔE^*_{SF}) in Experiment 4.

		Chromaticity contrast			
		0	1	2	3
Luminance contrast	0	0.12	1.72	3.28	4.82
	1	0.70	1.80	3.32	4.82
	2	1.52	2.11	3.45	4.97
	3	2.28	2.77	3.79	5.04

Procedure. The procedure was essentially the same as previous experiment, with the exception that during each experimental session the chromatic/luminance contrasts were taken from one of the five sets. Each of the five sets was used in separate experimental sessions for an individual observer, and the order was counterbalanced across observers. Thus, in each of 5 experimental sessions (Days 1 through 5) an observer completed 64 trials: a factorial combination of the 16 chromatic/luminance contrasts between squares (a factorial combination of four chromatic contrasts and four luminance contrasts), the two directions of the comparison bar (left or right), and the two rates for the standard bar (15 and 30 changes per second). The order of these trials was randomly determined.

Results

Accuracy. As in Experiment 3, all scores in which the final rate of the comparison bar was greater than the refresh rate of the screen (66 Hz) were not considered in the analyses. Of the 2,240 total scores 26 scores (1.16%) were removed using this criterion. The resulting scores were averaged across the two directions of the comparison bar resulting in a total of 32 scores per observer per experimental session. An 3 x 4 x 4 x 2 repeated-measures ANOVA was performed on these scores. The main effects of luminance, $F(3,18) = 7.10$, $p < .01$, chromaticity, $F(3,18) = 5.41$, $p < .03$, and rate, $F(1,6) = 19.57$, $p < .005$ were significant. The interaction between chromaticity and luminance was significant with the original degrees of freedom, $F(9,54) = 2.44$, $p < .02$, but failed to reach significance ($p < .12$) with the adjusted degrees of freedom required because of the correlation of repeated measurements associated with the effect. All other effects were not significant.

The means for the main effect of chromaticity and luminance are illustrated in Figure 16. Pair-wise comparisons for luminance indicated that matching performance for Luminance Contrast 0 was significantly less accurate than performance for all other luminance contrasts (Luminance Contrast 1, $F(1,18) = 15.43$, $p < .004$, Luminance

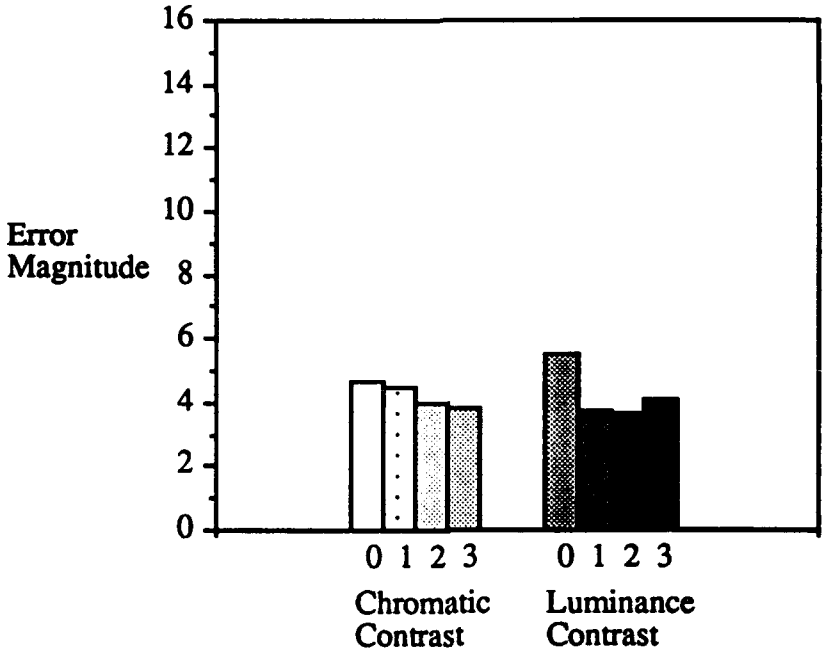
Contrast 2, $F(1,18) = 15.92$, $p < .004$, and Luminance Contrast 3, $F(1,18) = 9.88$, $p < .02$). All other comparisons between luminance contrasts were not significant.

The pair-wise comparisons for chromaticity indicated that performance for Chromatic Contrast 0 was significantly less accurate than for Chromatic Contrast 2, $F(1,18) = 7.96$, $p < .03$, and for Chromatic Contrast 3, $F(1,18) = 11.94$, $p < .01$. Performance for Chromatic Contrast 1 was marginally more accurate than performance for Chromatic Contrast 2, $F(1,18) = 4.26$, $p < .07$, and significantly less accurate than for Chromatic Contrast 3, $F(1,18) = 7.29$, $p < .03$. All other comparisons were not significant.

Although the interaction effect between chromaticity and luminance was not significant with the adjusted degrees of freedom, a very similar pattern of results was obtained as in Experiment 3 and the means for the effect are shown in Figure 17. The main effect of rate indicated that observers were more accurate when the standard pipe was moving at the slower rate (mean error magnitude = 3.43) than at the faster rate (mean = 5.38).

Latency. The 26 latency scores obtained when the rate of the comparison bar was greater than the refresh rate of the monitor were not considered in the analysis; the remaining latency scores were averaged across the two directions of the comparison bar. An 3 x 4 x 4 x 2 repeated-measures ANOVA was performed on these scores. The main effects of chromaticity, $F(3,18)=11.76$, $p < .009$, luminance, $F(3,18)=7.81$, $p < .03$, and the interaction effect between chromaticity and luminance, $F(9,54)=8.91$, $p < .002$ were significant. All other effects were not significant.

FIGURE 16. Accuracy means for chromaticity and luminance (Exp. 4).



The means for the main effects of chromaticity and luminance are illustrated in Figure 18. The pair-wise comparisons for luminance indicated that performance for Luminance Contrast 0 was significantly faster than Luminance Contrast 1, $F(1,18)=12.07$, $p< .02$, Luminance Contrast 2, $F(1,18)=17.25$, $p< .01$, and Luminance Contrast 3, $F(1,18)=16.72$, $p< .01$. All other comparisons were not significant. The pair-wise comparisons for chromaticity indicated that performance with Chromatic Contrast 0 was significantly faster than Chromatic Contrast 2, $F(1,18)=11.65$, $p< .02$, and than Chromatic Contrast 3, $F(1,18)=31.29$, $p< .003$. Performance with Chromatic Contrast 1 was significantly faster than for performance with Chromatic Contrast 3, $F(1,18)=16.94$, $p< .01$, and Chromatic Contrast 2 was marginally faster than Chromatic Contrast 3, $F(1,18)=4.75$, $p< .07$. All other comparisons were not significant.

FIGURE 17. Accuracy means for the luminance by chromaticity interaction (Exp. 4).

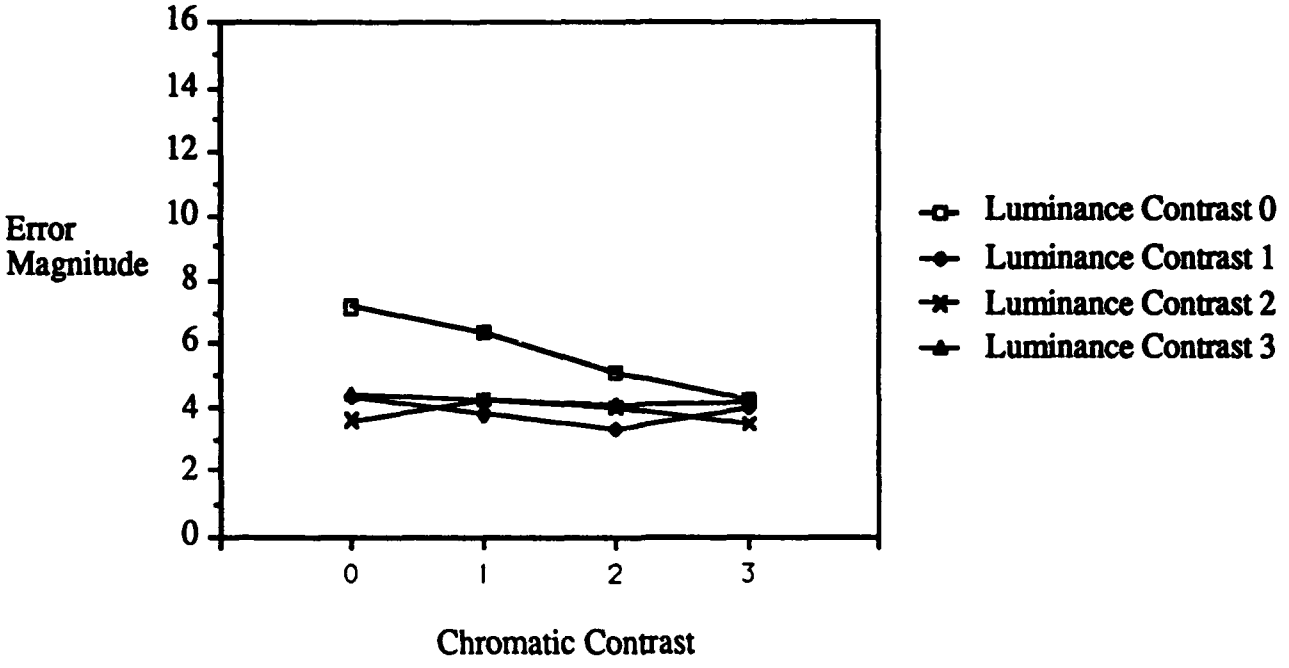
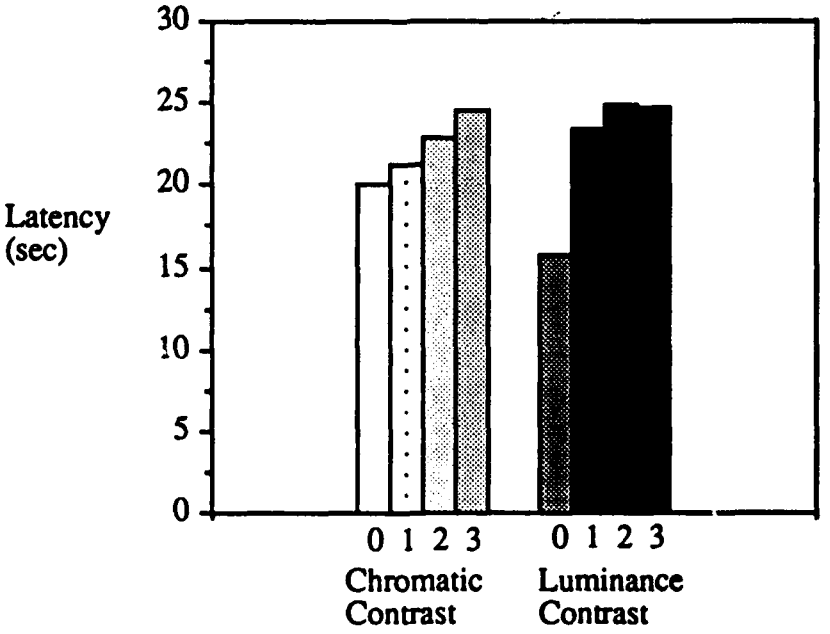


FIGURE 18. Latency means for chromaticity and luminance (Exp. 4).

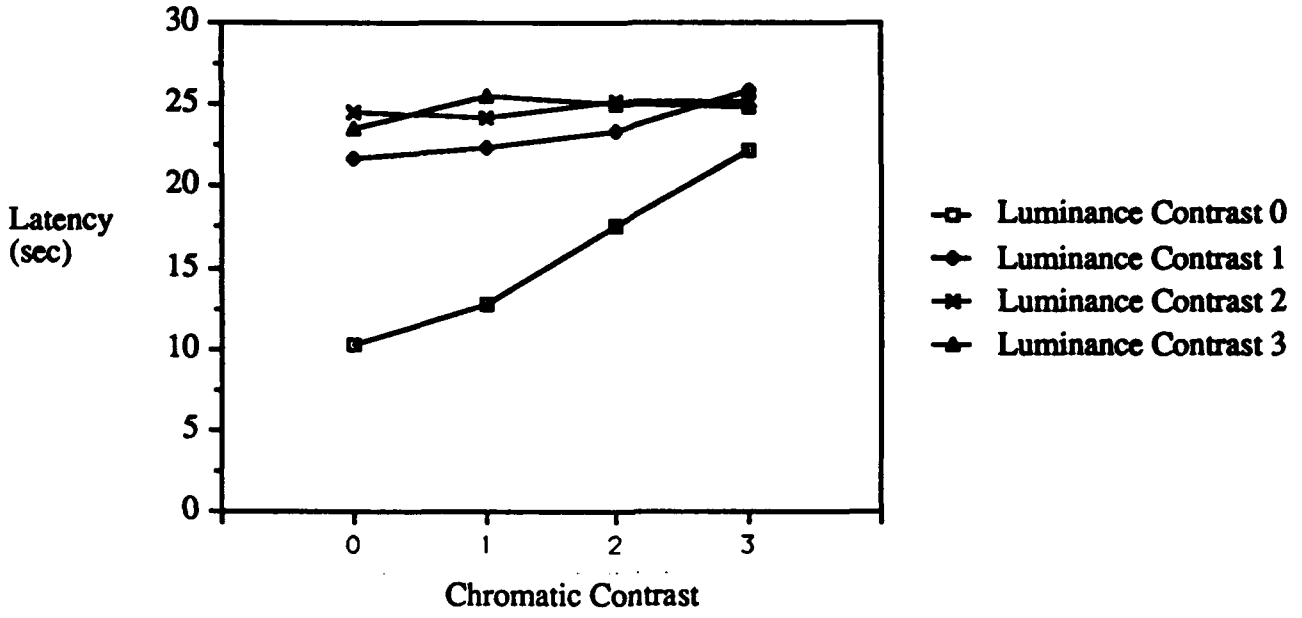


The significant interaction indicates that the latency of matching performance was dependent upon both chromatic and luminance contrasts. The means for this effect are illustrated in Figure 19. F-tests for simple interaction effects were examined by calculating the interaction for all pairs of luminance contrasts (e.g., Luminance Contrast 0 vs. 1). These tests revealed that all simple interactions between luminance and chromaticity were significant when Luminance Contrast 0 was included in the contrast (0 vs. 1, $F(3,54) = 10.17$, $p < .007$; 0 vs. 2, $F(3,54) = 19.59$, $p < .0008$; 0 vs. 3, $F(3,54) = 19.25$, $p < .0008$) while all simple interactions not involving Luminance Contrast 0 were not significant (1 vs. 3, $F(3,54) = 2.06$, $p < .17$; 1 vs. 2, $F(3,54) = 1.81$, $p < .20$; 2 vs. 3, $F(3,27) < 1.0$). Thus, these results indicate that the overall interaction effect between chromaticity and luminance was primarily due to the simple interaction between chromatic contrasts and Luminance Contrast 0. When no luminance contrast was available observers increased the amount of time to complete the matching task as the chromatic contrasts became larger. Additional F-tests for the simple effects of luminance at each of the chromatic contrasts support this conclusion.

Discussion

Overall, the accuracy for luminance contrast in Experiment 4 revealed a very similar pattern of results to those obtained in Experiment 3. Observers were able to accurately match the rate of motion in the standard and comparison bars when the squares inside these bars had luminance contrasts that were greater than approximately 0% (see Figure 16). As in Experiment 3, any non-zero change in luminance was sufficient for accurate performance of the matching task and there was no additional benefit to increasing the luminance contrast (as indicated by the lack of significant differences between conditions with non-zero percentage changes in luminance).

FIGURE 19. Latency means for the chromaticity by luminance interaction (Exp. 4).



The accuracy of performance for chromatic contrast revealed a similar, but slightly more positive pattern of results for Experiment 4 than for Experiment 3. Observers significantly improved the accuracy of matching performance with larger chromatic contrast (Chromatic Contrasts 2 and 3) relative to little (Chromatic Contrast 1) or no (Chromatic Contrast 0) chromatic contrasts (see Figure 16). In contrast to Experiment 3, the interaction effect between chromaticity and luminance failed to reach significance when the probability values were adjusted for the correlation of repeated measurements ($p<.14$). However, this interaction was significant with the original degrees of freedom ($p<.05$) and a very similar pattern of results (relative to those of Experiment 3) was obtained. Figure 17 indicates that chromaticity contrast had a clear impact on performance only when there was not sufficient luminance contrast to perform the task.

Although the results of Experiment 3 would seem to indicate that luminance contrast was more critical to matching performance than chromatic contrast, an alternative

possibility is that the observers were simply using the most salient information that was available. The size of the luminance contrasts that were used in Experiment 3 (0.49%, 11.82%, 25.98%, and 43.30%) resulted in contrast in luminance that was more salient perceptually than the corresponding contrast in chromaticity, as indicated in Table 3. In Experiment 4 size of the luminance contrasts were lowered considerably (0.39%, 2.80%, 6.21%, and 9.79%) resulting in a reversal of the relative perceptual salience between chromaticity and luminance (see Table 6). Despite this reversal, a similar pattern of results arose, indicating that luminance contrast is critical for accurate performance of the matching task. The size of the luminance contrasts employed in Experiment 4 results suggests that the processes underlying the perception of motion are extremely sensitive to luminance contrast.

The results of Experiments 3 and 4 would appear to indicate that chromatic information does have an impact on the perception of motion, although it plays a secondary role to luminance information. Chromatic contrast improved performance at the matching task, provided that sufficient luminance contrast was not available and that the chromatic contrast was sufficiently large. However, an alternative explanation of these results must be considered. Is it possible that observers were able to use the very small amount of luminance contrast that was available in the Luminance Contrast 0 condition for performance of the task? The results of Experiment 4 indicate that an average luminance contrast of 2.88% was sufficient for performance of the matching task. In both experiments the squares in Luminance Contrast 0 were not isoluminant: the luminance contrast increased with additional chromaticity. The increase in luminance contrast ranged from 0.0 to 0.92% cd/m^2 in Experiment 3 and from 0.11 to 0.67% cd/m^2 in Experiment 4. Both Campbell and Robson (1968) and Kelly (1979b) found that a luminance contrast of less than 1% could be sufficient for observers to detect the presence of sine-wave gratings. Whether or not this low level of luminance contrast was sufficient depended upon several factors, including overall luminance levels, spatial and temporal frequencies of the gratings, and the form of the gratings. There are too many differences between those studies and the present to assume that the increased performance attributed

to chromatic information was in fact due to the small amounts of luminance contrast present in the Luminance Contrast 0 condition. However, it is also a possibility that cannot be ruled out.

In contrast to Experiment 3, Experiment 4 revealed significant latency effects. These results mirror the pattern of accuracy results, except that the direction of performance was reversed. Consideration of both the latency and accuracy results suggest a speed-accuracy trade-off: although observers were able to improve the accuracy of their matching performance, they did so at a cost in latency. However, an alternative, and more likely, interpretation of these data is that the decrease in latency scores were a result of the fact that observers simply did not have the information required to complete the task, rather than trading speed for accuracy. In the Luminance Contrast 0 and Chromatic Contrast 0 condition there were extremely small perceptual differences between adjacent squares (if any) with the result that no apparent motion could be seen. Under these circumstances an observer could not possibly match the rates of apparent motion with any accuracy, and a natural response was to quickly alternate between control inputs until eight reversals were completed and the trial ended. This is illustrated in Figure 19 which shows that observers averaged only approximately 10 seconds in this condition to complete the task.

Under these circumstances it is clear that the decrease in latency scores resulted from a simple lack of information required to complete the task, rather than a trading of latency for accuracy. A similar argument can be made for the decrease in latency scores in the remaining Luminance Contrast 0 conditions. It is clear from the accuracy results that luminance contrast, and not chromatic contrast, was the primary factor determining the success of matching performance. Without luminance contrast the task was considerably more difficult, and observers were more likely to complete the task quickly. It should also be noted that accuracy performance was indirectly stressed: although observers received standard instructions to complete each trial "as quickly and as accurately as possible", only feedback regarding the accuracy of responses was provided on a trial-by-trial basis. Thus, it is reasonable to assume that the reduced latency scores reflect a natural observer

response to experimental conditions in which the information necessary to complete the task was either not available, or severely degraded, rather than a speed-accuracy trade-off.

General discussion

When interacting in complex, dynamic domains an individual must consider both higher-level, abstract issues (often determined by complex relationships between variables) and low-level information that is more directly tied to the physical implementation of the system (Rasmussen, 1986, Woods and Roth, 1988). Configural displays are a particularly promising graphic format to provide the high-level information (Bennett, Toms, and Woods, in preparation; Flach and Vicente, in press; Wickens and Andre, 1990). However, these displays do not provide explicit information about system resources that can be used to prevent or recover from trouble. This information is located at a lower level of abstraction that is more closely tied to the physical function of the system, *including a description of the alternative system resources that are available* and the physical connections between these alternative system resources.

Animated mimic or pictorial displays explicitly provide this information. They illustrate the important system components, the physical connections between them, and the flow of information or resources between system components. Animated mimic displays may play a special role in training where they can facilitate the formation of appropriate mental models of complex systems. These displays also have the potential to support an individual in a variety of cognitive tasks that occur during interaction with complex, dynamic domains. Despite the potential to improve both on-line and off-line (training) performance, little, if any, research has addressed fundamental issues in the design of animated mimic displays.

The present study investigated issues involved in producing apparent motion by alternating the perceptual characteristics of graphical elements inside the causal connections between system resources. Although there are other methods to produce

animation this method has certain advantages. It produces a satisfactory subjective impression of motion, a large body of basic vision research exists that can provide a basis for design decisions, and it is quite efficient computationally. Rather than re-drawing entire sections of screen, or copying and moving portions of the display (activities which are demanding computationally) apparent motion can be produced quite efficiently through successive calls to change the computer color table. This computational economy will be particularly advantageous when complex displays are involved, or when less powerful computers are desired.

From the perspective of display design outlined in the introduction section, a central issue in graphic displays is to encode information from the domain into a graphical representation that allows this information to be easily decoded, or extracted by the individual. Cleveland (1985, p. 229) states "When a graph is constructed, quantitative and categorical information is encoded by symbols, geometry, and color. Graphical perception is the visual decoding of this encoded information. Graphical perception is the vital link, the raison d'etre, of the graph. No matter how intelligent the choice of information, no matter how ingenious the encoding of the information, and no matter how technologically impressive the production, a graph is a failure if the visual decoding fails." The present study studied investigated the role of two fundamental perceptual dimensions that can be used to encode information concerning rates of flow in animated mimic displays, chromaticity and luminance.

Luminance and chromaticity in animated mimic displays

Previous research has indicated that luminance contrast plays a critical role in the perception of motion, and the results of the present experiments are no exception. In both experiments a luminance contrast of greater than approximately 0% between the repeating squares allowed successful performance of the rate matching task. An interesting aspect of the results is that a minimal level of luminance contrast was sufficient to perform the

task, and additional increases in luminance contrast did not result in incremental improvements to performance. The smallest level of effective luminance contrast was C = 2.80%, and it is possible that the minimal level of luminance contrast lies below this value.

This level of luminance contrast falls between estimates obtained in other studies. Travis, Bowles, Seton, and Peppe (1990) investigated both chromatic and luminance contrasts required to differentiate words from nonsense words. They found that luminance contrasts ranging from 6.2% to 7.8% were sufficient, while performance dropped off sharply for lower luminance contrasts. Campbell and Robson (1968) and Kelly (1979a; 1979b) investigated the level of luminance contrast required to detect the presence of spatial gratings. All three studies found that under certain conditions luminance contrast thresholds of less than 1% were sufficient. Kelly (1979b) found that the addition of motion to sine-wave gratings reduced the luminance contrast requirements by a factor of 2. Thus, it appears that the human visual system is extremely sensitive to luminance contrast, and that motion increases this sensitivity.

Previous research on the perception of motion has indicated that the effectiveness of chromatic contrast depends upon the perceptual process being investigated (short- or long-range processes) and the nature of the task requirements (Cavanagh et al., 1984; Gorea and Papathomas, 1989; Ramachandran and Gregory, 1978; Simpson, 1990). In the present experiments the visual angle that the squares inside the bar subtended were such that long-range processes of motion were being investigated, which should have increased the usefulness of chromatic information. In Experiment 4 the relative perceptual salience of the chromatic contrast was approximately double that of the luminance contrast. The results of both Experiments 3 and 4 indicate that chromatic information could be used to perform the rate-matching task, but that it played a secondary role to luminance contrast. Chromatic contrast improved performance only when sufficient luminance contrast was not available and when the chromatic contrast was relatively large. Perhaps Simpson (1990, p. 1421) most appropriately summarizes the comparative roles of chromaticity and luminance for the encoding of rate information in stating that "It might be that hue

[chromaticity] is not an especially *bad* feature for motion but that luminance is especially *good.*"

Design guidelines

As with display design guidelines in general, guidelines for animated mimic displays are difficult to formulate. Each design decision must be considered in the broader context of the overall display goals: what is the nature of information that is to be conveyed and what is its importance relative to other information in the display. A critical aspect of effective display design is to provide a hierarchical nesting of perceptual salience among the graphical elements of a display corresponding to the relative importance of the information being displayed. Thus, Tufte (1990) describes "layering" and "separation" as techniques to organize information into hierarchical levels and Cleveland (1985, p.26) recommends that "visually prominent graphical elements" be used to represent the most important information (data) in statistical graphs.

From the results of the present study it is clear that luminance contrast should be a key consideration for the graphical elements that provide apparent motion in animated mimic displays. In some cases a high degree of luminance contrast might be appropriate. For example, in a training system such as STEAMER (Hutchins et al., 1987; 1984) the causal explanations provided by the flow of information or resources between system components is perhaps the most important information for an individual who is attempting to understand the workings of a complex system. On the other hand, for status displays used in real-time control of a system rate of flow information might be less critical (for example, spot-checks to ensure that system resources are being distributed in an appropriate fashion) and lower levels of luminance contrast between the graphical elements would be more appropriate. Travis et al. (1990) recommend a luminance contrast of 50% between text and background. However, animation, in and of itself, provides visual information that is highly perceptually salient. Thus, for the design of

animated mimic displays luminance contrasts ranging from 10% to 50% between the repeating graphical elements are recommended. The choice of a particular luminance contrast will depend upon the relative importance of animated information.

Although the critical role of luminance contrast in producing accurate apparent motion has been highlighted, it should be emphasized that chromaticity can also play an important role in the design of animated functional mimic displays. The general literature on display design indicates that color-coding can be used to encode both quantitative and qualitative information (Reichman, 1986; Stokes, Wickens, and Kite, 1990; Tufte, 1990). In animated functional mimic displays chromaticity may be especially useful in representing the changes that information or resources undergo as they flow through the various components of the system. One example is an animated functional mimic display that was designed for a process control application (Bennett, in press). This display portrays the high-level system components relevant to a particularly difficult task, the manual control of feedwater in a nuclear power plant. Energy (in the form of heat) is removed from the reactor core as the coolant flows through the steam generator. The temperature of coolant is higher when it enters the steam generator (the hot leg) than when it leaves the steam generator (the cold leg). In the animated mimic display the change in temperature is represented by gradual decrements in the chromaticity (hue and saturation) of the graphical elements in the portion of the coolant loop inside the steam generator.

These considerations lead to a second design guideline for animated functional mimic displays. When information or resources undergo qualitative or quantitative changes as they flow through the components of the system color-coding can, and should, be used to represent these changes. However, when information does not go through changes a single level of chromatic contrast (hue and saturation) should be used, along with changes in luminance contrast to ensure accurate perception of motion. There is abundant experimental evidence indicating that inappropriate use of color can actually degrade, rather than improve the extraction of information from a display (e.g., Carter and Cahill, 1979). In the worst case, inappropriate use of color can result in a graphic display that resembles "a grim parody of a video game" (Tufte, 1990, p.88). The proposed design

solution will result in an emphasis of the qualitative or quantitative changes when appropriate, and will not result in an increase of visual clutter and confusion in the display when inappropriate.

The present investigation has studied only one issue in the design of animated configural displays. The basic literature on the perception of motion indicates that several other factors may have an influence on the effectiveness of these displays. The perceived rate of motion of sine-wave gratings has been shown to be dependent upon both the spatial and the temporal frequency of stimuli (Kelly, 1979b). What are the optimal combinations of spatial frequency (the number and size of the graphical elements) and temporal frequency (the speed at which the perceptual characteristics of the graphical elements are changed). Orientation has also been shown to influence the perception of motion (Simpson, 1990). When the flow rates of two segments of physical connections need to be compared are there particular visual angles that should be avoided, or angles that are particularly advantageous?

An alternative approach (or perhaps a complimentary approach) is to provide redundant perceptual cues that assist the observer in determining the rate of motion. Redundant visual contours that explicitly illustrate both the direction and rate of flow could be provided in animated functional mimic displays. In the present study the graphical elements in the physical connections between system components were squares. One way to provide redundant coding would be to allow the shape of these repeating elements to change as a function of the rate and direction of flow. This is illustrated in Figure 20. Figures 20a, 20b, and 20c represent the visual appearance of the repeating elements when there is no flow, a medium rate of flow from left to right, and a high rate of flow from left to right. Are there any performance advantages to encoding information concerning rate of flow information in this fashion?

An additional design question arises concerning the nature of the borders between the repeating elements. In the present experiment these borders were implicitly defined by the perceptual characteristics of the graphical elements. What effect on the perception of

motion would result if the borders were explicitly defined? On the one hand explicitly providing the borders could provide redundant coding that improves the ability to perceive rate of flow information. On the other hand, the provision of explicit visual contours between the graphical elements could impair the subjective impression of apparent motion. Studies to investigate these issues are currently underway.

FIGURE 20. Redundant coding of rate and direction of flow.

15a. No flow

15b. Medium rate of flow

15b. High rate of flow

IV. RECOMMENDATIONS:

One technique to improve computerized training and real-time decision support that is often overlooked is representational aiding, where machine power is used to create and manipulate representations of the domain. The quality of performance in complex, dynamic domains is dependent upon the relationship between three interactive and mutually constraining components: the cognitive demands produced by the domain of

interest, the cognitive agent(s) that meet those demands, and the representation of the domain through which the agent experiences and interacts with the domain (Woods and Roth, 1988). The characteristics of each component and the interactions between them determine the ease or difficulty of problem solving. With respect to the interface Woods (Woods and Roth, 1988) and Rasmussen (1986) have stressed that there can be no neutral representation: any representation that is chosen will necessarily emphasize certain aspects of the domain at the expense of others. When designed appropriately, graphic displays can be used to help the human problem solver find the relevant data in a dynamic environment, to visualize the semantics of the domain, and to restructure their view of the problem. This will be especially important during training and instruction, since an individual is explicitly learning about the domain. While technological developments have provided powerful capabilities to generate computer graphics, a clear understanding of how these capabilities can be used to support human cognition is needed.

There are several theoretical perspectives that can be used to guide the development of graphic representations. Cleveland and his colleagues (e.g., Cleveland, 1985) have investigated the visual system's effectiveness in extracting information that has been mapped into various graphical forms (e.g., area of a circle vs. length of a line, etc.). Wickens and his colleagues (Wickens and Andre, 1990; Wickens, Kramer, Barnett, Carswell, Fracker, Goettl, and Harwood, 1985) have investigated the relationship between the general information-processing capabilities of an individual, the general demands of the task, and the implications for display design. Hutchins, Hollan, and Norman (1986) describe a general theory of interface design that emphasizes the role of direct manipulation (the capability to effect changes in the domain by directly acting upon objects of interest).

A number of researchers have been investigating an alternative approach to display design for complex, dynamic domains. Although the theoretical orientations that they approach the problem from are different, and the specific conclusions and recommendations may differ slightly, they all share very similar basic beliefs. For these researchers the design of a graphic display critically depends upon matching specific perceptual and cognitive capabilities of an individual with specific characteristics of the

domain (Bennett, Toms, and Woods, in preparation; Flach and Vicente, in press; Rasmussen, 1986; Vicente and Rasmussen, 1990; Woods and Roth, 1988). In particular, the semantics of those domains (the critical variables, the relationships between these variables, and the relevant goals and constraints) must be mapped into the static appearance and dynamic behavior of the graphic displays so that critical information can be easily extracted or decoded by the individual. The research on the configural display and the animated functional mimic display was conducted from this perspective.

Experiments 1 and 2. For the extraction of high-level information regarding high-level properties the configural display significantly increased accuracy with no cost in latency. For low-level data there were no differences in accuracy, but a significant decrement in latency associated with the configural display. However, this performance decrement was dependent upon both experience and system state. These results suggest that configural displays can be designed to support the extraction of both high-level properties and low-level data in complex, dynamic domains. To support the extraction of information for high-level properties the emergent features produced by a configural display must reflect the critical data relationships that are present in the domain. To support the extraction of low-level data the graphical elements of the display must be made more salient perceptually through emphasis of scale, spatial separation, or color-coding.

Experiments 3 and 4. The perceptual characteristics of graphical elements in two bars were altered to produce apparent motion; the observers' task was to match the rate of flow in the two bars. The perceptual characteristics of the graphical elements were systematically altered, factorially combining levels of luminance and chromatic contrast. In both experiments it was found that observers relied primarily upon luminance contrast, even though in Experiment 2 the perceptual salience of the luminance contrast was lower than that of the chromatic contrast. From the results of these studies it is clear that luminance contrast should be a key consideration for the graphical elements that provide apparent motion in animated mimic displays. Thus, for the design of animated mimic displays luminance contrasts ranging from 10% to 50% between the repeating graphical elements are recommended. The choice of a particular luminance contrast will depend

upon the relative importance of animated information. However, when information or resources undergo qualitative or quantitative changes as they flow through the components of the system col oding can, and should, be used to represent these changes.

The principles that these researchers use to guide the design of graphic displays for skilled performance are applicable to the design of graphic displays for the acquisition of cognitive skills. However, the design of graphic displays for automated instruction places additional requirements for integrated sets of displays. Displays also need to be designed to facilitate the transition from an initial understanding of the domain semantics (e.g., animated functional mimics) to a more advanced conceptualization that approximates that of expert domain practitioners (e.g., configural displays). In addition, when the graphic displays in the training system are not available on the target system, sets of displays need to be designed that facilitate the transfer of training to the target system. Advances in the design of graphic displays will contribute significantly to the effectiveness of computer-based training and real-time decision support.

REFERENCES

Anstis, S. M. (1980). The perception of apparent movement. Phil. Trans. R. Soc. Lond., B(290), 153-168.

Andre, A. D., and Wickens, C. D. (1988). The interaction of spatial and color proximity in aircraft stability information displays. In Proceedings of the Human Factors Society 32th Annual Meeting (pp. 1371-1375). Santa Monica, CA: Human Factors Society.

Barnett, B. J., and Wickens, C. D. (1988). Display proximity in multicue information integration: The benefits of boxes. Human Factors, 30, 15-24.

Bennett, K. B. (In press). The use of on-line guidance, graphic displays, and discovery learning to improve the effectiveness of simulation training. In W. Regian and V. Shute (Eds.), Cognitive Approaches to Automated Instruction. Hillsdale, N.J.: Lawrence Earlbaum Associates.

Bennett, K. B., Toms, M. L., and Woods, D. D. Emergent features and configural elements: Designing more effective configural displays. Manuscript submitted for publication.

Bennett, K. B., Woods, D. D., Roth, E. M., and Haley, P. H. (1986). Predictor displays for complex, dynamic tasks: A preliminary investigation. In Proceedings of the 30th Annual Meeting of the Human Factors Society, 684-688.

Braddick, O. J. (1974). A short-range process in apparent motion. Vision Research, 14, 519-527.

Braddick, O. J. (1980). Low-level and high-level processes in apparent motion. Phil. Trans. R. Soc. Lond., B(290), 137-151.

Brown, R. L. (1985). Methods for graphic representation of systems of simulated data. Ergonomics, 28, 1439-1454.

Burton, R. R. (1988). The environment module of intelligent tutoring systems. In M. C. Polson, and J. J. Richardson (Eds.), Foundations of Intelligent Tutoring Systems (pp.

109-142). Hillsdale, N. J.: Lawrence Earlbaum Associates.

Campbell, F. W., and Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. Journal of Physiology, 197, 551-566.

Carney, T., Shadlen, M., and Switkes, E. (1987). Parallel processing of motion and colour information. Nature, 328, 647-649.

Carswell, C. M. (1988). Integral, configural, and unitary graphs. In Proceedings of the Human Factors Society 32th Annual Meeting (pp. 1345-1349). Santa Monica, CA: Human Factors Society.

Carswell, C. M., and Wickens, C. D. (1987). Information integration and the object display. Ergonomics, 30, 511-527.

Carter, R.C., and Cahill, M.C. (1979). Regression models of search time for color-coded information displays. Human Factors, 21, 293-302.

Casey, E. J., and Wickens, C. D. (1986). Visual display representation of multidimensional systems: The effect of information correlation and display integrality (Tech. Report CPL-86-2). Urbana-Champaign: Cognitive Psychophysiology Laboratory, University of Illinois.

Cavanagh, P., Tyler, C. W., and Favreau, O. E. (1984). Perceived velocity of moving chromatic gratings. Journal of the Optical Society of America, 1(8), 893-899.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association, 68, 361-368.

Cleveland, W. S. (1985). The elements of graphing data. Belmont, Ca.: Wadsworth.

Coury, B. G., Boulette, M. D., and Smith, R. A. (1989). Effect of uncertainty and diagnosticity on classification of multidimensional data with integral and separable displays of system status. Human Factors, 31(5), 551-570.

Crossman, E. R. F. W., and Cooke, J. E. (1974). Manual control of slow-response systems. In E. Edwards and F. P. Lees (Eds.), The Human Operator in Process Control. London:

Taylor and Francis Ltd.

Flach, J. M., and Vicente, K. J. (in press). Complexity, difficulty, direct manipulation and direct perception. International Journal of Man-Machine Studies.

Goettl, B. P., Kramer, A. F., and Wickens, C. D. (1986). Display format and the perception of numerical data. In Proceedings of the Human Factors Society 30th Annual Meeting (pp. 450-454). Santa Monica, CA: Human Factors Society.

Goldsmith, T. E., and Schvaneveldt, R. W. (1984). Facilitating multiple-cue judgements with integral information displays. In J. C. Thomas and M. L. Schneider (Eds.), Human Factors in Computer Systems. Norwood, NJ: Erlbaum.

Goodstein, L. (1981). Discriminative display support for process operators. In J. Rasmussen and W. B. Rouse (Eds.), Human Detection and Diagnosis of System Failures. New York: Plenum.

Gorea, A., and Papathomas, T.V. (1989). Motion processing by chromatic and achromatic visual pathways. Journal of the Optical Society of America A, 6, 590-602.

Hawkins, J.S., Reising, J.M., and Gilmore, J.D. (1983). Pictorial format display evaluation. In Proceedings of the National Aerospace and Electronics Conference (NAECON) 1983, 1132-1138. New York: IEEE.

Hollan, J. D., Hutchins, E. L., and Weitzman, L. (1987). Steamer: An interactive inspectable simulation-based training system. In G. Kearsley (Ed.), Artificial Intelligence & Instruction: Applications and Methods. Reading, Ma: Addison-Wesley.

Hollan, J. D., Hutchins, E. L., and Weitzman, L. (1984). Steamer: An interactive inspectable simulation-based training system. The AI Magazine, Summer, 15-27.

Hollnagel , E. and Woods, D. D. (1983). Cognitive systems engineering: New wine in new bottles. International Journal of Man-Machine Studies, 18,583-600.

Ivry, R. B., and Cohen, A. (1990). Dissociation of short- and long-range apparent motion in visual search. Journal of Experimental Psychology: Human Perception and Perfor-

mance, 16(2), 317-331.

Kahneman, D., and Henik, A. (1977). Effects of visual grouping on immediate recall and selective attention. In S. Dornic (Ed.), Attention and performance, VI. Hillsdale, N. J.: Erlbaum.

Kahneman, D. and Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman and D. R. Davies (Eds.), Varieties of Attention. New York, N.Y.: Academic Press.

Kelly, D. H. (1979a). Motion and vision. I. Stabilized images of stationary gratings. Journal of the Optical Society of America, 69, 1266-1274.

Kelly, D. H. (1979b). Motion and vision. II. Stabilized spatio-temporal threshold surface. Journal of the Optical Society of America, 69, 1340-1349.

Kleiner, B., and Hartigan, J. A. (1981). Representing points in many dimensions by trees and castles. Journal of the American Statistical Association, 76, 260-269.

Lindsey, D. T., and Teller, D. Y. (1990). Motion at isoluminance: Discrimination/detection ratios for moving isoluminant gratings. Vision Research, 30(11), 1751-1761.

Livingstone, M. S., and Hubel, D. H. (1987). Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. Journal of Neuroscience, 7(11), 3416-3468.

MacGregor, D., and Slovic, P. (1986). Graphic representation of judgmental information. Human-Computer Interaction, 2, 179-200.

Merrifield, R. M., and Silverstein, L. D. (1986). The development and evaluation of color systems for airborne applications: Fundamental visual, perceptual, and display systems considerations (NADC-86011-60). Warminster, PA: Naval Air Development Center.

Moray, N., Loowsteen, P., and Pajak, J. (1986). Acquisition of process control skills. IEEE Transactions on Systems, Man, and Cybernetics, SMC-16, 497-504.

Moray, N. (1987) Intelligent aids, mental models and the theory of machines. International

Journal of Man-Machine Studies, 27.

Naveh-Benjamin, M., and Pachella, R. G. (1982). The effect of complexity on interpreting 'Chernoff' faces. Human Factors, 24, 11-18.

Norman, D. A. (1986). Cognitive engineering. In D. A. Norman, and S. W. Draper (Eds.), User centered system design. Hillsdale, N. J.: Lawrence Earlbaum Associates.

Pomerantz, J. R. (1986). Visual form perception: An overview. In H. C. Nusbaum and E. C. Schwab (Eds.), Pattern Recognition by Humans and Machines (Vol. 2 Visual Perception). Orlando, Florida: Academic Press.

Ramachandran, V.S., and Gregory, R.L. (1978). Does color provide an input to human motion perception? Nature, 275(7), 55-56.

Rasmussen, J. (1986). Information processing and human-machine interaction: An approach to cognitive engineering. New York: North Holland.

Reichman, R. (1986). Communication paradigms for a window system. In D. A. Norman, and S. W. Draper (Eds.), User centered system design. Hillsdale, N. J.: Lawrence Earlbaum Associates.

Roth, E. M., and Woods, D. D. (1988). Aiding human performance: I. Cognitive analysis. Le Travail Humain, 51(1), 39-64.

Sanderson, P. M., Flach, J. M., Buttigieg, M. A., and Casey, E. J. (1989). Object displays do not always support better integrated task performance. Human Factors, 31(2), 183-198.

Simpson, W. A. (1990). The use of different features by the matching process in short-range motion. Vision Research, 30(10), 1421-1428.

Stokes, A., Wickens, C., Kite, K. (1990). Display technology- human factors concepts. Warrendale, PA: Society of Automotive Engineers, Inc.

Travis, D.S., Bowles, S., Seton, J., and Peppe, R. (1990). Reading from color displays: A psychophysical model. Human Factors, 32(2), 147-156.

Treisman, A. M., and Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12, 97-136.

Tufte, E. R. (1990). Envisioning Information. Chesire, Connecticut: Graphics Press.

Vicente, K. J., and Rasmussen, J. (1990). The ecology of human-machine systems II: Mediating "direct perception" in complex work domains. Ecological Psychology, 2(3), 207-249.

Wickens, C. D., and Andre, A. D. (1988). Proximity compatibility and the object display. In Proceedings of the Human Factors Society 32th Annual Meeting (pp. 1335-1339). Santa Monica, CA: Human Factors Society.

Wickens, C. D., and Andre, A. D. (1990). Proximity compatibility and information display: Effects of color, space, and objectness on information integration. Human Factors, 32, 61-78.

Wickens, C. D., Kramer, A., Barnett, B., Carswell, M., Fracker, L., Goettl, B., and Harwood, K. (1985). Display-cognitive interface: The effect of information integration requirements on display formatting for C3 displays (Tech. Report EPL-85-3). Urbana-Champaign: Engineering Psychology Research Laboratory and Aviation Research Laboratory, University of Illinois.

Woods, D. D. (1991). The cognitive engineering of problem representations. In G. R. S. Weir and J. L. Alty (Eds.), Human-Computer Interaction and Complex Systems. London: Academic Press.

Woods, D. D. and Hollnagel, E. (1987). Mapping cognitive demands in complex problem-solving worlds. International Journal of Man-Machine Studies, 26, 257-275.

Woods, D. D., and Roth E. M. (1988). Cognitive systems engineering. In M. Helander (Ed.), Handbook of human-computer interaction. New York: North-Holland.

Woods, D. D., Wise, J. A., and Hanes, L. F. (1981). An evaluation of nuclear power plant safety parameter display systems. In Proceedings of the Human Factors Society 25th Annual Meeting. Santa Monica, CA: Human Factors Society.

Zachary, W. (1986). A cognitively based functional taxonomy of decision support techniques. Human-Computer Interaction, 2, 25-63.

Report # 112
760-7MG-100
Prof. Ronna Dillon
Report Not Publishable

Report # 113
760-6MG-134
Prof. Stephen Loy
No Report Submitted

1989 RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

ADVANCING USER INTERFACE CAPABILITIES IN AN INTEGRATED INFORMATION ENVIRONMENT: A FISHEYE BROWSER

Prepared by:	Deborah A. Mitta
Academic Rank:	Assistant Professor
Department and	Industrial Engineering
University:	Texas A&M University
USAF Laboratory:	AFHRL/LRC Wright-Patterson AFB Dayton, OH 45433
USAF Researchers:	Lieutenant Cher Wynkoop David Gunning Robert Johnson Donald Thomas
Date:	31 December 1990
Contract No.	F49620-88-C-0053/SB5881-037B

ADVANCING USER INTERFACE CAPABILITIES IN AN INTEGRATED INFORMATION ENVIRONMENT: A FISHEYE BROWSER

by

Deborah A. Mitta

ABSTRACT

Human-computer interface issues associated with the Integrated Maintenance Information System (IMIS) are of significant importance. One interest is in improving the quality of human-computer interaction within the IMIS environment. Of primary concern is the presentation of information via electronic media, in particular, presentation techniques that enhance the electronic display of graphics-based aircraft maintenance. A presentation technique known as the fisheye lens viewing strategy has been considered as a mechanism for abbreviating unnecessary details associated with graphics-based aircraft maintenance data. It allows information closely associated with an item of interest (focus point) to be presented and enables a viewer to gain perspective on the focus point with respect to the larger system of which it is a part.

This report documents the development of the concept of focus relationship selection. Information selection of this type enables relationships between informational database elements to be specified and subsequent fisheye views to be displayed. In addition this type of information selection and presentation is suggested as a means of browsing graphics-based aircraft maintenance data. The research issue is to establish a presentation strategy that is (1) based upon focus relationship selection and (2) analogous to the fisheye concept associated with focus point selection. The focus relationship selection concept is demonstrated on a subset of maintenance data. A user interface facilitating the selection of focus relationships and the subsequent browsing of maintenance information is discussed, as is a prototype system incorporating the selected data subset and the interface design.

ACKNOWLEDGMENTS

I wish to thank the Air Force Systems Command, Air Force Office of Scientific Research, for sponsorship of this research. I also thank Mr. Bertram Cream, Technical Director of the Logistics and Human Factors Division, Air Force Human Resources Laboratory, Wright-Patterson AFB, for considering my 1989 summer research effort of sufficient merit to warrant further support.

The efforts of many individuals within the Combat Logistics Branch have not gone unnoticed. I thank Mr. David Gunning for his continued interest in my research efforts; his comments, suggestions, and support have been greatly appreciated. I also thank Lieutenant Cher Wynkoop for her help in answering many of my questions concerning the logistical details of this contract. A special word of thanks goes to Mr. Robert C. Johnson (Branch Chief) and Dr. Donald L. Thomas (Senior Scientist) for their commitment to the research concepts presented in this report.

I. INTRODUCTION:

The Air Force Human Resources Laboratory is currently developing an integrated computer-based information system to aid in tasks associated with aircraft maintenance. This system is known as the Integrated Maintenance Information System (IMIS); its purpose is to provide a comprehensive information system such that existing aircraft maintenance information systems and databases are consolidated. IMIS will provide a maintenance technician with a direct link to various maintenance information systems and databases such as supply data, historical databases, and automated technical orders. IMIS will provide diagnostic/troubleshooting recommendations, test procedures, appropriate graphics (e.g. locator diagrams, schematics) and enable a technician to obtain fault data from built-in tests. Eventually IMIS will provide specialized data for aircraft battle damage assessment tasks, enable technicians to order parts from supply, and feature an automated training capability.

Its developers suggest that IMIS will benefit Air Force maintenance personnel as a result of its abilities to (1) enable ready and easy access to a comprehensive set of data, including management data and job aids, (2) enhance the diagnostic capabilities of maintenance technicians, and (3) reduce the amount of training required for technicians to learn aircraft maintenance tasks. Thus, researchers and developers of the IMIS concept propose that IMIS will enhance technician productivity.

Combat Logistics Branch personnel (Logistics and Human Factors Division) recognize that human-computer interface issues associated with IMIS are of significant importance, and one interest is in improving the quality of human-computer interaction. One human-computer interaction issue of primary concern is information presentation, in particular, presentation techniques that enhance the display of graphics-based aircraft maintenance information. With respect to information presentation, two interaction scenarios are of particular importance. These scenarios occur when (1) the size of the display medium restricts the amount of information that can be displayed and (2) information contains inappropriate levels of detail. Two traditional interface design approaches for data access in these scenarios are "scrolling" and "zoom lens" facilities. One problem, however, is that

while scrolling and zooming actions enable access to detailed information, the views resulting from these actions provide no overall perspective.

One technique recently developed as a means of preserving global perspective while allowing information abbreviation or filtering is known as the fisheye lens viewing strategy (Furnas, 1982, 1986). This technique allows detailed information associated with a particular item of interest (focus point) to be presented; it also allows a viewer to gain perspective on the focus point with respect to the larger system of which it is a part.
In order for information to be presented from a fisheye perspective, the underlying structure of the associated informational database must be represented in network form. In such a network, nodes define information elements within the database, and branches identify relationships between database elements. Each network node is assigned a metric known as its degree of interest (DOI). The magnitude of a DOI reflects a presentation value or priority assigned to the information contained at the respective node and indicates if the information is of sufficient importance for presentation on a display medium. Furnas (1986) demonstrates the fisheye strategy on hierarchical tree graphs and bases the viewing strategy upon selection of a single focus point.

The initial interest in examining the fisheye strategy is in establishing a mechanism for abbreviating information and filtering detail from maintenance data. The fisheye presentation strategy has recently been considered as a mechanism for filtering details associated with graphics-based aircraft maintenance data. Additionally, several extensions of the original concept are reported (Mitta, 1989, 1990a, 1990b). The first extension allows the fisheye technique to be applied to any type of informational network (rather than solely to tree graphs); the second extension illustrates that fisheye views resulting from the selection of multiple focus points are possible.

In order to build fisheye presentations from any type of network structure and incorporate multiple focus point selection, the following function is required. For a network consisting of the set of nodes $N = \{y_1, y_2, \ldots, y_{n-1}, y_n\}$, a presentation value for any node in N is determined according to the following function:

$$V_k = I_k - \sum_j D_{k,j}, \quad (1)$$

where

V_k = presentation value of node y_k

I_k = importance rating of node y_k

$D_{k,j}$ = minimum path distance between node y_k and focus point y_j .

Equation (1) represents a slight modification of Furnas' original DOI function (Mitta, 1989, 1990a, 1990b). The variable V_k is analogous to the degree of interest metric and remains a function of importance and distance; however, under the conditions of multiple focus point selection, the minimum path distance to each focus point must be considered.

Note that the information abbreviation concepts addressed by this research are also addressed by the Aerospace Industries Association's recent initiative to simplify the content of graphics used in technical documentation. The Aerospace Industries Association (1989) contends that the simplification of detailed, graphics-based information will result in a savings in creating, storing, and transmitting graphics but will not deter from the utility of these graphics.

II. OBJECTIVES OF THE RESEARCH EFFORT:

The original fisheye concept revolves around the selection of focus points (network nodes). Based on the selected focus points and a given DOI (or V_k) threshold, information abbreviation occurs. The next question one might ask is, "Suppose the selection of network arcs, rather than selection of network nodes (focus points), is desirable?" In other words the information of interest during an interaction scenario is a set of relationships between database elements. In fact one might envision an IMIS troubleshooting scenario in which a technician suspects that a particular chip on a circuit board is causing a problem. In order to understand the extent of this chip's influence with respect to the remaining circuitry, he might select an IS_CONNECTED_TO relationship.

Recall that aircraft maintenance data is represented as a relational database. In addition to providing a mechanism for the selection of information elements (network nodes), an

accompanying fisheye interface would enable users to specify the links between information elements and in turn view respective fisheye presentations. Thus, at this stage of the research, another human-computer interaction issue meriting further examination is the selection of focus relationships.

The primary research objectives are to investigate (1) the concept of focus relationship selection and (2) the subsequent presentation of associated information. This type of information selection and presentation is suggested as a means of browsing graphics-based aircraft maintenance data. This final report documents the development of the focus relationship selection concept. Specifically, the concept was implemented on a subset of IMIS maintenance data. (A simple circuit schematic and the component diagram representing Hydraulic System 1 of the F/A-18 were selected as candidate data items.) An interface facilitating the selection of focus relationships and the subsequent browsing of maintenance information was designed. A prototype system incorporating the selected data subset and the interface design was developed.

III. FOCUS RELATIONSHIP SELECTION:

One side issue that must be addressed if a relationship selection feature is incorporated into a fisheye interface is network connectivity. Given a connected graph G consisting of p nodes and q branches, let its branch set be defined as $\mathbf{B} = \{b_1, b_2, \ldots, b_q\}$. Since G is connected, it contains at least one spanning tree (Behzad, Chartrand, and Lesniak-Foster, 1979). By definition a spanning tree has p - 1 branches; therefore, $q \geq p - 1$. Let a subset of **B**, $\mathbf{B}_s$, be defined such that $\mathbf{B}_s = \{b_{s_1}, b_{s_2}, \ldots, b_{s_n}\}$, and $n \leq q$. Since, by definition, a spanning tree must have p - 1 branches, if $n < p - 1$, the resultant graph containing p nodes and branches of $\mathbf{B}_s$ will be disconnected. For $n \geq p - 1$, the graph formed by p nodes and branches of $\mathbf{B}_s$ will be connected only under the condition that $\mathbf{B}_s$ contains a spanning tree. An examination of how the fisheye strategy should be implemented under conditions of network disconnectedness is a topic for further research. Note that the subgraph formed by $\mathbf{B}_s$ may also be disconnected.

To establish a set of initial conditions, consider a connected graph G′ with r′ types of

relationships. (A connected graph is one for which each node is linked to at least one other node via an arc.) The set of relationship types in G′ is defined as follows:

$$\mathbf{R} = \{r_1, r_2, \ldots, r_{r'}\}. \tag{2}$$

Let $\mathbf{R}_s$, a subset of **R**, be defined as follows:

$$\mathbf{R}_s = \{r_{s_1}, r_{s_2}, \ldots, r_{s_n}\}. \tag{3}$$

Here, $\mathbf{R}_s$ is the set containing n types of focus relationships. Focus relationship selection requires consideration of an important issue. Once a focus relationship set is identified, a connected subgraph resulting from this focus relationship set is not guaranteed. Consider, for example, the tree graph of Figure 1. Suppose this network represents a system hierarchy, where the relationships between subsystems and sub-subsystems (indicated with heavy line widths) are of interest, that is, they are the focus relationships. The resultant subgraph is disconnected. The research issue is to implement a presentation strategy that is (1) based upon focus relationship selection and (2) analogous to the fisheye concept associated with focus point selection.

A functional expression for establishing the presentation value of a network relationship is required, where presentation value is determined with respect to a set of focus relationships. Consider a focus relationship r_j and the respective set of subgraphs consisting of r_j. Let the presentation value of a given relationship r_k be defined in terms of (1) its importance and (2) the distance between its respective subgraph and an r_j subgraph. The presentation value function is given as follows:

$$P_k = I_k - \sum_j \sum_i D_{k_i, r_j}, \tag{4}$$

where

P_k = presentation value of relationship r_k

I_k = importance of relationship r_k

D_{k_i, r_j} = minimum path distance between the ith connected subgraph consisting of r_k and a subgraph consisting of focus relationship r_j.

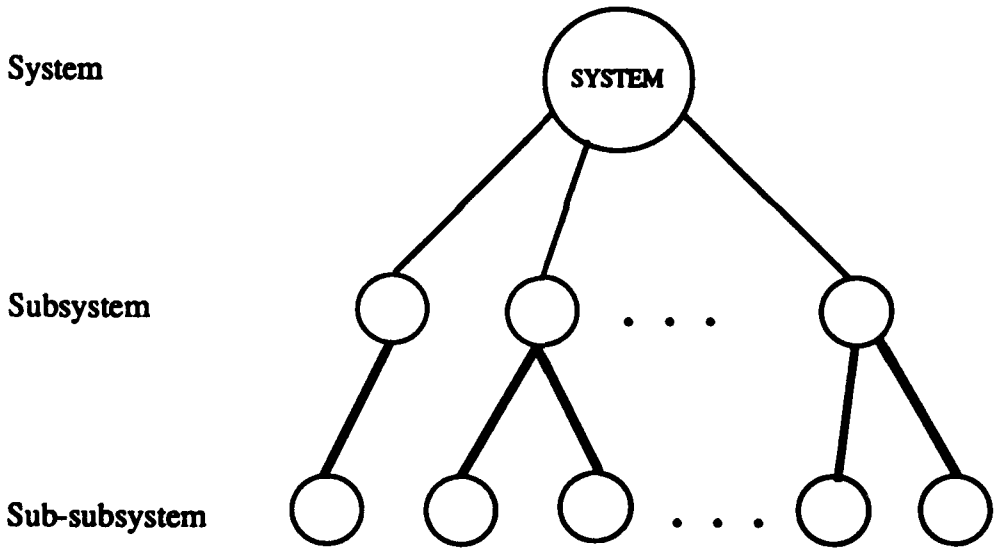


Figure 1. Tree graph representation of a system hierarchy.

Thus, as indicated by Equation (4) the presentation value of r_k increases with importance and decreases with distances between r_k subgraphs and focus relationship subgraphs.

At this point a comment on the assignment of importance ratings to network relationships is perhaps appropriate. In the examples provided in this report, hierarchical tree graphs are considered, and the parameter I_k is defined with respect to these types of acyclic network structures. The parameter I_k is weighted according to the number of node pairs spanned by relationship r_k at a given depth in the hierarchy d. Consider a general tree graph (Figure 2). The hierarchical structure associated with this graph has m levels. The root node is assigned to level m, and nodes having the greatest path distance from the root are assigned to level one. For a hierarchy consisting of m levels, $1 \leq d \leq m-1$. For a relationship spanning hierarchical levels one and two, $d = 1$; a relationship spanning hierarchical levels two and three has a depth of $d = 2$. In general, then, given a relationship spanning levels m' and $m'+1$, $d = m'$. For a hierarchical tree graph of m levels, relationship importance I_k is defined as follows:

$$I_k = \sum_{d=1}^{m-1} n_{k_d} d, \tag{5}$$

where

I_k = importance of relationship r_k

d = depth of r_k

n_{k_d} = number of node pairs spanned by r_k at depth d

Thus, the importance of a network relationship increases with its depth in the hierarchy and the number of node pairs it spans at a given depth. From Equation (5), Equation (4) can be rewritten as follows:

$$P_k = \sum_{d=1}^{m-1} n_{k_d} d - \sum_j \sum_i D_{k_i, r_j}. \tag{6}$$

Note that in general, relationship importance must be determined heuristically. In other words for a general network with no underlying hierarchical structure, relationship importance must be defined in terms other than hierarchical depth.

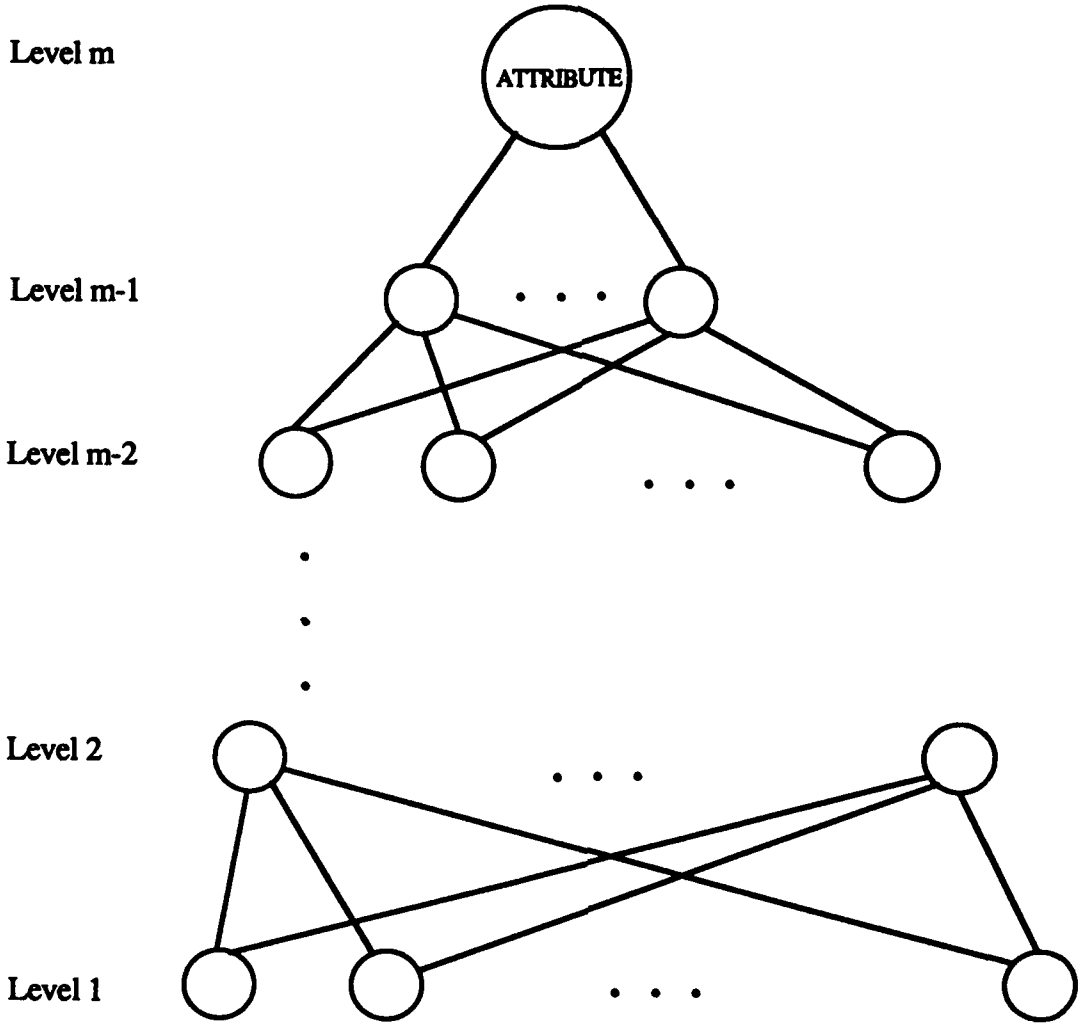


Figure 2. General tree graph.

Through the setting of P_k thresholds, fisheye views with varying degrees of information content can be provided, thereby enabling the filtering or abbreviation of information. For a given threshold level t, the network relationships satisfying the condition $P_k \geq t$ have the greatest presentation value. As t is decreased, greater amounts of information are displayed.

To demonstrate the methodology offered above, an example is provided. Consider the simple circuit schematic of Figure 3. The hierarchical structure underlying the system associated with this schematic is provided in Figure 4. Note that for the graph of Figure 4, $m = 5$ and $1 \leq d \leq 4$. As shown in Figure 4, seven hierarchical relationships are specified such that $\mathbf{R} = \{r_1, r_2, \ldots, r_7\}$. The relationships are defined as follows:

r_1: SUBSYSTEM(IN, circuit)
SUBSYSTEM(OUT, circuit)
SUBSYSTEM(internal circuitry, circuit)

r_2: IN_PORT(I_0, IN)
IN_PORT(I_1, IN)

r_3: CHIP(C_1, internal circuitry)

r_4: OUT_PORT(B_0,OUT)
OUT_PORT(B_1,OUT)

r_5: GATE(G_1, C_1)
GATE(G_2, C_1)

r_6: INPUT_PIN(i_{11}, G_1)
INPUT_PIN(i_{12}, G_1)
INPUT_PIN(i_{21}, G_2)
INPUT_PIN(i_{22}, G_2)

r_7: OUTPUT_PIN(o_1, G_1)
OUTPUT_PIN(o_2, G_2).

Consider a focus relationship set $\mathbf{R}_s = \{r_2, r_5\}$. Equation (6) is used to calculate the presentation value for each relationship: $P_1 = 11$, $P_2 = 3$, $P_3 = 1$, $P_4 = 1$, $P_5 = 1$, $P_6 = -4$,

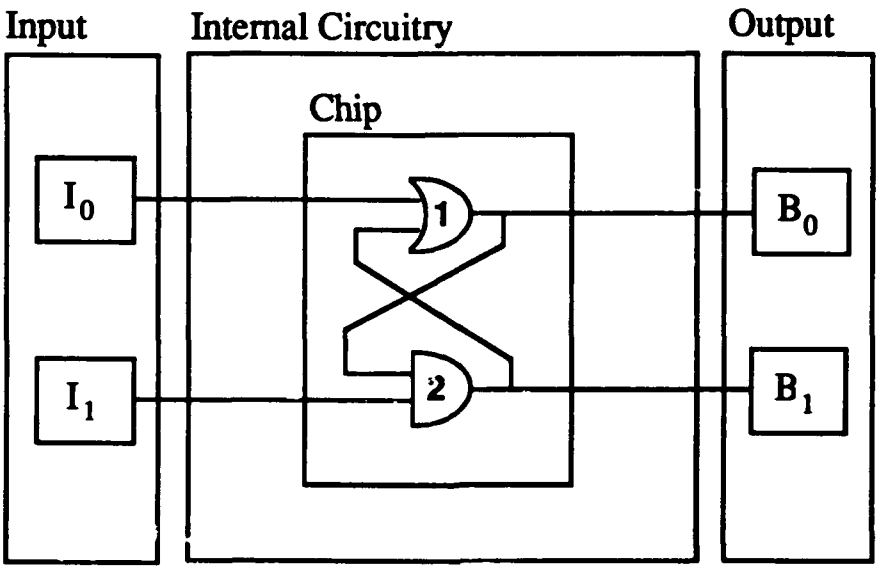


Figure 3. Circuit schematic.

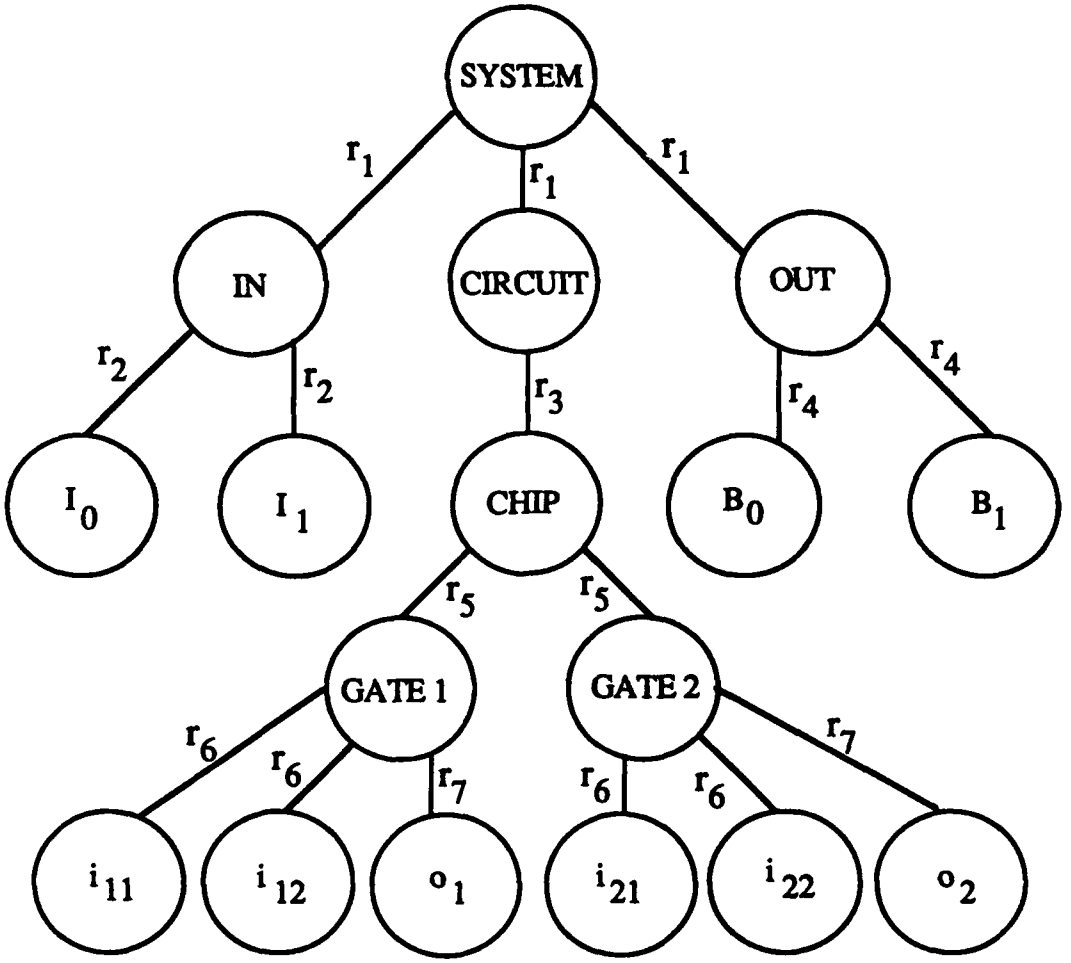


Figure 4. Hierarchical structure associated with circuit schematic.

and $P_7 = -6$. With respect to the focus relationships r_2 and r_5, r_1 has the largest presentation value (most important with respect to presentation), and r_7 has the lowest presentation value (least important with respect to presentation).

At this point, a brief explanation of several of the calculations is perhaps appropriate. Consider the presentation value of focus relationship r_2. A single connected subgraph ($i = 1$) containing r_2 exists at a depth $d = 3$. Additionally, r_2 spans two node pairs at $d = 3$; therefore, $n_{2_3} = 2$, while $n_{2_1} = n_{2_2} = n_{2_4} = 0$. Finally, the minimum path distance between the r_2 subgraph and the r_5 subgraph is three such that $D_{2_1,r_5} = 3$. Substitution into Equation (6) yields the following result:

$$P_2 = [n_{2_1}(1) + n_{2_2}(2) + n_{2_3}(3) + n_{2_4}(4)] - [D_{2_1,r_2} + D_{2_1,r_5}] \quad (7)$$

$$P_2 = [0(1) + 0(2) + 2(3) + 0(4)] - [0 + 3] = 3. \quad (8)$$

As a second example, consider the presentation value of relationship r_6. Two connected subgraphs ($i = 1, 2$) containing r_6 exist at a depth $d = 1$. Relationship r_6 spans four node pairs at $d = 1$ such that $n_{6_1} = 4$, and $n_{6_2} = n_{6_3} = n_{6_4} = 0$. The minimum path distance between each r_6 subgraph and the r_2 subgraph is equal to four ($D_{6_1,r_2} = D_{6_2,r_2} = 4$). Additionally, each r_6 subgraph is directly linked to the r_5 subgraph such that $D_{6_1,r_5} = D_{6_2,r_5} = 0$. Substitution into Equation (6) yields the following result:

$$P_6 = [n_{6_1}(1) + n_{6_2}(2) + n_{6_3}(3) + n_{6_4}(4)] - [(D_{6_1,r_2} + D_{6_2,r_2}) + (D_{6_1,r_5} + D_{6_2,r_5})] \quad (9)$$

$$P_6 = [4(1) + 0(2) + 0(3) + 0(4)] - [(4 + 4) + (0 + 0)] = -4. \quad (10)$$

Information abbreviation is enabled through the setting of thresholds. For $P_k \geq 1$, information associated with relationships r_1, r_2, r_3, r_4, and r_5 is of importance. As this threshold is successively decreased, information associated with r_6 and subsequently r_7 is of importance.

Sample fisheye views of the circuit schematic are developed in SuperCard™ 1.5 (Appleton and Poppitz, 1990) and provided in Figures 5, 6, 7, 8, and 9. Figure 5 represents

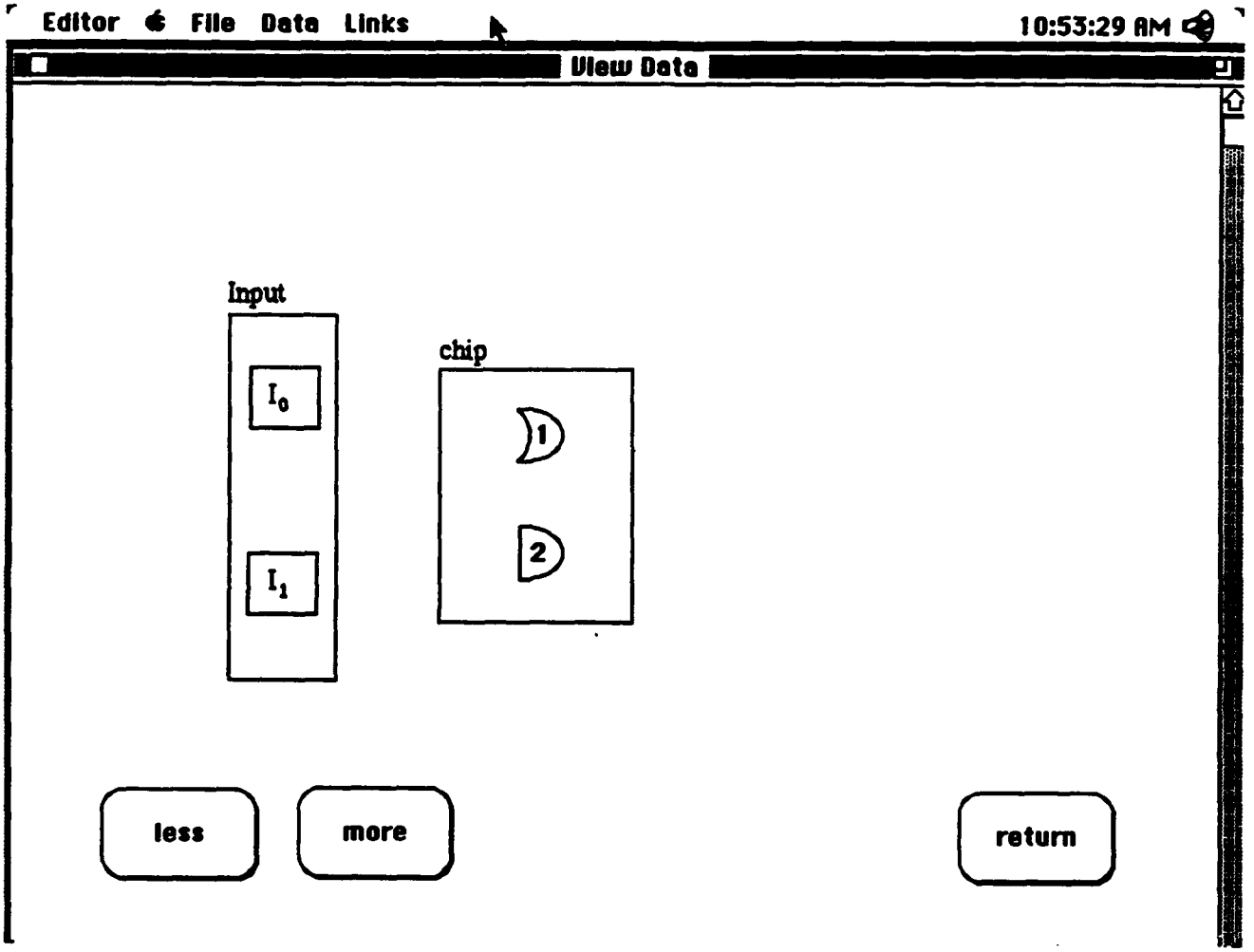


Figure 5. Information associated with focus relationships r_2 and r_5.

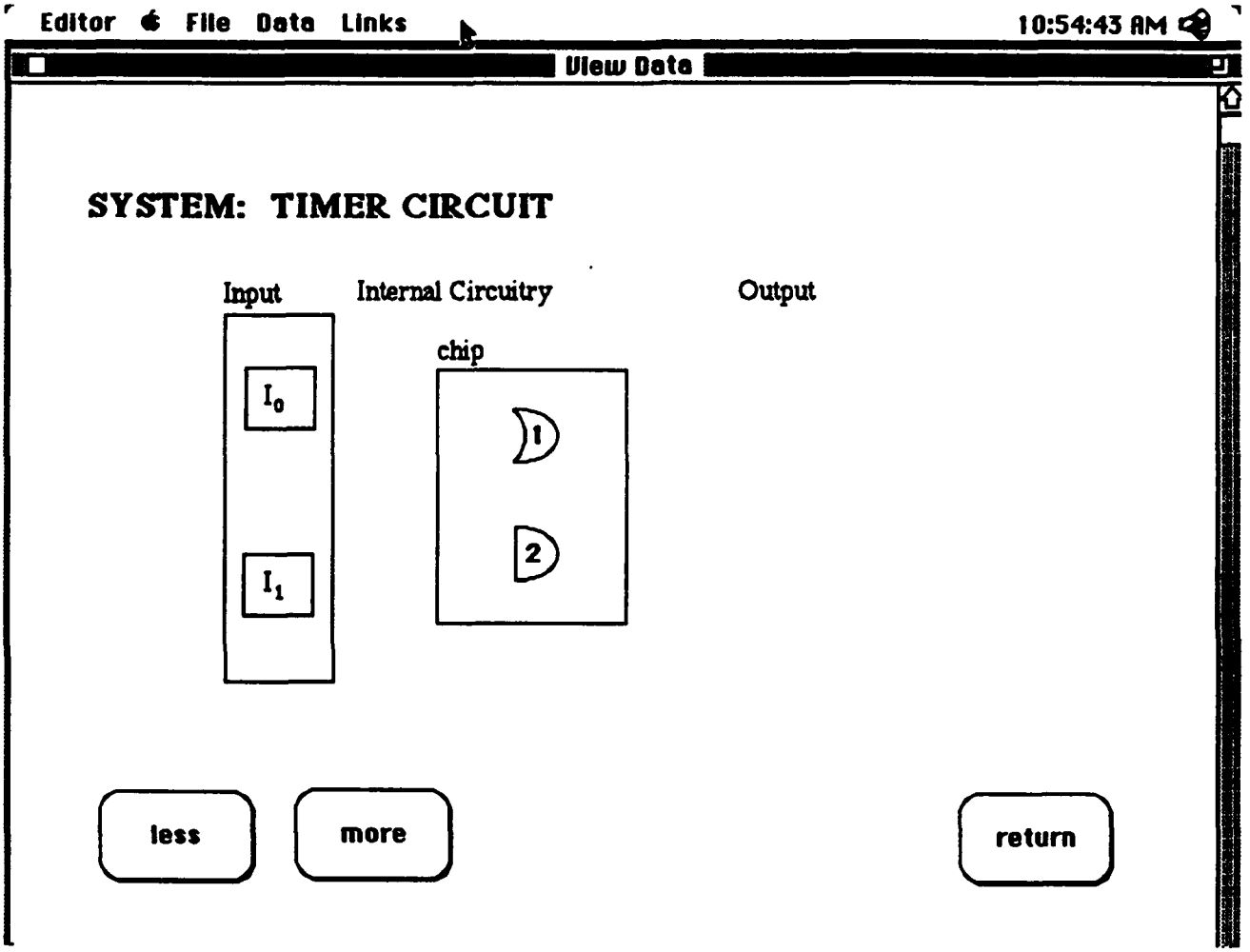


Figure 6. Information associated with a threshold $P_k \geq 11$.

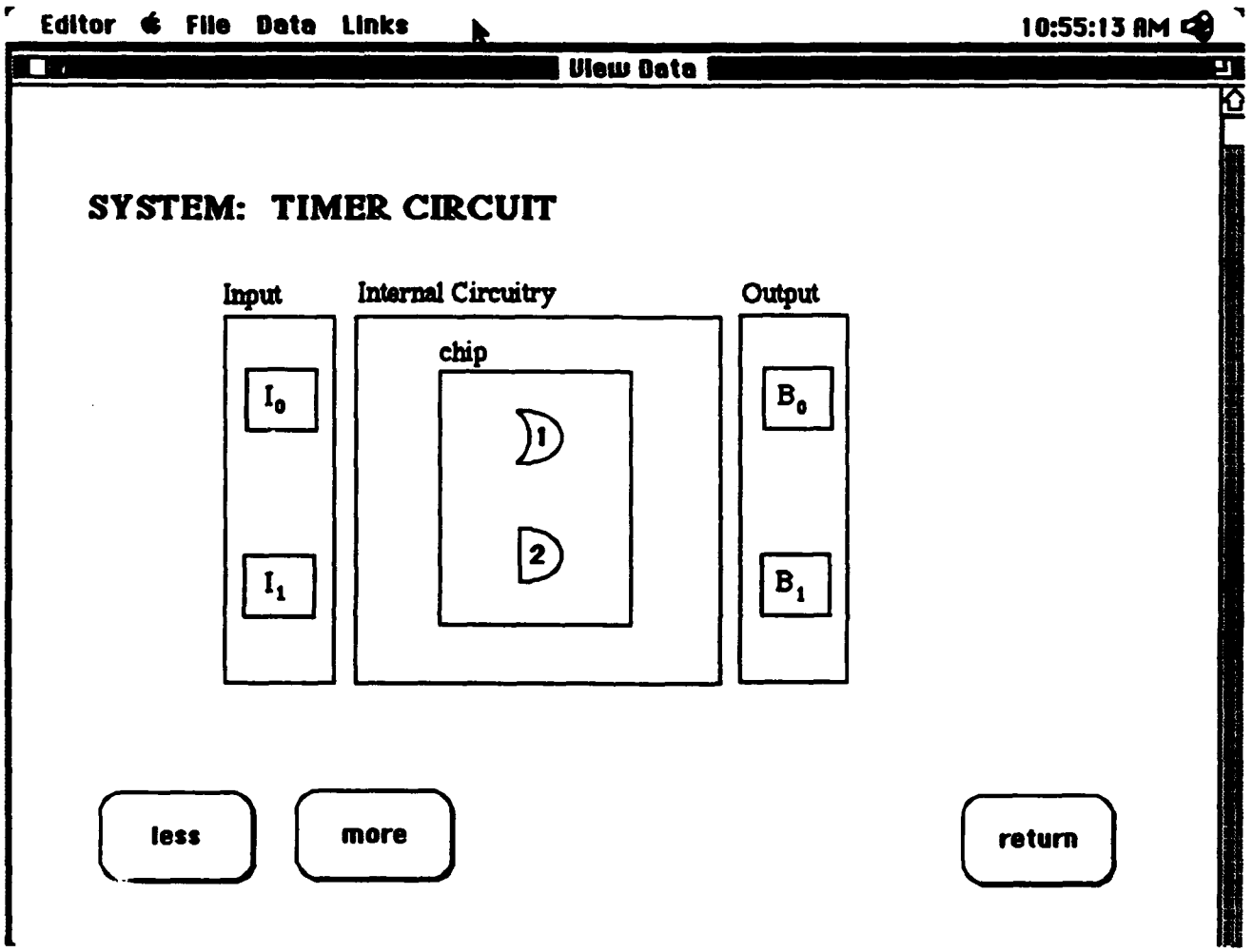


Figure 7. Information associated with a threshold $P_k \geq 1$.

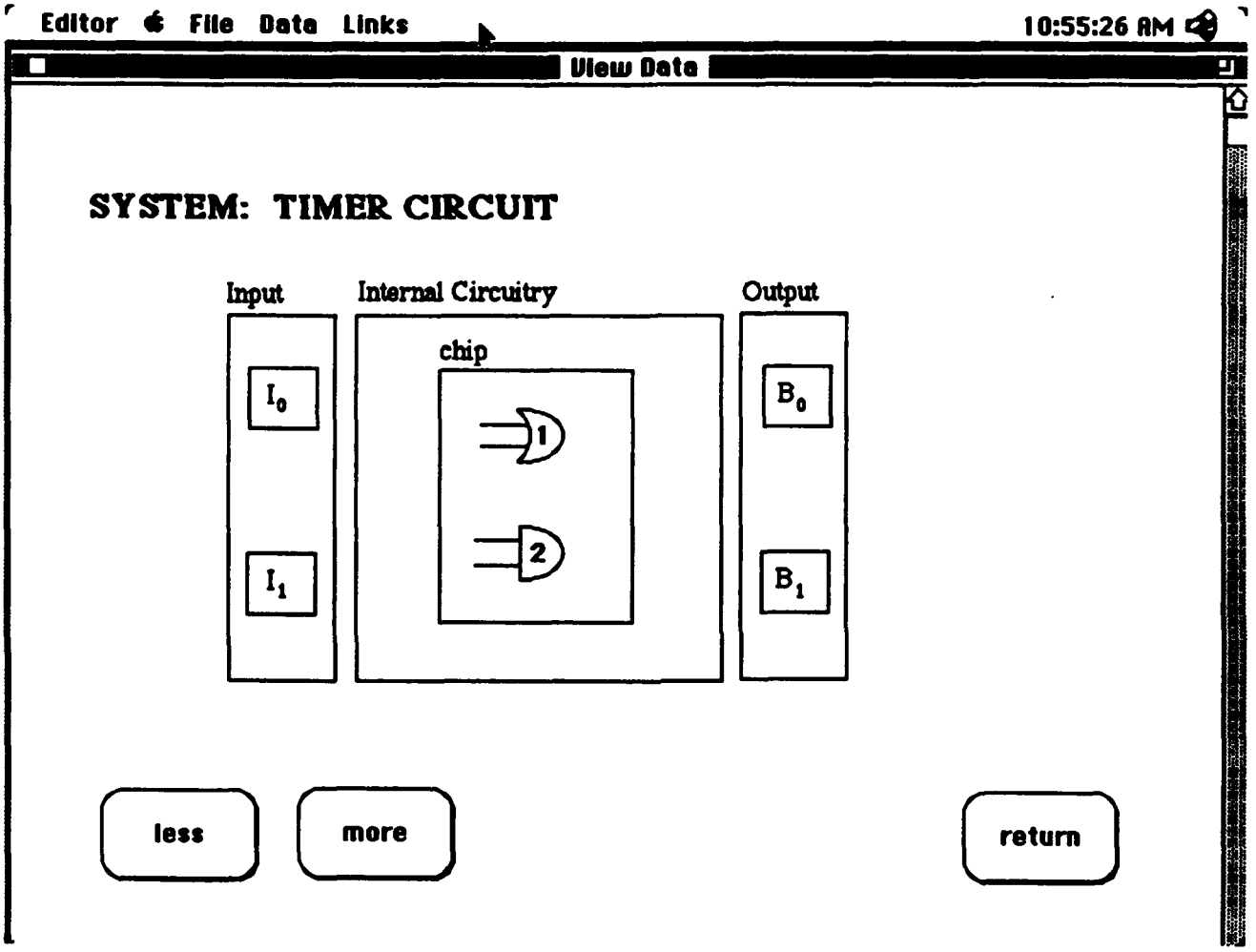


Figure 8. Information associated with a threshold $P_k \geq -4$.

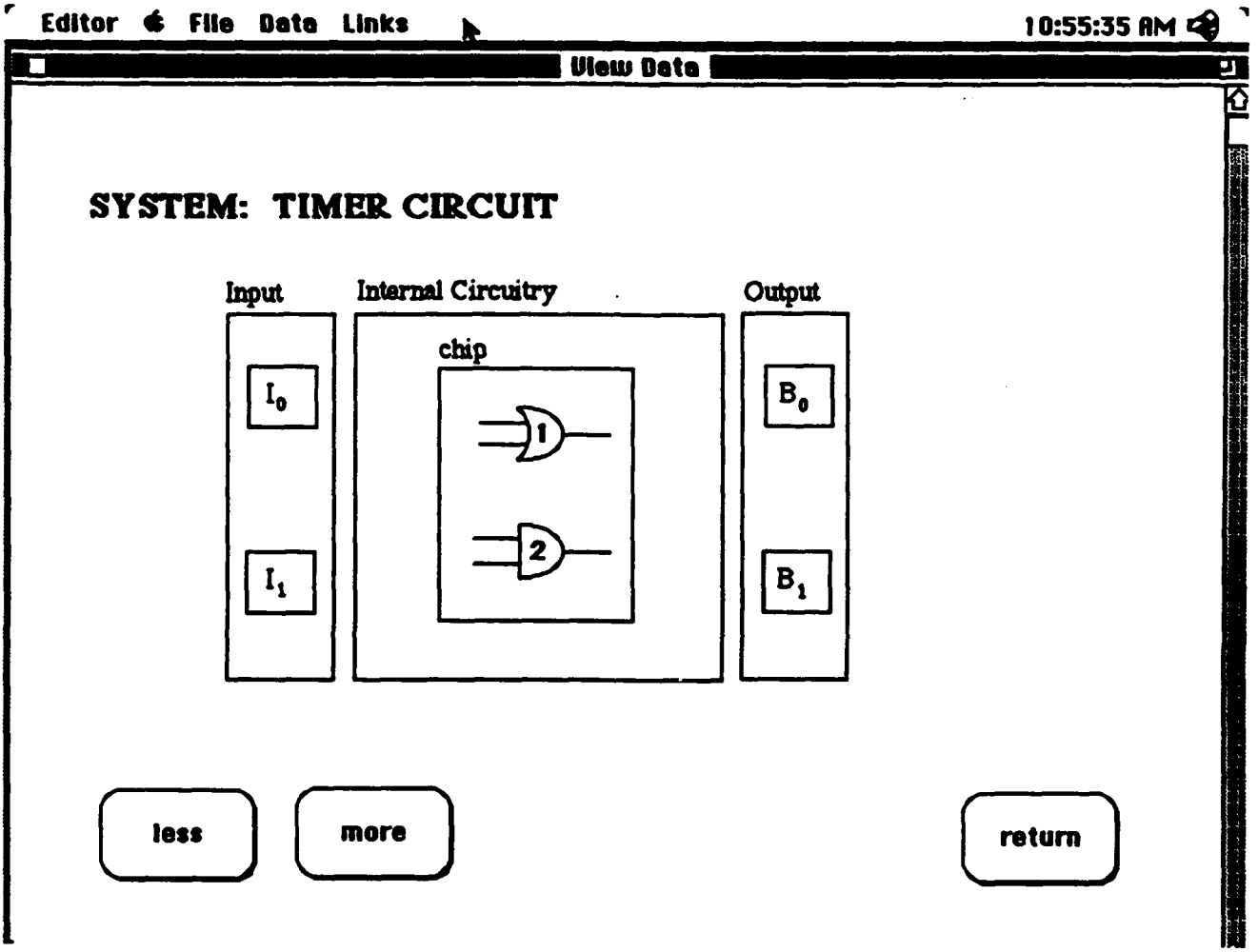


Figure 9. Information associated with a threshold $P_k \geq -6$.

information associated exclusively with the focus relationships r_2 (IN_PORT[I_0, IN], IN_PORT[I_1, IN]) and r_5 (GATE[G_1, C_1], GATE[G_2, C_1]) . Figures 6 ($P_k \geq 11$), 7 ($P_k \geq 1$), 8 ($P_k \geq -4$), and 9 ($P_k \geq -6$) demonstrate the concept of information abbreviation.

IV. APPLICATION TO HYDRAULIC SYSTEM 1:

Again, the abbreviation of graphics-based aircraft maintenance data is of interest. A component diagram of Hydraulic System 1 of the F/A-18 (Figure 10) has been selected for application of the procedure described in the previous section. The hierarchical structure underlying this system is of particular concern. Here, a portion of the hierarchical structure (a set of nine hierarchical relationships) is defined:

r_1: UNIT(fluid level indicator, hydraulic system 1)
UNIT(reservoir, hydraulic system 1)
UNIT(filter unit, hydraulic system 1)
UNIT(pressure transmitter, hydraulic system 1)

r_2: COMPONENT(piston, reservoir)
COMPONENT(valve, reservoir)
COMPONENT(switches, reservoir)

r_3: PISTON_TYPE(reservoir, piston)

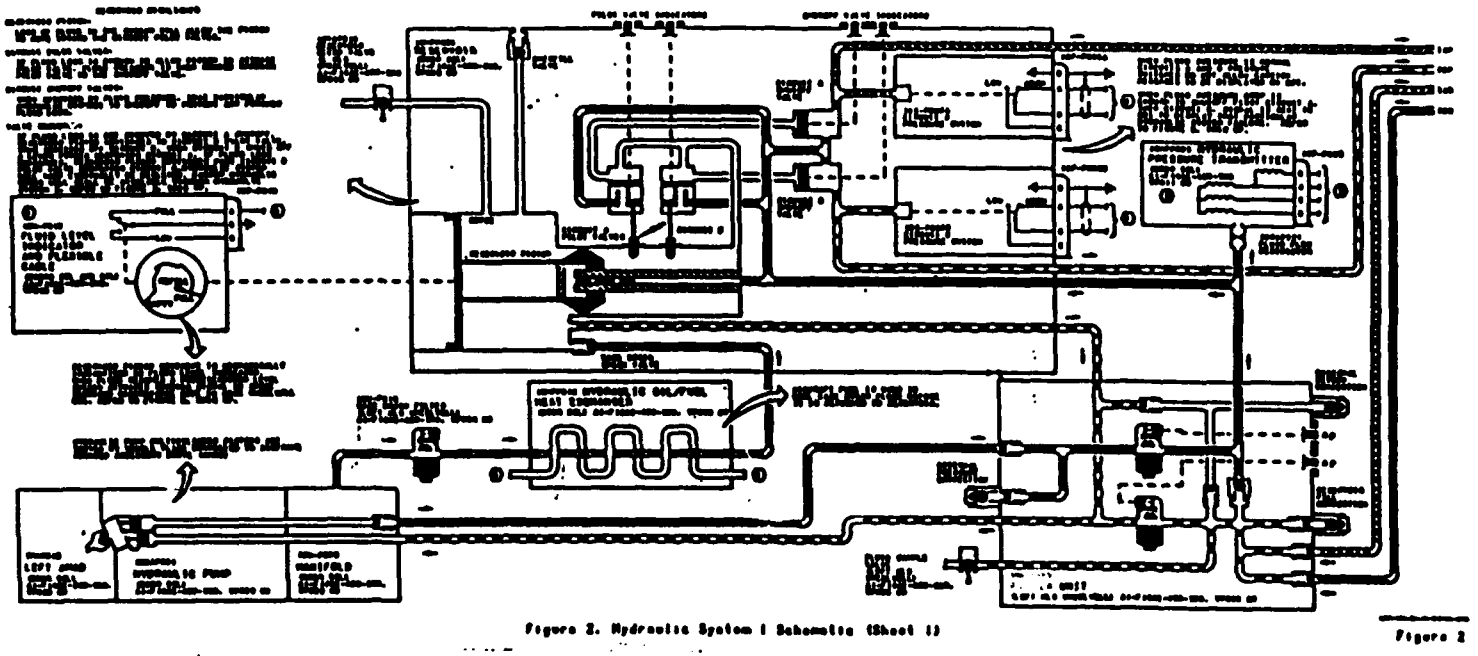
r_4: VALVE_TYPE(bleed, valve)
VALVE_TYPE(case drain check, valve)
VALVE_TYPE(overfill, valve)
VALVE_TYPE(pilot, valve)
VALVE_TYPE(shutoff, valve)

r_5: SWITCH_TYPE(pressure, switch)

r_6: PILOT_VALVE_CIRCUIT(A, pilot valve)
PILOT_VALVE_CIRCUIT(B, pilot valve)

r_7: SHUTOFF_VALVE_CIRCUIT(A, shutoff valve)
SHUTOFF_VALVE_CIRCUIT(B, shutoff valve)

r_8: PRESSURE_SWITCH_CIRCUIT(A, pressure switch)
PRESSURE_SWITCH_CIRCUIT(B, pressure switch)



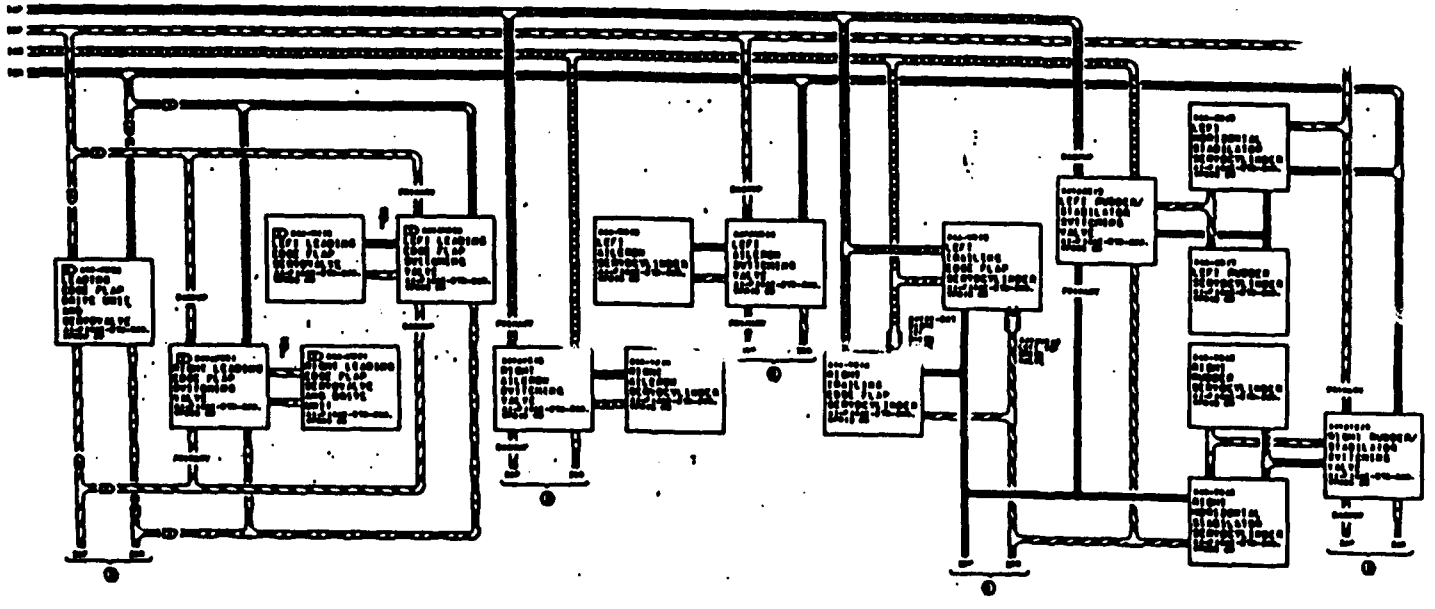


Figure 10. Hydraulic system one: F/A-18.

r_9: INDICATOR(A, pilot valve)
INDICATOR(B, pilot valve)
INDICATOR(A, shutoff valve)
INDICATOR(B, shutoff valve).

This hierarchical network structure represents a graph with levels $m = 1, 2, \ldots, 6$ (Figure 11). Relationship r_1 (UNIT) spans levels five and six ($d = 5$), and r_2 (COMPONENT) spans levels four and five ($d = 4$). Relationships r_3 (PISTON_TYPE), r_4 (VALVE_TYPE), and r_5 (SWITCH_TYPE) are located at a depth $d = 3$. Relationships r_6, r_7, and r_8 (PILOT_VALVE_CIRCUIT, SHUTOFF_VALVE_CIRCUIT, PRESSURE_SWITCH_CIRCUIT, respectively) are located at a depth $d = 2$. Finally, relationship r_9 (INDICATOR) spans levels one and two ($d = 1$).

An interface prototype is developed in SuperCard™ 1.5 (Appleton and Poppitz, 1990). This prototype supports interaction scenarios for which relationships associated with Hydraulic System 1 are to be selected. The menu bar (Figure 12) enables an end user to select from a set of relationships associated with the system hierarchy (**System Links**) and a set of relationships describing physical orientations of system components (**Operational Links**). Once a set of focus relationships has been selected, the **Data** menu allows the corresponding presentation of fisheye views.

The relationship set associated with the hierarchy of Hydraulic System 1 is defined as $R = \{r_1, r_2, \ldots, r_9\}$. Suppose relationships r_1 (UNIT) and r_5 (SWITCH_TYPE) are selected as focus relationships such that $R_f = \{r_1, r_5\}$. Note from Figure 11 that each focus relationship composes a single connected subgraph. The presentation value associated with each relationship is calculated according to Equation (6) such that $P_1 = 19$, $P_2 = 12$, $P_3 = 0$, $P_4 = 12$, $P_5 = 2$, $P_6 = -1$, $P_7 = -1$, $P_8 = 2$, and $P_9 = -24$.

Again, a brief explanation of several of the P_k calculations is perhaps appropriate. Consider focus relationship r_5 (SWITCH_TYPE). A single connected subgraph ($i = 1$) containing r_5 is located at a depth $d = 3$. Additionally, r_5 spans one node pair at $d = 3$; therefore, $n_{5_3} = 1$, and $n_{5_1} = n_{5_2} = n_{5_4} = n_{5_5} = 0$. Finally, the minimum path distance

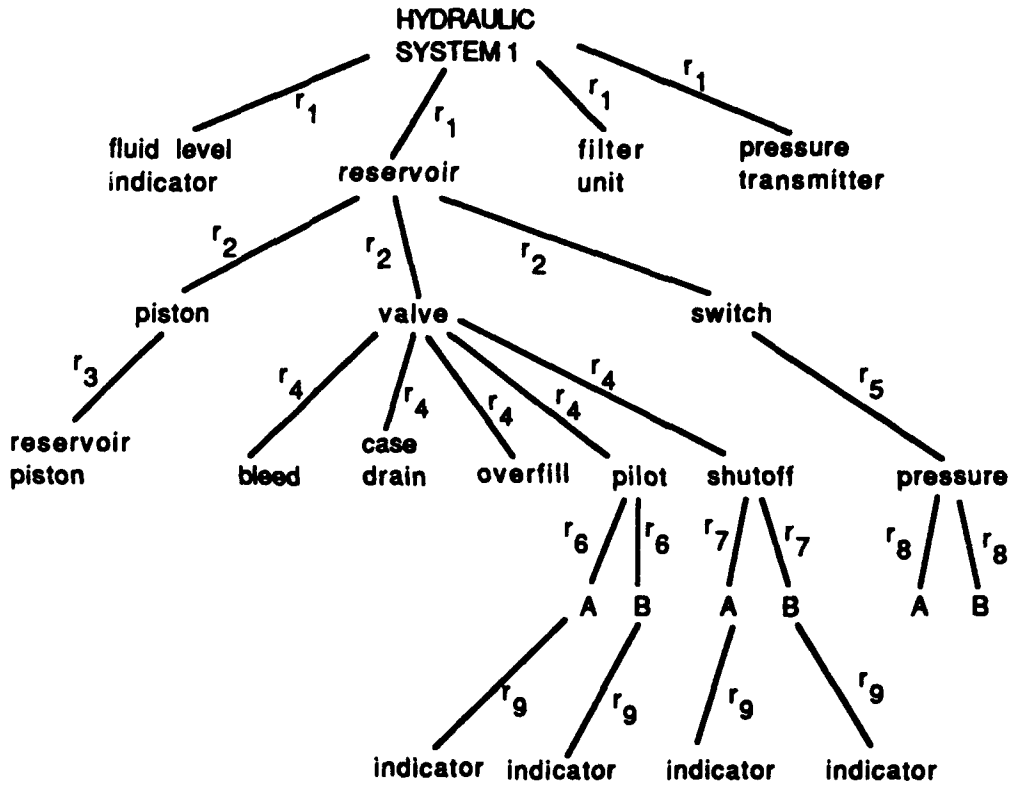


Figure 11. Hierarchical structure of hydraulic system one.

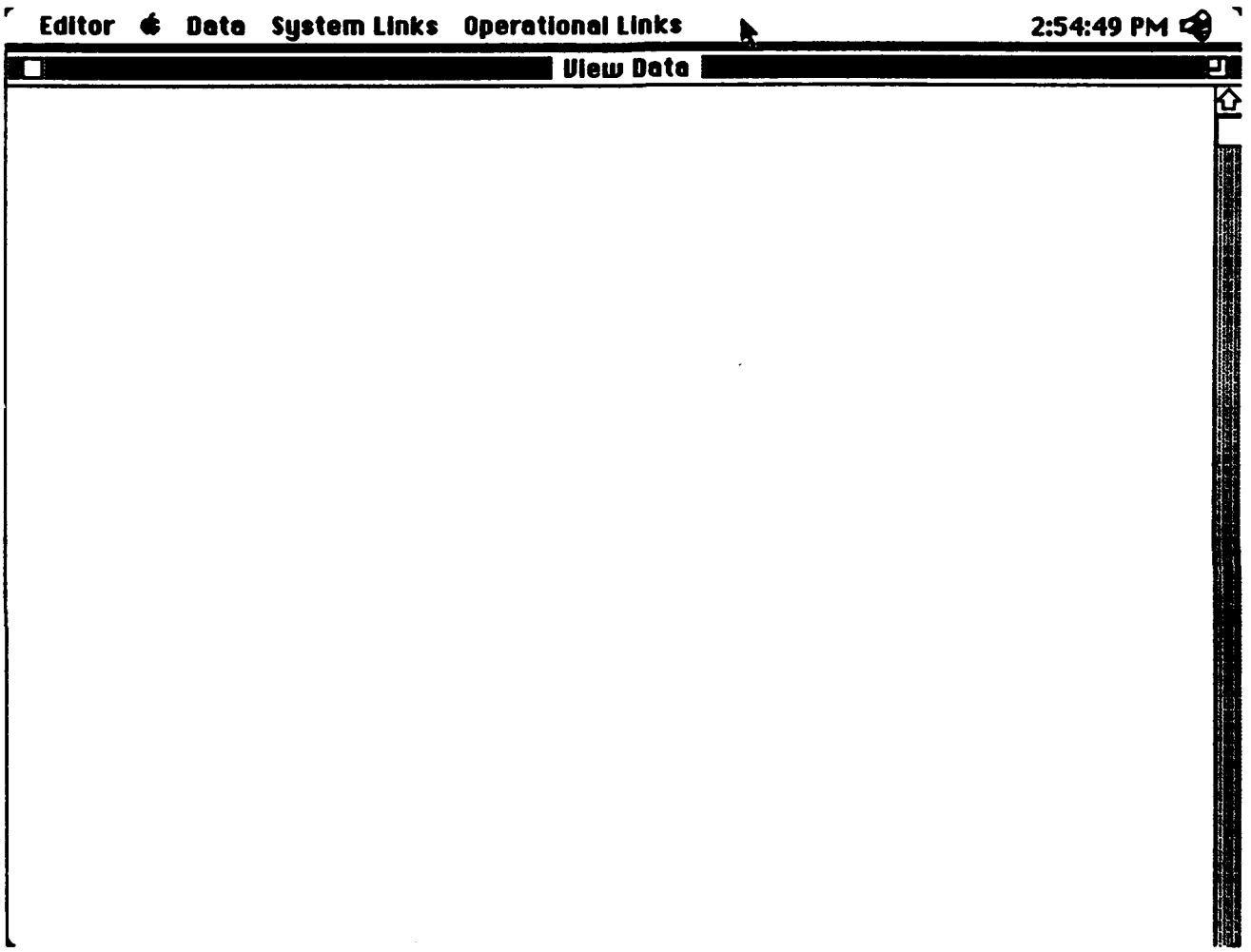


Figure 12. Interface prototype menu bar.

between the r_5 subgraph and the r_1 subgraph is one such that $D_{5_1,r_1} = 1$. Substitution into Equation (6) yields the following result:

$$P_5 = [n_{5_1}(1) + n_{5_2}(2) + n_{5_3}(3) + n_{5_4}(4) + n_{5_5}(5)] - [D_{5_1,r_1} + D_{5_1,r_5}] \quad (11)$$

$$P_5 = [0(1) + 0(2) + 1(3) + 0(4) + 0(5)] - [1 + 0] = 2. \quad (12)$$

As a second example, consider relationship r_9 (INDICATOR). Four connected INDICATOR subgraphs ($i = 1, 2, 3, 4$) are located at a depth $d = 1$. Each subgraph spans one node pair at $d = 1$, implying that $n_{9_1} = 4$, and $n_{9_2} = n_{9_3} = n_{9_4} = n_{9_5} = 0$. The minimum path distance between each r_9 subgraph and the r_1 subgraph is three ($D_{9_1,r_1} = D_{9_2,r_1} = D_{9_3,r_1} = D_{9_4,r_1} = 3$), and the minimum path distance between each r_9 subgraph and the r_5 subgraph is four ($D_{9_1,r_5} = D_{9_2,r_5} = D_{9_3,r_5} = D_{9_4,r_5} = 4$). Substitution into Equation (6) yields the following result:

$$P_9 = [n_{9_1}(1) + n_{9_2}(2) + n_{9_3}(3) + n_{9_4}(4) + n_{9_5}(5)] - [(D_{9_1,r_1} + D_{9_2,r_1} + D_{9_3,r_1} + D_{9_4,r_1}) + (D_{9_1,r_5} + D_{9_2,r_5} + D_{9_3,r_5} + D_{9_4,r_5})] \quad (13)$$

$$P_9 = [4(1) + 0(2) + 0(3) + 0(4) + 0(5)] - [4(3) + 4(4)] = -24. \quad (14)$$

Sample fisheye views of the hydraulic system are developed in SuperCard™ 1.5 (Appleton and Poppitz, 1990) and provided in Figures 13, 14, 15, 16, and 17. Figure 13 represents information associated exclusively with focus relationships r_1 and r_5:

r_1: UNIT(fluid level indicator, hydraulic system 1)
UNIT(reservoir, hydraulic system 1)
UNIT(filter unit, hydraulic system 1)
UNIT(pressure transmitter, hydraulic system 1)

r_5: SWITCH_TYPE(pressure, switch).

Figures 14 ($P_k \geq 2$), 15 ($P_k \geq 0$), 16 ($P_k \geq -1$), and 17 ($P_k \geq -24$) demonstrate the concept of information filtering: as P_k is decreased, additional graphics information is browsed.

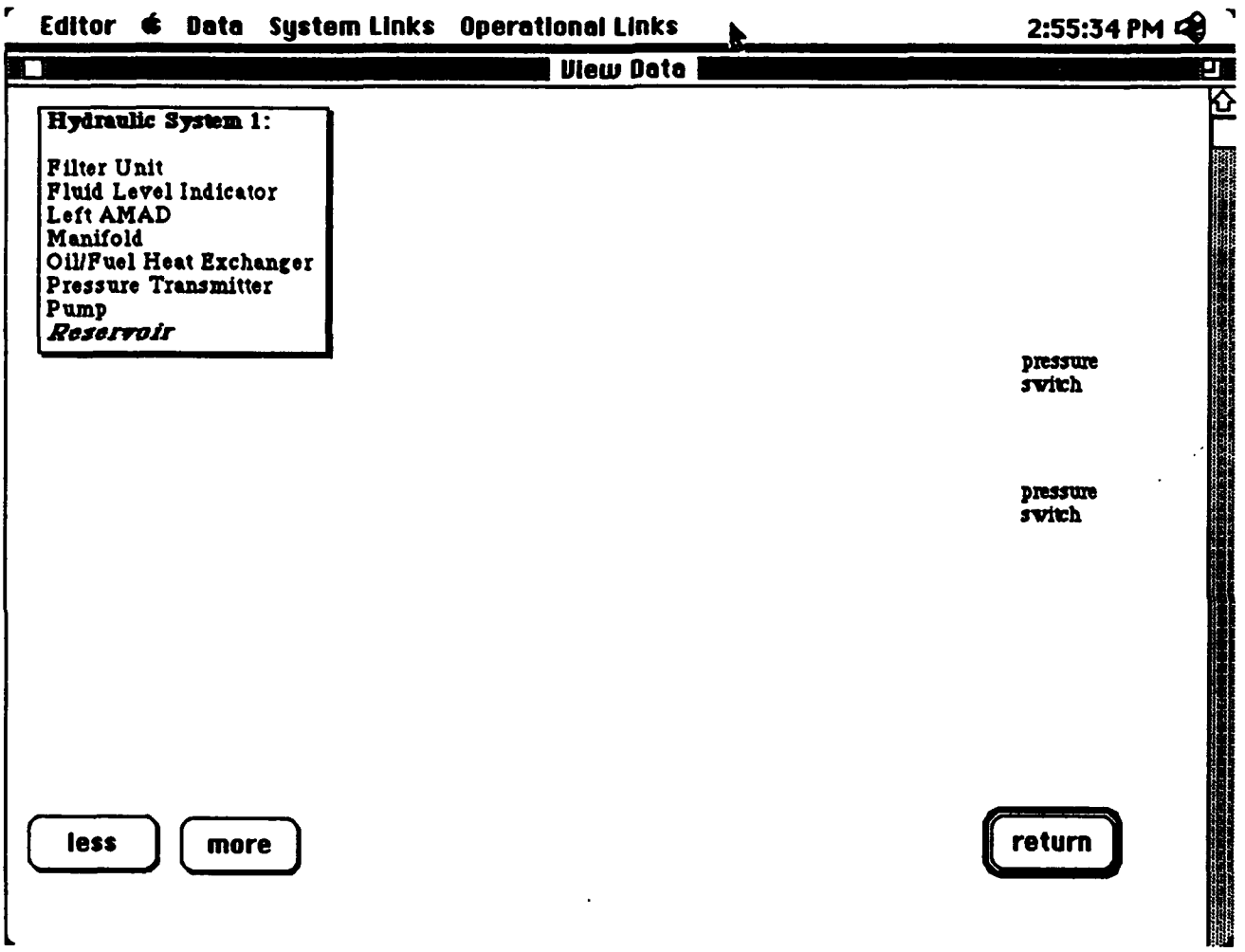


Figure 13. Information associated with focus relationships r_1 and r_5.

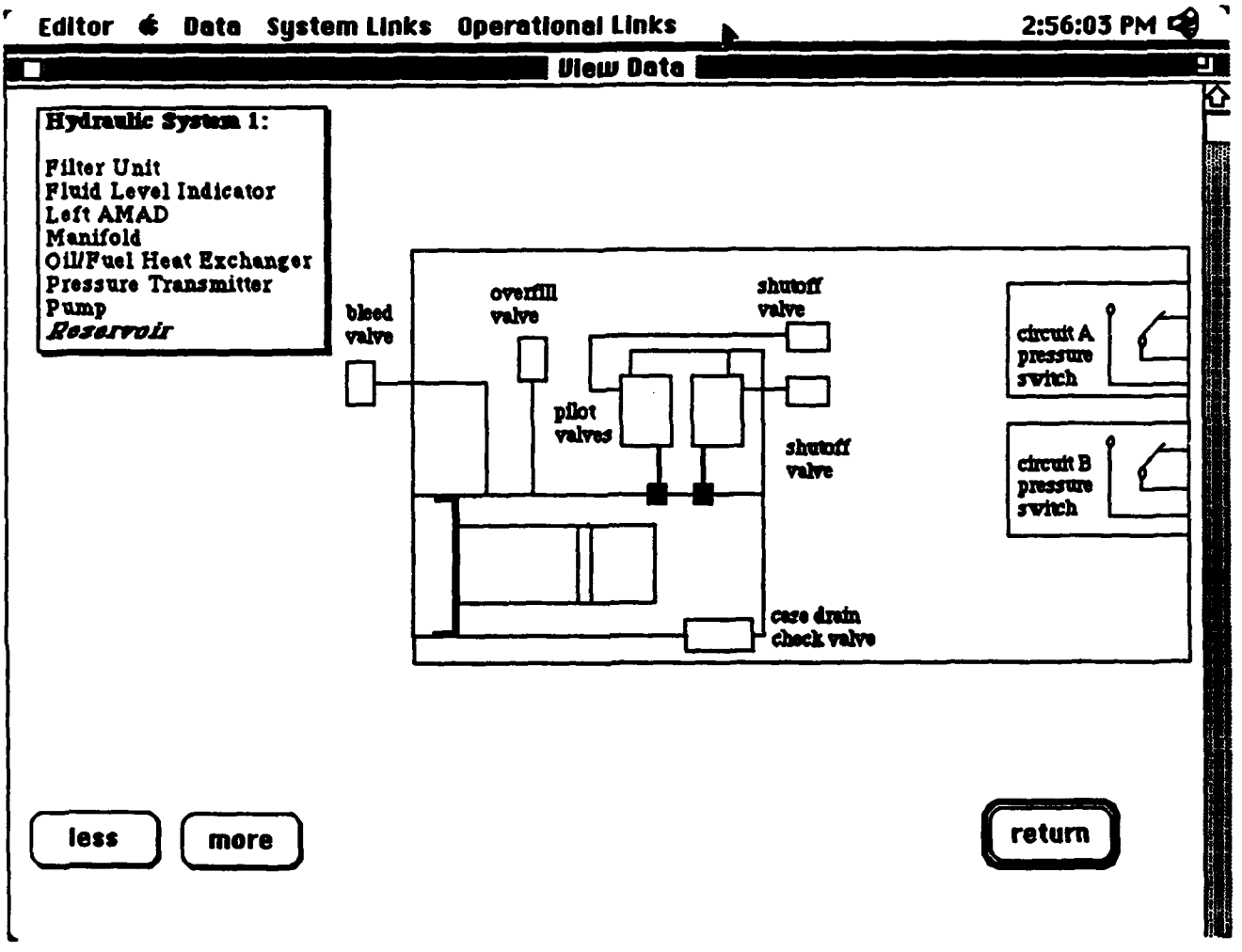


Figure 14. Information associated with a threshold $P_k \geq 2$.

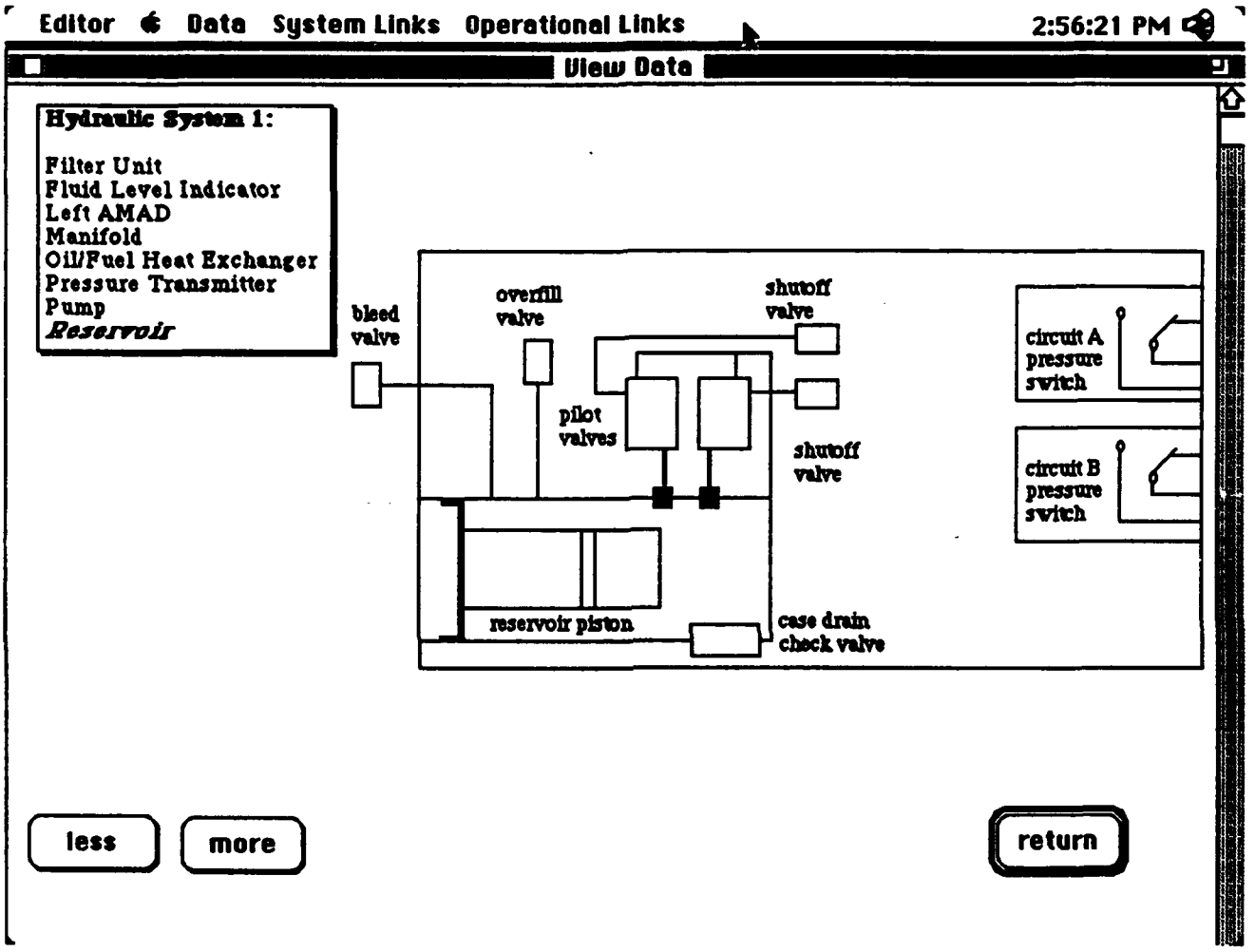


Figure 15. Information associated with a threshold $P_k \geq 0$.

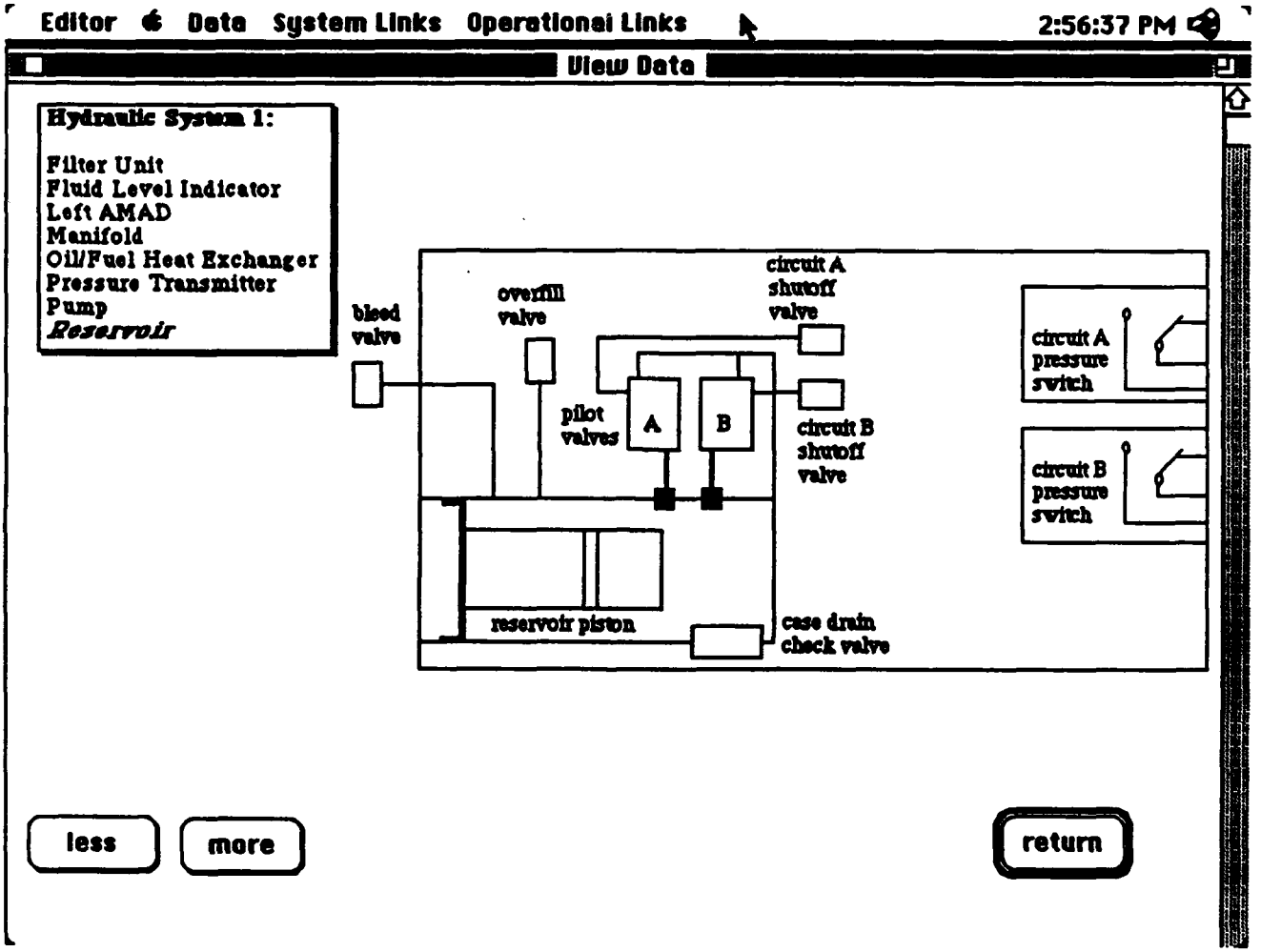


Figure 16. Information associated with a threshold $P_k \geq -1$.

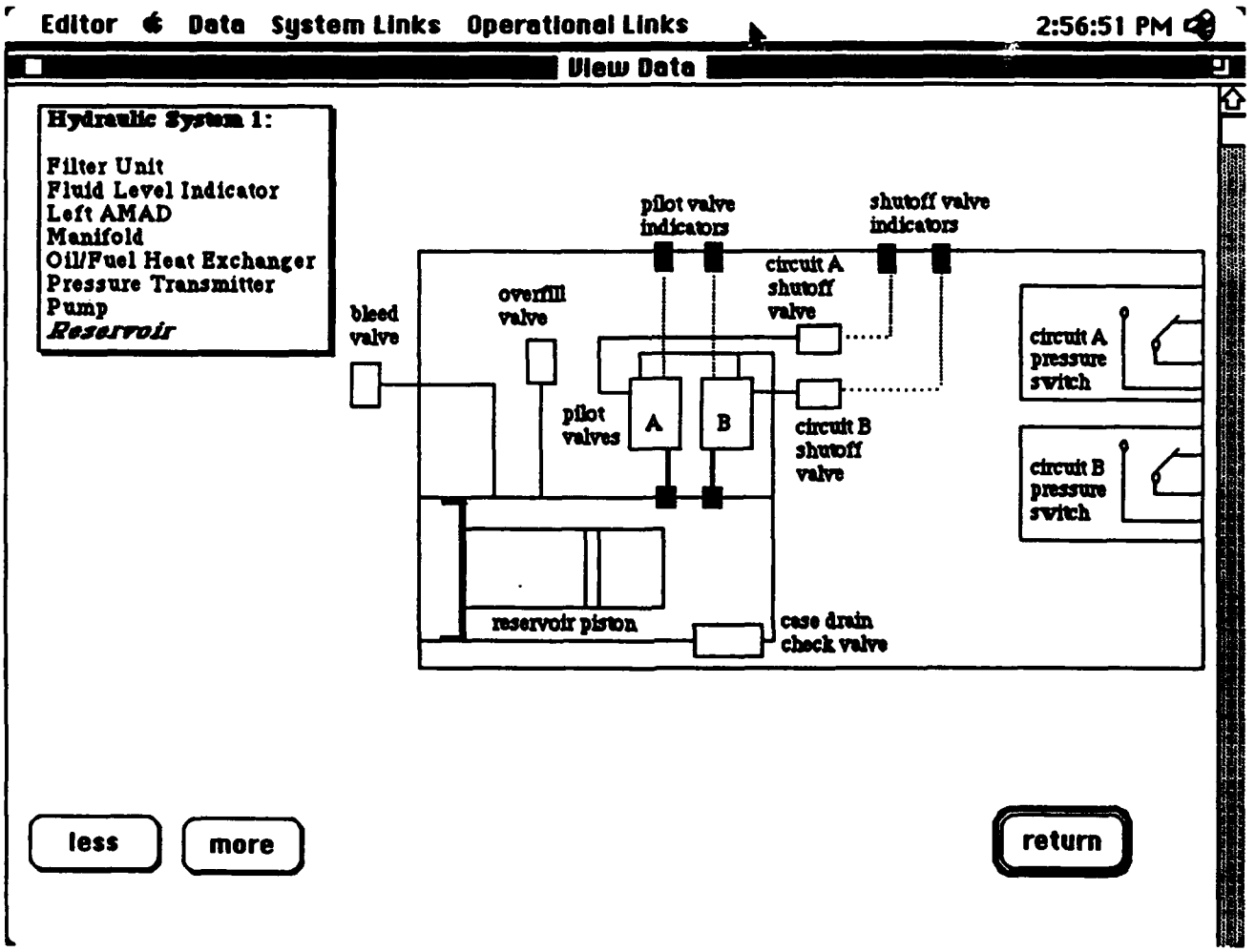


Figure 17. Information associated with a threshold $P_k \geq -24$.

V. RECOMMENDATIONS:

The results of this research demonstrate the concept of focus relationship selection and the subsequent presentation of associated database information. The concept of focus relationship selection and the information presentation strategy described in this report are suggested as a means of browsing graphics-based aircraft maintenance data. The methodology for specifying information content at various browsing stages (established by P_k thresholds) has been developed. This methodology is analogous to Furnas' (1982, 1986) original strategy for specifying information associated with a set of focus points. The presentation strategy coinciding with the selection of focus relationships has been implemented on a subset of IMIS maintenance data. An interface facilitating the selection of focus relationships and the subsequent browsing of maintenance information has been designed. A prototype system incorporating the selected data subset and the interface design was developed.

In order to support future IMIS field tests, a complete specification of Hydraulic System 1 (F/A-18) relationships is required. Currently, the specification of system relationships is being extended beyond the four units shown in Figure 11 (fluid level indicator, reservoir, filter unit, pressure transmitter). The final relationship specification should include both system (hierarchical) relationships as well as I/O (connectivity) relationships.

Recall that in general, relationship importance I_k must be determined heuristically. In other words, for a network without an underlying hierarchical structure, relationship importance must be defined in terms other than hierarchical depth. An example of such a network is one that might represent I/O relationships of Hydraulic System 1. One suggestion for heuristically defining importance of nonhierarchical relationships might be to weight each relationship on the basis of its relevance within a given troubleshooting or diagnostic scenario. For example, in a fault detection scenario for which a technician must determine whether a system component has failed, a connectivity relationship (indicating the extent of this component's influence with respect to other system components) is likely to receive more weight (greater importance) than a relationship describing the physical orientations or positioning of system components.

Flight control components and associated system relationships have been incorporated into the network structure of Figure 11. An initial specification of I/O relationships has been completed. As the underlying network structure of Hydraulic System 1 is finalized, additional examples of fisheye views based upon focus relationship selection can be provided.

REFERENCES

Aerospace Industries Association (1989, November). Guide to simplified graphics.

Appleton, B., and Poppitz, G. (1990). SuperCard™, Version 1.5, San Diego, CA: Silicon Beach Software.

Behzad, M., Chartrand, G., and Lesniak-Foster, L. (1979). Graphs and digraphs. Belmont, CA: Wadsworth International Group.

Furnas, G. W. (1982). The FISHEYE view: A new look at structured files (Technical Memo TM 82-11221-22). Bell Laboratories.

Furnas, G. W. (1986). "Generalized fisheye views." In Proceedings CHI '86 Human Factors in Computing Systems, ACM Special Interest Group on Computer and Human Interaction (pp. 16-23).

Mitta, D. A. (1989, September). Fisheye representation of information: IMIS user interface (Technical Report Air Force Office of Scientific Research/AFSC, 1989 USAF-UES Summer Faculty Research Program, Contract F49620-88-C-0053). Wright-Patterson Air Force Base, OH: Human Resources Laboratory.

Mitta, D. A. (1990a). "Presentation of Computer-Based System Information: Enhancing Human-Computer Interaction." In 1990 Advances in Bioengineering (pp. 203-204), BED-Vol. 17. New York, NY: The American Society of Mechanical Engineers.

Mitta, D. A. (1990b). "A Fisheye Presentation Strategy: Aircraft Maintenance Data." In D. Diaper, G. Cockton, D. Gilmore, and B. Shackel (Eds.), Human-Computer Interaction - INTERACT '90, Proceedings of the IFIP TC 13 Third International Conference on Human-Computer Interactio. ,pp. 875-880). North-Holland: Elsevier Science Publishers B. V.

FINAL REPORT
SUBMITTED TO

UNIVERSAL ENERGY SYSTEMS
FOR CONTRACT No.
F49620-85-C-0013/SB5851-0360

Proposal Title

"An Intelligent Teacher's Associate for Network Theory Based on the Heuristic of Polya"

Date March 31, 1988

by

Philip D. Olivier, Ph.D., P.E.
Assistant Professor of Electrical Engineering

ABSTRACT

The primary object of the supported work was to continue the development of the an Intelligent Computer Aided Tutor for Network Analysis.

During the course of the contract several tasks were performed. These included the developement of an expert system for generating a multiple choice test, the evaluation of several software products for the implementation of the tutor, and coding the tutor. The conclusion based on the work is that Polya's Heuristic can be used to guide a student through Network theory, however, it is necessary to have an "expert" that is capable of determining whether or not a mathematical expression that is the final solution of a problem is consistent with the constraints of the problem. This is the key. It appears that the framework of such an "expert" has been developed and is called "Equational Programming". However these systems only run on VAX mini computers and the goal of this project was a Micro Computer based system.

OBJECT OF THE RESEARCH

The object of the supported work was to continue the developement of an Intelligent Computer Aided Tutor to aid in mastering Network Theory at a level normally taught to Sophomores in an Electrical Engineering curriculum. Several other tasks were also performed.

ASVAP TEST GENERATOR

The Armed Services Vocational Aptitude Test (ASVAP test) is the tool that the US Military services use to assign individuals to carreers. Each version of the Mathematics portion of this test is meets certain guidelines. These guidelines were coded into rules that were run on the Expert System system CLIPS. Candidate problems were generated at random. Each candidate problem was then "filtered" by the CLIPS rules and the unsuitable ones were deleted. Once enough problems were generated, the program listed the problems. This system would allow for a suitable test to be generated for each test administration, and therefore reduce the security requirements.

SOFTWARE PRODUCTS EVALUATED

Several software products were evaluated during the course of the research. The goal was to find a software environment that would allow the manipulation and evaluation of mathematical expressions as well as an Expert System environment.

It appeared that MUMATH would be the ideal environment. MUMATH is written in MUSIMP, which a version of FranzLisp. MUMATH provides functions that allow for the symbolic manipulation of mathematical expressions, and since it is written in a version of LISP, expert systems could be built in the environment as well. However, the internal representation of the mathematical formulae used in MUMATH is not consistent with that of the human. It became obvious that a complete re-writing of the mathematical manipulation procedures would be required. This appeared very clumsy in the MUSIMP language.

CLIPS is an expert system shell developed by NASA based on the RETE algorithm. The version that I had turned out to be good for the development of expert systems but not for the mathematical manipulations portion.

Turbo Prolog and XLISP are general purpose AI progams and it was impossible to write the requried system from scratch in these environments.

Toward the end of the research, I became aware of "Equational Programming" systems. These allow the manipulation of equatons in a declarative way, and are written in LISP. Unfortunately, these systems were not available on micro computers, such as a Zenith 248 or similar platform.

THE TUTORIAL SYSTEM

The tutorial system is described in the earlier report and is included here.

CONCLUSIONS

It is my conclusion that full implementation of an Intelligent tutorial system for Network Theory requires a more powerful software environment thatn is currently present.

REFERENCES

1. Anderson, J. R. and B. J. Reiser, "The LISP Tutor", BYTE, pp. 159-175, April 1985.

2. Betz, D., "XLISP: An Experimental Object Oriented Language", Version 1.4, 1984.

3. Joobbani, R. and Talukdar, S., "An expert system for understanding Expressions from Electric Circuit Analysis", Proceedings of the 9th International Joint Conference on Artificial Intellegence, 18-23 August 1985, Los Angeles, CA.

4. Newell, A., "The Heuristic of George Polya and its Relation to Artificial Intelligence", International Symposium on the Methods of Heuristic, University of Bern, Bern Switzerland, September 1980.

5. Polya, G., HOW TO SOLVE IT, Princeton, New Jersey, Princeton University Press, 1973.

6. Strum, R. D., and J. R. Ward, ELECTRIC CIRCUITS AND NETWORKS Englewood Cliffs, New Jersey, Prentice-Hall, 1985.

7. Woolf, B., "Context Dependent Planning in a Machine Tutor", Dissertation, University of Massachusetts, 1984.

```
                                    'METHODS currentsolution))))
                                   (print message)
                                  ))))

print '(ANSWER?))
setq answer (read))
setq history (append history (list (list 'ANSWER answer))))
if (member answer (assoc 'ANSWER currentsolution))
        't
       (and (setq message '(please check for typographical errors))
               (print message)
               (setq answer (read))
               (setq history (append history (list (list 'ANSWER answer))))
               (if (member answer (assoc 'ANSWER currentsolution))
                       't
                       (and (setq message '(possible answers are))
                               (setq message (append message (cdr (assoc
                                   'ANSWER currentsolution))))
                               (print message)))))

:rint '(UNITS?))
etq answer (read))
etq history (append history (list (list 'UNITS answer))))
f (member answer (assoc 'UNITS currentsolution))
       't
       (and (setq message '(please check for spelling errors))
               (print message)
               (setq answer (read))
               (setq history (append history (list (list 'UNITS answer))))
               (if (member answer (assoc 'ANSWER currentsolution))
                        't
                       (and (setq message '(possible units are))
                               (setq message (append message (cdr (assoc
                                'UNITS currentsolution))))
                               (print message)))))

olist (x history)
       (print x))

ne following questions need to be asked in the debugging process
O YOU NEED TO FORMALIZE AND SOLVE AN INTERMEDIATE PROBLEM?
O YOU KNOW A PROBLEM RELATED TO THIS ONE?
O YOU NEED TO MAKE ADDITIONAL ASSUMPTIONS? [DO NO UNLESS ABSOLUTELY
ECESSARY
AVE YOU USED ALL THE INFORMATION PROVIDED?
OOK AT THE UNKNOWN!!!!
 RY OUT YOUR PLAN
    J CHECK YOU ANSWER ON A SIMPLER SITUATION?
```

```
rint '(GOAL?))
etq goal (read))
etq history (append history (list (list 'GOAL goal))))
f (member goal (assoc 'GOAL currentsolution))
      't
      (and (print '(please re-read the problem and answer again))
             (setq goal (read))
             (setq history (append history (list (list 'GOAL goal))))
             (if (member goal (assoc 'GOAL currentsolution)) 't
                    (and (setq message '(the correct goal is))
                           (print (append message (cdr (assoc 'GOAL
                                  currentsolution)))))))))

rint '(GIVEN?))
etq given (read))
etq history (append history (list (list 'GIVEN given))))
f (member given (assoc 'GIVEN currentsolution))
      't
      (and (print '(please re-read the problem and anwer again))
             (setq given (read))
             (setq history (append history (list (list 'GIVEN given))))
             (if (member given (assoc 'GIVEN currentsolution)) 't
                    (and (setq message '(the given information is))
                           (print (append message (cdr (assoc 'GIVEN
                                  currentsolution)))))))))

rint '(EQUATIONS NEEDED))
etq equationsneeded (read))
etq history (append history (list (list 'EQUATIONSNEEDED equationsneeded))))
ond ((equal equationsneeded '?)
      (and (quantityequation? equations goal)
             (quantityequation? equations given)
             (componentequation? equations goal)
             (componentequation? equations given)
             (print '(EQUATIONS NEEDED))
             (setq equationsneeded (read))
             (setq history (append history (list (list 'EQUATIONSNEEDED
             equationsneeded))))))))
f (member equationsneeded (assoc 'EQUATIONS currentsolution))
      't
      (and (setq message '(the needed equation is))
             (print (append message (cdr (assoc 'EQUATIONS
                           currentsolution))))
           ))

rint '(METHODS NEEDED))
etq method (read))
 q history (append history (list (list 'METHODS method))))
  (member method (assoc 'METHODS currentsolution))
      't
      (and (setq message '(how do you relate))
             (setq message (append message (cdr (assoc 'GOAL
             currentsolution)) '(and) (cdr (assoc 'GIVEN currentsolution))))
             (print message)
             (setq method (read))
             (setq history (append history (list (list 'METHODS method))))
             (if (member method (assoc 'METHODS currentsolution))
                    't
                    (and (setq message '(Appropriate methods are))
                           (setq message (append message (cdr (assoc
```

```
;this function finds the equations that have a given validity
(defun validityequation? (equations validity)
       (setq newequations 'nil)
       (setq message (list 'the 'following 'equations 'are validity))
       (print message)
       (dolist (x equations)
              (cond ((member validity (car (nthcdr 3 x)))
                     (print (caar x))
                     (setq newequations (append newequations x))
       )       )      )

the following function prints a problem from the list of problems in
the chapter
(defun printequations (equations)
       (print '(the equations in this chapter are:))
       (dolist (x equations)
              (print (car x))))

(defun printproblem (x)
       (print '(-----------------------------------------------------------
       (print x)
       (print '(-----------------------------------------------------------


;HE MAIN PROGRAM STARTS HERE

;rint '(Hi I will be helping you today.  What is your name?))
;etq name (read))
;eeds to check student base for existing record to append this record to
;etq history (list (list 'NAME name)))

;he following question should be answered by analyzing the student model
;rint '(what chapter do you wish to work on?))
;rint '(1: CHAPTER 1))
;rint '(2: CHAPTER 2 [NOT YET AVAILABLE]))
;rint '(3: CHAPTER 3 [NOT YET AVAILABLE]))
;rint '(4: CHAPTER 4 [NOT YET AVAILABLE]))
;rint '(TYPE YOUR CHOICE 1 THROUGH 4))
;etq ans (read))
;etq history (append history (list (list 'CHAPTER ans))))
;ond ((= 1 ans) (load "chapter1"))
       ((= 2 ans) (load "chapter2"))
       ((= 3 ans) (load "chapter3"))
       ((= 4 ans) (load "chapter4"))

;olist (x problems)
;rint '(your next problem is))
 setq x (car problems))
;rintproblem x)
;etq currentproblem x)
;etq history (append history (list (list (car x)))))
;etq currentso. ution (car solutions))
;etq solutions (cdr solutions))
(setq problems (cdr problems))

print '(YOU WILL NOW BE ASKED SOME QUESTIONS ABOUT THE PROBLEM:))
print '(IF YOU DO NOT KNOW THE ANSWER THEN TYPE ?))

setq answer 'nil)
```

```
                    (print (append message (car (nthcdr 2 x))))
    )       )       )

he following function finds the range of validity for the given
quation
efun equationvalidity? (equations equation)
      (setq newequations 'nil)
      (dolist (x equations validity)
              (cond ((equal equation (caar x))
                     (setq newequations (append newequations x))
                     (setq message '(has validity type))
                     (print (append message (car (nthcdr 3 x))))
      )       )     )

he following function finds the type of component that the equation
efers to
efun equationcomponent? (equations equation)
      (setq newequations 'nil)
      (dolist (x equations type)
              (cond ((equal equation (caar x))
                     (setq newequations (append newequations x))
                     (setq message '(this equation refers to the following
                                     component[s]))
                     (print (append message (car (nthcdr 4 x))))
      )       )     )

his function finds the equation given the equation number
efun numberequation? (equations number)
      (setq newequations 'nil)
      (setq message '(the equation[s] with the given number
      is [are]))
      (dolist (x equations newequations)
              (cond ((equal number (caar (cdr x)))
                     (setq newequations (append newequations x))
                     (print (append message (list (car x))))
      )       )     )

his function finds the equations that involve a given quantity
efun quantityequation? (equations quantity)
      (setq newequations 'nil)
      (setq message '(the equations that refer to))
      (setq message (append message (list quantity) '(are:)))
      (print message)
      (dolist (x equations newequations)
              (cond ((member quantity (car (nthcdr 2 x)))
                     (setq newequations (append newequations x))
                     (print (caar x))
      )       )     )

his function finds the equations that pertain to a given component
efun componentequation? (equations component)
      (setq newequations 'nil)
      (setq message (list '(the following equations pertain to)))
      (setq message (append message component '(are:)))
      (print message)
      (dolist (x equations)
              (cond ((member component (car (nthcdr 4 x)))
                     (print (caar x))
                     (setq newequations (append newequations x))
      )       )     )
```

```
;7-8-86

print '(----------------------------------NETWORK TUTOR-----------------------
print '(-----------------------------------------by-------------------------------
print '(---------------------------------Philip D. Olivier----------------------
print '(-------------------------------------June 1986---------------------------

print '(This program is an Intelligent Tutorial system designed to help
tudents learn basic network theory on the level of Strum and Ward s book.  The
tudent is assumed to have read the appropriate chapter in the textbook.  It is
ot assumed that the student has comple tely mastered nor memorized the
nformation.  This program will pose problems for the student to solve and
onitor the students answers to questions about the problems.  Based on a
omparison of the students work to that of an expert the program will attempt
o diagnose the cause of the students errors as well as alter its own
uestioning strategy.  The questioning strategy is taken from that of Polya s
euristic))

'or each chapte· the domain specific knowledge is contained in the
.ist EQUATIONS · nich is made of sublists according to the following format:))

 quation:equation#:quantities:validity:component))
 where))
 quation stands for the equation))
 quation# is the equation number in the text [sometimes this is NONE]))
 uantities is a list that contains the names of the circuit variables
 nd or parameters that are related by the equation))
 ----possible entries are))
 ------- CURRENT  CHARGE  ENERGY  VOLTAGE  POWER  RESISTANCE))
 ------- INDUCTANCE  CAPACITANCE  FLUX  PERIOD  SECONDARY*VOLTAGE))
 ------- PRIMARY*CURRENT  MUTUAL*INDUCTANCE  FREQUENCY HERTZ))
 ------- FREQUENCY RADIAN  EFFECTIVE*VALUE  RMS*VALUE))
 ------- UNIT*STEP*RESPONSE  UNIT*RAMP*RESPONSE  UNIT*DELTA*FUNCTION))
 ------- UNIT*RECTANGULAR*PULSE))
 alidity refers to the range of validity of the equation possible
 ntries are))
 ------- DEFINITIONS  THEOREMS  LINEAR  NONLINEAR  ALL  TIME*VARYING))
 omponent usually refers to the type of circuit componet ie ))
 ------- RESISTORS  INDUCTORS  CAPACITORS  TRANSFORMERS  OPAMPS))
 owever it can be FUNCTIONS if the equation defines a function))

 he following function finds the equation number for a given
 quation

 efun equationnumber? (equations equation)
       (setq newequations 'nil)
       (dolist (x equations number)
               (cond ((equal equation (caar x))
                      (setq newequations (append newequations x))
                      (print (car (cdr (x))))
       )       )      )

 he following function finds the quantities that are related by the
 iven equation
 efun equationquantities? (equations equation)
       (setq newequations 'nil)
       (dolist (x equations quantities)
               (cond ((equal equation (caar x))
                      (setq newequations (append newequations x))
                      (setq message '(relates))
```

Research Initiative Program (RIP)

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

Final Report

An Assessment of the Effects of CONFER: A Text-Based Intelligent Tutoring System Designed to Enact Tutorial Conversation and to Increase a Student's Sense of Intertextuality

Prepared by:	William L. Smith, Ph.D.
Academic Rank:	Professor
Department and University:	English Department University of Pittsburgh Pittsburgh, PA 15260
Phone:	(412) 624-6559
USAF Contact:	Major James Parlett
Date:	25 March 1991

Even a cursory examination of recent newspapers, magazines, and television newscasts, let alone the professional journals in education, reveals that there is a serious problem, if not a crisis, in literacy and critical thinking skills and abilities. Scholastic Aptitude Test (SAT) scores have declined; students graduating from high school lack higher order intellectual skills; colleges and universities have increased the number of remedial and basic skills courses, especially in the areas of mathematics, reading, and writing; business, industry, and the military are spending millions of dollars on remedial education.

These problems are compounded by the fact that high school graduates, whether they enroll in a college, enter the workplace, or join a branch of the military, encounter sophisticated, computer-based equipment which demands that users have not only knowledge but highly internalized knowledge to work a speeds previous technologies did not require. These recent advances in technology have created an information-rich environment, but at a price: Coping with and managing this environment requires literacy and critical thinking skills required of only a few in the past.

The new Air Force recruit, assigned to an aircraft maintenance squadron, faces an impossibly complex task. He or she must learn in short order to trouble-shoot and maintain some of the most highly sophisticated weapons systems ever devised. Moreover, to do so, that recruit must also master computer based trouble-shooting technologies--also more complex than anything he or she has ever encountered in high school or other earlier experiences. This recruit is confronted with massive technical manuals which seem to change

frequently. Furthermore, the recruit now feels the pressure of speed, for he or she knows the immense cost of the weapons system and thus feels the pressures such costs bring to bear.

The Air Force currently offers over 6,000 technical training course each year and graduates over 100,000 recruits from its six technical training centers. The cost per graduate ranges from $5,000 to $28,000. This cost, given the current demographic trends on literacy and critical thinking, is certain to escalate, as is the human cost, the sense that one is capable and of worth.

In such high-technology areas as intelligent tutoring/training systems, where computers are used to teach skills, the cost of low literacy and critical thinking skills is equally evident.

Although some intelligent tutoring systems (ITSs) use no written language (e.g., INFLITE), most use it to varying degrees. For example, ORBITAL MECHANICS uses written language only in its help and information windows whereas CHALLENGER requires users to read what amounts to a textbook. The ability to interact with written language is, therefore, a prerequisite for using ITSs (Hull, 1989). ITS users must understand what they read and comprehend it at functional levels. That is, they must appropriate the text, making the necessary connections among ideas presented in the text (intra-textual connections) and making connections to their prior knowledge and to other texts (inter-textual connections). Since prior knowledge can be seen as a type of text, making connections to prior knowledge is also making inter-textual connections. Inter-textuality is extremely important, for creating such relationships is necessary for long term

memory, for generalization, and for problem-solving given new information or situations.

Venezky, Kaestle, and Sum (1987), in their review of the National Assessment of Educational Progress data, note that the respondents in that study were able to perform well when asked identification questions when the language of the question matched the language of the text the respondents had read. However, those respondents could not perform well when required to make inferences from the text. They were ineffective problem solvers because they had not made the necessary intra- and inter-textual connections.

The help and information windows commonly used in ITSs typically contain explanatory text, and, thus, reading ability becomes a factor. Although others have stated that the help feature may be problematic, Hull (1989) studied the use of help on MINA and found that students used it "infrequently and poorly." When they did use the help feature, "they misread or misinterpreted much information" (Hull, 1989, p. 148). In other words, the students did not, or were not able to, make intra- and inter-textual connections. They could not relate the explanations of MINA's functions to those functions, nor could they relate the explanations about grammatical problems to what they they had previously been taught about grammar.

Traditional methods for assessing reading ability and comprehension rely on reproduction of factual (or declarative) knowledge by using multiple choice, sentence completion, or cloze methods. However, during the past decade, reading researchers (e.g., Tierney and Pearson, 1983; Tierney and Cunningham, 1984) have claimed that the best method for testing comprehension is by having

students write a response in which they connect the reading material to their own knowledge. (Advocates of "writing to learn" (e.g., Emig, 1977) claim that writing is a uniquely powerful method for learning because writing requires, according to these advocates, self-reflective thinking and because writing allows one to see and read one's ideas thus promoting revision and new ideas.) Although the strength of the correlation between writing ability and reading ability is still debated, reading/writing researchers (cf Applebee, 1984; Kucer, 1985; Langer, 1984) claim that writing tasks provide the best evidence of students' comprehension and of their ability to connect what they read to their prior knowledge (i.e., to create inter-textual connections).

It is generally believed that learning from one's reading is further enhanced by discussing the content and one's writing about that content (and relationships to prior knowledge and other texts) with peers and teachers, but especially with one's teacher because that teacher has the authority of an expert in the content and in the relation of the content to the objectives of the course. However, for most teachers, time for such conferences presents a problem. The choice is either to have few conferences so that the material can be discussed in depth (and to allow the student the time for fumbling to make connections with prior knowledge--an additional problem for most teachers because they tend to preemptively provide "correct answers" thus thwarting induction) or to have many short conferences during which only a few points can be discussed.

One solution to this dilemma is to use the power of an ITS to help the students make intra- and inter-textual connections in an environment which is safe (ITSs are not judgmental as is a teacher),

patient (teachers have a low "wait time"--about 3 seconds, cf Tobin, 1987, before they provide an answer or ask another question), and available at the students' convenience. One such ITS currently exists: CONFER (Parlett, 1987). CONFER presents an appropriate alternative to teacher-student conferences because it has an expert's knowledge of the material the student has read, because it enacts an appropriate pedagogy, and because it allows the student to manage the focus of the tutorial. Although Parlett labels CONFER as ICAI (Intelligent Computer Aided Instruction), CONFER is an ITS; its components (e.g., the student model and the expert) have flexibility which allows the program to be authorable, that is, to respond to unique situations and to be modified by the author. ITSs do not require the author to explicitly anticipate all possible contingencies. The intelligence of the ITS allows it to deal with diversity.

Although there is evidence that CONFER can function well with students and is appropriate to their needs and expectations (Parlett, 1987), CONFER has not been field-tested. More importantly, CONFER, nor any other ITS, has not been used as a test-bed in a sequence of studies which lead to both information about how CONFER would work with diverse domains, genres, and populations of students and about how ITSs in general work, especially when [illegible] or more ITSs are sequenced. Thus, research on CONFER would not only allow us to learn about the effectiveness of CONFER but about the possibilities and problems of ITSs in general.

It is important for any piece of assessment research to be embedded with a continuum of assessment research, moving, as Steuck and Fleming (1989) indicate, from pre-experimental to field tests and,

at each point, examining issues of functionality, effectiveness, and cost. The continuum of such assessment research is long and broad, and much of what is to be learned about ITSs, about how students learn from ITSs, and about how to assess students ITSs is yet to come. As Baker (1989) states, "The design of seriously planned embedded assessment systems that includes the full range of input, process, and outcome data, such as individual differences, process, trainee performance, and transfer data could be undertaken in a *long-term study* ." [italics mine] Yet Littman and Soloway (1988) state that "No one we know, including ourselves, has yet carried out [internal and external evaluations] evaluations in an elegant, comprehensive way." Overall, only about twenty ITSs have been evaluated, yet none of these have been evaluated extensively. Instead, the evaluations are piecemeal. Other evaluations seem to be underway, but it will be some time before the results are made public.

There is one example of an attempt to examine several different evaluations of ITSs. Baker and her colleagues at UCLA (Baker 1989) are, apparently, attempting meta-analyses of ITSs, but those results have not been published. Even more problematic is the quality of the outcome data produced in the assessments. According to Baker (1989), "Relatively few studies of intelligent systems use outcome data of any sort...." In particular, extant assessments of ITSs have not examined students' learning beyond declarative knowledge and have not been designed to collect fine-grained data.

The research conducted thus far on CONFER has focused on the functional aspects, specifically, whether a student would be able to use the program easily and whether the program adequately represents

what it purports to represent. For this research, CONFER was tested by professors who have used, in their composition courses, the essay which CONFER helps students with. The results from the first tests showed that there were some bugs in the program, but later tests indicated that the revised program worked well (Parlett, 1987). The professors, all of whom were practiced computer users (primarily for word processing) but none of whom had experience using an ITS, found the interface to be friendly and were able to use all of the windows with ease. Furthermore, these professors confirmed that the resulting transcript (between CONFER and the "student") read like a conference they would have with a student, thus establishing a form of face validity.

Objectives

The purpose of this research was to determine the effectiveness of CONFER by triangulating three axes: whether students can use CONFER with ease, whether they feel that CONFER is of value, and whether CONFER has the desired effect on their thinking and writing. There were also two secondary purposes, both crucial to subsequent research: to determine which measures provide the most information with the least investment and to establish the base-line data necessary for determining delayed effects. Finally, it was hoped that this research might provide the base-line data necessary for follow-up studies on the subjects throughout their college careers, studies designed to examine delayed and long-term effects.

Procedure

All faculty teaching the Basic Writing and the General Writing courses (the two most populated composition course) were invited to participate. In the University of Pittsburgh composition program, teachers create their own sequence of writing assignments and reading assignments on which the writing in based. Thus, in any term, only a few teachers will include the Percy essay in their sequence.

Nine teachers accepted. Their classes were randomly divided into two groups, those who would use CONFER and those who would not. For convenience, the term "Experimental" will be used to refer to the group which used CONFER and "Control" for the group which did not.

Not all students in each teacher's class were included in the final Experimental and Control groups. Only those from whom all materials (tests and essays) were submitted were included. Furthermore, a few students in the Experimental group did not keep their appointments to use CONFER. These students were either placed into the Control group (if all other materials were submitted) or were excluded. A total of 71 students comprised the Experimental group, and 72 the "Control" group.

Method

Using CONFER

Students used CONFER at their own schedule. The computer housing CONFER was located in a small room adjacent to the University of Pittsburgh Writing Workshop. The students made appointments at the Writing Workshop or made arrangements with my graduate assistant. When the students showed up for their appointment, they were escorted

to the small room and shown how to use CONFER. If the student was to be tape recorded (for the TAP data), the tape recorder was turned on and the student was asked to turn it off after the session.

Although CONFER allows a student to have a multi-session tutorial, no students used this option. All tutorials were completed in one session.

Knowledge tests

Three types of tests were used to assess prior, declarative, and domain knowledge. (See Figure 1, Design) All three were used as a package. Thus, students did not take one without taking the others. All tests were kept short to ensure that the students would have time to do all of them and to all students ample time to consider each question. The tests were administered by the teachers during class time. The first test included five multiple choice questions about the content of the essay. This test was scored by counting the number of correct answers.

The second test included three sentence-starter stems which the students completed. Sufficient space was allowed so that they could write as much as they wished. The focus of this test was on concepts central to the essay. Only one question required having read the essay. This test was scored by rating each answer on a three point categorical but linear scale. A rating of "A" indicated that the answer showed a firm grasp of the concept, as evidenced by warrants supporting claims. A rating of "B" indicated some grasp of the concept. A rating of "C" indicated no grasp of the concept.

The third test, the association test (cf. Langer & Nicholich, 1980; Langer, 1984; and especially Newell & MacAdam, 1987 and Greene, 1990), included three concepts central to the essay, each concept written on a separate sheet of paper. Student were asked to write "whatever comes to your mind" about each concept. The students were told they could use words, phrases, sentences, or paragraphs. The students' responses were analyzed using three ordered categories of knowledge organization based on a taxonomy created by Newell and MacAdam (1987) and refined by Greene (1990). Highly organized knowledge (HIGH) is evidenced by superordinate concepts, definitions, and analogies. Partially organized knowledge (PART) is evidenced by examples, attributes and defining characteristics. Diffusely (DIFF) organized knowledge is evidenced by list-associations and personal experience. Following Greene's procedure, HIGH was given a rating of 3, PART a rating of 2, and DIFF a rating of 1. The three ratings for each student (i.e., the ratings for each of the three probes) were summed. Thus, for each student, the final score could range from 3 to 9.

Think-Aloud Protocols

Selected Experimental students were asked to think aloud into a tape recorder as they used CONFER. They were told to relate what they were doing and what they were thinking. The usual procedure for such protocols is to ask the student to talk constantly and to have someone monitor the session in order to prompt the student to keep talking whenever the student fell silent. This procedure was not used because silence would be an indication of non-dialogic behaviors, e.g., reading or thinking. A monitor was present and recorded what the student was

doing during periods of silence. If the students simply forgot to talk into the microphone, the monitor prompted them.

Interviews

Selected Experimental students were interviewed after they had used CONFER and/or after they had finished writing their essay. The interviews included standard questions designed to elicit information about the students' perceptions of CONFER, but the interviews were not restricted to those questions.

Statistical Analyses of Data

For the multiple choice test and the test of association on central concepts, both of which yielded quantitative data, ANOVAs of pre-set groups (determined by the design and by the question being asked) were performed. For the short answer tests, Chi Square tests of pre-set groups were used.

Research Questions

This research addressed, concurrently, three sets of questions, each examining the effects of CONFER from a different perspective. The first set (Question 1) addressed students' facility in using CONFER (e.g., using the windows, maintaining and advancing the dialog), for if students have trouble using it, the overall effectiveness will, almost certainly, be diminished. The second set (Questions 2 through 4) addressed students' perceptions of CONFER, especially whether they felt the program helped them. Research on human tutors indicates that when a student comes to the tutorial believing that the tutor can help,

the results of the tutorial are always positive, even when the tutor thinks the tutorial was less than satisfactory. However, if students don't think CONFER is helpful, three things will happen--all negative: 1) The students won't be receptive to learning, 2) they won't want to use CONFER again (e.g., an alternate form of CONFER with a different essay), and 3) the word will spread to other students. Consequently, those other students will come to CONFER with a negative mind-set and won't be receptive to learning, thus repeating the cycle.

The third set of questions (Questions 5 and 6) focused directly on the effectiveness of CONFER, on determining the degree to which the students evidence change in knowledge about the essay and the degree to which students change in the connections they make within the essay and between their prior knowledge and the essay. In other words, this set focused on whether using CONFER helps students develop their ideas and whether they revise those ideas or develop new ideas as a result of using CONFER.

Specific Research Questions

QUESTION 1: Do students have any problems using CONFER? If they do, what are the problems and what training, etc. is needed?

QUESTION 2: Do students say that CONFER helped them to understand the passage?

QUESTION 3: Would students want to use CONFER again on a different passage?

QUESTION 4: a) Do students say that using CONFER helped them fulfill their expectations about their texts? b) Do other data support what they say?

QUESTION 5: Are the texts written by the CONFER students more fully developed than those written by the students who did not use CONFER (specifically, are paragraphs more developed)?

QUESTION 6: Do students evidence increased knowledge of content of the essay?

Results and Discussion

Six questions guided this research. For ease of presentation and discussion, the results will be presented by research question.

Question 1: Do students have any problems using CONFER? If they do, what are the problems and what training, etc. is needed?

The observational records kept by the tutors in the University of Pittsburgh Writing Workshop and by the graduate assistants who worked with the individual students consistently noted that the students had little trouble getting started in their CONFER tutorials. The single most consistent problem was that the students had either not carefully read the Percy essay or, in the case of a few students, had not read it at all. Thus, when CONFER asked them to specify a part of the essay which troubled them, the students either quickly skimmed through the essay or selected, in what seemed to be a random manner, a paragraph.

The TAP data showed that the students indeed had not adequately prepared for the tutorial. Here is one student's transcript. It is an adequate representation of all students who evidenced this lack of preparation in all ways but one. This student admitted not reading

carefully. (Note: Here, and in all subsequent TAP transcripts, the pound sign "#" will be used to represent abnormal pauses. For each additional five seconds, a pound will be added. Thus, "##" represents a pause of at least 10 seconds. In many of the long pause cases, the student was reading silently.)

"Uh. Uh. ### Oh God. It's asking me what I want to start talking about. # I don't know. There's a lot. It wants a paragraph. What one? I didn't read it [the essay] that way. I thought the computer would tell me."

In the interview after the tutorial, students were asked about this apparent problem. In general, they said that they assumed that the computer would dictate the flow of the tutorial, picking the topics and answering questions. When they learned that CONFER would not answer questions (would only ask them) and would allow them to initiate the conversation, they were taken aback. Because of their assumption, many had not bothered to read the essay or had read it only cursorily. Thus, they had not prepared for CONFER's opening question. The student whose TAP transcript is cited above said that he had heard about the smart computer from other students and "just assumed that [he] could get some good ideas for the essay [he] had to write." Therefore, he had only skimmed the essay and talked briefly with one classmate about it.

After this initial problem, however, the students evidenced little problem in advancing the the dialog. After they had entered into the conversation, they became facile in responding. However, many noted in the interviews that CONFER sometimes repeated itself, and sometimes did not seem to be responding to what they had said.

Students made little use of CONFER's windows. They had printed copy of essay available, so that may have limited their use of the essay window, but few used the notes window, and only a couple used the dictionary window. Thus, although the windows feature could be important, it had little impact on the students and, thus, little impact on the data.

Questions 2 and 3: Do students say that CONFER helped them to understand the passage? Would students want to use CONFER again on a different passage?

Of the students who used CONFER, 70% said that they would like to be able to use a program like CONFER for other assignments in their composition course and in other courses. And about 30% of these students asked if they could redo their computer conference on the Percy essay. They felt they could learn even more the next time. The interviews with the students provided the reason for this desire: they viewed CONFER as being less judgmental than their teachers, as being more tolerant, and as being friendlier.

One student had the misfortune of suffering through three "crashes" of the program. Thus, he had to begin anew each time. He did so willingly, stating that for the first time in his academic life, he was afforded the opportunity to think before answering questions. He said he would much prefer to work with the computer than his teachers because the computer was patient and didn't put pressure on him. He also said that he liked the computer because it didn't "get mad" if he said something wrong or funny or off-track. He felt he could take risks with the machine--risks he could not take with a teacher or with other students around.

The transcript of his conference with CONFER supported his contention. He often waited as much as several minutes before answering questions, and he also paused for long periods between words or sentences while he was answering a question. And he often re-read some of the essay before or while answering. His answers to CONFER's questions showed clear evidence of risk-taking. He made assertions which he couldn't support and then later, after an exchange of several questions, he would go back and change his earlier assertions. He was clearly learning as he went.

Question 4: a) Do students say that using CONFER helped them fulfill their expectations about their texts? b) Do other data support what they say?

In the interviews, a majority (55%) of the students in the Experimental group said that CONFER did help them in the planning of their texts, but many more (83%) said that CONFER helped them to better understand the Percy essay. The consensus opinion among those who didn't find CONFER helpful in their understanding was that they thought CONFER wasn't as helpful as they wished because CONFER would not answer their questions about the essay. (CONFER was purposefully designed as a Socratic tutor. That is, it asks questions but refrains from answering.)

The students' comments provided a means for creating two sub-groups which valued CONFER differently. Sixteen of the students (23%) said that they had begun thinking about and planning what they would write before they used CONFER. Of these students, only 19% said CONFER was helpful in their writing and 62% said CONFER helped them increase their understanding of the essay. These students came to

CONFER with more specific problems in mind, i.e., problems specific to the text they were planning, and they were disappointed because CONFER would not address those specific concerns, usually questions about how to present their case in their text.

The students who said that they had not begun planning of their text (77%), were more positive in their response to CONFER. Sixty-five per cent said CONFER helped them with their text and 89% said the tutorial helped them better understand the essay. Some of these students (18%) admitted that they had not read the essay before the tutorial, and it is likely that others also had not read carefully even though they did not admit it. Thus, this "not yet in the planning stage" group can be further divided into those who had read carefully and those who had not. Since CONFER is designed to help students who have done their homework, it is not surprising that almost all of the students who hadn't done the reading would not find CONFER helpful.

Table 1 depicts the distributions of students who thought they were helped in writing their texts and helped in understanding the essay, grouped by those who planed and those who had read the essay before using CONFER. Chi Square comparisons show that there are significant differences in the distributions, except where the "did not read" group is involved. This lack of significance, however, can be attributed to the very small number of students who said that they did not read the essay.

Thus, the answer to the first part of this question is a qualified yes. Although CONFER was designed to help students in the incubation stage of the writing process, the fact that some students were already well into this stage may lessen the effectiveness of CONFER. And it

stands to reason that students who have not read the essay before beginning the tutorial might have different opinions about CONFER.

The second part of the question requires using other data sources. Since the interviews showed that the students could be grouped according to who had read and who was already planning, the knowledge test data, particularly from the multiple-choice and the association tests, could be analyzed using these groups.

The results of these analyses (Tables 2 and 3) are very similar, and confirm what the students said in the interviews. The scores of the students who did not read the essay before doing the tutorial were significantly lower ($p=.0197$ for the Association; $p=.016$ for the Multiple-Choice) than those who read it. However, the students who didn't read the essay apparently gained considerable knowledge during the tutorial, for their Multiple-Choice scores were significantly higher ($p=.003$) on the Post-CONFER test. The was also a gain on the Association test, but the significance level was not as high ($p=.119$). It appears, then, that CONFER provides a setting in which even students who haven't done their homework can learn. Those students, according to their TAP data and the interviews, read much of the essay during the tutorial. Consequently, their tutorials were much longer in time but not in the number of exchanges between CONFER and the students.

The comparisons of the Plan and Did Not Plan groups show that there were no significant differences on either the Post-Reading or Post-CONFER tests. However, if the students who did not read the essay before the tutorial--students found only in the Did Not Plan group--are removed, then there is a significant difference on the Post-CONFER Association test. The Did Not Plan group's performance

exceeded the Plan group. Furthermore, only the Did Not Plan group (excluding the Did Not Read students) significantly increase their Association scores on the Post-CONFER test.

Thus, it appears that, at least in terms of test scores, CONFER has a much greater impact on the students who read the essay but have not yet begun to plan their texts. This makes intuitive sense, for those students who have begun to plan would have pre-set notions about what they wished to discuss during the tutorial. And, as these students said in the interviews, they wanted to ask specific questions. Since CONFER cannot answer questions, they were less impressed (see Table 1) with CONFER. A tutorial with a knowledgeable human (e.g., their teacher or a tutor in the Writing Workshop) would probably be more effective.

Question 5: Are the texts written by the CONFER students more fully developed than those written by the students who did not use CONFER (specifically, are paragraphs more developed)?

The students in both the Experimental and Control groups wrote texts which ranged widely in the number of paragraphs (Experimental: 4 to 18; Control: 4 to 21) and in the number of clausal propositions (Experimental: 60 to 348; Control: 61 to 321). Thus, for both groups the variance in propositions per paragraph was large. The mean depth for the Experimental group was 33.15. For the Control group, 29.99. The standard deviations for the groups were 14.93 and 14.24, respectively. Thus, even though the Experimental group's mean was larger, that mean was not significantly different from the Control mean ($F=1.68$; $df=1,141$; $p=.19$). Therefore, no claim can be made that using CONFER altered the students' writing in terms of the depth of their texts.

Question 6: Do students evidence increased knowledge of content of the essay?

Since there were several type of measures used to answer this question, each measure will be discussed separately.

Multiple Choice

The Multiple Choice test assessed knowledge about the Percy essay. This test was given prior to the students' reading the essay (in order to determine prior knowledge), after the students read the essay (to assess knowledge gained from reading alone), and after the students used CONFER. (See Figure 1 for design)

The results (Table 4) show that there is no significant difference between the Experimental and Control groups on either the Pre-Reading or Post-Reading scores, but for both groups, the Post-Reading score is significantly ($p<.001$) higher than the Pre-Reading score. The Post-CONFER score for the Experimental group is higher than that group's Post-Reading score ($F=3.58$, $df=1,68$, $p=.0596$).

Short Answer

The short answer question test assessed knowledge about the Percy essay. This test was given prior to the students' reading the essay (in order to determine prior knowledge), after the students read the essay (to assess knowledge gained from reading alone), and after the students used CONFER. (See Figure 1 for design)

The results (Table 5) indicate that, as expected, the students in both the Experimental and Control groups had little knowledge of the specific content of the essay before they read it. The Post-Reading

scores show that half of the students in the Experimental group and forty-six percent of the Control group could fully capture the concepts. The Experimental and Control distributions are not significantly different (Chi Square=3.12, $p>.10$). After the tutorial with CONFER, the Experimental group fully captured the concepts in 84% of their answers. The Experimental group's Post-Reading and Post-CONFER distributions are significantly different (Chi Square=29.79, $p<.001$). Similarly, the Control group's Post-Reading is significantly different from the Experimental group's Post-CONFER (Chi Square=37.78, $p<.001$).

Association

The association test assessed knowledge about three concepts addressed in the Percy essay. Unlike the previous two tests, knowledge of the essay itself is not the issue. Rather, the issue is the way one marshalls knowledge about the concepts. This test was given prior to the students' reading the essay (in order to determine prior knowledge), after the students read the essay (to assess knowledge gained from reading alone), and after the students used CONFER.

The results (See Table 6) were quite similar to the results from the other two tests. The Experimental and Control groups' performance on the Pre-Reading was not significantly different ($p=.6228$), nor was there a significant difference between the groups on the Post-Reading tests ($p=.5759$). As expected, both groups' scores on the Post-Reading test were significantly higher than on the Pre-Reading test (Experimental Group: $p<.0001$; Control Group: $p<.0001$).

The Experimental Group's Post-CONFER score was significantly higher than their Post-Reading score ($p=.0089$) and significantly higher than the Control Group's Post-Reading score ($p=.0185$).

The first two tests, the multiple choice and the short answer, probed the students' knowledge of the content specific to the essay. The consistent finding that using CONFER improved scores can be interpreted in two ways. First, CONFER itself was not the reason for the improvement in the scores, but, rather, was an indirect agent. It might be the case that although the students did not read carefully (as shown by their Post-Reading scores and the interview data), using CONFER caused them to read again the material. The interview and TAP data provide some support for this interpretation. Thus, it could not be claimed that CONFER did anything more than provide a context for re-reading. To test this interpretation, it would be necessary to provide other, alternative agents, such as peer groups in which students would ask and answer questions (to prompt re-reading) or in-class tests on the essay (to prompt closer reading). If the increase in scores from such activities were the same as from using CONFER, then CONFER could not be considered a true cause of the increase.

Under the second interpretation, CONFER is the cause, and this interpretation is supported by both interview and TAP data. According to the students, CONFER allowed them to focus on the concepts which they felt they needed help with and allowed them time to think about those concepts. In short, CONFER did what Parlett claimed: it provoked incubation. Many, but certainly not all, students were surprised that a computer could sustain what they considered to be an intelligent conversation. They were also surprised that CONFER began by asking

them to select a problematic paragraph from the essay. One student's comment illustrates what others said or implied:

> I didn't expect to be allowed to pick my own problems. I thought the computer would be like a teacher. My teacher said that it [CONFER] was modelled on a real teacher. So I expected it to tell me what to think. That's what my teachers always did. So it had to quickly read over the essay again to find the place where I had trouble understanding what Percy was talking about. There were several places, but I didn't mark them beforehand; I had to find them quickly. But I couldn't. And that didn't bother the computer. I think that's what I liked best about using it [CONFER]. After a while, I didn't feel rushed like I do with teachers. I could stop for as long as I wanted and read or think or do anything. So, I think I got a lot more out of it [the tutorial] because I didn't feel pressured. I really liked that. I wish there were computers like this for every subject, especially the ones I'm not very quick at.

Conclusions and Implications

Because only one version of CONFER exists, and thus only one essay is addressed, it is not possible to fully determine the effectiveness of CONFER. Nevertheless, the results from this study indicate that CONFER has considerable promise. The students liked using CONFER and would like to have it available for other assignments and other courses, and the students who used CONFER outperformed the students who didn't in important ways.

Thus, the primary conclusions are that CONFER is a valuable resource for students, that further research on CONFER will be profitable (particularly research on which students profit most/least from using CONFER), and that more versions of CONFER should be created and tested.

While the results presented above were fairly straight-forward (they were delimited by the design), they were not the only results from this research. One unexpected, but potentially very important, finding was that teachers who were teaching "Loss of the Creature" for the first time benefitted from a conference with CONFER. They were invited to use CONFER to see how it worked and to get a sense of what their students might learn. From this relatively brief exposure (usually less than one hour), they reported to us that they got new insights into the essay and into ways to conference with their students. Thus, we may find that CONFER-like computer tutors are of value to teachers even if the value to students is less than what we might hope for.

REFERENCES

Baker, E. L. (1989). Technology assessment: Policy and methodological issues. Proceeding of the 2nd intelligent tutoring systems research forum, San Antonio, TX.

Hull, G. A. (1989). Literacy as prerequisite knowledge. In J. J. Richardson & M. C. Polson (Eds.), Proceedings of the air force forum for intelligent tutoring systems, AFHRL-TR-88-41, Brooks Air Force Base, TX, pp 143-150.

Kyllonen, P. C., & Shute, V. J. (1989). Taxonomy of learning skills. In P.L. Ackerman, R.J., Sternberg & R. Glaser (Eds), Learning and Individual Differences. San Francisco: Freeman.

Littman, D., & Soloway, E. (1988). Evaluating ITSs: The cognitive science perspective. In M. C. Polson & J. J. Richardson (Eds.), Foundations of intelligent tutoring systems, Hillsdale, NJ: Lawrence Erlbaum Associates, pp 209-242.

Barlott, J. W. (1987). CONFER: An ICAI System for Prewriting and Reflective Inquiry. Unpublished dissertation, University of Pittsburgh, Pittsburgh, PA.

Spivey, N. N. (1984). Discourse synthesis: Constructing texts in reading and writing. Newark, DE: International Reading Association.

Steuck, K., & Fleming, J.L. (1989). Issues in the evaluation of intelligent tutoring systems. Paper presented at Annual Meeting of the American Educational Research Association, San Francisco.

Tierney, R., & Cunningham, J.W. (1984). Research on teaching reading comprehension. In P.D. Pearson (Ed.), Handbook of reading research (609-654). New York: Longman.

Tierney, R., & Pearson, D., 1983. Toward a composing process model of reading. Language Arts, 60, 568-580.

Venezky, R.L., Kaestle, C.F., & Sum, A.M. (1987). The subtle danger: Reflections on the literacy abilities of America's young adults. Princeton, NJ: Center for the Study of Educational Progress.

Figure 1. Design

GROUP	TEACHER MAKES ASSIGNMENT		STUDENT READS ESSAY		STUDENT DOES CONFER		STUDENT WRITES ESSAY	
E1		T				INT		INT
E2				T	TAP	INT		
E3						T		INT
E4		T		T	TAP	INT		
E5		T				T		INT
E6				T		T+INT		
C1		T						
C2				T				
C3		T		T				

NOTES:

CONFER used only by E Groups
T: The three tests
TAP: Think-Aloud Protocol
INT: Interview

Table 1

Distribution of Student Opinions about the Helpfulness of CONFER in preparing their Text and in Understanding the Essay, by those who Planned and did not Plan before Using CONFER and those who Read and did not Read the Essay before Using CONFER

	Writing Text		Understanding Essay	
	Helped	Did Not Help	Helped	Did Not Help
All	39	32	59	12
Planned	3	13	10	6
Did Not Plan	36	19	49	6
Read	32	13	43	2
Did Not Read	4	6	6	4

Table 2

Summary Statistics of Association Knowledge Test:
Students who Planned and did not Plan before Using CONFER
and those who Read and did not Read the Essay before Using CONFER

	Post-Reading			Post-CONFER		
	N	Mean	S.D.	N	Mean	S.D.
Planned	10	6.90	.88	9	7.00	.71
Did Not Plan	24	6.25	1.70	27	7.52	1.60
Read	20	6.60	1.64	20	8.25	1.02
Did Not Read	4	4.50	.58	7	5.43	.98

Table 3

Summary Statistics of Multiple-Choice Knowledge Test:
Students who Planned and did not Plan before Using CONFER
and those who Read and did not Read the Essay before Using CONFER

	Post-Reading			Post-CONFER		
	N	Mean	S.D.	N	Mean	S.D.
Planned	10	4.50	.71	9	4.78	.44
Did Not Plan	24	4.04	1.46	27	4.59	.75
Read	20	4.35	1.39	20	4.75	.72
Did Not Read	4	2.50	.58	7	4.14	.69

Table 4

Multiple Choice Knowledge Test: Summary Statistics

By Sub-Group

Group	N	Pre-Reading Mean	Pre-Reading S.D.	Post-Reading Mean	Post-Reading S.D.	Post-CONFER Mean	Post-CONFER S.D.
E1	13	.46	.52				
E2	10			4.00	1.70		
E3	13					4.62	.65
E4	12	.58	.90	4.33	1.15		
E5	11	.36	.50			4.82	.40
E6	12			4.17	1.11	4.50	.90
C1	26	.46	.76				
C2	22			4.18	.85		
C3	24	.62	.92	4.04	1.32		

By Group

	Pre-Reading			Post-Reading			Post-CONFER		
Group	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
E	36	.47	.69	34	4.18	1.29	36	4.64	.68
C	50	.54	.84	46	4.11	1.11			

Table 5

Short Answer Knowledge Test:
Sum Responses, by Answer Category

By Sub-Group

		Pre-Reading			Post-Reading			Post-CONFER		
Group	N	A	B	C	A	B	C	A	B	C
E1	13	38	1	0						
E2	10				7	13	10			
E3	13							0	5	34
E4	12	35	1	0	4	12	20			
E5	11	32	1	0				1	8	24
E6	12				7	8	21	0	5	31
C1	26	76	2	0						
C2	22				10	27	29			
C3	24	69	2	1	7	32	33			

By Group

	Pre-Reading				Post-Reading				Post-CONFER			
Group	N	A	B	C	N	A	B	C	N	A	B	C
E	36	105	3	0	34	18	33	51	36	1	18	89
C	50	145	4	1	46	17	59	62				

Table 6

Association Knowledge Test: Summary Statistics

By Sub-Group

Group	N	Pre-Reading Mean	Pre-Reading S.D.	Post-Reading Mean	Post-Reading S.D.	Post-CONFER Mean	Post-CONFER S.D.
E1	13	3.31	1.75				
E2	10			6.60	1.78		
E3	13					7.38	.87
E4	12	4.17	1.03	6.66	1.30		
E5	11	4.55	1.29			8.00	1.18
E6	12			6.08	1.56	6.83	1.95
C1	26	4.34	1.20				
C2	22			6.95	1.25		
C3	24	4.08	1.02	6.33	1.52		

By Group

	Pre-Reading			Post-Reading			Post-CONFER		
Group	N	Mean	S.D.	N	Mean	S.D.	N	Mean	S.D.
E	36	3.97	1.46	34	6.44	1.52	36	7.39	1.43
C	50	4.22	1.11	46	6.63	1.42			

1990 USAF-UES FOLLOW-ON GRANT RESEARCH PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

THE EFFECT OF STUDENT-INSTRUCTOR INTERACTION ON ACHIEVEMENT IN COMPUTER-BASED TRAINING

Prepared by: Stanley D. Stephenson, Ph.D.,
Principle Investigator

Academic Rank: Associate Professor

Department and
University: Department of CIS/ADS
Southwest Texas State University
San Marcos, TX 78666

Date: 25 Feb 1991

Contract No: F49620-88-C-0053/SB5881-0378

The Effect of Student-Instructor Interaction on Achievement in Computer-Based Training

by

Stanley D. Stephenson

ABSTRACT

The question of what an instructor should do in CBT has not been answered nor studied. However, research on the role of the instructor in traditional instruction (TI) has shown that an effective TI instructor can influence achievement. It would appear logical to therefore assume that the behavior of the CBT instructor could likewise influence achievement. This report covers the results of two studies conducted on the role of the instructor in CBT. Ss worked a spreadsheet tutorial over a three day period. On the third day, Ss were given an exercise to complete using the spreadsheet software. Experimental dimensions studied were student-instructor interaction (present or absent) and CBT configuration; Ss worked either individually or in pairs. Results of the two studies were that instructor interaction positively influenced achievement when Ss worked CBT individually. However, if Ss worked CBT in pairs, instructor interaction had no effect. Evidently, the partner in a study team provides the social functions usually provided by an instructor in an individual configuration. These and other results are discussed.

ACKNOWLEDGEMENTS

I would like to thank the Department of Computer Information Systems/Administrative Sciences and the School of Business at Southwest Texas State University for the excellent support they have provided. In addition, the efforts of two student computer laboratory assistants were essential to the success of this effort. Without the help of Mr. Karl Kampschroeder and Mr. James Pennington, I simply could not have conducted these experiments. I am also very grateful to the United States Air Force Human Resources Laboratory and in particular the Technology Development Branch of the Training Systems Division for providing me the opportunity to conduct this research.

INTRODUCTION

The use of the computer has been steadily increasing within both industrial training and academic education. Computer-based training (CBT) appears to be more efficient, more thorough, and cheaper than traditional instructional techniques.

CBT research has typically focused on comparing a CBT course with a corresponding traditional instruction (TI) course. Compared to a similar TI course, CBT generally produces increases in learning and retention while concurrently requiring less learning time than TI (Fletcher & Rockway, 1986; Goodwin et al., 1986; Kulik & Kulik, 1986, 1987; McCombs, et al., 1984; O'Neil, 1986). However, CBT results have not always been positive: there are instances in which CBT did not produce increases in performance or decreases in learning time (Goodwin et al., 1986; McCombs et al., 1984).

In general, there has been very little research on maximizing performance within a CBT system (Gillingham & Guthrie, 1987). Conversely, there is a long history of research on variables which influence achievement in TI systems. One of the most researched TI variables is instructor behavior. TI research has produced a relatively high degree of consensus as to what an effective instructor does versus what a not-so-effective instructor does, with effective being defined in terms of academic achievement

(Brophy, 1986; Brophy & Good, 1986; Rosenshine, 1983). Yet, CBT research has neglected the role of the instructor (Moore, 1988). Little is known about the influence of instructor behavior on achievement in CBT (Stephenson, 1990).

In one of the few studies which did examine the role of the CBT instructor, Moore (1988) found that students who had teachers with positive attitudes scored higher than those in classes with teachers with negative attitudes. In a more general paper McCombs et al. (1984) reviewed a variety of CBT courses and found that two factors were critical to the success of the CBT courses. These were: (a) adequate opportunities for student-instructor interactions, and (b) the incorporation of group activities with individualized training.

As noted from McCombs et al. (1984), student-instructor interaction is a critical factor with regard to success of a CBT system. This is a significant finding since one of the most consistently reported positive TI instructor behaviors is frequent but short student-instructor interactions; i.e., an increase in student-instructor interaction produces an increase in achievement (Brophy, 1986; Brophy & Good, 1986; Rosenshine, 1983). Therefore, a TI instructor behavior which may also be a factor in CBT is student-instructor interaction. The second McCombs et al. (1984) dimension, group activities, is a dimension frequently not

present in CBT, perhaps due to the fact that CBT is typically conducted in a one student-one terminal environment.

To further define the role of the instructor in CBT, two experiments were conducted. The purpose of the first study was to examined the effect of student-instructor interaction in CBT. Based on the TI instructor literature, it was hypothesized that increased student-instructor interaction would produce increased achievement. The objective of the second experiment was to examine the effect of group activities in CBT. Both experiments were conducted in a field study environment.

EXPERIMENT 1

Method

Subjects

Subjects were 25 (15 female and 10 male) college juniors and seniors enrolled in a Business Statistics class. As part of a project designed to teach students how to use computer spreadsheet software to perform statistical computations, Ss volunteered to participate in a CBT spreadsheet tutorial for extra credit. The extra credit was awarded for project completion, not for project performance. All Ss completed a survey to assess their personal computer (PC) and spreadsheet experience.

Experimental Materials

The spreadsheet tutorial was part of a larger commercial software tutorial package designed for an integrated

spreadsheet-word processing-database program. The tutorial is basically linear and learner-controlled but does provide the capability to repeat a lesson.

For this study, the larger tutorial was modified to include just the introduction to the integrated package plus that portion of the tutorial software devoted to the use of the spreadsheet. The introduction portion (Part A) contained four lessons, and the spreadsheet portion (Part B) contained eight lessons. The tutorials were run on Tandy 1000SX PCs.

An exercise designed to evaluate mastery of the spreadsheet tutorial was added to the experimental software. Since the students were volunteers from a Business Statistics class, the exercise used statistical calculations as the vehicle for evaluating spreadsheet mastery. In sum, the experimental material consisted of a CBT spreadsheet tutorial modified to include a statistics-based exercise. The statistics exercise was also conducted on the computer.

Procedure

Ss were randomly assigned by spreadsheet/PC experience to one of two student-instructor interaction modes. Group I (n=13) had essentially no instructor-initiated interactions. All Group I interactions were initiated by the student and consisted of requests by the students for help in overcoming an obstacle in the tutorial. Group II (n=12) experienced the same type of student-initiated interactions

experienced by Group I. In addition, Group II was exposed to multiple instructor-initiated interactions. All Ss worked individually on both the tutorial and the exercise.

Both groups worked the CBT tutorial in three sessions. In session one, all Ss started on lesson 1A and worked in the tutorial for 90 minutes. In the second session, all Ss started on lesson B1 and worked through the last lesson, B8. In the third session, all Ss started on lesson B3 and again worked through the last lesson, lesson B8. Therefore, all Ss had a single exposure to lessons A1 though A4 and repeated exposure to lessons B1 through B8, the spreadsheet portions of the tutorial. Since each S went at his/her own speed, Ss' total time on task varied. At the completion of lesson B8 on day 3, all Ss were given an exercise designed to evaluate their mastery of the tutorial material. Ss had 30 minutes to work on the exercise.

During the startup period of the project (i.e., the first 15 minutes of the first session), the instructor responded to all questions in both groups to insure that the Ss were properly logged into the tutorial. For both groups, the instructor also responded to all subsequent student-initiated interactions with one or more of three responses: (1) "Try pushing the [ESCAPE] key;" (2) "Try pushing the [SPACE] bar;" or (3) "Re-boot the system and start over." These suggestions were given in sequence; e.g., if "Try pushing the [ESCAPE] key," did not correct

the problem, then the S was told to "Try pushing the [SPACE] bar." For Group I Ss, these suggestions were the only instructor-initiated interactions experienced after the first 15 minutes of session one.

In addition to the interactions listed above, Group II Ss also experienced instructor-initiated interactions. In the first session, the instructor initiated four interactions with each S. In sessions two and three, the instructor initiated three and one interactions, respectively. These interactions were related to location of keys on the Tandy keyboard. E.g., shortly before the Back Slash (\) key was needed in the tutorial, the instructor would tell the students where that key was located on the Tandy keyboard. Key location was explained and diagrammed in printed instructions given to all Ss, but for most Ss key location on the Tandy keyboard was a minor problem due to previous exposure to an IBM keyboard. Instructor-initiated interactions lasted between 5 and 10 seconds.

It should be noted that in no instance did the instructor provide information which was not available to all Ss elsewhere in the instructional materials. Also, in no instance did the instructor comment, provide feedback, or give praise on any S's performance on the tutorial.

Dependent Measures

Two dependent measures were recorded. First, Ss' performances on the exercise were scored. Second, Ss also

recorded the spreadsheet commands they actually used while working on the exercise. Since most spreadsheet procedures can be performed in more than one way (e.g., a cell entry can be changed via an EDIT command or by simply re-typing the entry), this second measure was recorded to assess how many different spreadsheet commands were actually used during the exercise.

Results

Means and standard deviations for Spreadsheet Performance and Use of Spreadsheet Commands are given in Tables 1 and 2. Due to the small sample sizes (and possible problems with the assumption of normality) and the question of equality of variance, the Mann-Whitney U non-parametric test statistic was used to analyze differences between Group I (no instructor-initiated interaction) Ss and Group II (instructor-initiated interaction) Ss.

Exercise Performance

Group II (instructor-initiated interaction) Ss significantly out performed Group I (no instructor-initiated interaction) Ss (Mann-Whitney $U = 34.50$, $p < .017$).

Use of Spreadsheet Commands

There was no difference in command usage between Group I Ss and Group II Ss; (Mann-Whitney $U = 82.00$, $p < .824$).

Sex Differences

Sex differences were not significant (for Spreadsheet Performance, Mann-Whitney $U = 56.00$, $p < .289$; for Use of

Table 1

Spreadsheet Performance

		Mean	StDev
Gp I	(No Inter) (n=13)	58.00	18.28
Gp II	(Inter) (n=12)	72.42	7.40

Table 2

Use of Spreadsheet Commands

		Mean	StDev
Gp I	(No Inter) (n=13)	32.31	7.25
Gp II	(Inter) (n=12)	30.83	7.40

Spreadsheet Commands, Mann-Whitney U = 69.50, $p < .755$).

Although the small sample sizes prevented a statistical test, the interaction between Group (Instructor Interaction/No Interaction) and Spreadsheet Experience Level (High/Low) should be mentioned. The impact of instructor interaction was not consistent across experience level. Low spreadsheet experience Ss who interacted with the instructor scored higher on the exercise than did low experience Ss who did not interact with the instructor.

Discussion

The hypothesis that increased student-instructor interaction would lead to increased achievement was supported. Given the limited length of the CBT program used in this experiment, the degree of difference in achievement between

the two groups was surprising. For some reason, having the instructor interact with/take notice of/care about the student affected the student to the point where it increased his/her achievement. The underlying cause for the difference in achievement did not seem to be knowledge. All Ss seemed to "learn" the commands presented in the tutorial; there was no difference between groups in the number of commands used to work the exercise. The difference was in how well the commands were used.

Nor was the difference in achievement due to praise or feedback, neither of which was given. Unless relatively brief human interaction is defined as praise, praise was not a factor in this study. Extra credit for higher performance on the exercise also was not a factor; all Ss received the same amount of class credit regardless of their performance.

A clue as to why Group I Ss did not perform as well as Group II Ss comes from observations made by the experimental instructor. It seemed that Group I Ss, especially low spreadsheet experience Ss, used the spacebar more frequently than did Group II Ss. In this study's tutorial, Ss had the capability to literally spacebar their way through the tutorial. I.e., rather than actually performing the requested tutorial action, Ss could depress the spacebar and sequence through the tutorial.

Although not measured, Group I (no interaction) Ss

seemed to take this approach more frequently. While both groups were equally exposed to the material, Group II Ss seem to actually perform the tutorial more. If in fact Group I Ss did spent less time on task, the spacebar behavior could account for the difference in achievement. The large standard deviation difference between the two groups could also be a result of the differing amounts of time spent on task.

If the explanation offered above is accurate, it suggests that brief human interaction serves to keep students on task more so than no human interaction. If no one is aware of what I am doing, I am more likely to try to ease my way through the CBT course, especially if I am not sure of what to do. However, if someone is aware of what I am doing, irrespective of whether or not that someone gives me praise or feedback, then I had best stay on task.

The data also suggest that low skill level Ss benefit more from interacting with the instructor. This would certainly make sense in that these Ss should be the ones who have more need for knowledge, guidance, and social support. High skill level Ss, on the other hand, know what they are about and do not require external support to master the material.

Due to the manner in which the Group II interactions occurred, instructor monitoring of the students was confounded with interaction. For the instructor to know when

to interact with an appropriate comment, the instructor had to know when a student was approaching a particular point in the tutorial. In order to know this, the instructor had to monitor the students' progress. Consequently, while the Group I instructor sat at a desk and waited for students to request assistance, the Group II instructor was constantly walking around the room and visually checking on the Ss' progress. Therefore, it may be that monitoring, and not interaction, was the basis for Group II's higher achievement.

These results add to the results reported by Moore (1988) who found that CBT teachers with positive attitudes produced higher achievement than teachers with negative attitudes. Evidently, student-instructor interaction can also affect achievement. Whether or not the interaction needs to be tied to course content is unknown.

EXPERIMENT 2

Method

Except for the two differences discussed below, the procedure in experiment 2 was identical to the procedure in experiment 1. I.e., Ss worked a spreadsheet CBT tutorial in a field study environment and were tested on their performance on a statistics exercise and their use of spreadsheet commands. In experiment 2, 41 Business Statistics students were assigned to one of two groups. Group I (n=19) received instructor interaction, and Group II (n=22)

did not receive instructor interaction.

The primary difference between experiment 1 and experiment 2 was the manner in which the Ss worked the tutorial. In experiment 2, Ss were paired by sex, grade point average, and spreadsheet experience and worked the tutorial in dyad versus individually as in experiment 1. However, as in experiment 1, experiment 2 testing was done individually. A second difference was the amount of time available during each session. Due to a change in field conditions, experiment 2 Ss had 15 fewer minutes per session than had experiment 1 Ss.

The purpose of running the Ss in dyads was to create opportunities for group activities, a dimension suggested by McCombs et al. (1984) as being critical to the success of any CBT program. Justification for this configuration of the CBT environment comes from two other sources. First, there is a body of TI literature on the effect of students working in groups versus working individually. Much of the group work has focused on the effect of cooperation versus competition. The general consensus is that students working in small groups produce higher achievement than students working alone, especially in a cooperative setting (Johnson et al., 1985; Warring et al., 1985; Yager et al., 1985). The optimum size seems to be either two or three (Cox & Berger, 1985; Trowbridge & Durnin, 1984; Webb, 1987). There is also a general consensus that paired stu-

dents should be like-gender and have similar abilities (Dalton, 1990; Dossett & Hulvershorn, 1983; Hooper et al., 1989; Johnson et al., 1985).

The second justification for arranging students in groups in CBT comes from recent CBT research. This research shows that achievement of students working CBT in dyads or triads is equal to or surpasses achievement of students working alone (Carrier & Sales, 1987; Cox & Berger, 1985; Dalton, 1990; Dalton et al., 1989; Dossett & Hulvershorn, 1983; Hmelo, 1989; Johnson et al., 1986; Justen et al., 1990; Shull, 1990; Trowbridge & Durnin, 1984; Webb, 1987). "No study has reported significantly greater learning when students work alone (Webb, 1987, p. 195)."

Results

Means and standard deviations for the two dependent measures are given in tables 3 and 4. With all Ss working the tutorial in dyads, there were no differences between the two groups (interaction/no interaction) on the spreadsheet statistics exercise or on the use of spreadsheet commands. There were also no sex differences nor were there differences in main effect interactions. The only statistical difference was for spreadsheet experience level on spreadsheet performance where Ss with prior experience in the use of spreadsheets outperformed those Ss without prior experience. The analyses of variance for the two dependent measures are given in tables 5 and 6.

Table 3

Spreadsheet Performance

Means and Standard Deviations

	Interaction	No Interaction
Low Experience	56.15/19.06	51.25/20.13
	(n=13)	(n=12)
High Experience	73.33/9.83	73.00/9.49
	(n=6)	(n=10)

Table 4

Use of Commands

Means and Standard Deviations

	Interaction	No Interaction
Low Experience	28.08/9.02	22.92/8.91
	(n=13)	(n=12)
High Experience	30.00/5.00	32.78/8.33
	(n=5)*	(n=9)*

* 2 Ss did not complete a Use of Commands form

Discussion

The student-instructor interaction effect produced in experiment 1 was not found in experiment 2. Evidently, many of the social functions usually performed by the CBT instructor when Ss work individually are taken over and performed by the dyad partner. Moreover, instructor interaction did not have a statistically significant, disproportionately larger effect on those Ss with little spreadsheet experience. However, raw mean scores were in the direction

Table 5

Spreadsheet Performance by Group and Experience and Sex

Analysis of Variance

Source	Sum Sq	DF	Mean Sq	F	Prob
Group	21.21	1	21.21	0.08	0.784
Experience	1867.11	1	1867.11	6.70	0.014
Sex	285.99	1	285.99	1.03	0.318
Group x Experience	70.61	1	70.61	0.25	0.618
Group x Sex	403.03	1	403.03	1.45	0.238
Experience x Sex	298.01	1	298.01	1.07	0.309
Group x Experience x Sex	198.77	1	298.77	0.71	0.404
Error	9198.06	33	278.73		

reported in experiment 1. Overall, it appears that even in the dyad teams composed of low experience Ss, the dyad partner evidently provides the feedback, support, and social facilitation usually provided by the instructor in a more traditional classroom setting.

Experiment 2 results re-emphasize the social nature of learning. For some students, learning is simply a social event. In the traditional classroom the instructor may

Table 6

Use of Commands by Group and Experience and Sex

Analysis of Variance

Source	Sum Sq	DF	Mean Sq	F	Prob
Group	7.23	1	7.231	0.11	0.748
Experience	174.27	1	174.27	2.53	0.122
Sex	59.12	1	59.12	0.86	0.362
Group x Experience	43.50	1	43.50	0.63	0.433
Group x Sex	0.49	1	0.49	0.01	0.933
Experience x Sex	134.43	1	134.43	1.95	0.173
Group x Experience x Sex	10.26	1	10.26	0.15	0.703
Error	2139.73	31*	10.26		

* 2 S did not complete a Use of Commands form

provide most of the social functions. In individual CBT situations the computer can not provide these functions; consequently, interaction with a human instructor has a measurable effect. However, when the social functions can be provided by a team partner, the need for interacting with an instructor is reduced.

These results question a frequent justification for

CBT; i.e., CBT's potential for 1:1 interaction. It may be that, due to the lack of social interaction, a 1 student:1 computer environment is not comparable to a 1 student:1 instructor environment. Instead, it may be that, because of the social interaction potential, a 2 students:1 computer situation is more comparable to the traditionally accepted ideal of 1 student:1 human instructor.

COMBINED DISCUSSION

The relatively short-term nature of the tutorial used in these two experiments obviously limits the generalization of these results. That limitation not withstanding, the specific conclusion from experiment 1 is that brief instructor-initiated interactions can increase achievement when Ss work CBT individually. However, instructor monitoring without interaction may produce the same result.

It may also be true that in the individual setting the CBT instructor can influence group achievement most by spending relatively more time with low skill students. Skill could be defined as, for example, selection scores (SAT, GPA, proficiency test, etc.) or by initial student performance on the training package. E.g., in the first hour of training the CBT software could be programmed to identify those students with whom the instructor might want to allocate more of his/her time. The capability to allocate time in this fashion is, of course, one of the advantages of CBT.

Experiment 2 suggests that with regard to achievement the best CBT environment may be the one in which students work CBT in pairs. The positive effect of social facilitation on learning is served quite well by the partner in a study team and, in fact, may be served quicker than by an instructor.

Since the role of the instructor in CBT is frequently undefined, the results from these two experiments give some direction as to what a CBT instructor can do to influence achievement. Moreover, since instructor-initiated interactions are controlled by the instructor, these interactions can be both built into the larger learning system (which includes the CBT subsystem) and also included into the instructor evaluation system.

A larger implication from these studies is that social interaction, either with an instructor or another student, does seem to influence achievement in CBT. The results obviously support Moore's research (1988) and McCombs et al. (1984) suggestions. There is simply something about having another human around and aware of your actions that alters your behavior. Rather than trying to design a CBT system which does away with social interaction (or designing a system which essentially ignores such interaction), CBT developers should try to find ways to incorporate the results of these and other studies into the overall CBT system.

The results of these two studies suggest that further research should be conducted on the role of the instructor in CBT. The question which appears to have the most immediacy is whether instructor interaction with low ability students can increase group achievement, even in CBT situations where the students work in pairs. This and other questions related to the role of the CBT instructor remain to be answered.

REFERENCES

Brophy, J. E. (1986). Teacher influences on student achievement. American Psychologist, October, 1069-1077.

Brophy, J. E. & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), Third Handbook of research on teaching: 328-375. New York: Macmillian.

Carrier, C. A., & Sales, G. C. (1987). Pair versus individual work on the acquisition of concepts in a computer-based instructional lesson. Journal of Computer-Based Instruction, 14, 11-17.

Cox, D. A., & Berger, C. F. (1985). The importance of group size in the use of problem-solving skills on a microcomputer. Journal of Educational Computing Research, 1, 459-468.

Dalton, D. W. (1990). The effects of cooperative learning strategies on achievement and attitudes during interactive video. Journal of Computer-Based Instruction, 17,

8-16.

Dalton, D. W., Hannafin, M. J., & Hooper, S. (1989). Effects of individual and cooperative computer-assisted instruction on student performance and attitudes. Educational Technology Research and Development, 37, 15-24.

Dossett, D. L., & Hulvershorn, P. (1983). Increasing technical training efficiency: Peer training via computer-assisted instruction. Journal of Applied Psychology, 68, 552-558.

Hmelo, C. E. (1989). Computer-assisted instruction in health professions education: A review of the published literature. Journal of Educational Technology Systems, 18, 83-101.

Fletcher, J. D., & Rockway, M. R. (1986). Computer-based education in the military. In J. A. Ellis (Ed.), Military contributions to instructional technology (pp. 175-222). New York: Praeger.

Gillingham, M. G., & Guthrie, J. T. (1987). Relationships between CBT and research on teaching. Contemporary Educational Psychology, 12, 189-199.

Goodwin, L. D., Goodwin, W. L., Nansel, A., & Helms, C. P. (1986). Cognitive and affective effects of various types of microcomputer use by preschoolers. American Educational Research Journal, 23, 348-356.

Hooper, S., Ward, T. J., Hannafin, M. J., & Clark, H. T. (1989). The effects of aptitude composition on achieve-

ment during small group learning. Journal of Computer-Based Instruction, 16, 102-109.

Johnson, R. T., Johnson, D. W. & Stanne, M. B. (1985). Effect of cooperative, competitive, and individualistic goal structures on computer-assisted instruction. Journal of Educational Psychology, 77, 668-677.

Johnson, R. T., Johnson, D. W. & Stanne, M. B. (1986). Comparison of computer-assisted cooperative, competitive, and individualistic learning,. American Educational Research Journal, 23, 382-392.

Justen, J. E., Waldrop, P. B., & Adams, T. M. (1990). Effects of paired versus individual user computer-assisted instruction and type of feedback on student achievement. Educational Technology, July, pp. 51-53.

Kulik, C. C., & Kulik, J. A. (1986). Effectiveness of computer-based education in colleges. AEDS Journal, Winter/Spring, 81-108.

Kulik, J. A., & Kulik, C. C. (1987). Review of recent research literature on computer-based instruction. Contemporary Educational Psychology, 12, 222-230.

McCombs, B. L. (1985). Instructor and group process roles in computer-based training. Educational Communication and Technology Journal, 33, 159-167.

McCombs, B. L., Back, S. M., & West, A. S. (1984). Self-paced instruction: Factors critical to implementation in Air Force technical training - A preliminary inquiry.

(AFHRL-TP-84-23). Lowery Air Force, Base, CO: Air Force Human Resources Laboratory, Training Systems Division.

Moore, B. M. (1988). Achievement in basic math skills for low performing students: A study of teachers' affect and CAI. The Journal of Experimental Education, 5, 38-44.

O'Neil, H. F., Anderson, C. L., & Freeman, J. A. (1986). Research in teaching in the Armed Forces. In M. C. Wittrock (Ed.), Third handbook of research on teaching: 971-987. New York: Macmillian.

Rosenshine, B. (1983). Teaching functions in instructional programs. The Elementary School Journal, 83, 335-351.

Shull, B. D. (1990). Your terminal or mine? Performance & Instruction, May/June, pp. 37-39.

Stephenson, S. D. (1990). Design of computer based education software: Maximizing achievement. A paper presented at the 1990 Meeting of the Southwest Decision Sciences Association.

Trowbridge, D., & Durnin, R. (1984). Results from an investigation of groups working at the computer (NSF Grant No. SED-8112633). Irvine, CA: California University, Irvine, Educational Technology Center. (ERIC Document Reproduction Service No. ED 238 724).

Warring, D., Johnson, D. W., Maruyama, G., & Johnson, R. (1985). Impact of different types of cooperative learning on cross-ethnic and cross-sex relationships. Journal of

Educational Psychology, 77, 53-59.

Webb, N. M. (1987). Peer interaction and learning with computer in small groups. Computers in Human Behavior, 3, 193-209.

Yager, S., Johnson, D. W., & Johnson, R. T. (1985). Oral discussion, group-to-individual transfer, and achievement in cooperative learning groups. Journal of Educational Psychology, 77, 60-66.

Report # 118
760-0MG-030
Prof. Christian Wagner
No Report Submitted

Research Initiation Program

FINAL REPORT

An Evaluation of Stereoscopic and Other Depth Cues in Computer Display

UES Project 210
210-10MG-112
Contract # F49620-88-C-0053

Submitted by: John Williamson
Wayne Shebilske
Texas A&M University
Department of Psychology

Submitted to: Susan K. Espy
UES, Inc.

Date: 21 December 1990

Abstract

Three experiments were conducted in an attempt to determine if accuracy and reaction time performance in ITSs may be enhanced with additional depth cues such as hidden line removal and retinal disparity (stereoscopic vision). Using existing technology, these experiments provided a preliminary cost-benefit and feasibility analysis of incorporating existing stereoscopic technology in Intelligent Tutoring Systems (ITSs). In addition, they shed some light on the underlying perceptual and cognitive processes that are used to encode and manipulate three-dimensional spatial data.

It was found that stereoscopic 3D can be successfully incorporated into existing ITS. Overall, the addition of stereoscopic 3D cues was found to be most beneficial when the stimuli are more ambiguous, either through the absence of monocular cues or incomplete/conflicting cues. This was shown in both a reduction in reaction time and error rates. Evidence for the generalization of a new Gestalt law of enclosure was found. Several new studies are proposed to expand on the results obtained.

1.0 INTRODUCTION:

While the principles for creating stereoscopic illustrations have been known for over 150 years, past applied research has not given conclusive evidence that stereoscopic, 3D presented materials improve accuracy or reaction time performance when compared to traditional, flat, 2D presentations. Even so, there are very strong industry claims that stereoscopic 3D presentation is beneficial, and there is a rapidly growing market for computer displays which can produce stereoscopic 3D images. These current technologies and the potential of virtual realities have sparked an interest in determining the most realistic training environment necessary in computer based Intelligent Tutoring Systems (ITSs).

The Intelligent Systems Division of the Human Research Laboratory at Brooks Air Force Base examines a variety of technologies which can be applied to intelligent tutoring systems (ITSs). In the past, these have included speech recognition, neural networks, expert systems and interactive graphics. One of the goals of the branch is to develop ITSs to a point where they provide the most realistic simulation necessary for accurate training. In addition to computer intelligence using either neural networks or expert systems, a successful interface must be designed to enable the most efficient communication between the student and the ITS.

While a great deal of research has been conducted by this branch in the development of intelligence in the ITS, considerably less has been conducted in regard to the interface. A large number of the domains which ITSs may be created for in the near future will involve spatial tasks (aircraft maintenance, map reading, etc.) A well-designed interface and the ability to accurately and quickly impart spatial information to a student is very important in such an ITS.

In the near future, this may also include virtual realities either for telerobotics (Figure 1) or training which is currently done on a traditional computer display. Virtual realities are a relatively recent development in computer/human interaction. One enters a virtual world, which may exist only on the computer, by wearing a lightweight headset that contains two small computer monitors. The headset is connected to a 6D motion detector which can determine not only the viewer's head orientation, but his/her location as well.

The viewer may enter a computer generated house, for example. When the viewer tilts his/her head toward the floor, the view in the computer monitors changes accordingly. If the viewer turns to look out a window, the view again changes dependent entirely upon how far the viewer moves his/her head.

The viewer may wear a "dataglove" which also has a 6D motion detector. This allows his/her hand position and finger articulation to be reported to the computer. Using this glove, the viewer may manipulate objects which appear on the monitors in

the headset, much in the same way he/she may manipulate real objects outside of the virtual reality.

Recent estimates show that the cost of virtual reality hardware may be reduced by as much as 40% if stereoscopic depth cues are omitted. Rather than each eye receiving a slightly different view generated from two different view points in space (stereoscopic vision), each eye receives an identical image generated from the same view point in space. This eliminates half of the processing required to animate virtual realities.

However, it is not known if this cost savings outweighs any performance degradation that may result. Determining the role of stereoscopic depth cues and their interaction with monocular cues (shading, interposition, etc.) in these novel environments in, addition to traditional computer displays with stereoscopic modifications, requires new research initiatives.

The current ability to adequately represent a three-dimensional object on a traditional flat, two-dimensional computer screen is rather poor. Several approaches can be taken in an attempt to overcome this shortcoming; multiple views of an object can be presented (Figure 2), the object can be made more "life-like" through shading or the removal of hidden lines (Figure 3) or the object may be animated and rotated upon the student's request.

While all of these methods may be effective to a certain degree, each places restrictions on the amount of information that can be presented on the screen. This may be due to the

additional space required to present multiple views or the additional computational time required to present animated or shaded illustrations. As a result, the lesson may be slowed down.

More importantly, the student may be forced to exert considerable effort in understanding the two-dimensional representation of the three-dimensional object. This may further impede the progress of the lesson. Valuable time may be spent manipulating the object until the student understands the image represented. Worse yet, the student may be unable or unwilling to understand the spatial relationship of a particular item and will simply move on to the next topic. This may lead to both decreased performance and frustration. In addition, because the student may work with physical, three-dimensional objects in the field after the lesson is finished, there may be a negative carry-over effect after having been trained with flat, two-dimensional lessons.

Different materials or topics covered by ITSs may require the use of task-specific combinations of depth cues to successfully impart all of their spatial information. This may further depend upon the type of display used (traditional monitor or virtual world), the viewer's orientation and distance from the display, and the stress level of the viewer's environment. The selection of the most effective technique for an individual task may lead to better retention and comprehension during the lesson itself and better performance on the task in the field, while

possibly reducing training cost with the elimination of unnecessary or redundant cues.

1.3 DEFINITION OF STEREOSCOPIC 3D

Although the principles for creating stereoscopic illustrations have been known since Charles Wheatstone invented the stereoscope in 1838 (Brewster, 1856), past applied research has not given conclusive evidence that stereoscopic, 3D presented materials improve accuracy or reaction time performance when compared to traditional, flat, 2D presentations. Prior to Julesz's random dot stereogram experiments, stereoscopic depth cues were of questionable importance in depth localization (Julesz, 1971). It was assumed that monocular cues, such as texture gradient or linear perspective were more important.

Stereoscopic depth cues are binocular; that is, to perceive them requires two eyes. Because our eyes are separated horizontally by approximately 2 1/2", each eye sees the world from a slightly different point in space. This disparity results in the retina of each eye receiving a slightly different image of the same scene. This retinal disparity is the stereoscopic cue for depth. The two disparate images are merged through a process still not fully understood and depth is perceived.

This stereoscopic depth cue is different from monocular cues such as shading or perspective, which require only one eye to be seen. While these monocular depth cues are effective at giving us depth sensations and do allow us to represent three-dimensional space in two dimensions with great accuracy, there

are times when two dimensional cues may not accurately represent the three-dimensional object.

1.2 OBJECTIVE:

Three main objectives were addressed by this grant. First, this grant examined whether the addition of stereoscopic 3D and other depth cues enhances performance in spatial tasks. Second, this grant began to look at whether these depth cue enhancements are task specific. Third we examined if spatial material is encoded differently dependent upon the type of depth cues available. In addition, the work done to program the display for the experiments can be seen as a preliminary step in determining the feasibility of incorporating stereoscopic liquid crystal shutter technology into micro-computer based ITSs,

1.3 REVIEW OF PAST RESEARCH:

Even though experimental research has been conducted on determining the benefits of stereoscopic depth cues for over 60 years (Freeman, 1927), very little support has been found for the hypothesis that the added depth cue provided in stereoscopic displays should improve performance. Past studies have examined stereoscopic benefits using a wide variety of displays. These have included ε ereocards (Freeman, 1927), polarized slides (Damron, 1951) and movies (Cogswell, 1952), anaglyph illustrations (Hatley, 1969), and liquid crystal shutter displays (Draper et al, 1988).

The lack of experimental evidence that shows a performance advantage for stereoscopic 3D displays may be explained in part by the experimental design used. Past applied studies which have examined the benefits of stereoscopic 3D presentation have tended to use accuracy as the dependent variable and have used stimuli that are rich in monocular depth information. With correct response rates as high as 95% (Kaushall & Parsons, 1981), these designs may have obtained a ceiling effect which may have obscured any differences between the monocular and stereoscopic groups.

While it may not be possible to ever be sure if subjects are using the stereoscopic depth cues provided, past studies have at least addressed this issue to a degree by having subjects take a stereoblindness test. This test determines if the subject has the ability to perceive stereoscopic depth and exists in a variety of formats. However, determining if a subject can see stereoscopic depth cues and if they actually use them in the experiment are two sperate issues. This is particularly important in designs which require the subject to exert considerable effort in seeing stereoscopic depth cues. Such presentations as traditional stereoscopes, free viewing or anaglyphic illustrations place a considerable burden on the viewer. Modern displays, such as holograms and liquid crystal shutter displays require considerably less effort.

It may be that the subjects simply do not use the stereoscopic information provided because of the extra effort

required. One way to help to assure that the subject at least initially processes the stereoscopic depth information is to have them identify at the onset of each trial if the image is in stereoscopic 3D or not (Erwin, 1978). If the subject responds correctly no better than chance, then that subject may be assumed to not be using the stereoscopic depth cues in the task itself. This procedure may also be used to possibly subtract the amount of time required for image fusion from the remainder of trial. Fusion time is estimate to be about 50msec (Julesz, 1971) for simple random dot stereogram, as determined by a masking paradigm. The total time to respond as determined by the yes/no, forced choice stereoscopic design described above has been found to be 800 msec (Erwin, 1978).

Time has been used as dependent variable in remote arm manipulation tasks. These studies examined the overall time required to complete a series of tasks with a r motely controlled "robot" arm (Fujita et al, 1986, Draper et al, 1988, Crooks and Coan, 1977). This is a different measure than looking at reaction time for discrete, individual stimuli such as target detection, and recognition (Erwin, 1978). Because of the amount of time that occurs between each step in a series of remote arm manipulation task, any benefits that may exist for stereoscopic depth cues may be overshadowed. In addition, the number of salient monocular cues, which require only one eye (interposition, shading, linear perspective) may also help to overshadow any stereoscopic depth cue benefits.

Only one stereoscopic study using discrete, individual stimuli examined reaction time as a dependent variable (Kaushall & Parsons, 1981). It should be noted that this study used actual, physical objects, not computer generated illustrations in a mental rotation paradigm. Subjects were either shown the models with both eyes open, or with their non-dominant eye covered. They were required to make a same/different discrimination on two models. Either each Shepard-Metzler model was identical or they were mirror images (isomorphs) of one another (Figure 4). They reported no significant difference in mental rotation time or accuracy between the monoscopic and stereoscopic conditions.

However, before the benefits of stereoscopic vision can be discounted, it must be noted that this design probably did not isolate the effects of stereoscopic vision. The richness of monocular cues available in this design may have overshadowed any benefits of stereoscopic depth cues. These monocular cues included shading, interposition, linear perspective. In addition, motion parallax may have also been an available cue, as the subjects may have been able to move their head from side to side. These are cues which are often not available in computer displays.

Because every improvement in computer display realism is more expensive, both in computer hardware and computational time, there is a growing interest in trying to determine if a cut-off point exists between cost and performance (Foley, 1988, Barfield

et al, 1987). While not examining the question of stereoscopic depth cues directly, Barfield et al (1987) did find some results based on predictions which may lead to similar findings for stereoscopic displays.

Barfield found that mental rotation time was significantly faster for shaded drawings than for the same drawings presented as wire frames (approximately 9.8 sec vs 8.2 sec). The Shepard and Metzler mental rotation task involves the comparison of two objects. Traditionally one of these is rotated in either the picture plane (Figure 5) or the depth plane (Figure 6) by an angle ranging from 0 degrees to 180 degrees in reference to the other object. The two objects are either identical in all respects except for angle of rotation, or a different pair. In a different pair one object is a mirror image of the other (Figure 7). This helps to assure that the subjects will perform a mental rotation of the objects in an attempt to bring them into congruence and not discover that they are different by some distinctive feature which only one of the objects possess.

Shepard and Metzler (1971) found that there was no difference between rotations that occurred in the depth plane compared to those which occurred in the picture plane. The primary finding from their paradigm was that a linear relationship exists between the time to make a same/different choice and the angle separating the two objects. Shepard and Metzler found the rate of rotation to be approximately 60 degrees per second.

Barfield found that additional processing to make the images more "realistic" (further smoothing and additional light sources) did not result in any a significant further improvement in reaction time. There was no significant difference in accuracy, as in most mental rotation tasks the success rate was above 90% (Shepard & Metzler , 1971, Shepard & Judd, 1976). Though a trend did exists as the accuracy rate was 92.5% and 96.0% for the least realistic (wireframe) and most realistic (shaded) images respectively.

It should be noted that the reaction times Barfield recorded were considerably slower than those found by Shepard and Metzler (1971). Where Barfield found reaction times ranging from approximately 4.5 sec for 0 degree rotation to approximately 12.3 sec for 180 degree rotation (approximately 23 degrees per second), Shepard and Metzler found reaction times ranging from approximately 1.2 sec at 0 degrees rotation to 4.2 sec for 180 degree rotation. While Barfield did not comment on this discrepancy, several possible explanations exist.

First, Shepard and Metzler's subject were exposed to 1600 trials, and Barfield's subjects were only shown 80 trials indicating a possible training effect. Second Barfield used pipe-shaped objects rather than the block-shaped images used by Shepard and Metzler.

Third, Barfield's stimuli were photographs of a computer monitor while Shepard and Metzler's stimuli were line drawings. This possibly resulted in Shepard and Metzler's stimuli having

greater contrast, and far less grainy images. "Graininess" defined by screen resolution has been shown to have a detrimental effect on reading speed (Nakamura, 1990) and target detection (Booth, Bryden, Cowan, Morgan and Plante, 1987). In addition, simple typographic variations have been demonstrated to have an effect on reading speed (Shebilske and Fisher, 1983). Finally, the size of the images presented by Barfield were larger than those presented by Shepard and Metzler. This may have forced Barfield's subjects to make several visual scans before they could encode the image.

Barfield based his hypothesis on the belief that cues which improve realism should reduce the time it takes for images to be encoded and manipulated, subsequently improving mental rotation time. He feels that such a finding supports the propositional rather than the analogue theory of mental imagery. The propositional theory states that images are encoded as sets of propositions in such a way that they are structured according to a set of rules. These propositions for an image are either true or false. The analogue theory states that there is a one-to-one correspondence between the image and its encoded mental representation. The original analogue theory predicts that no differences should exist in mental rotation time between either simple and more complex images or increased realism cues, so long as the image is discriminable (Barfield, 1988).

Before dismissing the benefits of stereoscopic depth cues in computer displays, we felt they should be examined in a manner

which isolated them from other sources of depth information. Additional tests are important because there are stimuli which cannot be presented with other monocular cues either because of the nature of the stimuli itself or the amount of time and money required to do so. For example, depending upon the computer used, it may take as much as 5-10 times longer to draw one hidden line drawing than the two views necessary for a stereoscopic image as wireframes.

The first hurdle in doing additional tests was to develop appropriate instrumentation. The next section will begin, therefore, with a report on the development of the general instrumentation used in all our experiments.

2.0 EXPERIMENTS:

One of John Williamson's assignments as a graduate student in the 1989 USAF-UES Summer Graduate Student Research Program and the first objective of this series of experiments was to review the equipment available for presenting stereoscopic computer generated graphics. Several demonstrations were arranged of the most promising equipment. Based on performance, availability and price it was recommended that the best current method is the liquid crystal shutter technology.

The liquid crystal shutter technology has several advantages; it can be used on any inexpensive 80286, 80386 or 80486 based IBM PC or compatible, the stereoscopic images may be either drawn using the computer itself or input through scanners

or video cameras, up to 256 colors may be used, both wire frame and solid/shaded images can be presented, and images can be manipulated in real time.

The liquid crystal shutter technology has two major variations. In the earliest devices, the liquid crystal shutters were worn as goggles in front of the eyes (this technique as been revived by StereoGraphics-Crystal Eyes product for PCs and minicomputers, SEGA-home entertainment system, and Toshiba-3D video camcorder). Each shutter alternately flashes on and off in sync with the computer display. While the left eye view is on the computer screen the right eye shutter is closed. While the right eye view is on the computer screen, the left eye shutter is closed. This assures that each eye sees only one image. Because the shutters open and close 60 times a second, there is no noticeable flicker. The only disadvantage is that only viewers who wear the $2,000 glasses can see the image in 3D.

The alternative approach is to place the liquid crystal in front of the computer screen itself. Rather than have each shutter completely block one view, the shutters polarizes the light from each image. The shutters are arranged so that the left eye image is polarized at a 90 degree angle from the right eye view. The viewer wears passive polarized glasses that resemble ordinary sunglasses. Again, only viewers with proper glasses may see the image in 3D, but the glasses cost as little as $5. In addition, there is less chance of a subject accidentally damaging the liquid crystal shutters as they do not

ever come in direct contact with them.

Two companies currently make the second stereoscopic liquid crystal shutter display discussed; Tektronix and StereoGraphics. Both packages included the necessary hi-resolution, multi-sync monitor, video card, polarizing unit and glasses and are currently priced at $9,000,

A Tektronix SGS421 Stereoscopic Display system was purchased rather than the StereoGraphics Crystal Eyes primarily because of the expense of the glasses. Due to the high number of subjects involved and possibility of accidents, the inexpensive glasses of the Tektronix system were deemed to be more suitable for a laboratory setting. The Tektronix display is more cost efficient for some studies as it allows a greater number of people to see the display simultaneously for less investment.

The Tektronix SGS421 Display System includes a 16" monitor and a Tektronix Stereoscopic Graphics Adapter Card. With this configuration it is capable of displaying a resolution of 512 X 512 pixels in a maximum of 256 colors. In addition, the Adapter Card has graphics processor on board which speeds up the drawing time of graphics primitives.

The Tektronix SGS421 Display system performed very well. The images were displayed with virtually no flicker. The glare and reflection off of the monitor were substantial however. This is to be expected to some extent as non glare coatings cannot be applied to the monitor without disrupting the polarizing effect of the liquid crystal.

This glare was further accentuated by the method in which the stimuli was presented. To increase contrast and reduce the time to draw the images, the stimuli were presented against a black background. The glare was eliminated during the experiments by turning off the lights in the room except for one small red light above and behind the monitor.

In order to conduct the experiments over 15,000 lines of code were written in C, BASIC and Assembler in the development of programs to draw and manipulate the stimuli and record the subject's responses. Traditional IBM PC algorithms needed to be modified because of the unusual nature of the Tektronix's screen and number of images required to presented stereoscopic 3D images. The resolution on the Tektronix monitor is 512 X 512 pixels which has no corresponding screen mode in the IBM PC. In addition, several of the more traditional algorithms proved to be too slow for the presentation of data in the experimental setting required by the 3D monitor. Also, to display images in stereoscopic 3D, each image must be drawn twice (in most experiments, they were drawn 6 times). This further accentuated the sluggishness of several algorithms.

The lack of appropriate, inexpensive or public domain software for this system lead to the development of a wide variety of programs in house In essence, a Stereoscopic 3D CAD (Computer Aided Design) package, in addition to several other smaller packages had to be written essentially from scratch. These programs represent the majority of the time invested on

this grant and have demonstrated an excellent potential for their inclusion in additional studies.

The programs written for the following experiments were all written such that they can easily be modified to present a wide range of stimuli, from Shepard Metzler blocks to contour maps, to skylines and building plans. The majority of these can be presented in a variety of ways, wire-frame, hidden-line, shaded, orthographic, perspective, scaled, rotated to any angle, in a wide variety of colors and in either stereoscopic 3D or 2D.

The majority of foreseeable stimuli will require no modification to the programs. Current modifications to the programs have included animation of wire frame objects in stereoscopic 3D in near real-time, a simple stereoscopic 3D flight simulator, the conversion of the output to display anaglyph drawings, and panning images to examine motion parallax. Future modifications and those currently under way include a better shading algorithm, faster animation through the use of Assembler routines, more efficient use of the math co-processor, a flight simulator with better collision detection and the possible integration of an eye tracker and "dataglove" type interface.

The proposed experiments initially were to have included the use of shaded images, both in stereoscopic 3D and 2D. The large amount of time required to draw a shaded image with available algorithms and the poor shaded image produced resulted in this part of the study being removed. Several pilot studies had been

conducted prior to the removal of these conditions.

The shaded images took on average 2 1/2 minutes to produce the images necessary for the experiment. While relatively fast considering that 6 views of the object were required, the delay between presentations of the images after the subject responded lead all of the pilot subjects to complain. For comparison, the wire frame and hidden line images only took 3 sec between pairs. Because the Tektronix display alternately presents left and right views, the 2D images required that two views be drawn for each image, for the 2D images, these views were identical. For the 3D images, these views are slightly different.

The shaded images were rather poor in appearance. Because of the shape of Shepard Metzler figures, and a simple, disperse, one-light source algorithm used only 3 colors (as only 3 sides of any cube are visible) would be drawn on each Shepard Metzler figure for shading. If curved surfaces, or multiple light sources or ray tracing could have been incorporated, far more colors could have been used. However, this would have dramatically increased the drawing time.

The images could not have been drawn before hand and then saved to disk because of the excessive amount of disk space this would have required for the large number of images needed. Each image would require two large files, one for the left eye, and one for the right eye. The time to load and display these files was found to be excessively slow for this experimental design.

In addition to the boredom reported by the subjects waiting

for the shaded images to be drawn, the long delay also made it impractical because of the large number of trials required in addition to the training. The first Shepard Metzler experiment usually took slightly over an hour with 160 trials which took 3 sec to display. To have made this a 240 trial experiment with 2 1/2 minutes between trials would have been excessive.

Three pilot subjects were run and no difference was found between the shaded vs hidden line conditions. This differs from Barfield (1987) who did find significant reaction time differences between the hidden line and shaded drawings. However, he was able to rapidly present the trials using a slide projector and he used rounded objects resulting in more colors for the shaded images.

By examining a variety of depth cues we were able to address our first question, determining if the addition of stereoscopic 3D and other depth cues enhances performance. Additionally, we used two different tasks and two different orientations of objects within one of those tasks. This allowed us to examine our second question, are depth cue enhancements task specific? By using a mental rotation paradigm, we were able to gain insight into the encoding and processing of spatial material using different depth cues, our third question. In addition, another of our experiments has shown us that we may be able to generalize the application of a new law of organization.

Experiment 1

Subjects

Twenty-eight subjects were used, with a mean age of 21, 13 were female and 15 were male. All received class credit for participation in the hour long study.

Materials

An IBM compatible, 80286 based personal computer running at 12 MHz was used to administer the materials and collect data. The materials were presented on a Tektronix's SGS421 stereoscopic liquid crystal display system. All reaction times and responses were saved to disk.

Procedure

After running several pilot groups it was realized that two changes would have to be made. First, the subjects were taking far longer to complete the task than was anticipated and second, they were still unclear on the requirements of the mental rotation task and the distinction between 3D and 2D. Several changes were subsequently made which alleviated these problems.

When the subjects first arrived, they were asked if they had any gross visual deficiencies other than a need for glasses or contacts. Specifically, they were asked if they had vision in both eyes. They were given a brief overview of the task, how long it would take and were requested to sign an informed consent form.

Because of confusion expressed by subjects in the pilot study, the subjects were given a brief explanation of the mental

rotation task. This included both isomers (mirror images) and same pairs. As in Barfield's study, they were shown LEGO models of shapes similar, but not identical, to those which were to be presented on the computer screen. Identical pairs of models were shown to the subjects to illustrate the concept of "same", and isomers were shown to the subjects to illustrate the concept of "different." Both of these were presented at a variety of different angles until the subject was comfortable with the task.

Because of confusion in early pilot studies, the subjects were then given explanations of what was defined as two-dimensional (2D) and three-dimension (3D) drawings. In effect, two-dimensional meant monoscopic and three-dimensional meant stereoscopic. Prior to this demonstration, in the pilot study, subjects were responding as to whether the object illustrated was a three dimensional shape, not whether it was presented in stereoscopic 3D.

To demonstrate the stereoscopic 3D effect, subjects were shown a vectograph of the skyline of New York. A vectograph is a type of 3D print which requires polarized glasses to see depth. One eye was alternately covered and uncovered to illustrate how an image would appear to "pop out" when it was a 3D image.

The lights were dimmed and the subjects were next tested for stereoblindess using the Tektronix screen. In addition to testing for stereoblindness, this also allowed the subjects to become familiar with the keyboard and gave them time to dark adapt. The experiment was run in near darkness primarily to

reduce the glare and reflections on the monitor. It also served to give the subjects a better image on the monitor as the two sheets of polarizers and glasses reduce the illumination of the monitor considerably.

The monitor has two sheets of polarizing LCD in front of it, in addition the subjects wore polarized "sun glasses." The result was a dimmer image than one would like for experimental conditions. In addition, the background for the stimuli was black and this resulted in unacceptable reflections from the off-white walls and filling cabinets.

The subjects sat in a chair placed in front of the Tektronix monitor with the IBM computer Keyboard placed in their lap. The subjects were initially placed 28 inches away from the monitor. They were instructed that they could move closer or further away if needed. While virtually all of the subjects moved slightly, none of the subjects moved to a distance closer than 20 inches and none moved to a distance further than 38 inches.

The subjects were then given a modified stereoblindness test designed by John Williamson. The subjects were presented with 3 shapes, a circle, an oval and a square. Using the stereoscopic display the shapes would appear to be floating on top of one another, separated by 1/4 to 1/8 of an inch. The subject's task was to determine either which shape was closest to them or alternately which shape was furthest from them as prompted by the computer display. The only cues available to make this decision were stereoscopic depth cues. No monocular cues were provided.

Subjects were presented with 6 of these screens, one following another immediately after each keypress.

The subjects were then shown 9 sample mental rotation trials with the experimenter providing feedback. This was done to further familiarize the subjects with the stereoscopic 3D effect, the keyboard and the mental rotation task. The trials were illustrated with stimuli that were similar to but not identical to the Shepard Metzler figures used in the actual experiment and any questions the subject had were answered. Correct responses were provided when needed. The subjects then were shown 7 more sample trials without experimenter feedback. The stereoblindness test was then retaken.

The Shepard and Metzler figures presented could be of four types. Either in the traditional two-dimensional hidden line drawing used in the original experiments, a three-dimensional hidden line drawing, a two dimensional wire frame drawing or a three dimensional wire frame drawing. Two types of graphic images (wire frame, hidden line removal), two levels of depth (2D, 3D), 4 angles of rotation (0, 60, 120, 180 degrees), and two possible choices (same - identical images, different - mirror images) exist, giving 32 different combinations of stimuli. Five different Shepard Metzler figures were used for a total of 160 trials which were presented randomly to each subject.

The trials consisted of first presenting one Shepard Metzler figure on the left side of the screen. The subject was to respond if this image was presented in stereoscopic 3D or not.

Immediately following the subjects response to the question "Is this image in 3D?", the second Shepard Metzler figure would appear along with the question "Are these two shapes identical?". The subject's task was now to determine if the two figures were different rotations of the same object or two different objects. Due to the number of calculations required to present each stimuli, there was a 3 second delay between the conclusion of one stimulus pair and the presentation of the second.

All of the figures were presented at the same scale and perspective. All were rotated in the depth plane (if they were animated they would appear to rotate through the monitor). Shepard and Metzler reported no differences between picture and depth plane rotation. Each figure was drawn such that part of it would appear to float in front of the screen and part of it would be behind the screen when presented in 3D. Each figure consisted of 10 blocks with three right angle bends and took up a space of no more than 5 X 3 inches on the display monitor. This was the largest the images could be drawn without overlapping one another.

The trials were broken down into this two step process to make sure that the subjects saw each pair in 3D and to factor out any time required to fuse the two images and see the figure in 3D.

Upon completion of the experiment, the subjects were debriefed and any remaining question they had were answered.

Results

All of the subjects passed the stereoblindness test with at least 83 percent correct, or no more than two out of 12 wrong. An anaglyph version of this test had been used in a previous, unpublished study by John Williamson with very good results at determining stereoblindness. This absence of stereoblindness in the subjects could be possibly be explained by the relatively young age, and type of stereoscopic 3D presented and good health of the subjects.

The Tektronix 3D technique requires virtually no training to perceive stereoscopic depth. This is very different from anaglyph and traditional stereoscopic presentations which require training and color vision. While usually not explicitly stated, the majority of previous experiments did not use the LCD display to test subjects for stereoblindness. This instrumentation difference coupled with the possibility of an older subject pool may, help to account for their larger estimates of stereoblindness estimated between 2 and 10 percent of the population (Erwin, 1978) (Julesz, 1971).

Overall, the reaction times were slower for this study than the Shepard Metzler study; but, in agreement with their study, we found a significant trend as a function of angle of rotation ($F(1,117) = 1.54$, $p < .01$). For all types figures (2D wire, 3D wire, 2d hidden, 3D hidden) the mean reaction times were 5.7 sec for 0 degrees of rotation, 6.7 sec for 60 degrees of rotation, 7.6 sec for 120 degrees of rotation, 7.8 sec for 180 degrees of

rotation (Figure 8). Figure 9 shows that the mean reaction time for wire frame figures (7.8 sec) was higher than that for hidden line figures (6.2 sec) (F (1,39) = 1.99, p < .01). Figure 10 shows that, as predicted, the overall reaction time was lower for 3D (6.4 sec) than 2D (7.6 sec) (F(1,39) = 2.04, p < .01). As shown in Figure 11, the superiority of 3D over 2D was substantial for wire frame stimuli (6.8 sec vs 8.8 sec) and much smaller for hidden line stimuli (6.1 sec vs 6.3 sec) (F(1,117) = 1.76, p < .01).

Although the reaction times were higher than those obtained by Shepard and Metzler, they are comparable to those found by Barfield. Possible reasons for the discrepancy between the Shepard Metzler study and Barfield's were listed above in greater detail. To quickly review, first, Shepard and Metzler's subjects were exposed to 1600 trials, where Barfield's subjects were only shown 80 trials and these subjects were shown 160 trials indicating a possible training effect. The influence of training is supported in the present study by a decrease in reaction time across trials. Figure 12 shows that the first 40 trials had a mean of 9.4 sec, the second 40 trials had a mean of 7.0 sec, the third quartile had a mean of 5.7 sec, and the fourth 40 trials had a mean of 5.8 sec (F(1,117) = 2.12, p < .01).

Second, Barfield's stimuli were photographs of a computer monitor and these stimuli were presented on a computer monitor where Shepard and Metzler's stimuli were line drawings. This possibly resulted in Shepard and Metzler's stimuli having

greater contrast, and far less grainy images. Finally, the size of the image presented by this experiment and Barfield were larger than those presented by Shepard and Metzler. This may have forced both these subjects and Barfield's subjects to make several visual scans before they could encode the image.

There were substantial differences in accuracy as shown in Figure 13. The 2D wire frame had a correct response rate of 73%, the 3D wire frame had a correct response rate of 89%, the 2D hidden line 90% and the 3D hidden line 92% (CHI-SQUARE (9) = 251.2, $p < .01$). The last three are all in line with correct rates reported by Shepard and Metzler (1971) and Barfield (1987). The dramatic increase in accuracy between the 2D wire frame and 3D wire frame is very interesting. The only difference between these two groups is the addition of stereoscopic 3D.

Experiments 2 and 3

Another phase of this research investigate subjects' ability to locate figures in 2D and 3D presentations of contour maps and Shepard metzler figures. The original purpose of this study was to examine if the benefits of stereoscopic 3D where task and stimulus specific. Serendipity changed the focus of this experiment toward an evaluation of the laws of organization in determining the relative benefits of 2D and 3D displays.

Subjects

Ten male subjects were used. With a mean age of 19.3. All received class credit for participation in the hour long study.

Materials

The same computer and Tektronix display were used to present the stimuli. The stimuli in Experiment 2 were representations of contour maps. There were four types of figures presented for each condition: the stimuli were be either in stereoscopic 3D or 2D and the contour map were presented from a viewing angle of either directly above or at an angle from the side and slightly above.

The stimuli in Experiment 3 were four types of Shepard metzler figures presented for each condition, 2D hidden line, 3D hidden line, 2D wire frame, or 3D wire frame. The blocks were rotated in such a way as to have approximately half of each figure floating in front of the screen and half of each figure floating behind the screen when viewed in the stereoscopic 3D condition.

Procedure

In Experiment 2, the subjects were first informed of the purpose and scope of the study and were requested to sign an informed consent form. The subjects were given several examples of contour maps and the method of representing elevation by using concentric circles. The lighting was the same as in the first experiment. The subjects were presented with the stereoblindness test prior to the beginning the task.

Upon completion of the stereoblindness test, the subjects were instructed that they would be shown contour maps on the computer screen. Six practice trials were shown on the screen

with the experimenter providing feedback prior to the presentation of 40 experimental trials.

In each trial, the subject would first be shown a screen with a contour map. The contour map could be either viewed from directly above in 2D, directly above in 3D, viewed from the side and above in 2D or viewed from the side and above in 3D. The first screen asked the subject "Is this image in 3D?" After the subject had responded either yes or no, the second question "Is the marker on the same level as #x?" appeared. Each concentric ring was now labeled 1 through 5 with the number placed inside the ring. The marker ("$") would appear inside of a ring which would either be the same ring in the prompt or a ring one above or below. For example, if the marker ("$") appeared in the ring labeled 3,and the question read "Is the maker in the same level as #4?" the correct answer would be "no".

In Experiment 3, the subjects were first informed of the purpose and scope of the study and had already signed an informed consent form and had already taken the stereoblindness test.

The subjects were told that the next task would require them to make simple spatial location decisions about two points on figure. The first image appeared with the question "Is this Image in 3D?" Upon answering this question, two points on the figure would be labeled "A" and "B." The screen would either ask, "Which shape is closest to you" or "Which shape is furthest from you?" This modification of the proposed study was made to

keep the answering in a two option-forced choice paradigm rather than the 3 choices proposed. The subjects repeated this task 32 times.

Upon completion of this task the subjects were debriefed and any questions that they had on either part of the experiment were answered.

Results and Discussion

The results seemed disappointing at first because there were no differences between conditions in either Experiment 2 or 3. These results suggested that the utility of 3D presentations is indeed task specific. Our attention shifted, therefore, toward determining why 3D presentations had lost the substantial advantage they had shown in Experiment 1. Our disappointment vanished when we discovered that we could explain why 2D worked as well as 3D in Experiments 2 and 3, and we could do so in terms of laws of organization that promise to provide general guiding principles for predicting the relative effectiveness of 2D and 3D presentations.

A reexamination of or stimuli in light of our results indicated that our experiments enabled us to generalize the application of a new law of organization, the law of enclosure or common region, proposed by Rock and Palmer (1990). According to this law, observers have a tendency to group elements that are located in the same perceived region.

In the following example: (. .)(. .)(. .)(. .) equally spaced dots are grouped by perceived regions. This demonstration

and similar ones have been used to show that the law of enclosure holds for grouping objects in the picture plane.

We found that the law also holds for grouping in depth when stratified stimuli have regions perceived at different depths. In Figure 14a, the 2 and the $ are both seen at the same depth because there is a tendency to perceive both enclosed by region 2. In figure 14b, the 2 and the $ are seen at different depths because there is a tendency to see them enclosed by different strata regions. These tendencies also prevail in side views (Figure 14c).

These tendencies are very functional for contour maps. That is, the natural groupings in Figures 14a, 14b, and 14c are very likely to be correct if the figures represent stratified land formations, which was our original intent. Because of the law of enclosure, therefore, nearly perfect performance can be obtained without 3D presentations of land stratification. The same reasoning applies to locating symbols in Shepard Metzler figures.

This serendipitous finding changed the focus of our study to a different question: Can 3D displays overcome the organizational tendencies described by the law of enclosure? This question is important for representing situations in which the law of enclosure will be dysfunctional. Suppose, that figure 14a represented particles floating in a funnel. There is still a tendency to group both particles, the 2 and the $ together at the funnel's level 2. But now this grouping could be misleading, because the particles could be at levels 2, 3, or 4.

Can a stereoscopic 3D display overcome this tendency? Yes. referring again to Figure 14a, when we put the $ at level 2 and the $ at level 4 in a 3D presentation, subjects saw the locations quickly and accurately.

These findings suggest that the relative accuracy of 2D and 3D displays depend on whether or not the law of enclosure will lead to veridical perception. When materials are limited to stimuli for which the law of enclosure will always be veridical, 2D displays will be as accurate as 3D; when materials include stimuli for which the law of enclosure might cause misrepresentations, 3D displays will be much more accurate.

We are currently studying generalizations of this conclusion to other laws of organization that can lead to veridical perceptions in some situations and to misrepresentation in others. Consider, the Gestalt laws of proximity and similarity. Objects that are close to one another tend to be grouped, (e.g. ** ** ** **), and objects that are similar to one another tend to be grouped (e.g. * * @ @ + +).

Now imagine that these examples represent top views of stimuli located at two different depths. In the second example, the first one of each shape might be located on a lower plane than the others. Would a 3D presentation overcome the tendency to group the similar objects together? Based on our results with the law of enclosure, we predict that it will, not only for the law of similarity, but also for the law of proximity and other

laws. If these predictions are supported, the results will set the stage for establishing general principles for predicting the relative accuracy of 2D and 3D displays.

3.0 CONCLUSION:

Because of their nature and complexity, the creation of the materials required for the above experiments can be also be used as a preliminary step in examining the feasibility of incorporating stereoscopic 3D displays in ITSs. While the stereoscopic display does require modifications to existing algorithms, once they are made, the display can be programmed without great difficulty. In addition, because the stereo display acts as a separate unit from the monitor already attached to the computer, the amount of screen space to display information is effectively doubled. The stereoscopic display could be used to display 3D images, while the computer monitor is used to present accompanying text. Interactive video disk systems employ a similar design.

The images displayed on the stereoscopic system are very aesthetically pleasing and attention grabbing. Virtually all of the subjects were impressed with the 3D effect and several even asked to repeat the 3D portions of the experiment.

Overall, it was demonstrated that the more ambiguous the stimuli, the more likely stereoscopic 3D will have a beneficial effect. This was seen in the dramatic increase in accuracy and reaction time on the mental rotation task between 2D wireframe and 3D wireframe drawings. This finding has immediate applied

benefits. Because of the complexity and time required, nearly every CAD package presents its images as wireframe drawings. Only when requested will the package do hidden line removal or shading of the image. Becaus this image enhancement can represent a substantial amount of time, the CAD user will not do hidden line removal or shading except when they either need to be sure of the orientation of the objects in the image or to create a final image for presentation.

It is very easy to confuse the orientation of a 3D object on a 2D screen. Most CAD packages attempt to overcome this by presenting multiple views of the image. However, each of these views requires additional time to draw. The results of the first experiment can be seen as the first step in demonstrating that the addition of stereoscopic 3D to CAD packages could reduce the time to create an object by reducing the number of errors made in determining orientation, allowing for more efficient use of shading and hidden line removals.

The improvement in reaction times in the mental rotation experiment supports Barfield's hypothesis that increased realism cues will improve performance in mental rotation tasks. This can also be seen as support for the propositional model of mental rotation. The propositional model predicts that anything that can improve the formation of the propositions will decrease mental rotation time. The original analogue theory predicts that no differences will be found in mental rotation time between either simple and more complex images or increased realism cues.

Further insight into cognitive and perceptual organization questions were gained in experiments 2 and 3. Based on these experiments, we were able to generalize the new law of enclosure to stimuli separated in depth.

New studies are also being constructed to further test the generalizability to other laws of organization. These include the Gestalt laws of closure, proximity and enclosure. The results from these additional studies, if our predictions prove accurate, will allow us to begin to establish general rules to be used to predict the success, usefulness, and accuracy of 2D and 3D displays.

Among the laws of organization being reviewed, are those related to size and shape constancy especially in moving patterns (e.g. Perrone (1986). We are particularly interested in those laws as they relate to pseudo motion parallax, one of the aspects of the stereoscopic displays that adds to the realism of the images. When a person sits in front of the display, with their glasses on, and moves from side to side, the image also appears to move. This apparent motion of the image is similar to the motion parallax one experiences in a moving vehicle, except the points on the image move in exactly the opposite direction they would if it were a real object. Even with this "error", none of the subjects had any difficulty with the perception of the image. In fact, they usually felt that moving their head from side to side (motion parallax) made the image appear even more lifelike or 3D. We would like to examine why this error is not noticed by

the subjects and what the consequences of this unnoticed motion error might be for the perception of animated stimuli.

We would also like to investigate individual differences in this and other perceptions in stereodisplays including stereoblindness. Estimates of stereoblindess in the general population have be found to vary from two to ten percent. We now have the capability to examine if these differences in these estimates are caused, in part, by the type of the stereoscopic presentation used. Using the Tektronix Stereoscopic Display System and the software we have written we can display polarized stereo-pairs (the original use of the Tektronix display), anaglyph stereo-pairs (different colored lenses are used rather than polarizers) and side-by-side stereo-pairs (similar to those first designed by Wheatsone).

There are other unresolved differences in the literature that we plan to address with our system. For instance, the differences in overall reaction time in Barfield et. al. (1987) and our Experiment 1 compared to those found by Shepard and Metzler may be explained, in part, by the size of the images used. The resolution of the Tektronix monitor does not allow us to make the images as small as those used by Shepard and Metzler. The possibility that size differences caused different scan times brought to our attention the realization that people may attend to different parts of a figure if the image is in 3D or 2D. Different edges and contours may be scanned more heavily if stereoscopic depth is present or absent. Using an eye-tracking

device in conjunction with the Tektronix display and the software we have already developed, we could examine this question programmatically.

The findings from a study such as this could also be used to predict the results of another proposed study: examining the use of Stereoscopic 3D as an attention getting device when color or sound either cannot be used, or are already be used to the extent that additional cues of that nature could result in overload. Two possible applied designs for this research include an Air Traffic Controller (ATC) simulation and threat/target detection in a cockpit. An unpublished study by John Williamson, found that using an ATC simulation, subjects were more quickly able to distinguish relative altitudes among planes if stereoscopic 3D cues were available. Other than the altitude printed below each plane, no other cues were present. In a cockpit design, stereoscopic cues could be used to distinguish altitude of targets and or threats. Stereoscopic cues could alternately be used to represent threat severity rather than altitude. The greatest threat or best target could be made to float in front of the others.

We are also in the preliminary stages of incorporating an inexpensive dataglove to be used in conjunction with Tektronix monitor. On most CAD systems several input devices typically need to be adjusted to move in each of the three dimensions. For example, a mouse may be used to maneuver in the picture plane, the keyboard may be used to move in and out of the depth plane.

With the addition of the dataglove, only one device would be needed to move any direction. We would like to examine the effectiveness and efficiency in which subjects could learn to use this type of 3D interface with stereoscopic images over traditional 2D interfaces (mouse, keyboard) with stereoscopic images.

The implementation of this dataglove would also allow us to examine a variety of spatial location tasks and motor control in manual tracking. Previous research has established that stereoscopic displays do generally permit superior three-axis manual tracking when one joystick is used to control horizontal position on a grid and another joystick is used to control vertical position perpendicular to the grid plan (Kim, Ellis, Tyler, Hannaford, and Stark (1987). The same research found, however, that monoscopic displays allow equivalent performance in special conditions. A first step for this part of our research would be to replicate the Kim et. al. experiments using a data glove in place of the two joysticks. A next step would be to predict the relative effectiveness of stereoscopic and monoscopic displays from the laws of organization that we are currently studying in other tasks.

The majority of the effort expended on this grant was in the construction of the materials and programs used in the experiments. Every step was taken to make the resulting programs as expandable and flexible as possible, given the restraint of the need for fast drawing time. As a result, the programs, with

modification, are currently be used in the development of an expanded systematic evaluation of stereoscopic and other depth cues in computer displays.

Bibliography

Barfield, Woodrow, Sanford, James, Foley, James, The Mental Rotation and Perceived Realism of Computer-Generated Three-Dimensional Images, International Journal of Man-Machine Studies, 29, pp. 669-684, 1987.

Booth, K. S., Bryden, M. P., Cowan, W. B., Morgan, M. F. and Plante, B. L, On the Parameters of Human Visual Performance: AN Investigation of the Benefits of Antialiasing. Computer Human Interaction and Graphics Interface Proceedings, 13-19, 1987.

Brewster, Sir David, The Stereoscope:Its History, Theory and Construction, London, Morgan and Morgan, Inc. 1856.

Crooks, W. H. and Coans, P. P., TV Requirements for Manipulation in Space, Mechanism and Machine, 12, pp. 425-438, 1977.

Damron, D. F., A Possible Role for Two and Three Dimensional Slide Images When Used with Tachistoscopic Training Techniques in Instructing High School Football Defenses, Doctor's Thesis, Indiana University, August, 1951.

Draper, J. V., Handel, S. J., Herndon, J. N. and Fujita, Y., High Definition Television:Evaluation for Remote Task Performance, International Symposium on Teleoperation and Control, England, 1988.

Erwin, Donald E., The Importance of Providing Stereoscopic Vision in Training for Nap-Of-The-Earth (NOE) Flight, Proceedings of the Human Factors Society, 22nd Annual Meeting, 1978.

Foley, James, Interfaces for Advanced Control, Scientific American:Trends in Computing, 1, pp. 62-67, 1988.

Freeman, Frank N. (ed), Visual Education:A Comparative Study of Motion Picture and Other Methods of Instruction, Chicago, The University of Chicago Press, 1924.

Fujita, Y., Omori, E., Hayashi, S., Drapert, J. V., Herndon, J. 2N., High Definition Television Evaluation for Remote Handling Task Performance, Department of Energy Technical Report, DE87-00039.

Hatley, Jimmy David, The Development and Evaluation of Linear Stereoscopic Teaching Devices for College Descriptive Geometry, Doctors Thesis, Texas A&M University, August, 1969.

Julesz, Bela, Foundations of Cyclopean Perception, Chicago, The University of Chicago Press, 1971.

Kaushall & Parson, Optical Information and Practice in the Discrimination of 3-D Mirror-Reflected Objects, Perception, 10 (5), pp. 545-562, 1981.

Nakamura, Roxanna Li, Better CRT's Shown to Boost Reading Speeds, InfoWorld, September, 1990, p. 23. pp. 425-438, 1977.

Kim, W.S., Ellis, S.R., Tyler, M.E., Hannaford, B., and Stark, L.W. Quantitative Evaluation of Perspective and Stereoscopic Displays in Three-Axis Manual Tracking Tasks. IEEE Transactions on Systems, Man, and Cybernetics, 17, 61-72, 1987.

Perrone, J.A. Anisotropic responses to motion toward and away from the eye. Perception & Psychophysics, 39, 1-8, 1986.

Rock I. and Palmer, S. The Legacy of Gestalt Psychology. Scientific American, December, pp. 84-90, 1990.

Shebilske, Wayne, L. and Fisher, Dennis F., Eye Movements and Context Effects during Reading of Extend Discourse. In K. Rayner (Ed.) Eye Movements in Reading: Perceptual and Language Processes, pp. 153-178, 1983.

Shepard, R. and Metzler, J., Mental Rotation of Three-Dimensional Objects, Science, 171, pp. 701-703, 1971.

Shepard, R. and Judd, S., Perceptual Illusion of Rotation of Three-Dimensional Objects, Science, 176, pp. 952-954, 1976.

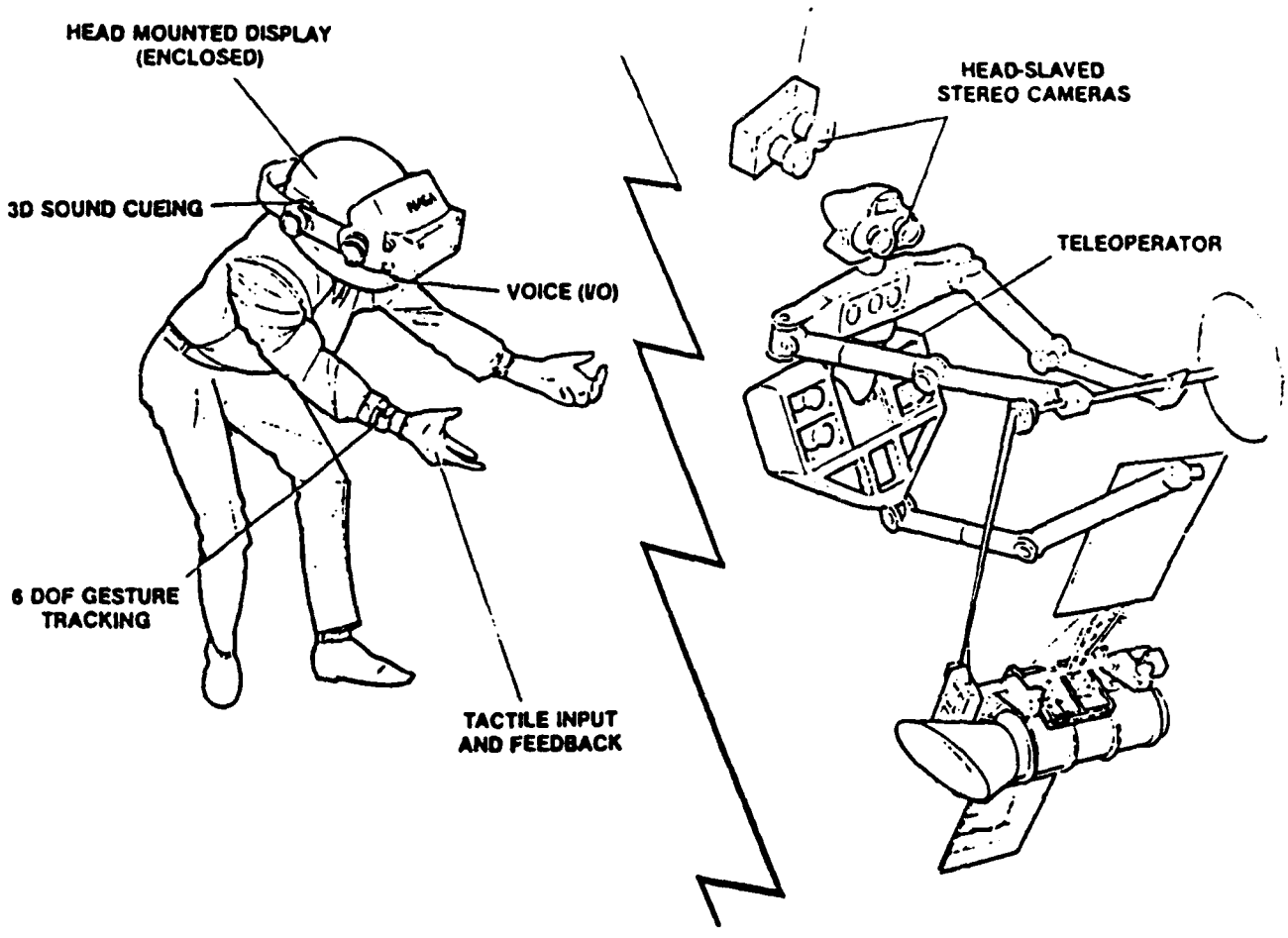
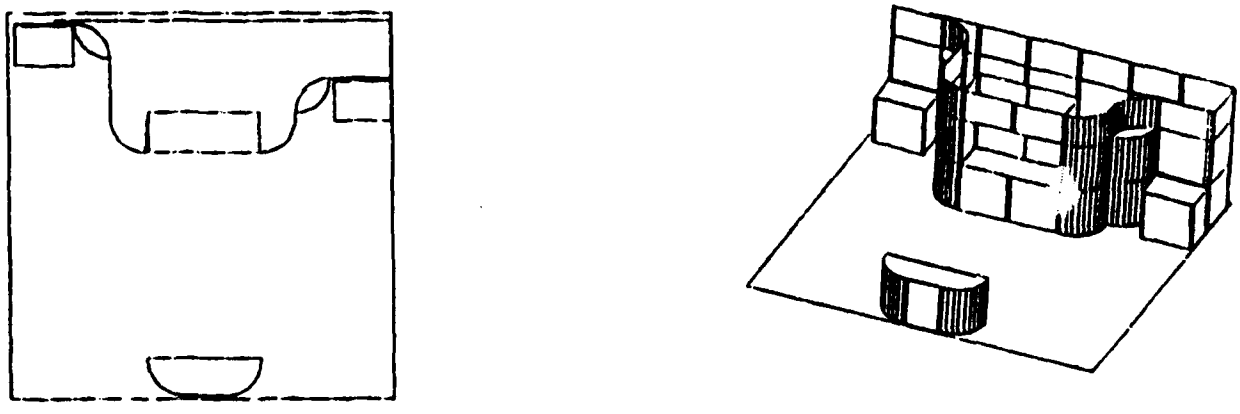


Figure 1.
Telerobotics



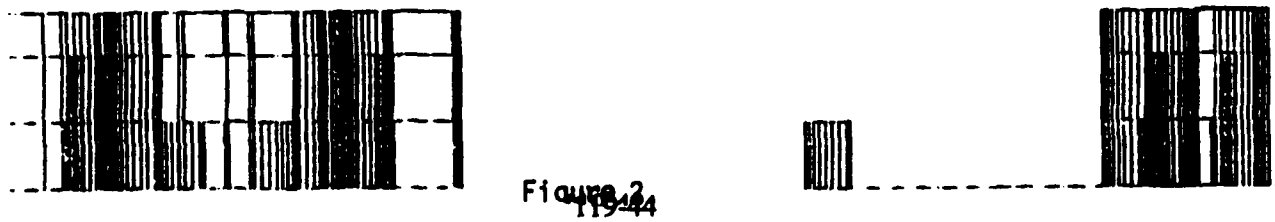


Figure 2
Multiple View Presentation

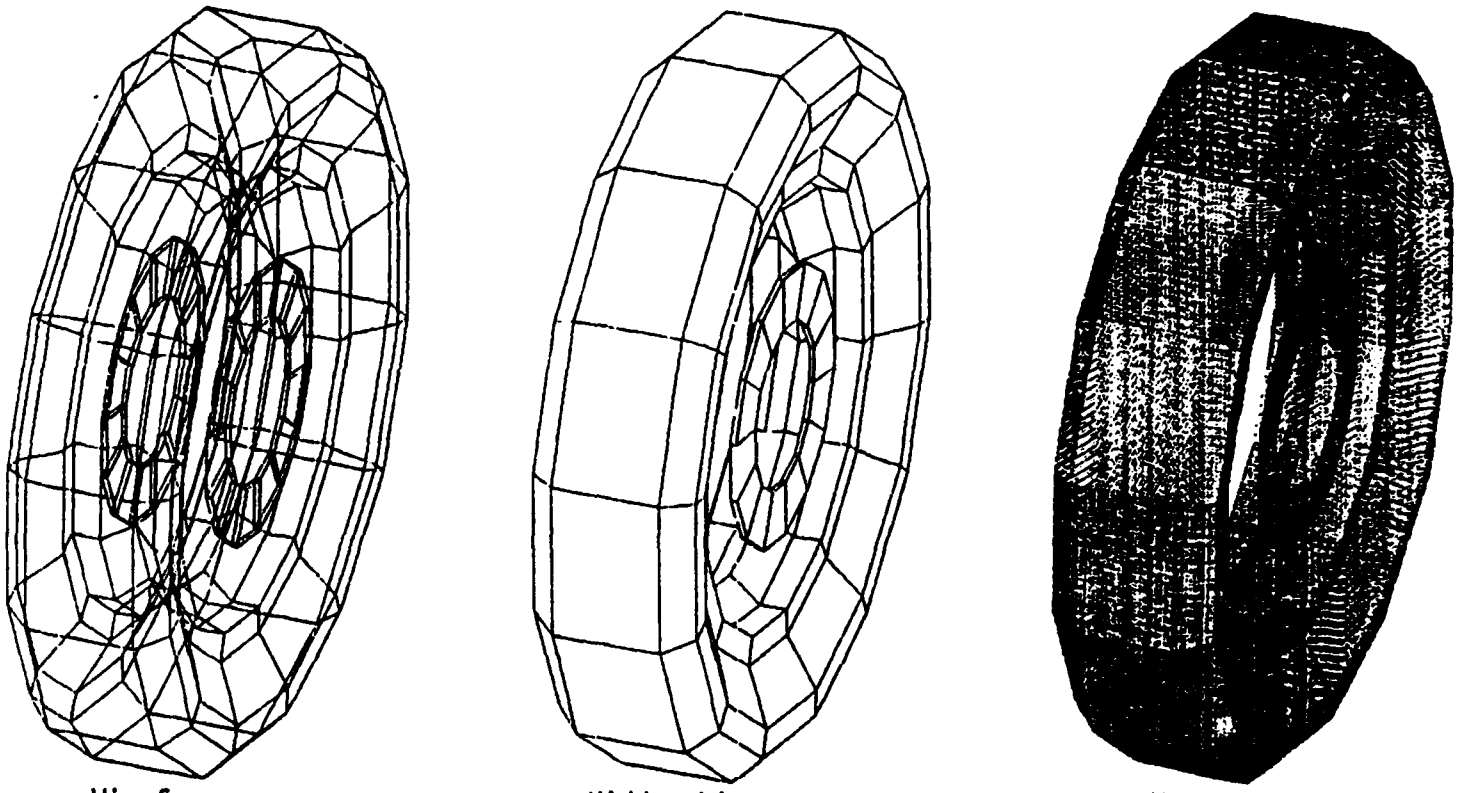


Figure 3.

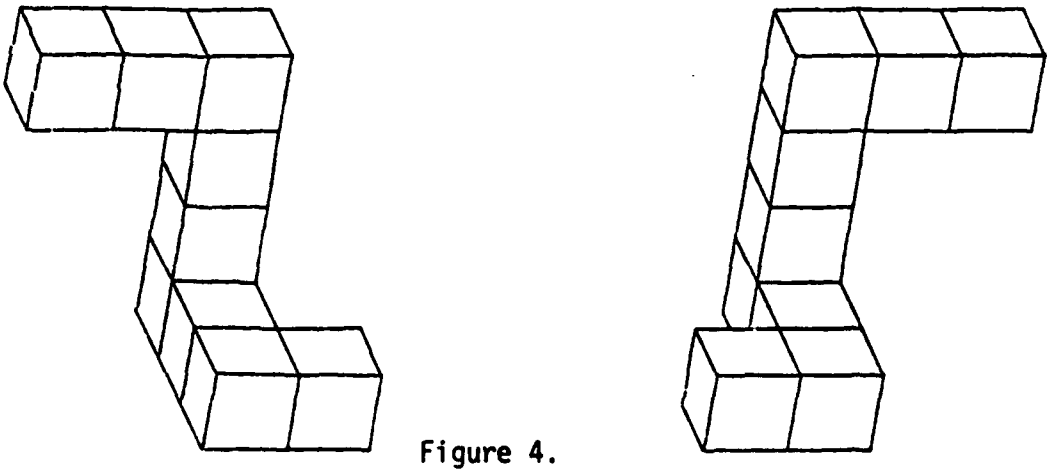


Figure 4.
Mirror Image (Isomorph)
Based on Shepard-Metzler Figures

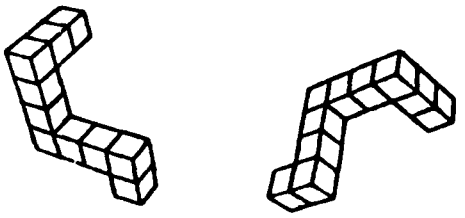


Figure 5. Picture Plane Rotation

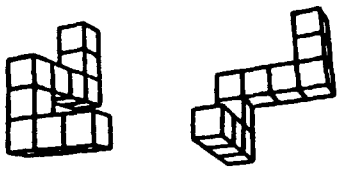


Figure 6. Depth Plane Rotation

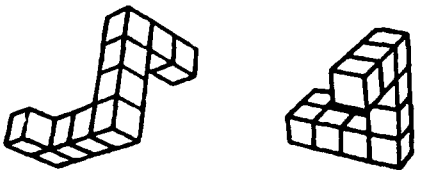


Figure 7. Different Pair Rotated in Depth

Reaction Time By Angle of Rotation
Seconds
10
8
6
4
2
0
0 Degrees
60 Degrees
120 Degrees
180 Degrees

Figure 8.

Reaction Time By Monoscopic Depth Cues
Seconds
0
2
4
6
8
10
Hidden Line Figure
Wire Frame Figure

Figure 9.

Reaction Time By Stereoscopic Depth Cues
Seconds
0
2
4
6
8
3D Figures
2D Figures

Figure 10.

Reaction Time By Type of Figure
Seconds
0
2
4
6
8
10
2D Wire Frame
3D Wire Frame
2D Hidden Line
3D Hidden Line
Type of Figure

Figure 11.

Reaction Time By Quartile
Seconds
0
2
4
6
8
10
Trials 1-40
Trials 41-80
Trials 81-120
Trials 121-160
Quartile

Figure 12.

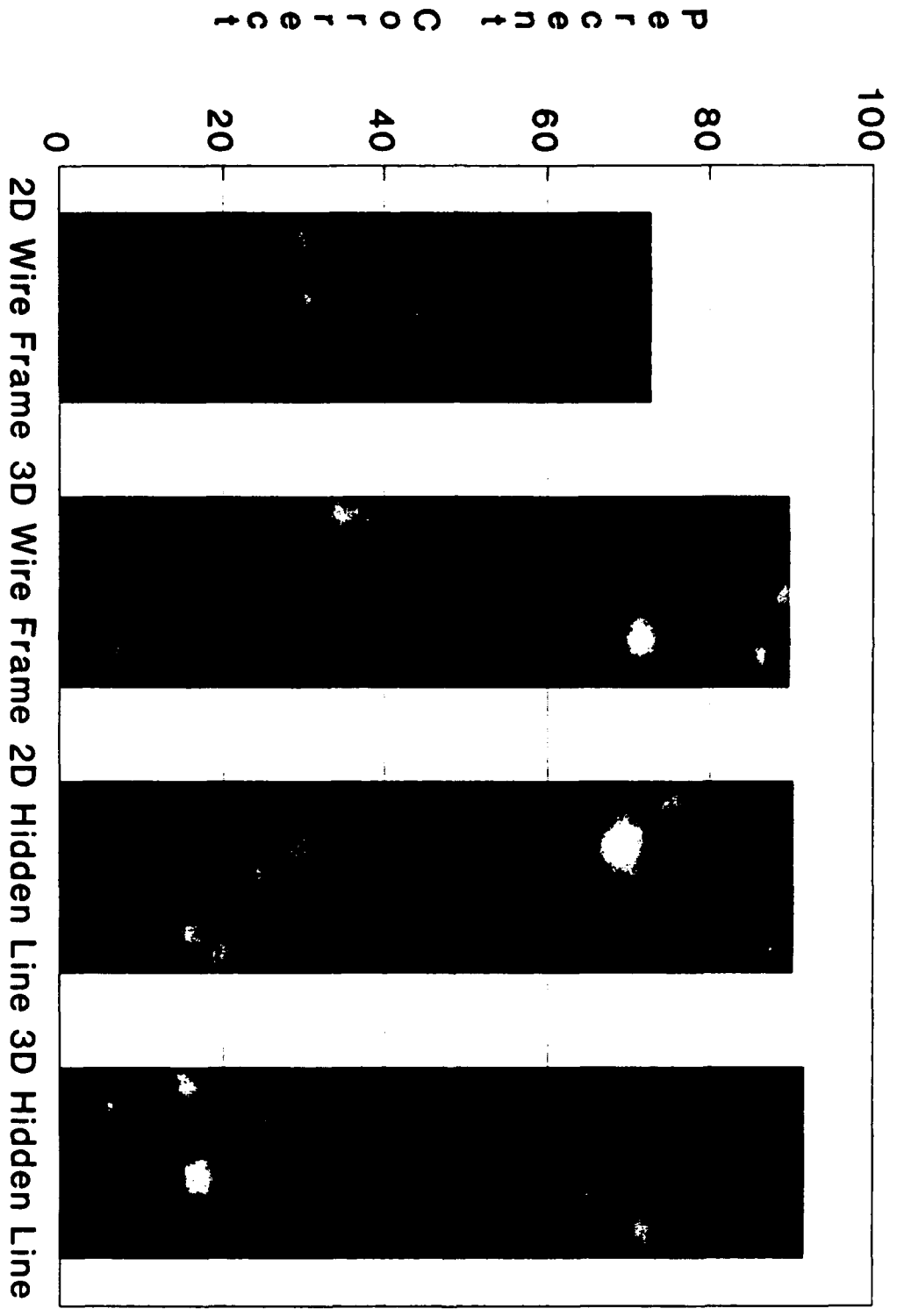

Accuracy By Type of Figure
Percent Correct
0
20
40
60
80
100
2D Wire Frame
3D Wire Frame
2D Hidden Line
3D Hidden Line

Figure 13.

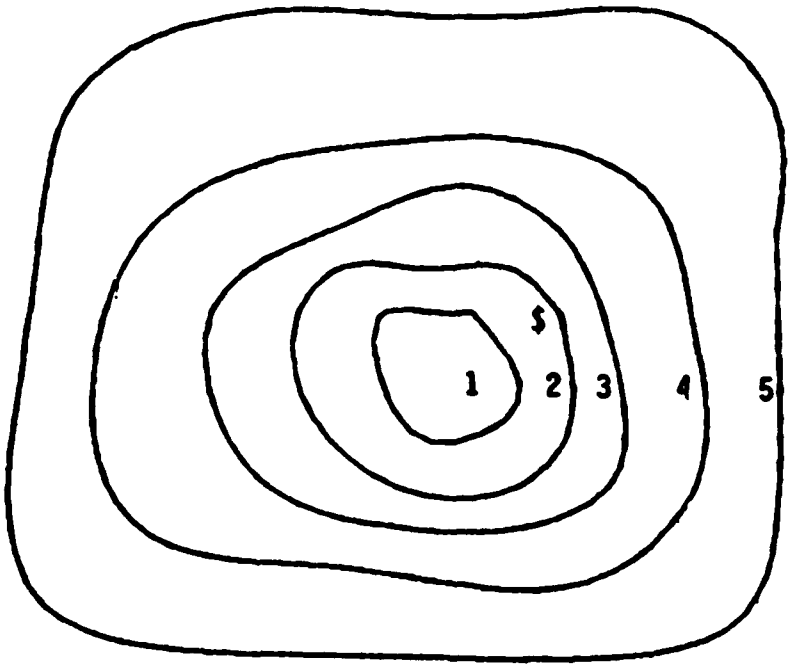


Figure 14a.

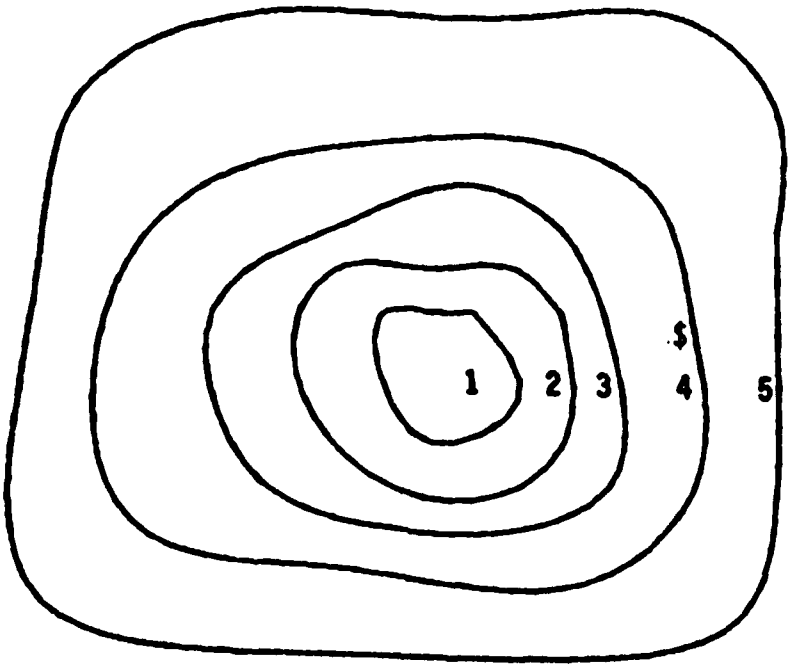


Figure 14b.

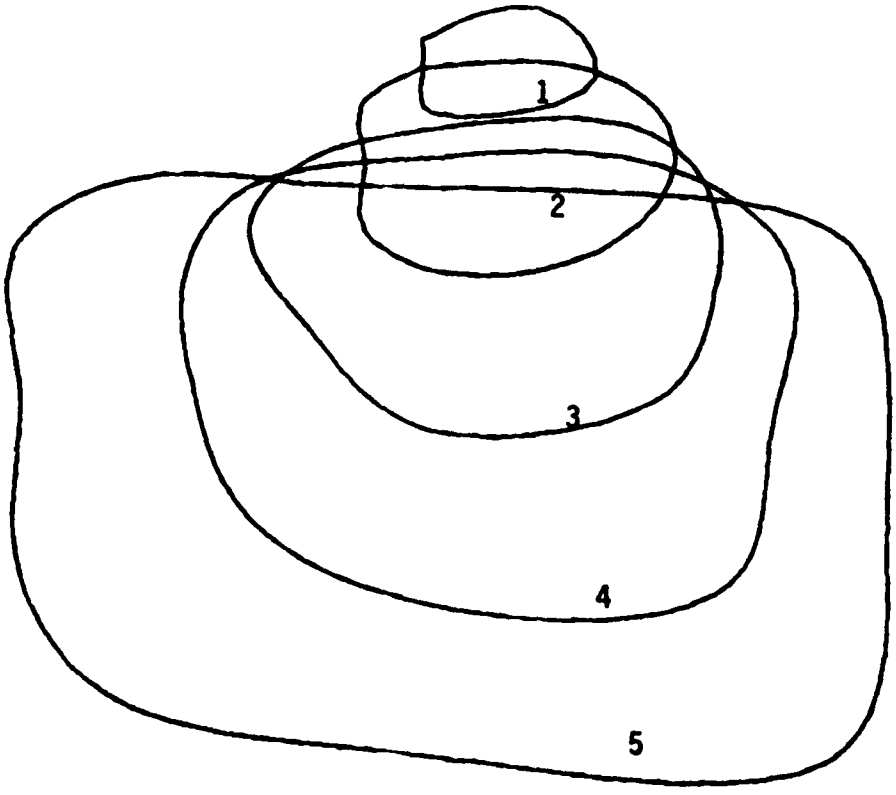


Figure 14c.

1989 USAF-UES RESEARCH INITIATION PROGRAM

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the

Universal Energy Systems, Inc.

FINAL REPORT

New Architectures for WISIWYSWIWSWYS

Prepared by:	Michael David Wolfe, Ph.D.
Academic Rank:	Assistant Professor
Department and University:	Department of Management West Virginia University
Research Location:	USAFHRL/LRL Wright-Patterson AFB Dayton, OH 45433
USAF Researcher:	CAPT Ray Hill
Date:	Monday, December 31, 1990
Contract No:	F49620-88-C-0053

ABSTRACT

The Development of a large, complex system involves designers, decision makers, and others, all with distinct points of view and frames of reference, and needing sophisticated software to support the combined group decision process. In this project, new architectures are examined for WISIWYSWIWSWYS, a system which goes beyond WYSIWIS to make the central, concrete concept of the system being designed available to all participants in their own paradigm and language.

This is accomplished with a new, "rising sun" architecture and a relational database management system. The framework so developed also supports the Japanese decision making style of nemawashi/ringi

ACKNOWLEDGEMENTS

I wish to thank the Air Force Systems Command and the Air Force Office of Scientific Research for sponsorship of this research, and the Logistics Systems Branch of the AFHRL for their cooperation.

Special thanks are due to CAPT Hill for his support in technical and administrative matters, and to Mr. Brian Smith for his assistance.

Finally, the interest of Dr. Duffy in my demonstrations and progress was appreciated.

I. INTRODUCTION

The development of a large, complex system involves a number of designers, engineers, and decision makers from different disciplines who must coordinate their efforts in order to achieve a system of high quality. Information technology (IT), in the form of Group Decision Support Systems (GDSS) can support such efforts. The Air Force Summer Research Program of 1989 culminated with a proposal for a prototype system, based on an extension of a concept called the House of Quality [Clausing and Hauser, 1988] and this research was continued as a Research Initiative Program through the calendar year 1990.

II. OBJECTIVES OF THE RESEARCH EFFORT:

The objective of this Research Initiation Program were to extend the research done as part of the summer research program. That research suggested that Hypertext is a useful paradigm with which to implement a GDSS to support the development of high quality systems Other researchers have also demonstrated the efficacy of Hypertext for data modelling [Conklin and Begeman,1988] and [Lai, Malone, and Yu,1988]; however, the current effort is intended to demonstrate its use over a much wider range of design problems. Specifically, the research was to examine possible architectures, and identify which would be most suitable for as platforms on which to implement a prototype. Cost, as well as performance, was a major consideration, as this system needs to be widely available, but simultaneously powerful and easy to use.

III. CONCEPTUAL AND PHYSICAL ARCHITECTURES FOR GROUP DECISION SUPPORT SYSTEMS

Three broad conceptual architectures have been suggested for GDSS [Dennis et al., 1988]. The process of bringing a group to consensus may be aided by one or more individuals acting in any of various roles, including that of coordinator, facilitator, chauffeur, etc.. The simplest conceptual IT architecture to support group decision making is, therefor, a single system devoted to helping this central figure. A

second conceptual architecture makes IT available to two or more units in the group, but in a serial fashion. The third architecture has members of the group using the IT concurrently. It is this third and most complex architecture that is the most interesting from a research standpoint.

An initial goal of the concurrent GDSS was the development of What You See is What I See, or WYSIWIS systems [Kraemer and King, 1988]. Such systems proved extremely effective for data modelling, but are basically inadequate for the systems engineering n[illegible] for more general systems development projects. In particular, area specialists are accustomed to seeing the system from their own disciplines' perspective, and have trouble relating to the Computer Assisted Dra[illegible] (CAD) from other disciplines. These specialists need to be looking at the same system simultaneously, but within the paradigm of their own speciality.

The idea for a system to support this level of inter-disciplinary communication was developed by a joint effort between CAPT R. Hill, USAF and myself. We called the concept "What I See Is What You See When You See What I Want You to See What You See (WISIWYSWIWSWYS)."

An additional consideration was to make the system as inexpensive as possible, so that it might be made readily available to a wide range of designers, engineers, and decision makers.

In proceeding to physical architectures, the first two conceptual architectures may be supported by a single computer system: CPU, I/O, and secondary storage, or a virtual single system on a time-sharing system. Such systems raise a number of research issues that will not be addressed here. Instead, we concentrate on inexpensive but effective solutions to the problems associated with the third conceptual architecture, that in which the group are concurrently supported in their decision making by the IT. The system developed here involves, at the software level, individual decision support systems (IDSS) for each user, as well as the group decision support system, a concept already proven effective [Bui and Jarke, 1986].

To support such a system, each user must have access to appropriate hardware for the IDSS. This includes a CPU, Input/Output

(I/O) device, and Storage. This can be done by providing each user with an actual computer, or by providing each user with a virtual machine using time sharing for the CPU, an individual input device, an individual display device, and a virtual mini-disk (to use the IBM terminology) or individual directory (to use the standard, non-IBM terminology). To support the group aspect, the system must also provide some sort of communication between and among the group members. There are three traditional topologies that might support such communication: the star topology, the ring topology, and the bus topology. The star topology is shown in Figure 1. In this topology, all communications is done by storing information on a shared device.

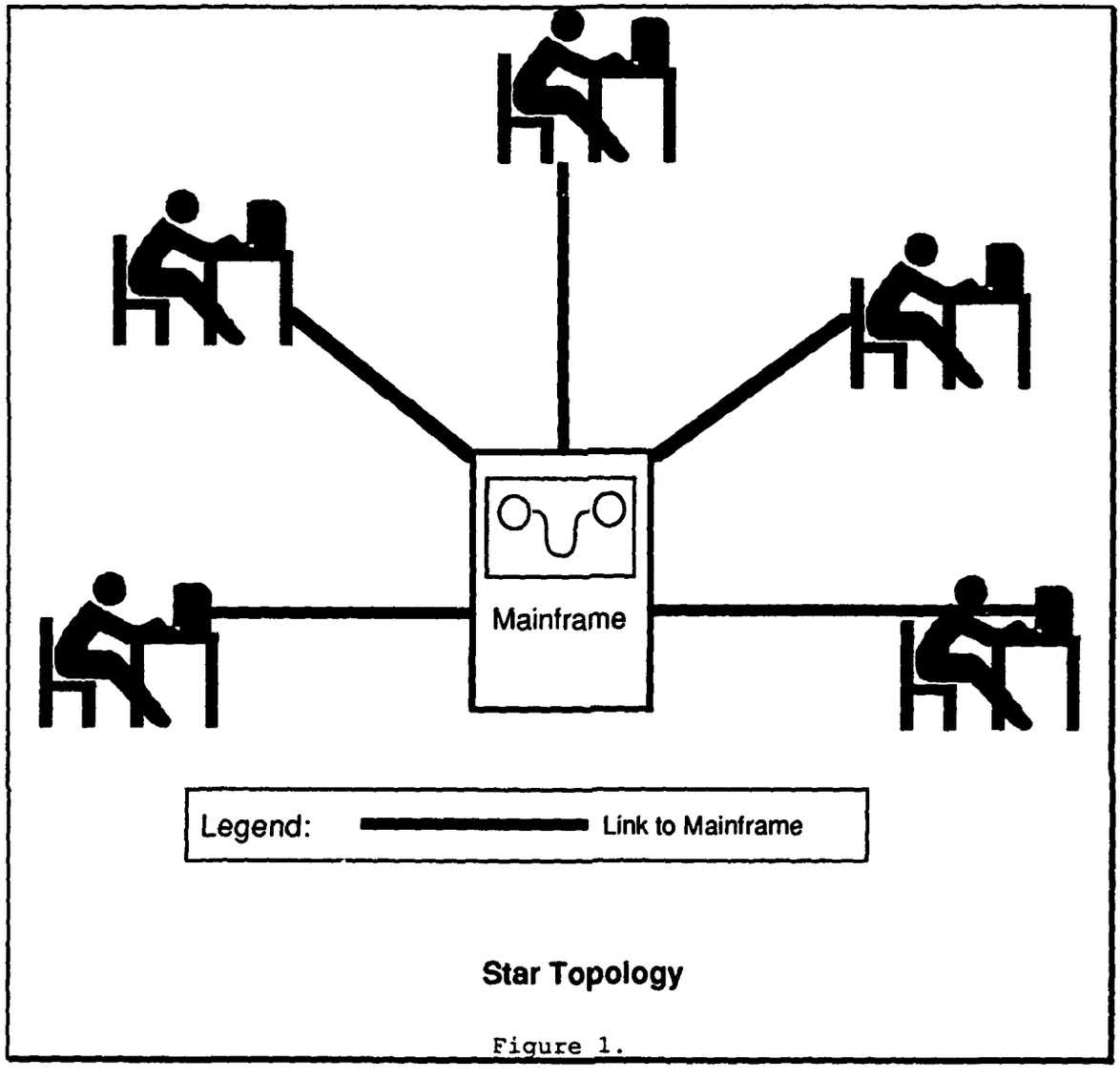


Figure 1.

The common storage used for the GDSS, even if physically just a partition of the storage unit which is also used for the IDSS, is conceptually much more complex. The IBM terminology for the possible access modes is given in Table 1. The table gives a fairly inclusive collection of access modes.

Theorem: Suppose each user has a correct IDSS or GDSS process. Then in R, W, or M access modes, the concurrent process is also correct.
Proof: Each individual process must terminate correctly between accesses, unless all users have access in read-only mode, and hence cannot corrupt each others' datasets.

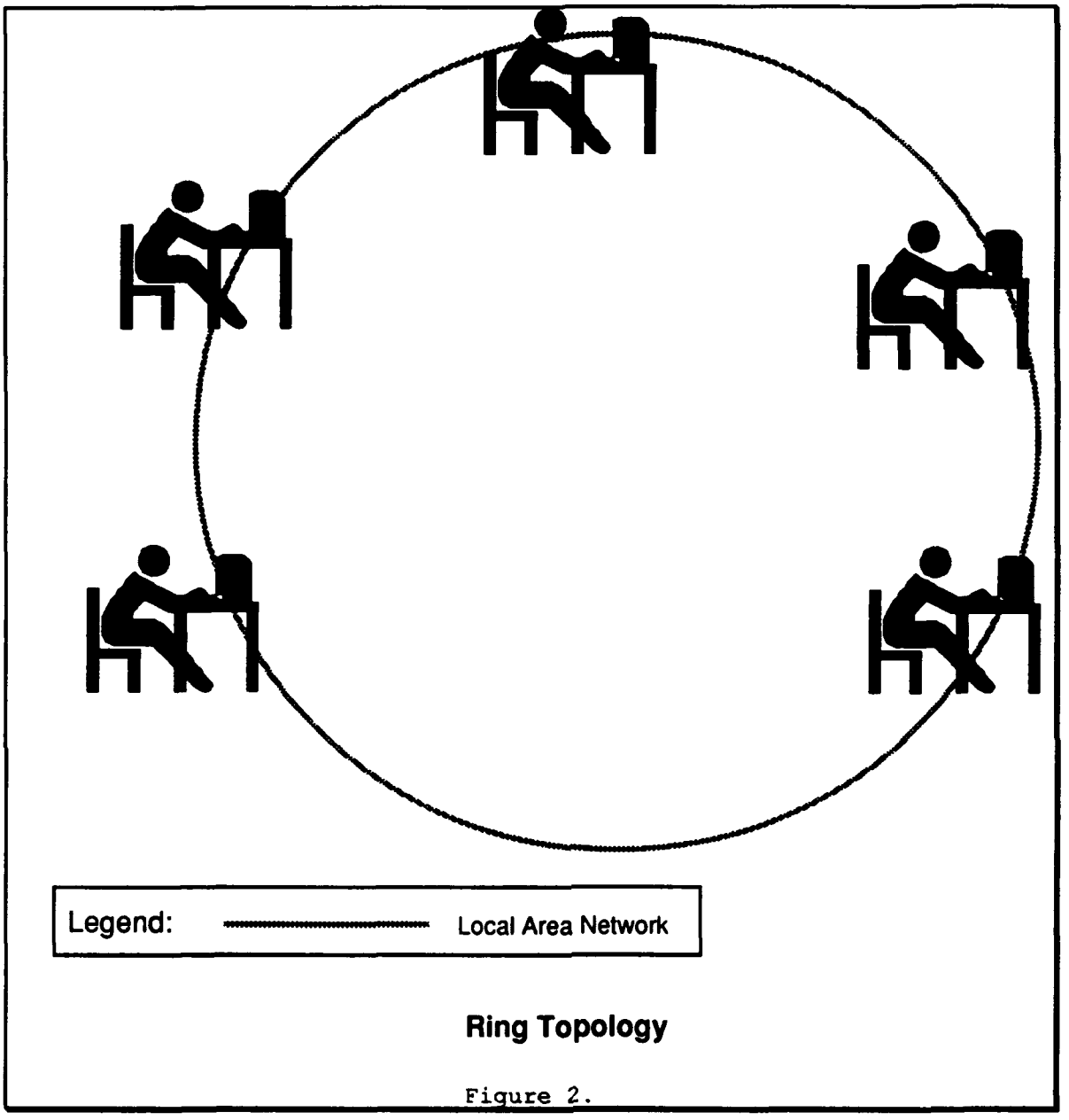
Caution: possible errors may occur in any of the other modes.

R	Read-only access. The link will not be established if any other user has a link to the shared data in write mode. Thus, this mode allows concurrent read-only access. If anyone has write-access to the shared data, sequential access only is allowed.
RR	Read-only access. The link will be established in all cases.
W	Write access. The link will not be established if any other user has a link to the shared data.
WR	Write access. If another user has a link to the shared data, read-only access will be established.
M	Multiple-write access. The link will not be established if any other user has a link to the shared data in write mode.
MR	Multiple-write access. If another user has a link to the shared data in write mode, read-only access will be established.
MW	Multiple-write access. This link will be established in all cases. This is the only mode that allows concurrent writing to the shared data.

Table 1. Possible Access Modes

The problem of access and ensuring that processes terminate correctly may be solved by using a relational database management system (RDBMS) which allows only "safe" access modes [Conklin and Begeman, 1988.] Connection to the shared device in a low cost architecture is a relatively slow process. While large amounts of data can be stored, it is undesirable to actually transfer such large amounts in a single operation. Ideally, individual operations would only involve small subsets of the data. Thus, each user may be expected to have a private storage device, and only rarely transfer limited amounts of information to or from the shared device.

The second method of establishing communication uses "messages" passed between the users. This may be done over a network using ring topology, as in Figure 2, or a bus topology, as in Figure 3.



Ring Topology

Figure 2.

From the user's perspective, little difference is observed between the bus and the ring, the real differences being at a rather low level where information is being physically transmitted. A ring network must be physically connected into a complete circle, while a bus network must never contain any closed loops, a consideration when setting up or modifying the physical layout.

Whether using the ring or bus topology, large datasets can be transmitted only at some expense. Since a goal of this project was to

study low cost architectures, messages sent in this way were limited to relatively short bursts of information. The ring or bus provides no storage directly, unless a storage device is connected to the ring or bus. This is possible, but tends to be a rather low performance system, as data is transferred over moderate cost networks at a rather low speed.

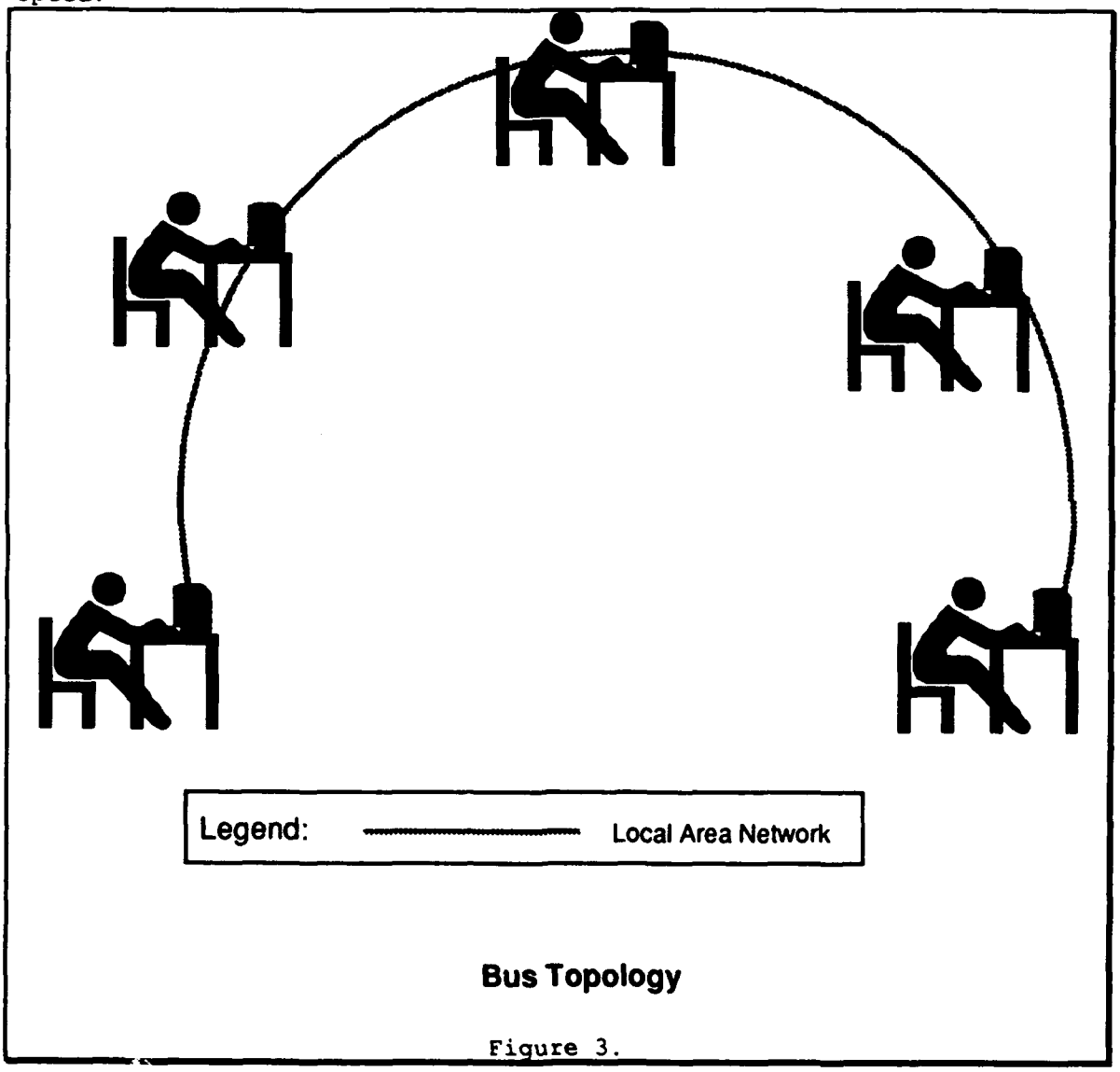


Figure 3.

None of the three topologies mentioned is really adequate for WISIWYSWIWSWYS, however, or even for low-cost WYSIWIS. In order to transmit the complex images used in systems design on a real-time basis, a very high-speed and extensive network would be required. This project

was developed without the need for such a network using a new topology, that of the rising sun, as in Figure 4.

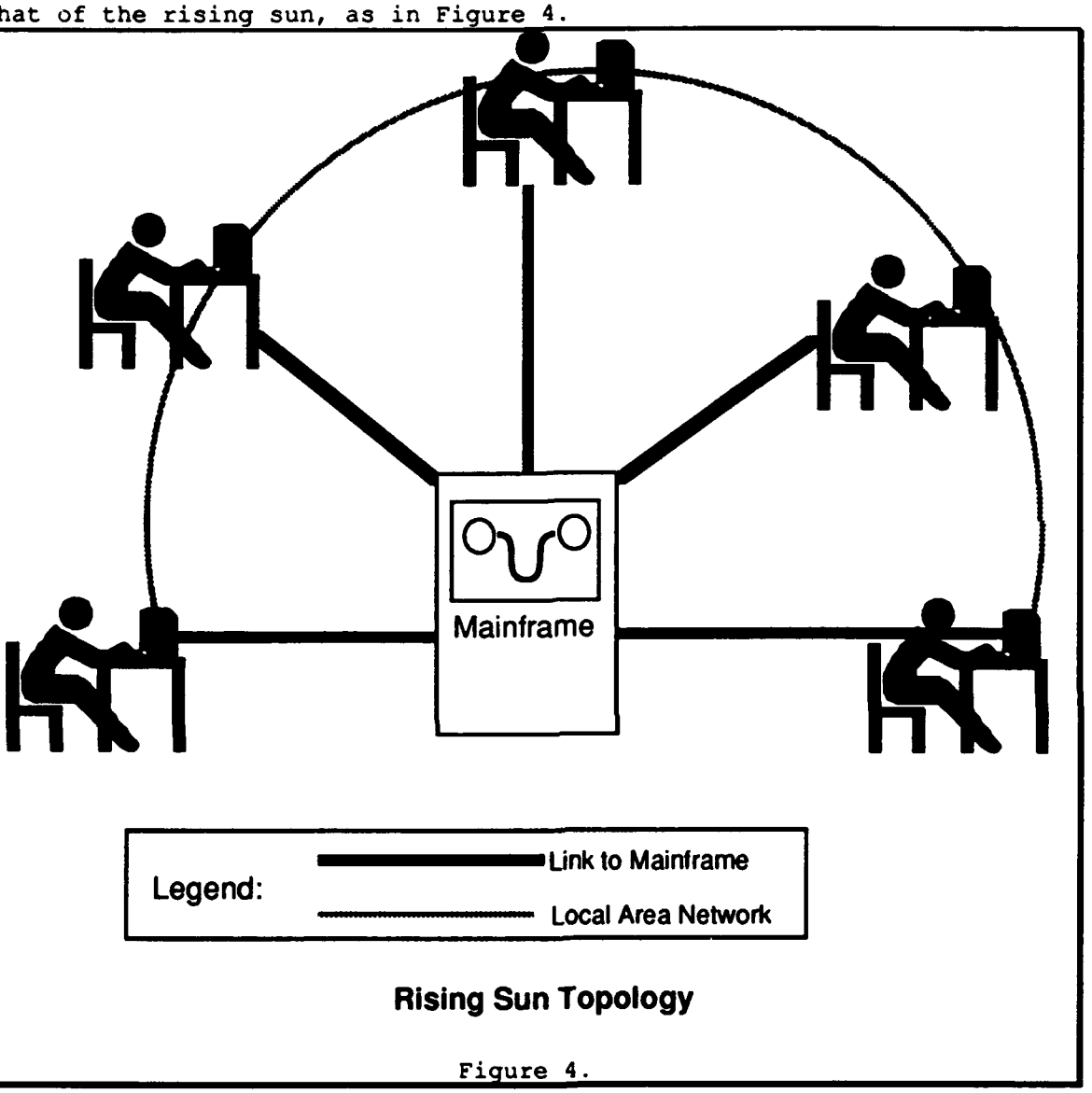


Figure 4.

This topology is predicated on using a bus for the network, in addition to the mainframe links in order to access the RDBMS. A Wagon Wheel topology would describe the conjunction of a ring topology with the star topology. Because we had access to a bus network, however, the actual architecture used in this project was the rising sun of Figure 4.

The actual physical architecture used consisted of complete computer systems for each virtual user, connected to each other via a bus network, and connected over a star network to a mainframe running a

RDBMS. (Just individual sub-directories on a shared disk are called virtual minidisks by IBM, groups who might sometimes required to share CPUs are called as virtual users.) The complete system included CPU, I/O, and private storage. This is the basic architecture to support a low-cost WYSIWIS. Initially, each system must be initialized to a state S_0. Users may only communicate with their IDSS and GDSS via messages to the system. These messages are interpreted by the IDSS or GDSS Language System (LS) which interprets the message and passes it to the Problem Processing System (PPS) which can draw on an extensive Knowledge System (KS) [Chang, Holsapple, and Whinston, 1988] Any virtual user u may originate a message $m \in M_u$; however, there may be some distinction between users (in particular, our system includes a facilitator, f, who has access to a special set of facilitator-only messages M_f)

The WYSIWIS subsystem consists of identical virtual finite state machines for each virtual user, each supporting an identical copy of the GDSS LS, PPS, and KS. The message is sent along the bus from the originator, and causes all the virtual machines to update, as in Figure 5.

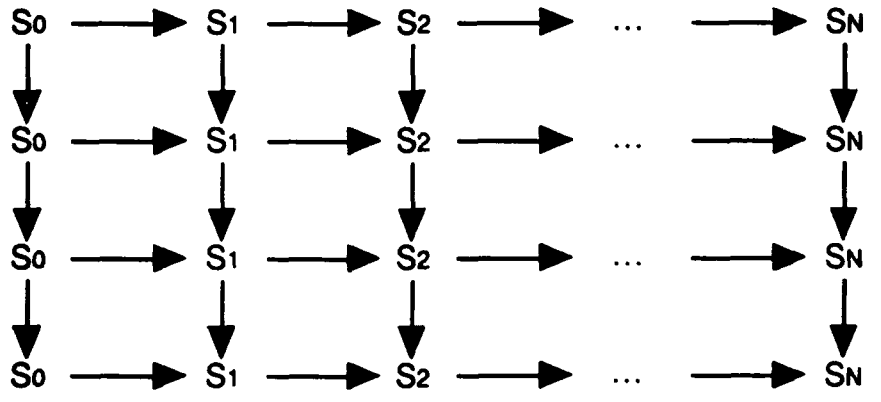


Figure 5. Conceptual Diagram for WYSIWIS

The figure is, unfortunately, somewhat misleading, in that messages are shown as originating from the top. In fact, messages could originate from any user in the network, and would be passed down the bus. The concurrent finite state machine for the group, $G \in \prod_u S_u$ where

S_u is the finite state machine associated with user u. WYSIWIS is just the restriction of G to the diagonal D of the product, S^N.

WYSIWIS is complicated by two possible failures: first, a virtual user might decide to make an individual excursion δ on a finite state machine that bypasses the message broadcast system, and second, the network may fail. The first situation is covered by the following:
Theorem: If an individual excursion loops back to the pre-excursion state before any message is transmitted over the network, then WYSIWIS is maintained.
Proof: The states have no memory, hence, the machine, looped back, is indistinguishable from the machine had no excursion taken place.

Failed or garbled messages, however, produce diversions ε from the diagonal which are unpredictable and not preventable by the virtual users. The net result is a state $D + \delta + \varepsilon$ which is not WYSIWIS, and which must be returned to state D by re-syncing all the virtual machines. This can be done by the facilitator at a re-syncing cost C_r. When WYSIWIS fails, there is a cost C_e from the error. The system must then minimize $\sum_{\text{all resyncs}} C_r + C_e$.

At the general, most conceptual level, implementation of WISIWYSWIWSWYS is only slightly more complicated that implementation of WYSIWIS. The messages must be transformed from the originator's DSS language to the recipients' DSS language, as shown in Figure 6. Alternatively, from the originator's language to a common language, then from the common language to the recipients' language. A number of problems can arise, however, when the specifics of implementation are confronted [Lee and Malone, 1990].

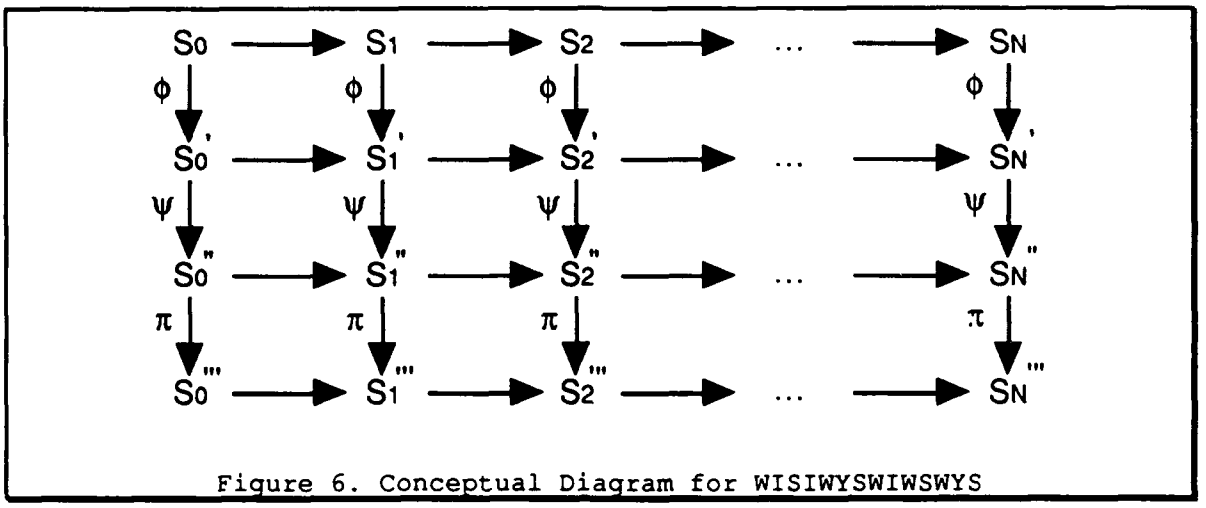


Figure 6. Conceptual Diagram for WISIWYSWIWSWYS

Where the mappings in Figure 5 were identities from $S_i \rightarrow S_i$, now the mappings are morphisms, i.e. structure preserving mappings, from $s_i \rightarrow S_i'$ where S_i' is another system supporting the conceptual needs of a different virtual user than S_i. Figure 6 shows the case where the message is transformed as it passes along the bus, although this is unlikely to be a common method. A more likely approach is the common language intermediate approach.

IV. CITY OF QUALITY TEMPLATE FOR GDSS

The framework selected for collecting the system design data is based on the model observed at Mitsubishi [Hauser and Clausing, 1988]. The system

1) Assists decision makers to establish functional requirements.

2) Provides a framework for translating functional requirements into formal design parameters;

3) Maintains a record of how trade-offs were resolved;

4) Reduces the probability that a functional requirement fails to be addressed in the final design.

The system that does this is shown in Figure 7, which shows the basic House. The left side of the house, labeled "Customer Attributes," contains the customer requirements and a column for their relative weights. The "attic" (just below the roof) labeled "Engineering Characteristics" is used to capture the design parameters, while the bottom of the house is used to list the target values, importance,

estimated difficulty and costs of these parameters. Also provided are spaces for comparison with existing designs.

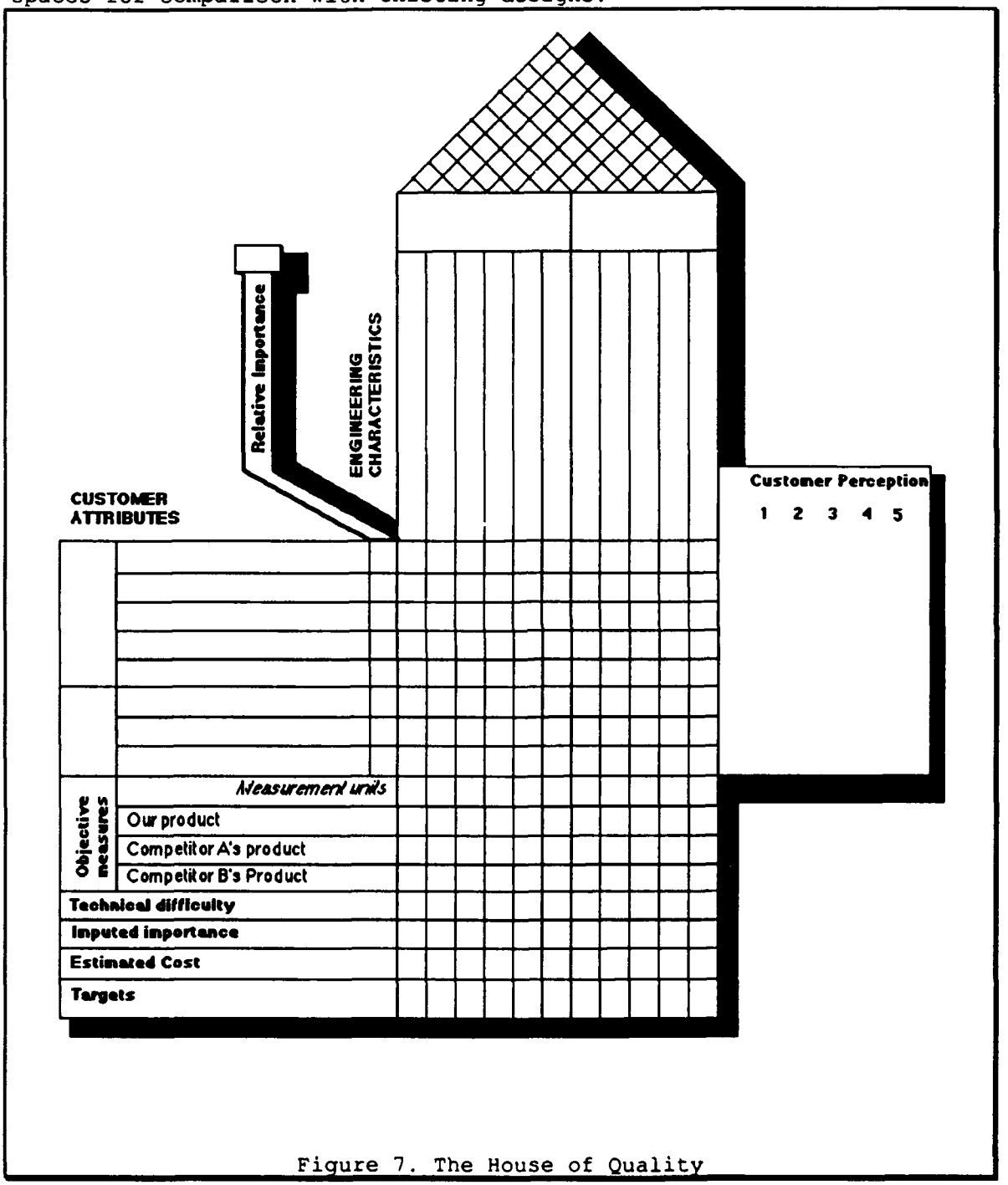


Figure 7. The House of Quality

The central portion of the house provides space to indicate which engineering specifications address (or degrade) a customer requirement. The roof indicates the synergies and conflicts among the engineering

specifications. Finally, a graph on the right hand side of the House indicates competitive position.

The initial state indicated above, S_0 is a blank house for every virtual user. Messages from decision makers consist of functional requirements and weights. Messages from marketing experts indicate the competitive positions. Finally, messages from system designers consist of parameters, as well as the target values, importance, estimated difficulty and costs of these parameters. In addition, system designers are responsible for indicating how each functional requirement is addressed by design parameters, and the trade-offs between each pair of design parameters via appropriate messages.

Where necessary, these messages may include simple models and formulae, or the messages may invoke complex models developed in FORTRAN or other conventional languages. These messages become part of the House, and hence may be reviewed by other members of the group. It is these supporting models, particularly CAD models, for which translation from one engineering discipline to another is needed.

The biggest problem with the House as depicted in the above Figure is that, in actual systems, the number of functional requirements and design parameters number at least in the hundreds, if not in the thousands or tens of thousands. Screen limitations make the display of such a house impossible. Thus, an important class of message is to move from one house to another. Different houses exist for different stages in the system life cycle, as shown in Figure 8. The overall ensemble of Houses is called the City of Quality (Coq).

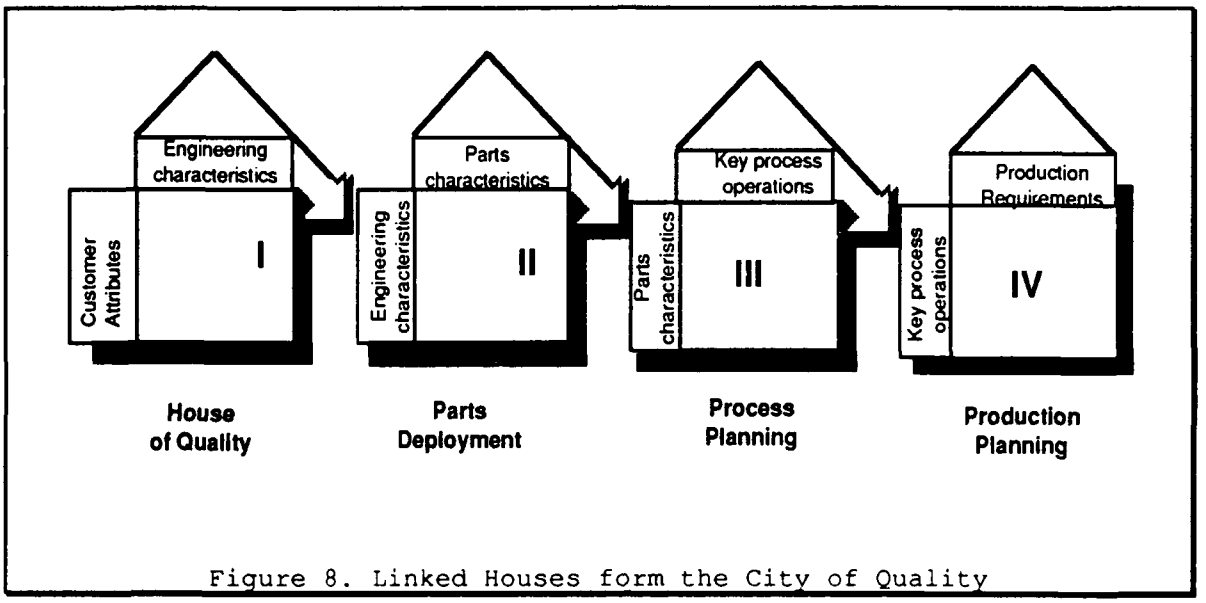


Figure 8. Linked Houses form the City of Quality

V. HYPERTEXT

Overview of Hypertext

The concept of hypertext is credited to Vannevar Bush, while the word is credited to Ted Nelson by [Smith and Weiss, 1988] and by [Vaughan, 1988.] Pure **hypertext** is a non-linear form of writing. Ordinary written documents may be visualized as a linear graph as in Figure 9. Hypertext, on the other hand, allows an arbitrary network as in Figure 10. Topologically speaking, the linear graph of ordinary text can be embedded in one dimensional space, while the network of hypertext requires at least two dimensions, and possibly three.

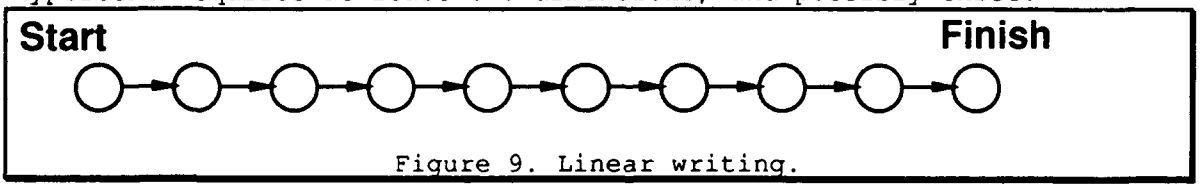


Figure 9. Linear writing.

Each node in the network represents some text, and is, perhaps, best visualized as a page in an encyclopedia or instruction manual. The reader, given a book or other paper document, is forced to move sequentially through the text, with random access tedious, and the time required for such access interrupting the reader's train of thought.

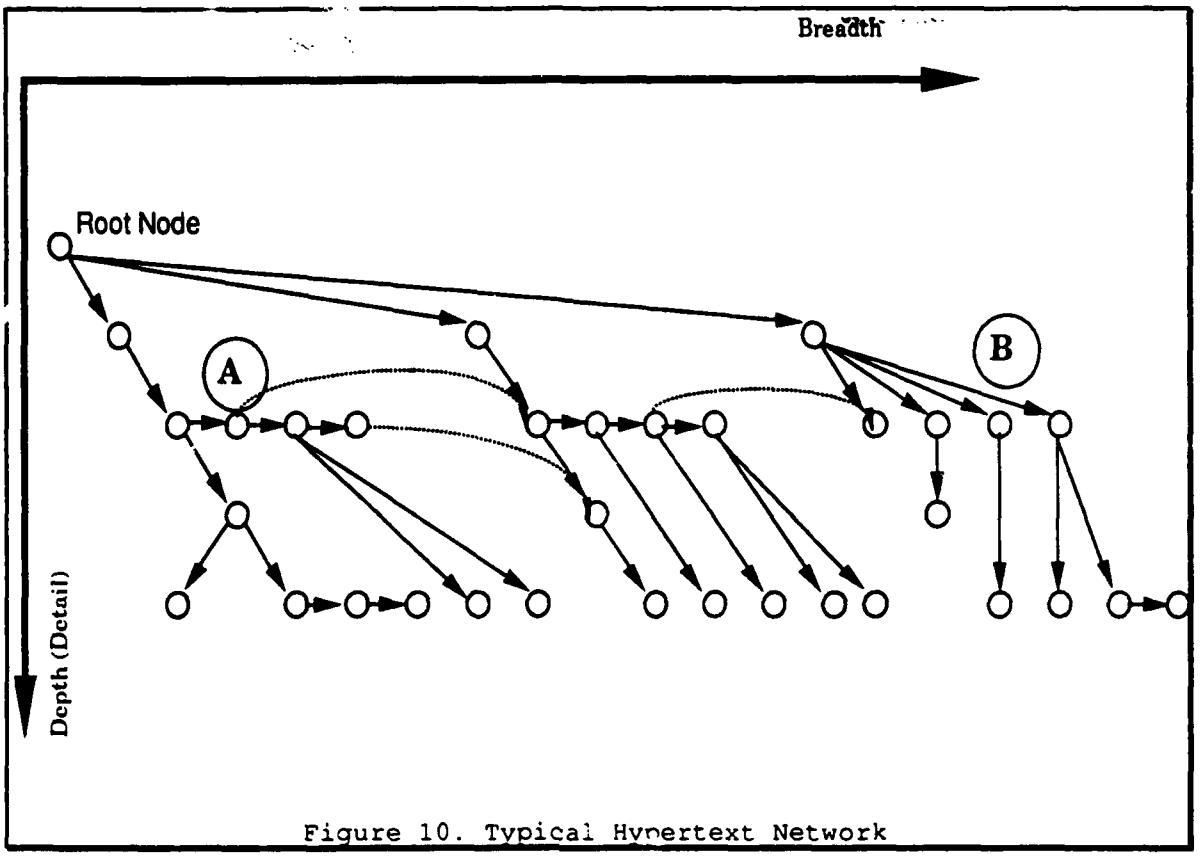


Figure 10. Typical Hypertext Network

The horizontal dimension represents breadth; the vertical dimension represents detail.

Hypermedia

While Bush probably envisioned pure hypertext, the storage of information on electronic media encourages much more creative methods for communicating that information. At a minimum, graphics are added if the hardware supports them; at the same conceptual level, digitized sound may be played at a node. The hypermedia concept is not, however, limited to static text and graphics. In its most general form, each node may be a complete program, which executes when the node is accessed. Many implementations of hypertext/hypermedia include a programming language. In this general form, hypermedia is an underlying structure connecting available electronic resources in whatever manner the author feels is most suitable for the presentation of information.

Specifically, in the prototype system developed in hypertext, nodes were Houses or data acquisition nodes, which prompted the users for the data needed to fill the house.

Synergy between Hypertext and Coq template The most important advantage of Hypertext is its ability to fit a conceptual model of information that has more than two dimensions. Let the conceptual dimensions of a general system be C_1, C_2, … , C_N. Ideally, a document would cover the space $\prod_{i=1}^{N} C_i$.

In particular system design includes dimensions of

1) functional requirements;

2) design parameters;

3) details indicating which specifications address which requirement, and to what extent;

4) trade-offs between the various design parameters;

5) the virtual users; and

6) time.

These six dimensions may be effectively captured in a hypertext document which allows the user to move freely along any of these dimensions.

VI. THE RELATIONAL DATABASE

The various data items mentioned above are best stored in a relational database on a mainframe. The tables needed include:

1) The users table, which maintains each virtual user's name, id, rank, influence, status, reputation, and persuasion difficulty;

2) The functional requirements master table, which maintains the requirement id, name, description, and originator.

3) The design parameter master table, which maintains the parameter id, name, description, and originator.

4) The functional requirements transaction table, with requirement id, user id, suggested weight, date, time, and rationale.

5) The design parameter transaction table, with transaction id, user id, suggested target value, importance, estimated difficulty and costs of these parameters, and rationale.

6) The existing designs parameter table, with design id, name, description, parameter id, target value, user id, date and time..

7) The existing designs ranking table, with design id, functional requirement id, rank, date, time, and rationale.

8) The parameter vs. requirement table, with parameter id, requirement id, the extent to which the parameter addresses the requirement, date, time and rationale.

9) The parameter trade-off table, with two parameter ids, code for synergy or antagonism, user id, date, time, and rationale.

The rationale field may include text, models, graphics, etc.

Together these 9 tables support the City of Quality and the data can be extracted for Japanese nemawashi/ringi decision making.

VII. RECOMMENDATIONS

The design of a new system architecture has been described; implementation has been started, but is not yet complete. The relational tables have been set up, but have never been populated. Communication between virtual users using the messaging system was demonstrated in principle, but not in practice. Procedures were developed to support the link between virtual uses and the RDBMS, prototypical amounts of data were transferred, but a real system was not designed using the prototype system as a design tool.

The next stage is to complete the implementation. This will require identification of a real system design project which will drive the development.

The most important direction for further research, then, is to actually field test the system: this will conclusively identify deficiencies, strengths and potentials.

VIII. REFERENCES

Bui, Tung; Jarke,Matthias. Communications Requirements for Group Decision Support Systems. Journal of Management Information Systems; 1986; 2(4); ISSN: 8–20.

Chang, Ai-Mei; Holsapple, Clyde W.; Whinston, Andrew B. Decision Support System Theory. West Lafayette, IN 47907: Management Information Research Center
Krannert Graduate School of Management
Purdue University; 1988.

Conklin, Jeff; Begeman, Michael L. gIBIS: A Hypertext Tool for Exploratory Policy Discussion. ACM Transactions on Office Information Systems; 1988; 6(4): 303–331.

Dennis, Alan R; George, Joey F; Jessup, Len M; Nunamaker, Jr., Jay F.; Vogel, Douglas R. Information Technology to Support Electronic Meetings. MIS Quarterly; 1988; 12(4): 591–619.

Hauser, John R.; Clausing, Don. The House of Quality. Harvard Business Review; 1988; 66(3): 63–73.

Kraemer, Kenneth L.; King, John Leslie. Computer-Based Systems for Cooperative Work and Group Decision Making. ACM Computing Surveys; 1988; 20(2).

Lai, Kum-Yew; Malone, Thomas W.; Yu, Ken-Chiang. Object Lens: A 'Spreadsheet' for Cooperative Work. ACM Transactions on Office Information Systems; 1988; 6(4): 332–353.

Lee, Jintae; Malone, Thomas W. Partially Shared Views: A Scheme for Communicating among Groups that Use Different Type Hierarchies. acm Transactions on Information Systems; 1990; 8(1): 1-26; ISSN: 1046-8188.

Smith, John B.; Weiss, Stephen F. An Overview of Hypertext. Communications of the ACM; 1988; 31(7): 816–819.

Vaughan, Tay. Using Hypercard®. Carmel, Indiana: Que® Corporation; 1988; ISBN: 0-88022-340-5.

1989-1990 AFOSR RESEARCH INITIATION PROGRAM

Sponsored by the
AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by the
Universal Energy Systems, Inc.

FINAL REPORT

VARIABLE RESOLUTION IMAGERY FOR FLIGHT SIMULATORS

Submitted by:	Yehoshua Y. Zeevi, Ph.D.
Academic Rank:	Professor
Department:	CAIP Center
University:	Rutgers University Piscataway, NJ, 08855
In cooperation with USAF Researchers:	Elizabeth L. Martin, Ph.D. George A. Geri, Ph.D.
USAF Location:	Williams AFB, AZ 85240
Program Period:	Dec. 1, 1989 - June 1, 1990

VARIABLE RESOLUTION IMAGERY FOR FLIGHT SIMULATORS

by
Yehoshua Y. Zeevi

ABSTRACT

In visual flight simulators, a vast amount of information has to be processed and displayed to the trainee, in real time. The dual requirement of high resolution, and wide field of view, renders the real time, high fidelity, image generation task practically impossible. A combination of two approaches are proposed, and tested in this study:

1. Accommodating the displayed information to the specific requirements of the human visual foveation process.
2. Organizing the data in a variable resolution, multidimensional structure. Using this structure, only a small portion of the data base has to be extracted and processed , to generate the images, in real time.

The proposed underlying theoretical approach has been elaborated, and a combination of the proposed solutions is implemented. The results indicate that the proposed technique for nonuniform representation of visual information is promising insofar as its application to flight simulators is concerned. A problem of hidden regions has been identified. This problem affects the implementation considerations. Another issue which deserves further investigation is related to the redundancy of the proposed visual data structure.

I. INTRODUCTION

The work conducted at the CAIP Center under the sponsoring of UES, during the period of 12/1/89 - 6/1/90, was concerned with development of techniques for efficiently distributing visual information (i.e with variable resolution) over a wide field-of-view in flight simulators. The applications of such techniques is in particular important in simulations of low altitude flights, where there is a major difference between the contents of the low and high portions of the projected image. The portion corresponding to the area projected from the further areas (at the high parts of the image) covers a big area, with low resolution, whereas the portion projected from lower areas entails higher resolution (or more detailed information,, and covers a small area.

Two strategies have been devised, in order to reduce the computation effort required of the simulator's CGI system:

1. Accommodating the displayed information to the foveation process of the human visual system.
2. Organizing the data base in such a way, that given any viewing conditions, only a small portion of the data base will have to be processed.

Both strategies use nonuniform filtering and sampling techniques which were presented by Clark et al in [1], and extended by Zeevi et al in [2] and [3]. With the aid of those techniques, a pyramidal structure is constructed, to represent the scenery data base. This structure is different from other structures which have been proposed so far (Burt and Crowley in [5] and [6], Mallat in [7], Porat and Zeevi in [4], Zeevi et al in [3]). It's unique feature is that each level of the pyramid contains several approximations of the data base.

These approximations are represented by filtered versions of the data base. Each approximation is filtered according to a certain "local bandwidth" pattern, which causes it to be "locally band limited". This local band limitedness enables to sample

those signals nonuniformly, and ensures complete reconstruction of the approximation from those samples. The number of sampling points of the approximations is much smaller than the number of sampling points of the original data base. In practice these approximations correspond to (grey level and depth) images of the scene, taken from different vantage points.

The data in the pyramid is extracted and processed to produce the image that will be displayed to the trainee. Several approximations are chosen, according to the simulated viewing conditions. A union of those approximation is found, and the required output image generated. Choosing the candidate approximations for the union is accomplished so that the "local bandwidth" content of the union will be greater than that of the output image, everywhere. In this stage, the nonuniform filtering and sampling technique is used to produce a foveated image of the scene. Using the radial, non uniform filter described by Zeevi et al in [3], a foveated image is produced. In this image, the resolution degrades as a function of the distance from the fovea. The local resolution requirements of the output image are directly translated to' "local bandwidth" requirements. The "local bandwidth" requirements influence the choice of the candidates for the union, and the amount of processing required to generate the output image.

The advantage of the above-discussed solution is that it "tailors" the representation of the data base, and defines the requirements of the output image, according to "local bandwidth" considerations. In this manner the pyramidal data base is constructed from approximations which will be sufficiently close, in the "local bandwidth" sense to the output images, given any viewing conditions. The number of sampling points, in each such set of approximations, is of the same order as the output image. Thus, the amount of processing is of the order of the effective (due to the foveation) number of points of the output image.

The disadvantage of this solution, which calls for further research, is, that the pyramid itself is redundant in the sense of it's information contents, compared to the original data base.

II. GENERAL THEORETICAL BACKGROUND

IMAGE FORMATION AS A SAMPLING PROCESS

Viewing a surface with a T.V camera can be described as a sampling process. Each ray, emerging from a pixel of the camera face plate, samples the surface at a specific point. The optics and the electronics of the camera perform some filtering operation on the optical data, which (hopefully) prevents aliasing.

This filtration and sampling process can be described in two alternative ways: (see figure 1)

1. Filtration and sampling at the camera's faceplate: Each point at the viewed surface is projected on the faceplate's surface. According to the (uniform) spacing of the pixels, the projected texture is (uniformly) filtered and sampled.
2. Filtration and sampling at the viewed surface: the spread function of each pixel is projected through the optics on the surface. The result is convolved with the surface texture, and sampled. In this case a nonuniform sampling has been performed.

NONUNIFORM SAMPLING

In section , it has been argued, that the image formation process can be described either as a uniform, or as a non uniform sampling process.

In analogy to this description, Clark et al [1] present their nonuniform sampling theorem:

A non-bandlimited signal $f(t)$ can be completely recovered from its unevenly spaced samples $f(t_n)$, if:

1. There exists a monotonic function $\gamma(t)$ (called the distortion function), for which $f(t) = g[\gamma(t)]$.
2. $g(t)$ is bandlimited.

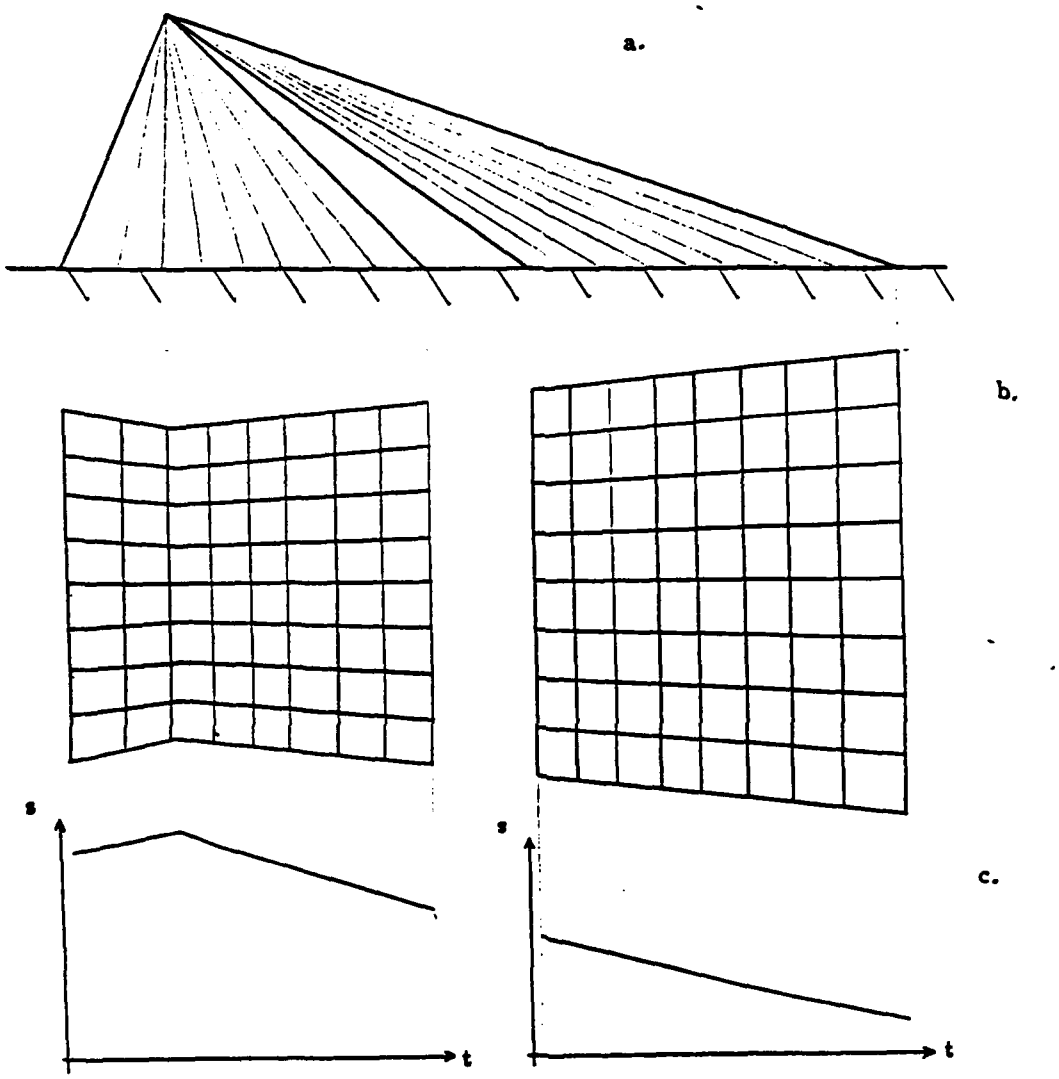


Figure 1: The sampling density patterns, induced by a T.V camera, viewing a flat surface. a. The camera setup. b. The induced sampling pattern. c. The induced sampling density pattern.

Sampling $g(t)$ at $t = n$ is equivalent to sampling $f(t)$ at $t_n = \gamma^{-1}(n)$.

$\gamma(t)$ is a function that distorts the scale of the independent variable t. In the image formation process, this function relates coordinates of points in the viewed surface to those of the corresponding points on the camera's image acquisition plane (e.g array of photodiodes or CCD). $\gamma(t)$ is determined by the viewing conditions and the surface structure.

The function $f(t)$ is related to the surface texture, and $g(t)$ to the image on the acquisition plane.

FUNCTION SPACES

It has been shown by Zeevi and Shlomot [2] that in case the conditions stated in section are satisfied, $f(t)$ belongs to a function space $B_{\gamma(t)}$, which is a reproducing kernel space. $f(t)$ can be represented as the sum of its projections on the spaces defined by the basis functions. This representation is unique.

In case a function $f(t)$ does not belong to a function space $B_{\gamma(t)}$, It can be projected to this space by finding it's projection on each of the basis functions of the space. The projection is accomplished by filtering and sampling f(t) according to a pattern defined by $\gamma(t)$. The outcome of this operation - $f'(t)$ will be an approximation of $f(t)$. In a distorted metrics, induced by the distortion function $\gamma(t)$, $f'(t)$ is the closest approximation to $f(t)$ among all the functions which belong to $B_{\gamma(t)}$.

LOCAL BANDWIDTH

Given a distortion function $\gamma(t)$, and defining:

$$W(t) = \frac{d\gamma(t)}{dt},$$

it can be shown that a function $f(t)$, which belongs to $B_{\gamma(t)}$ can be expressed by:

$$f(t) = \frac{1}{2\pi} \int_{-\pi W(t)}^{\pi W(t)} F(t, \omega) e^{j\omega t} d\omega \qquad (1)$$

Where $W(t)$ is called the local bandwidth of $f(t)$.

In this context, projecting a function to the function space $B_{\gamma(t)}$ can be described as locally band limiting the function. $B_{\gamma(t)}$ can be referred to as a space of locally bandlimited functions.

PROJECTION OF A FUNCTION FROM ONE FUNCTION SPACE TO ANOTHER

It can be shown that under some conditions, an approximation of a function can be obtained from another approximation by filtering it. This is equivalent to projecting the function from one function space (denoted here as $B_{\gamma_1(t)}$) to another ($B_{\gamma_2(t)}$). The above operation is feasible when the target space $B_{\gamma_2(t)}$ is fully "contained" by the source space $B_{\gamma_1(t)}$.The meaning of "contained" is that for every t the local bandwidth $W(t)$ of $B_{\gamma_2(t)}$ is lower or equal that of $B_{\gamma_1(t)}$ space.

The meaning of the above claim is, that an approximation of a signal obtained in this manner, will be identical to one obtained by projecting the original signal directly on the target space. In case the source approximation is small in size, compared to the original signal, there is an advantage in processing the source approximation instead of the original signal.

PERFORMING THE UNION OF SOME APPROXIMATIONS

The motivation for performing a union of several approximations is to obtain a new approximation, with higher local bandwidth contents than each source approximation. The local bandwidth of an approximation of a signal, belonging to a function space $B_{\gamma_i(t)}$ was defined by:

$$W_i(t) = \frac{d\gamma_i(t)}{dt}$$

By performing the union of two or more source approximations of the same signal (which belong to the function spaces $B_{\gamma_i(t)}$), we actually find the signal's projection to the target function space $B_{\gamma_u(t)}$ which is equal or bigger than each of $B_{\gamma_i(t)}$. This space contains all the basis functions of $B_{\gamma_i(t)}$.

The local bandwidth function $W_u(t)$ in the function space $B_{\gamma_u(t)}$ will acquire, for each t, the highest value of the individual values $W_i(t)$. The related distortion function $\gamma_u(t)$ can be derived by integrating $W_u(t)$ along t. (See figure 2).

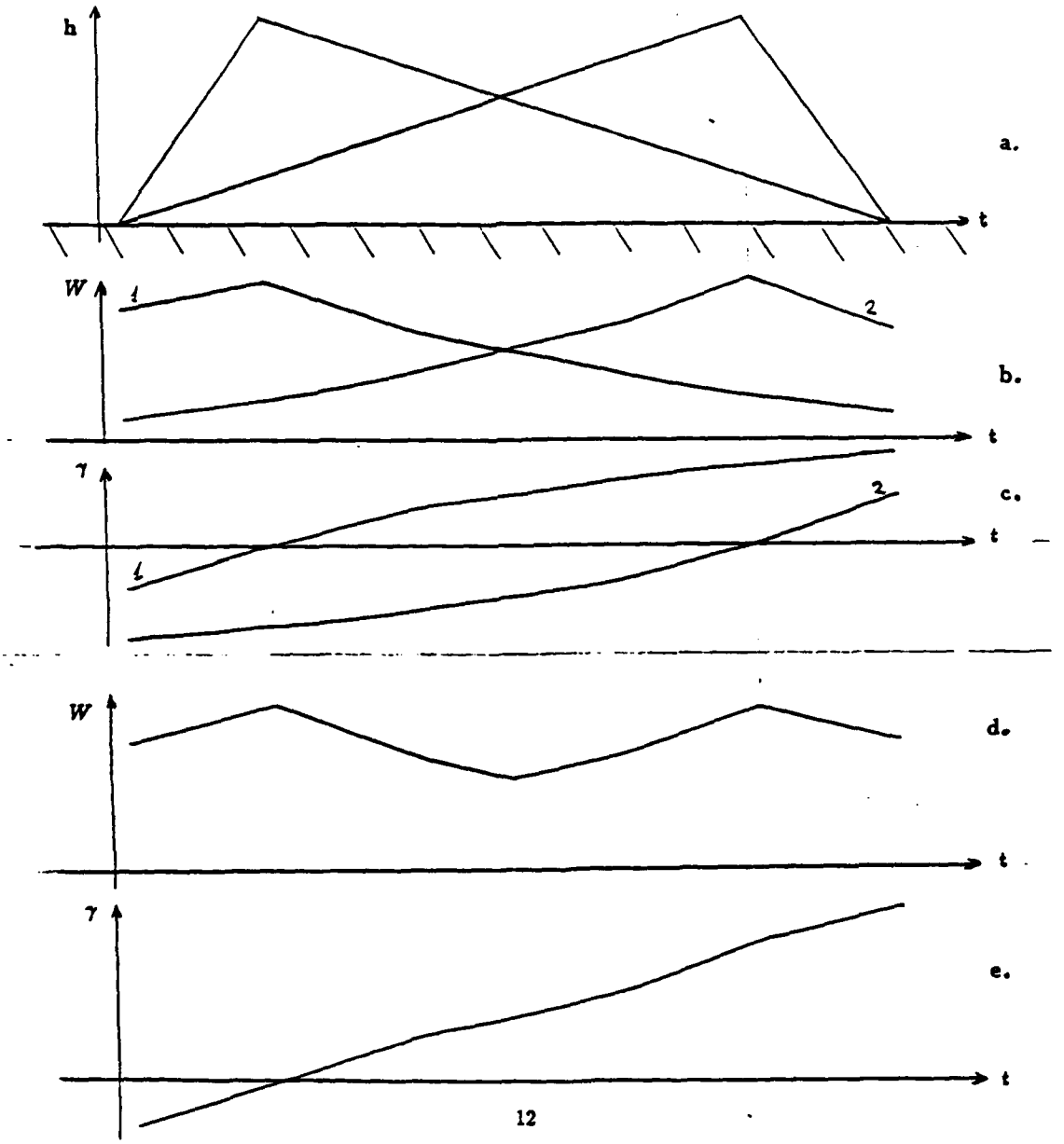


Figure 2: Union of two approximations of the texture of a planar surface. a. The cameras setup, whose outputs are the images of the approximations. b. The local bandwidth of the individual approximations. c. The distortion function, related to the individual approximations. d. The combined local bandwidth of the union. e. The combined distortion function of the union.

THE DISTORTION FUNCTION, INDUCED BY THE IMAGE FORMATION PROCESS.

It was stated in section that when trying to simulate the image formation process, the distortion function is induced by the viewing conditions and the shape of the viewed surface. [See figure 3 as an example of a distortion function, induced be the viewing conditions, and the surface shape. Note that the distortion function is the integral of the local bandwidth. Also note that the distortion functions and the local bandwidth, are plotted along the Cartesian axis x, and not the length of arc T. Two cases will be considered: a planar surface, and a non-planar surface.

VIEWING A PLANAR SURFACE

A simple case is when the viewed surface is planar. The location of a point on the surface, and that of the corresponding point on the camera's sensors plane are specified by two coordinates (x,y). In this case, a relatively simple function, relates the coordinates of the corresponding points on the two planes. This function is the two-variables distortion function $\gamma(x, y)$. The function is monotonic, and can be derived analytically, when the viewing conditions are given.

THE NON-PLANAR SURFACE CASE

In the more general case, when the surface is non planar, some complications arise:

1. The distortion function $\gamma(x, y)$ might turn out to be non monotonic.

The distortion function non-monotonicity problem is related to the existence of surface regions which are hidden from the viewing point. By forming a union of the

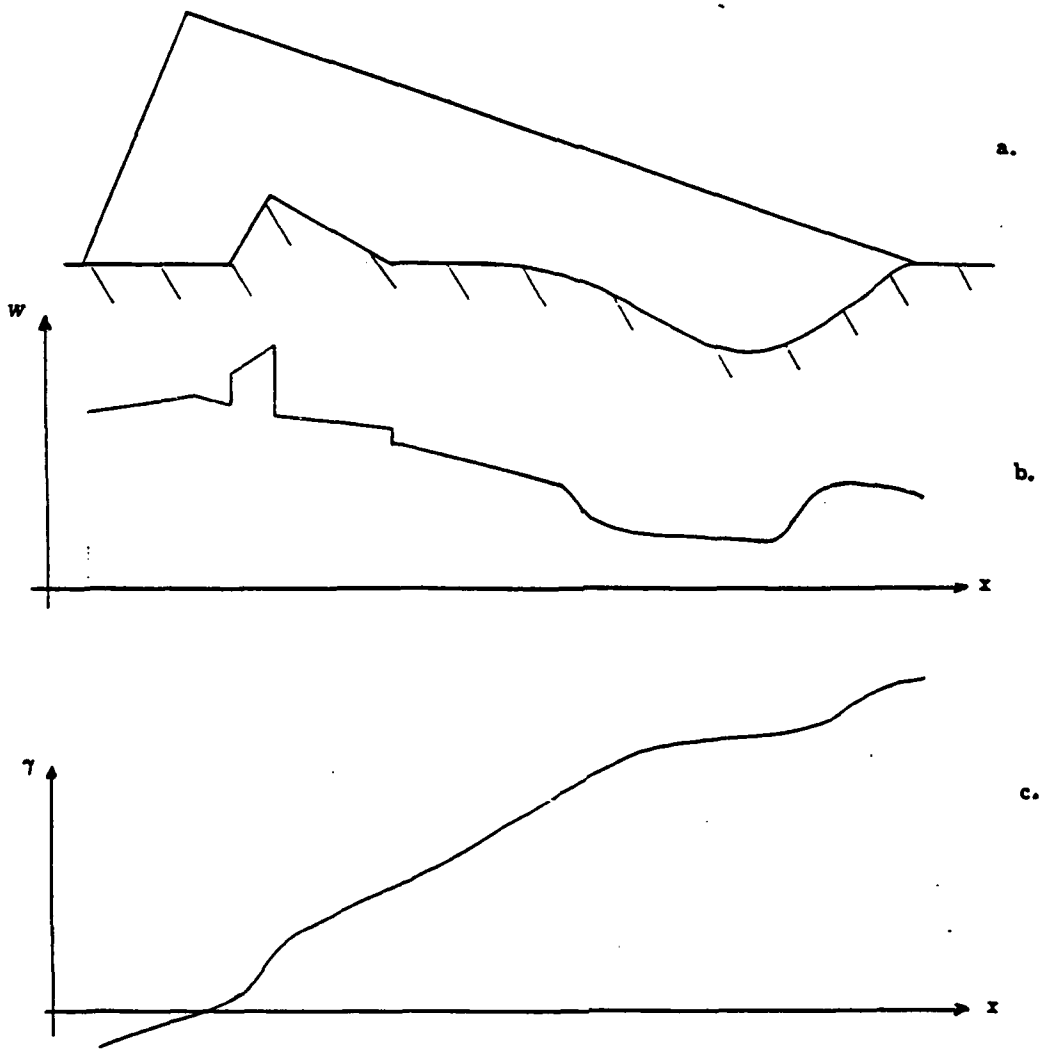


Figure 3: The local bandwidth and the distortion function, induced by viewing a non planar surface. a. The camera - surface setup. b. The local bandwidth. c. The distortion function.

different images, which cover all those hidden regions, a monotonic distortion function is created.

2. When the surface geometry is defined in a digitized form (as a table), it would be advantageous to represent it in a multi-resolution form. In this case, approximations of the geometry, and hence, approximations of the distortion functions will be used. This might have some influence on the quality of the output images.

III. THE IMPLEMENTATION OF THE PROPOSED METHOD

In this chapter, a specific implementation, which is appropriate for applications such as flight simulators is discussed.

In general, the method is implemented in two stages:

In the first stage, the multi-resolution structure is constructed. This is an off line stage.

In the second stage, which is the on-line or the real time stage, parts of this structure are processed and the required image is generated.

THE OFF-LINE STAGE

In this stage, a pyramid structure is constructed. This structure is different from the classic pyramidal structure. Here, each level of the pyramid contains more than one reduced resolution representation of the data. Each such representation can be regarded as a wide angle image of the surface, viewed from a specific point. Those view points are distributed over horizontal planes which correspond to the pyramid levels. As the levels of the pyramid grow higher, they contain less number of images. Within the pyramid, each such image can be addressed by specifying the location of it's related view point.

In addition to the texture information, the geometry of the surface is also kept in

a pyramidal structure. Essentially, from each viewpoint, corresponding to an image above. a depth or range image is taken. In this depth image, each "pixel" contains the value of the distance the corresponding ray travels between the camera and its intersection point with the surface.

It should be noted that each such texture or range image corresponds to a non-uniformly filtered and sampled approximation of the original data. Referring to the data as "images " seems more convenient because of the fact that in this case, the sampling is uniform, and the location of the sampling points is implicitly defined for the definition of images and approximations).

The images which constitute the pyramid, are created by projecting all the points in a source data base on the image planes, using a perspective view transformation. Then the projected data is interpolated and filtered, to reduce its band width, and to enable sampling. The sampling is done according to the image pixels resolution. Note: The source data base, referred to in the above paragraph, can be the original data base, or a lower pyramid level. In general, it is advantageous to use the immediate lower level of the pyramid as the source, because the reduced amount of required computations. However, because the pyramid levels are constructed from several images, a process similar to that performed in the on-line stage have to be carried out (see next section).

THE ON-LINE STAGE

In this stage an image of the surface is generated, according to arbitrary given viewing conditions. This is done in several steps:

1. According to the required viewing conditions, several images, from the pyrami-

dal structure are chosen. Those images have the property that their unions - related local bandwidth is higher than that of the required image. Chosing those candidate images is performed by finding the virtual location of the required image, within the pyramidal structure. The immediate surrounding images will be chosen from the pyramid.

2. A union of the chosen images is performed. This union is actually obtained by an orthogonalization process, and is similar to the Grahm-Shmidt process:

a. The first image is included in the union, as is.

b. Another image is picked up. All images already in the union are projected to the function space which the currently picked up image belongs to. The result of each projection is subtracted from picked up image.

c. The remainder, after all the subtractions have been accomplished, is included in the union (as an image).

d. Steps b and c are repeated until all the chosen images have been picked up.

The projection operation of step (b) defined above is a combination of transformation, filtering and sampling operations. This is equivalent to displaying the information contained in each image, which already belongs to the union, at the viewing conditions of the newly picked up image.

It should be noted, that except for the first picked up image, all the other images that are included in the union, are reduced-information-content versions of the original corresponding images. The union process yields an ensemble of images, all related to the union approximation. This approximation contains all the information in the individual candidate images.

3. The above union is once again projected (transformed, filtered and sampled), according to the required viewing conditions, to form the output image.

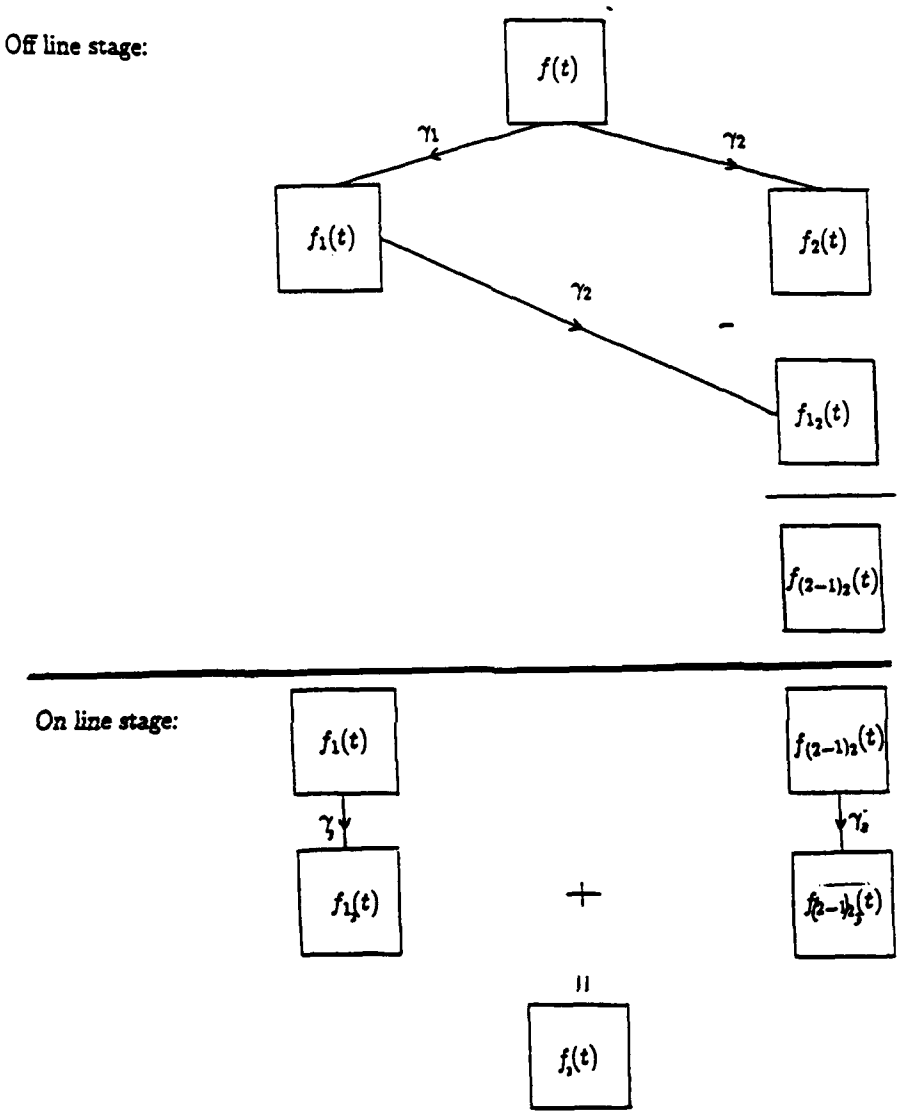


Figure 4: The implementation steps in the tests reported in this work

IV. CURRENT RESULTS AND CONCLUSIONS.

The proposed method have been implemented, at this stage, on planar, textured surfaces.

Figure 4 illustrates the implementation steps, performed in this experiment:

f_1 and f_2 are images representing the projections of the texture of the surface on B_{γ_1} and B_{γ_2} function spaces.

f_{12} is the projection of f_1 on B_{γ_2}.

$f_{(2-1)_2}$ is the difference of f_2 and f_{12}.

f_1 together with $f_{(2-1)_2}$ constitute the union of f_1 and f_2.

The required output image f_3 is generated by projecting f_1 and $f_{(2-1)_2}$ on B_{γ_3}, and adding the results.

The method was tested by applying it on a (planar) aerial photograph, as the source (figures 6 - 15).

In the result images, g_1, g_2 and g_3 correspond to images taken according to the viewing conditions demonstrated in figure 5.

Combined views of the steps in the off line and the on line stages are displayed, as well as individual, full size images of each one of the implementation steps.

Another, slightly different implementation:

A planar aerial photograph is filtered, using a radial foveation filter discussed in [3]. g_1 and g_2 are filtered versions of the original image, using the same filter, with different foveation centers. g_3, whose in-focus area is the combination of those of g_1 and g_2, is generated from g_1 and g_2, using the method proposed in this work. (figure 16).

DISCUSSION OF THE RESULTS.

In general, it can be said that the computational results, presented in figures 6 to 16, agree with the theoretical considerations, presented in the previous chapter.

The following Phenomenon, observed in the results, should be pointed out:

When examining the results carefully, some disturbance, in the shape of horizontal lines can be detected in the output image (g_3) (figure 15), of the second example (the aerial photograph). This disturbance can be traced back to g_{1_2} (figure 11). This disturbance is probably due to the fact that the local bandwidth requirements were not satisfied, when g_1 was projected to B_{γ_2}, and this disturbance is actually aliasing. Although the test results are not quantitative, they do provide further support for the theoretical framework and its implementation accomplished by this study. In particular, one can conclude from these results that the proposed approach of constructing unions of several signal approximations, and extracting new approximations from these unions, appears to work.

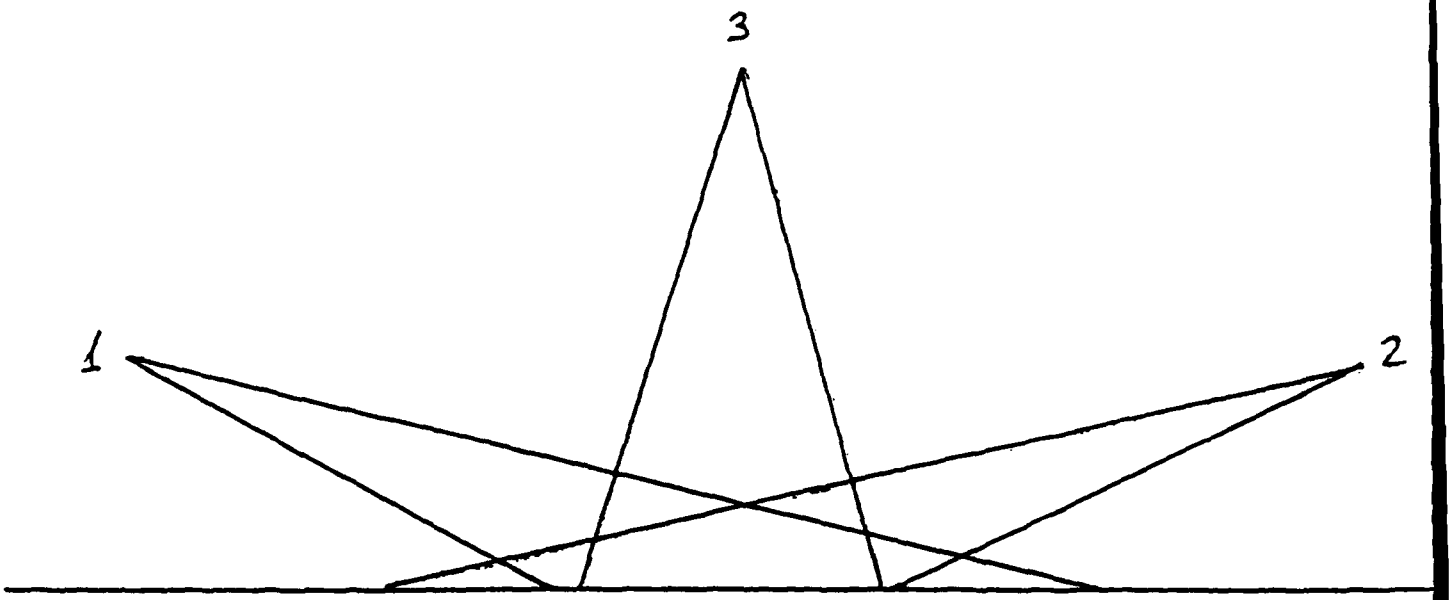


Figure 5: The simulated viewing conditions, related to the images g_1, g_2, g_3, in the following test.

V. ACKNOWLEDGEMENTS

I wish to thank the Air Force Office of Scientific Research, Universal Energy Systems and CAIP Center, for sponsoring this research. Universal Energy Systems provided excellent support and guidance.

The outstanding research facilities, instrumental in the application of the new concepts described in this report, were generally provided by CAIP Center. I would specifically like to thank Eyal Shlomot and Jacob Sharf for their excellent research, spanning both the theoretical and applied aspects of this project. I also wish to thank Dr. Elizabeth Martin and Dr. George Geri with whom I interacted extensively. This interdisciplinary interaction was a delightful and productive experience which enhanced my understanding of some subtle issues related to image representation, and flight simulators.

Last but not least, I wish to thank Ms. Barbara Daniels for typing this report.

References

[1] J. J. Clark, M. R. Palmer, P. D. Lawrence, "a Transformation Method for the Reconstruction of Functions from Nonuniform Spaced Samples", IEEE Trans. on ASSP, Vol ASSP-33, No. 4, pp. 1151-1165, October 1985.

[2] Y. Y. Zeevi, E. Shlomot, "Nonuniform Sampling And Representation Of Images Which Are Not Band limited", Technical Report CAIP-TR-118, CAIP-Center for Computer Aids for Industrial Productivity, Rutgers University, New Jersey, March 1990.

[3] Y. Y. Zeevi, N. Peterfreund, E. Shlomot "Pyramidal Image Representation in Nonuniform Systems", SPIE Vol 1001, Visual Communications and Image Processing 1988, pp. 563-569.

[4] M. Porat, Y. Y. Zeevi, "The Generalized Gabor Scheme of Image Representation In Biological and Machine Vision", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol PAMI-10, No. 4, pp. 452-468, July 1988.

[5] P. J. Burt, "Fast filter transforms for image processing" Comp. Graphics, Image Processing, Vol 16, pp. 20-51, 1981.

[6] J. Crowley, "A Multi Resolution Representation for Shape", IEEE Comp. Vision and Pattern Recognition, pp. 326, 1983.

[7] S. G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol PAMI-11, No. 4, pp. 674-693, July 1989.

APPENDIX

THE STUDY RESULTS

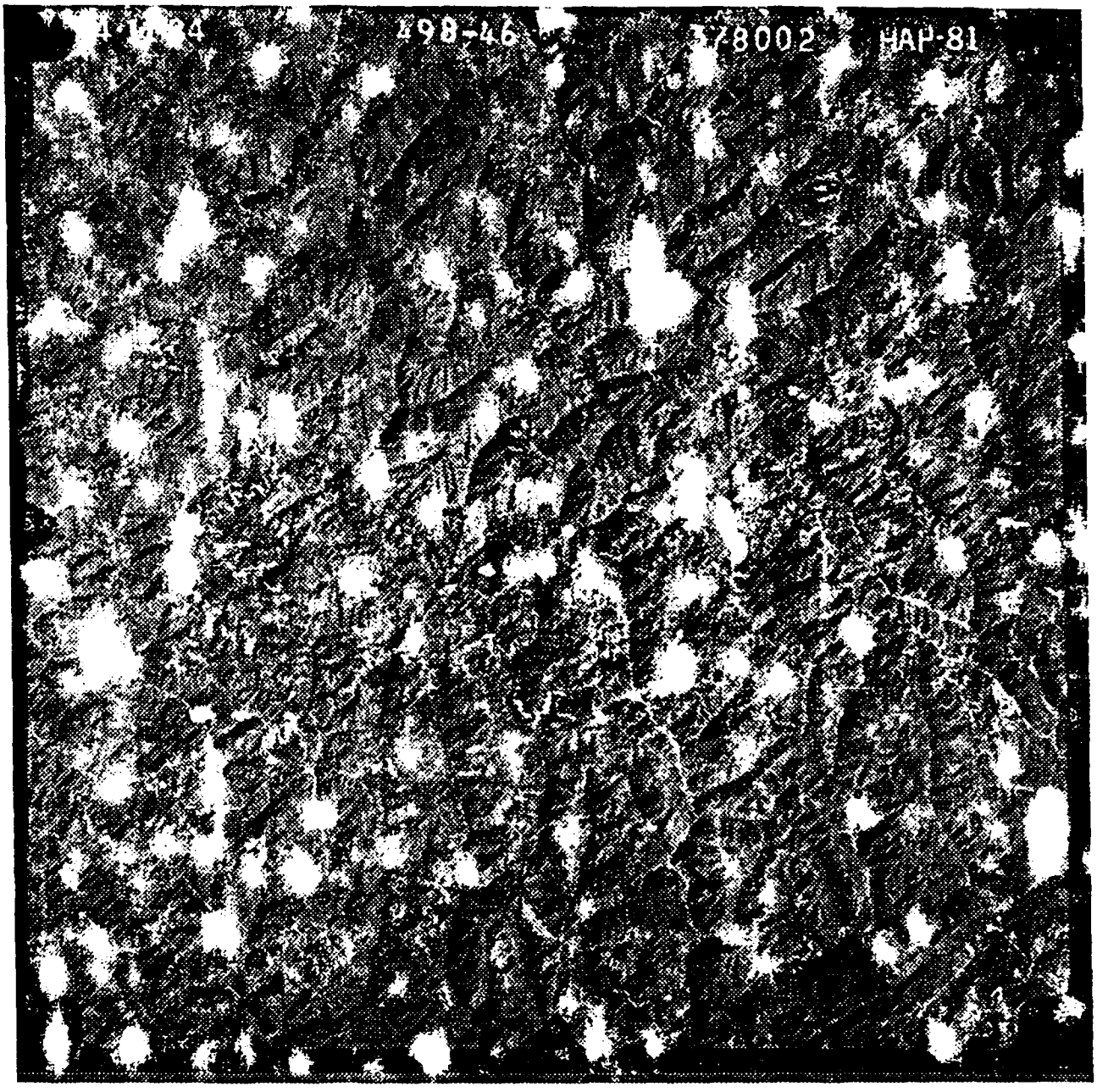


Figure 6: The original texture map (planar surface)

Figure 7 The off line steps. a g_{11}, b g_{12}, c g_{13}, d g_{21}, [illegible]

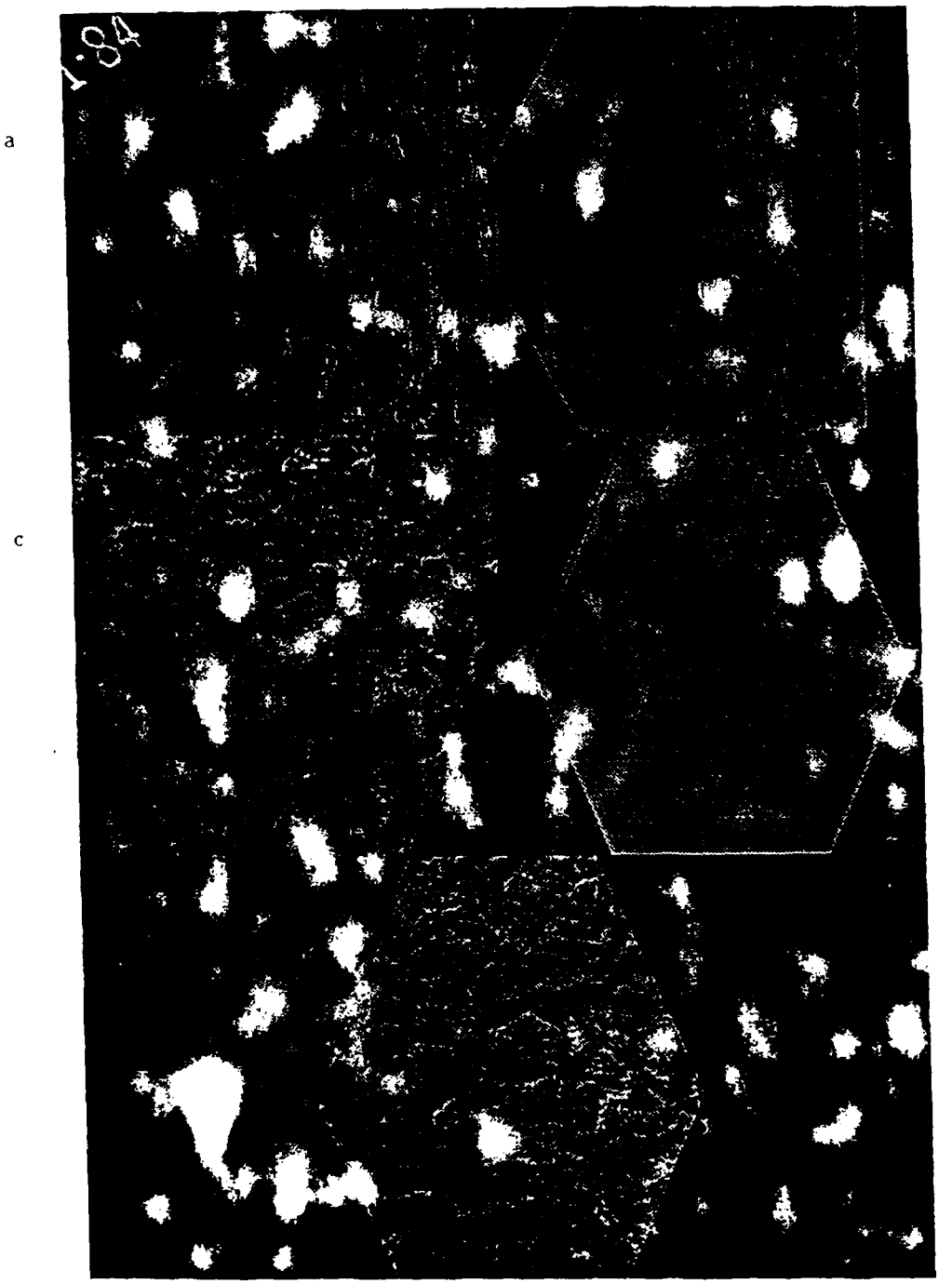


Figure 8: The on line steps. a.g_1, b.$g_{(2-1)_2}$ c.g_{1_3}, d.$g_{(2-1)_{2_3}}$ e.g_3

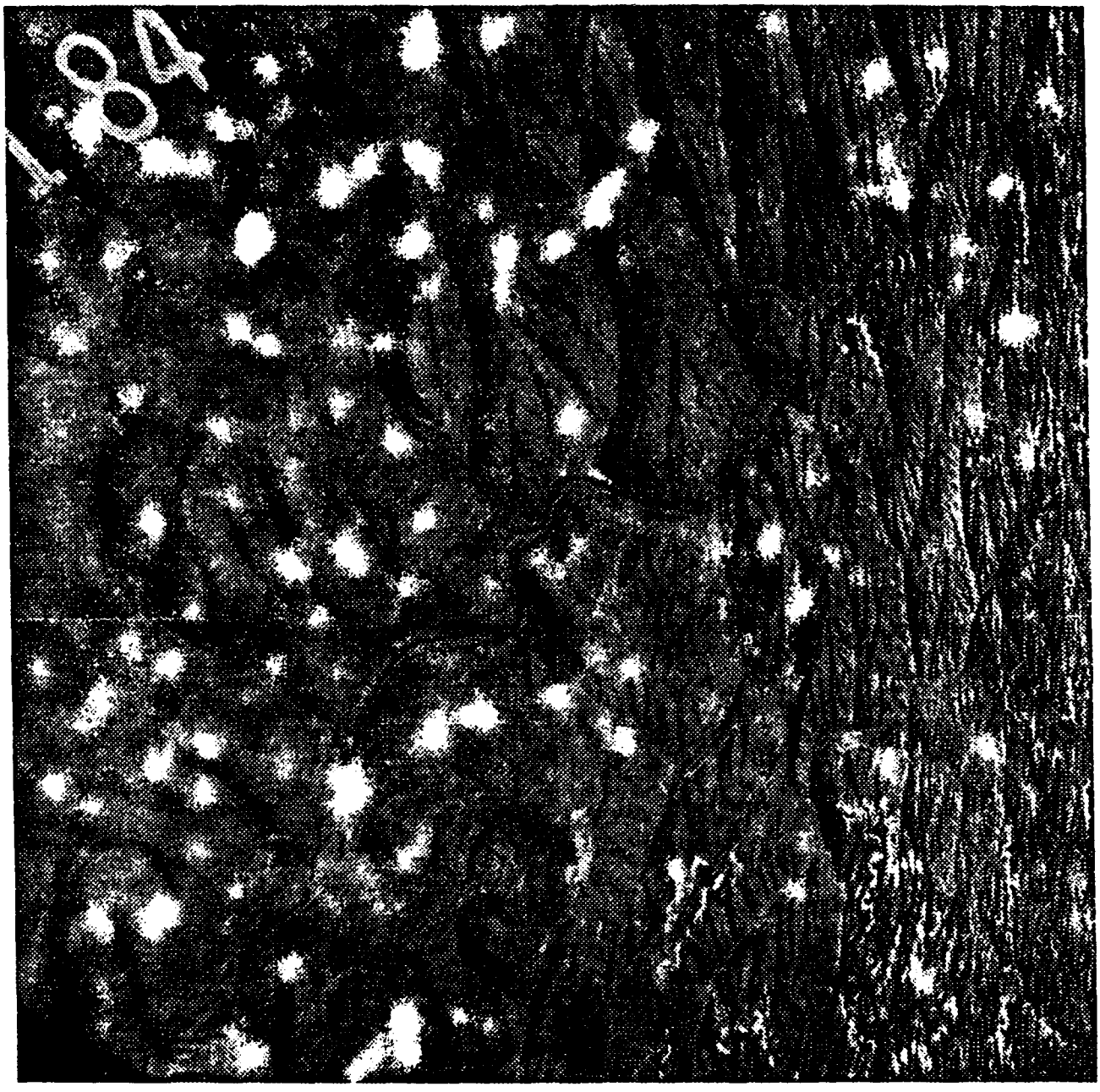


Figure 9: The first approximation, g_1

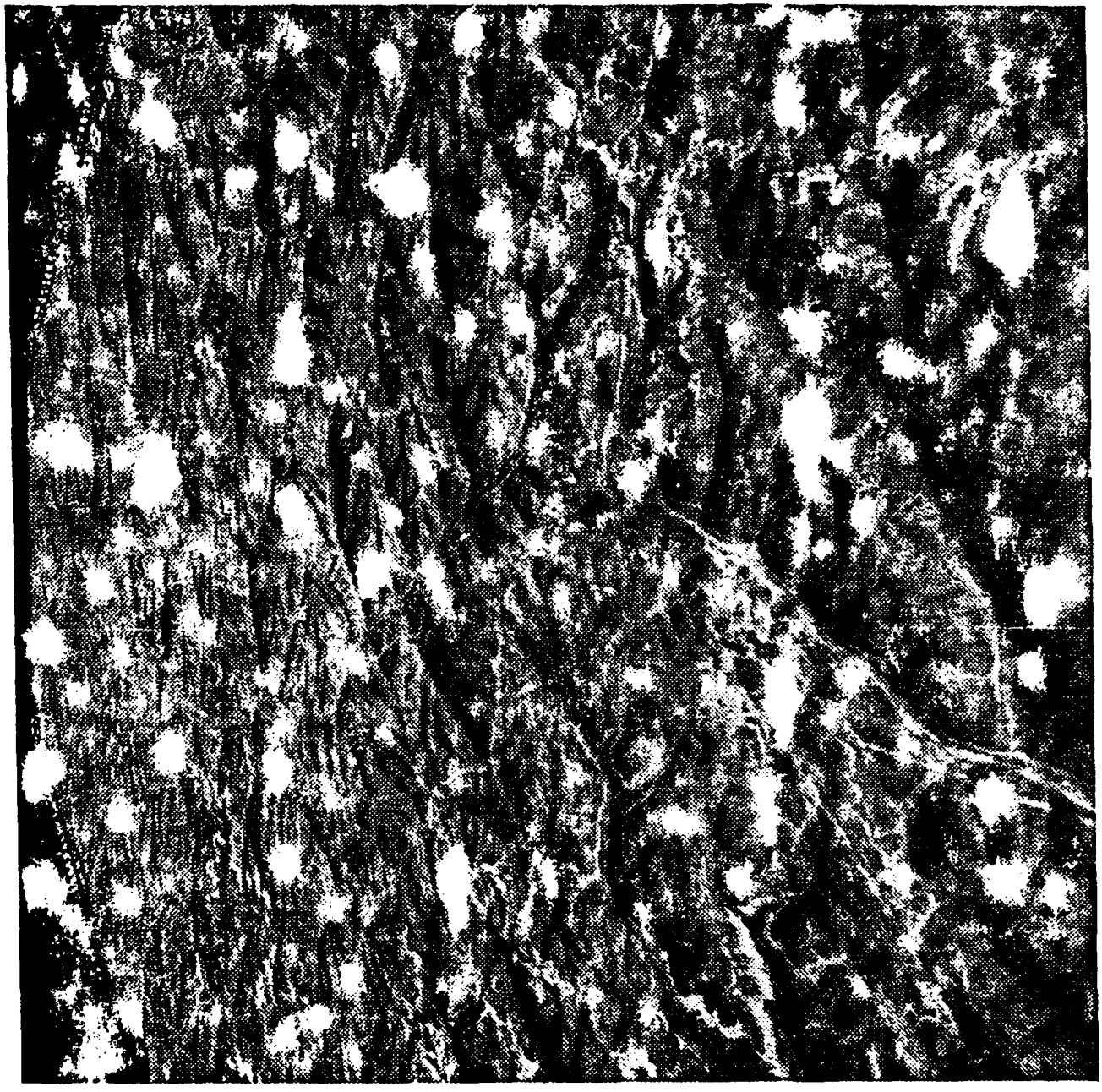


Figure 10: The second approximation, g_2

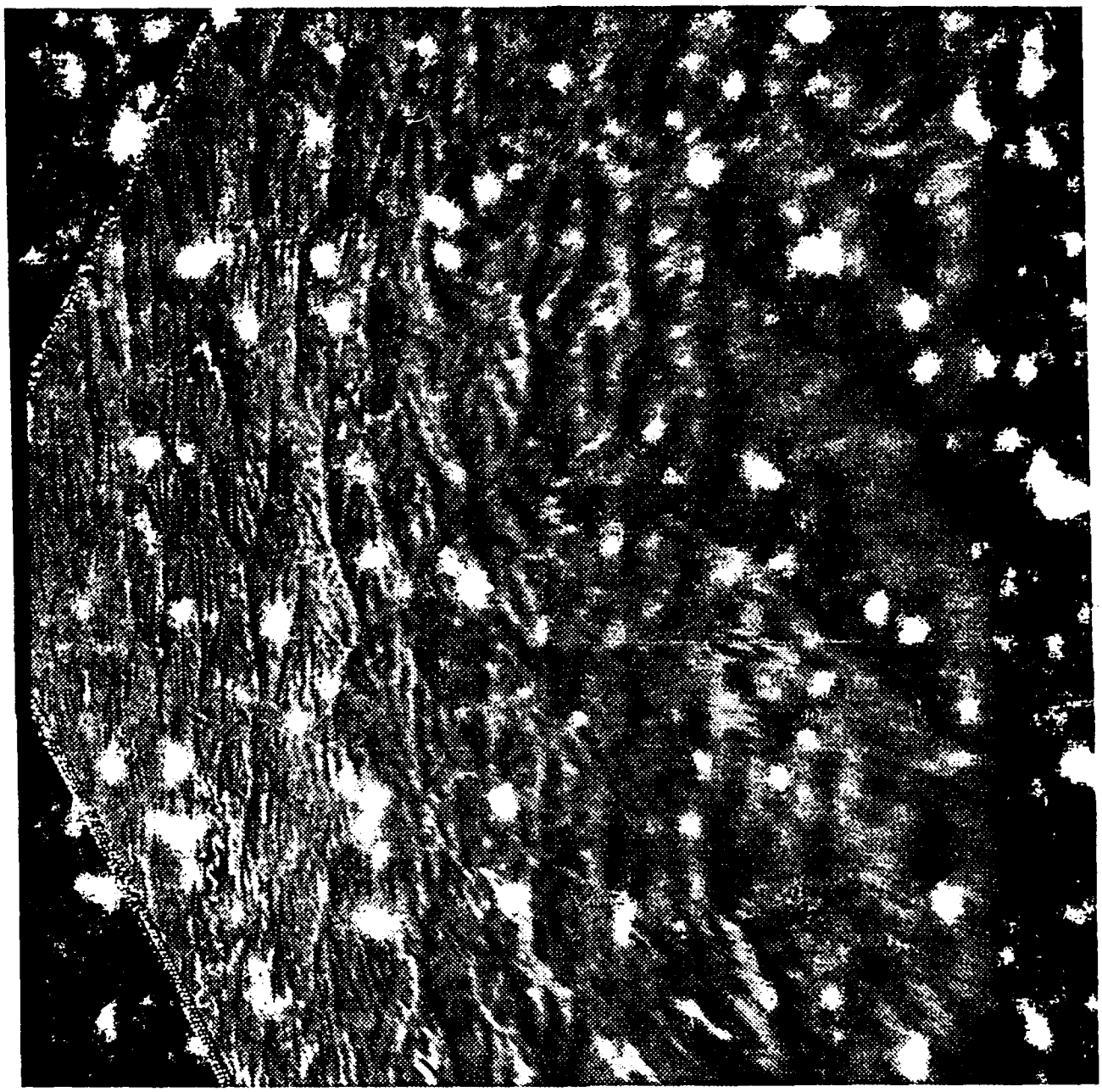


Figure 11: g_{Σ_1} — the first approximation projected on B_{Σ_1}

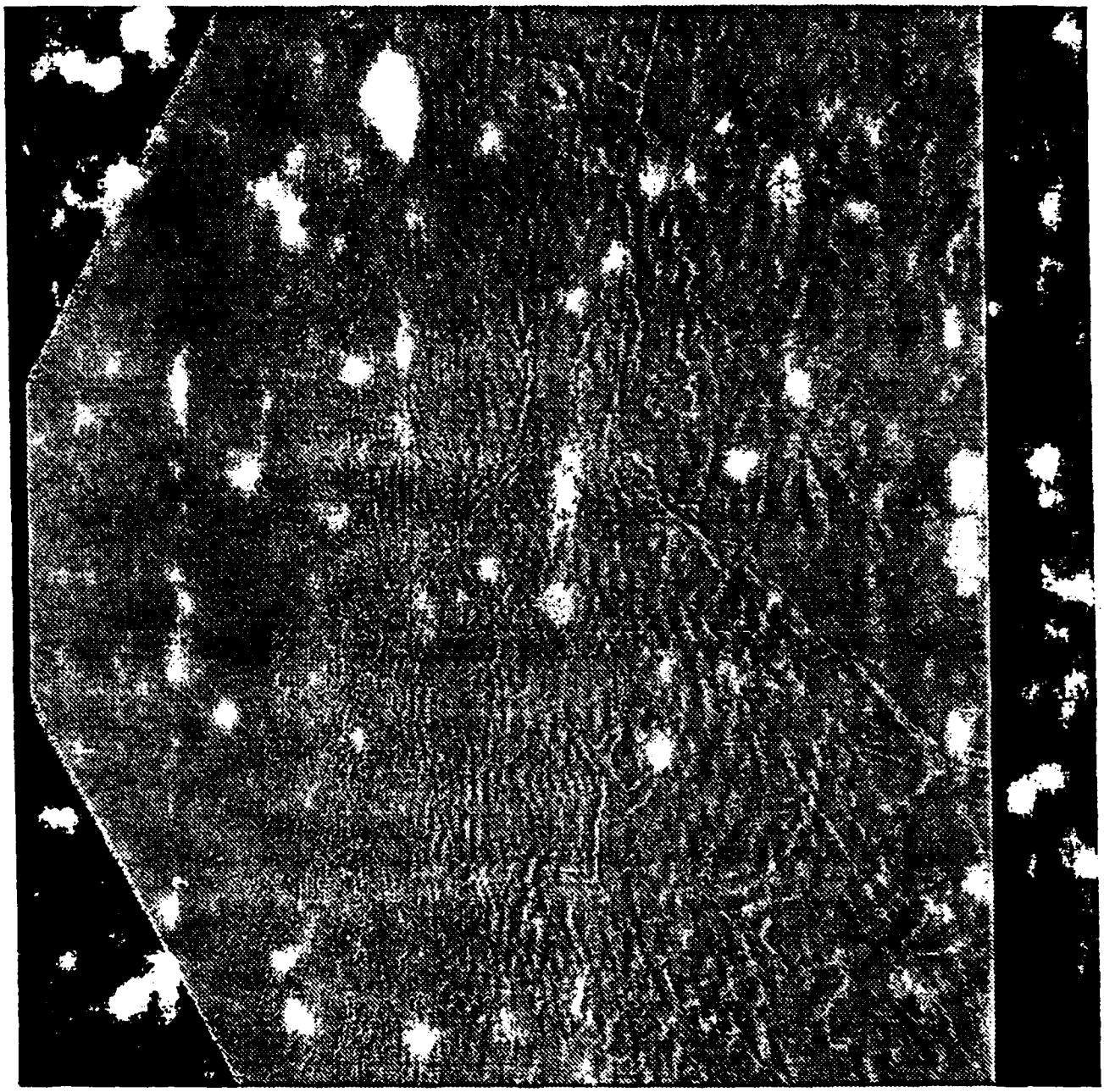
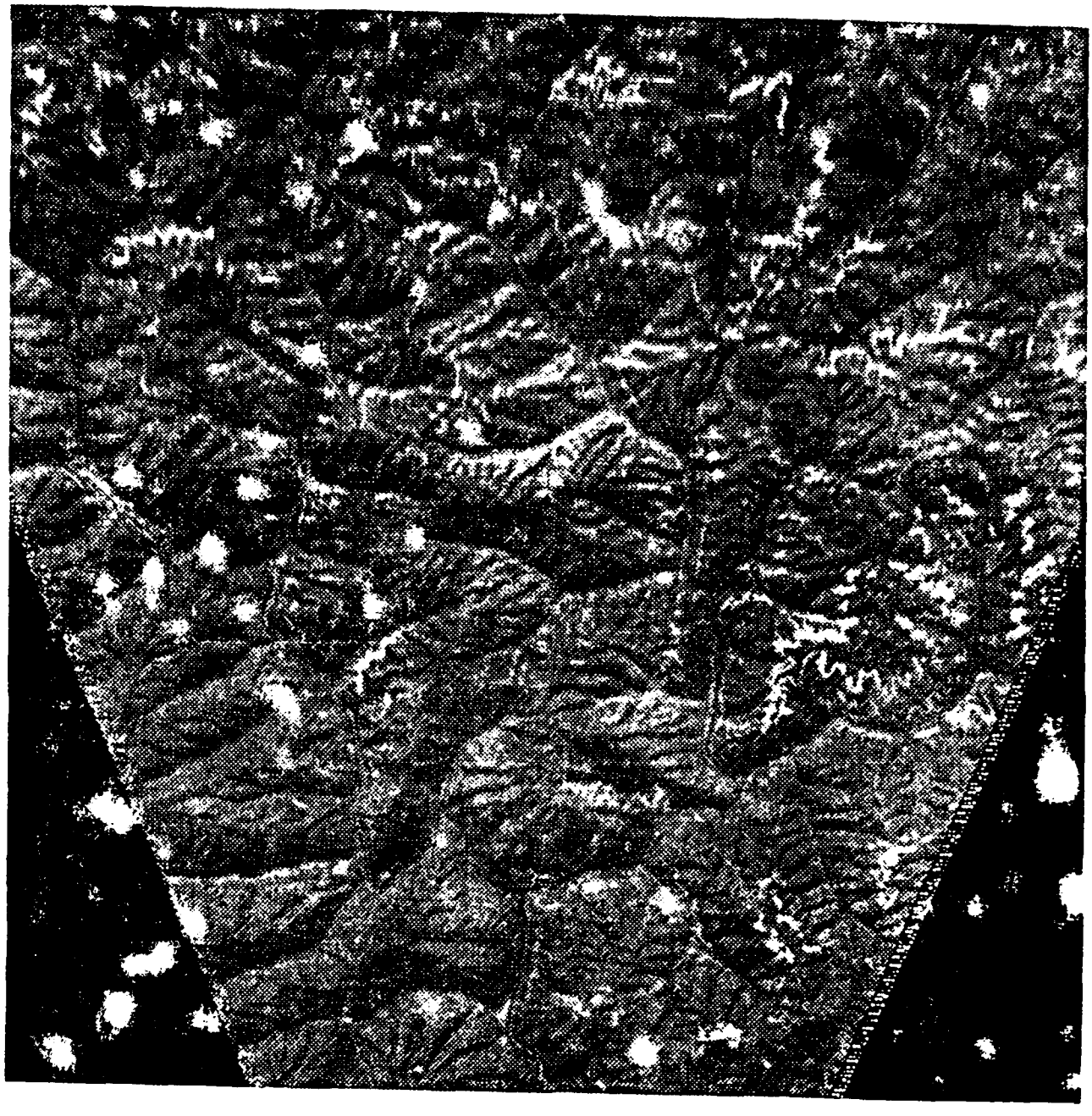


Figure 12: $g_{(2-1)_2}$ - the difference between g_1 and g_{1_2}



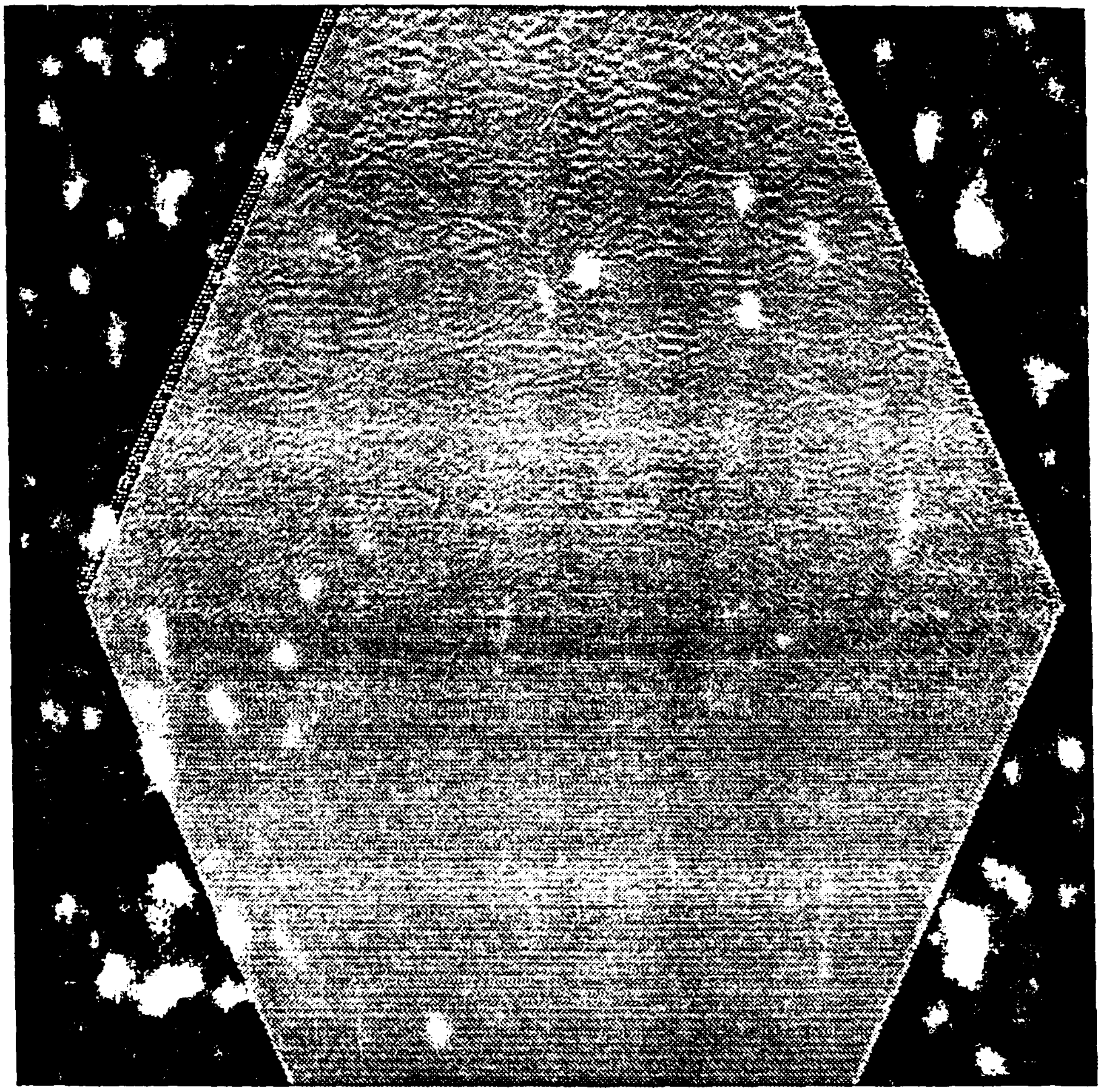
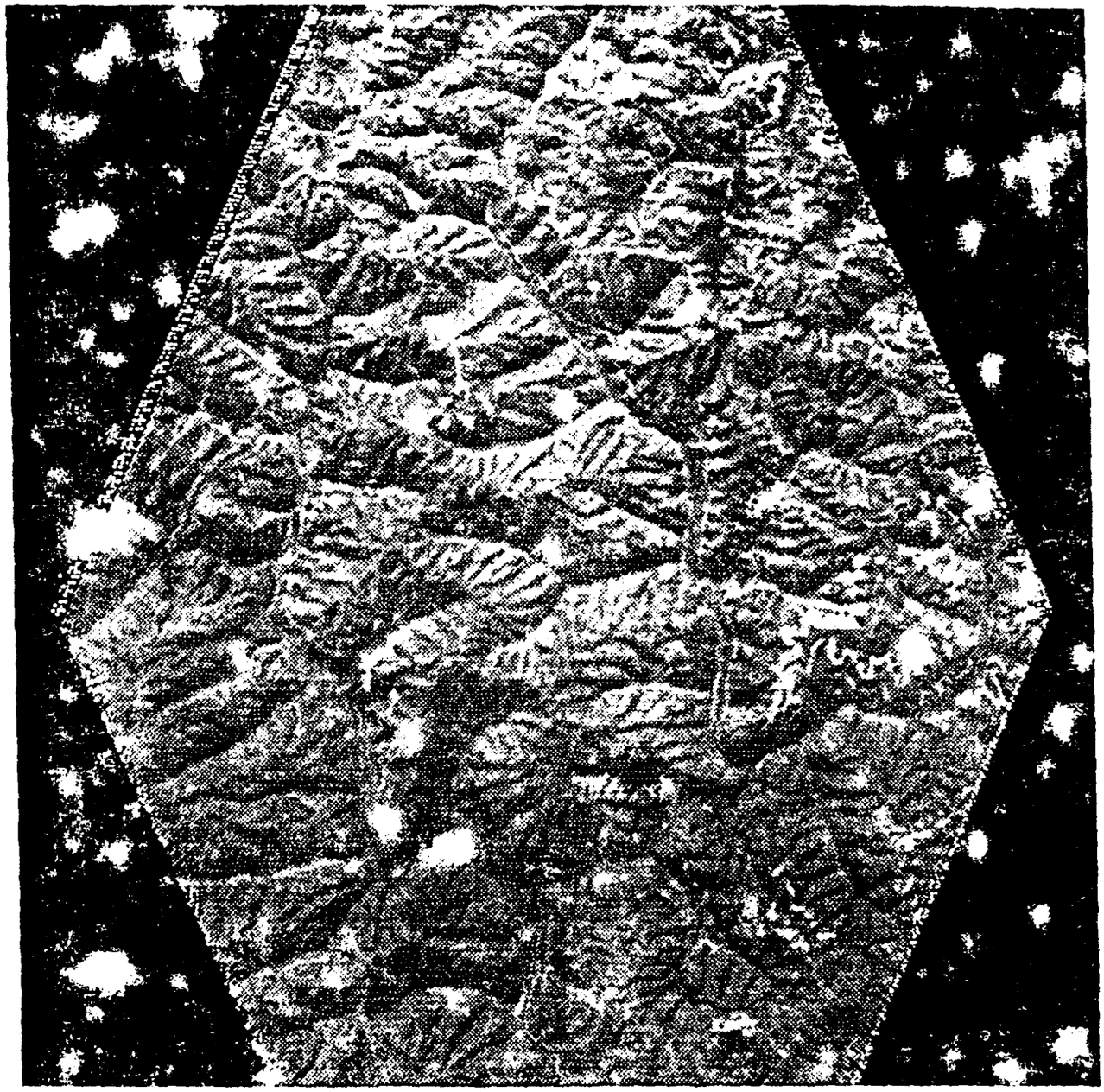


Figure 11 $g_{k2-1_{B_\gamma}}$ - the projection of $g_{k2-1,2}$ on B_γ



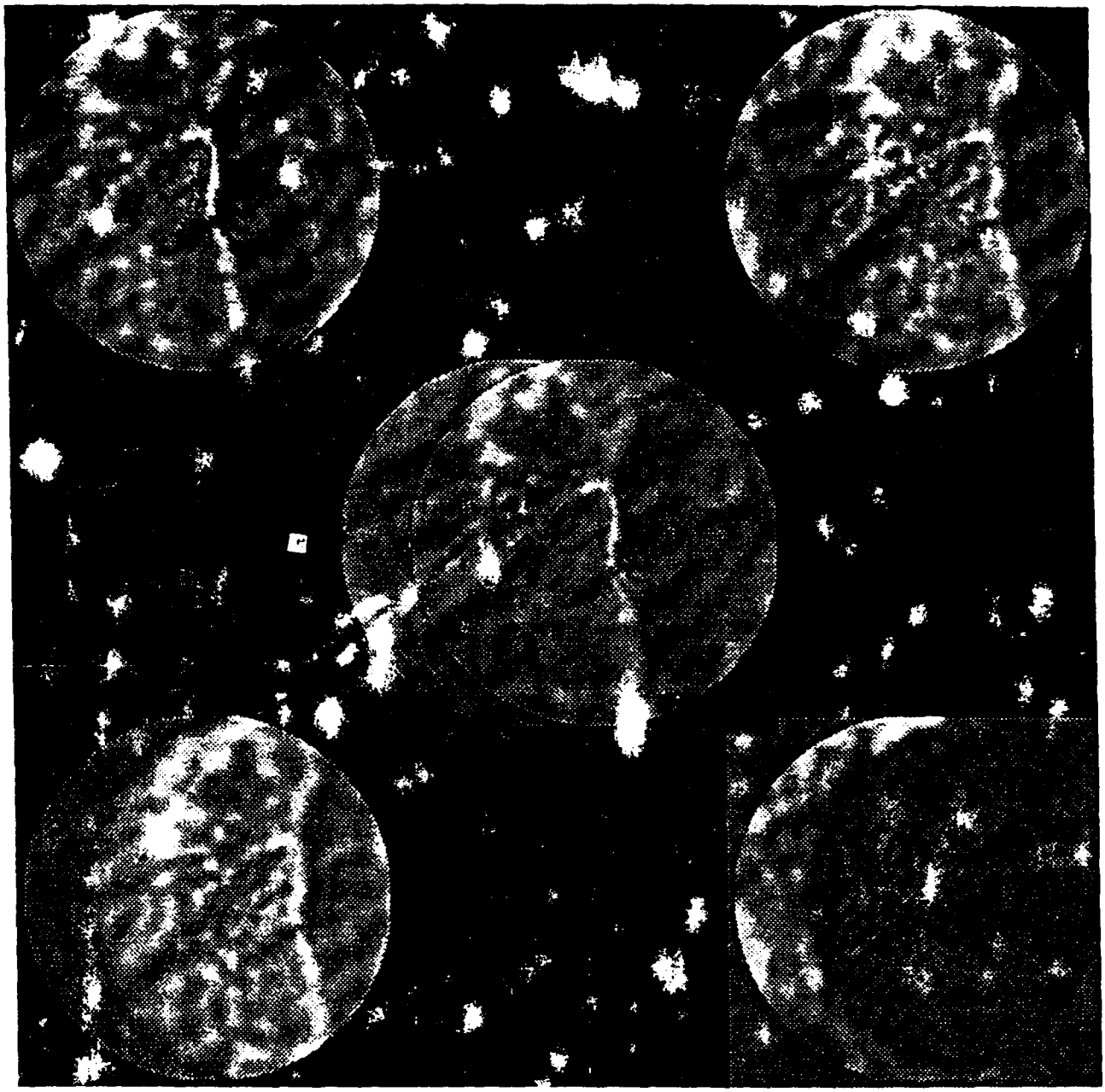


Figure 20: Combining two approximations, generated by a radial foveating filter

Technical Report may be obtained

from the author

or from UES

Research Initiative Program

Sponsored by the

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Conducted by

Universal Energy Systems, Inc.

Final Report

NEUROCOMPUTING IN INTELLIGENT TUTORS: STUDENT MODEL DIAGNOSIS

Principal Investigator: Robert A. Zerwekh, Ph.D.
Department of Computer Science
Northern Illinois University
DeKalb, Illinois 60115

Date: 14 December 1990

ABSTRACT

This report describes the results of using neurocomputing as a means of performing student model diagnosis in an intelligent tutoring system. Neurocomputing was investigated as an alternative to traditional methods of student model diagnosis because of its potential to alleviate some of the computing drawbacks associated with these conventional methods.

Two types of training algorithms for neural networks were explored. In one, a back-propagation neural network was to be taught to recognize various levels, or categories, of (language) competence in individuals who were learning a second language. When an individual began using a computer assisted language tutor, the network would be able to recognize his language competence level. This information could then be used to assist the instructional planner in offering more effective lessons for that individual.

The other training algorithm explored was a form of competitive learning known as adaptive resonance theory. With this type of network, one does not have to teach the network the categories of language competence beforehand. Rather, the network is allowed to create its own categories merely by reviewing input vectors of information. The input vectors represented the choices and actions (the behavior) of individuals who had used the foreign language tutor.

ACKNOWLEDGEMENTS

I wish to thank the Air Force Systems Command, the Air Force Office of Scientific Research, and the Human Resources Laboratory at Brooks Air Force Base for sponsoring this research. I also wish to acknowledge the support and assistance of Universal Energy Systems during the past year.

A number of individuals at Northern Illinois University have also helped me to have a very productive year of research. I would especially like to thank Jim Henry for his friendship, support, and marvelous insights, Patricia Henry and John Hartmann for their patience in tolerating my questions, and Rodney Angotti for his assistance and support. I would also like to extend my appreciation to the people in the Office of Sponsored Projects and to Robyn Bassing for helping me stay organized.

I. INTRODUCTION.

The research conducted during the 1989 Research Initiation Program was a continuation of work begun during the summer of 1989 at the Intelligent Systems Division of the Human Resources Laboratory, Brooks Air Force Base. This Division conducts research into principles and techniques of artificial intelligence that can be applied to the design and implementation of Intelligent Tutoring Systems (ITSs). These tutoring systems are used by the Air Force for training personnel for situations and tasks that would be very costly or highly risky if performed on real equipment. Although functioning ITSs are a relatively recent phenomena, enough success has been achieved in their design and deployment that it is possible to identify a system configuration, or architecture, that is common among them all. This architecture consists of five primary components: (1) an expert module that contains expert knowledge about the teaching domain; (2) an instructional module that consists of pedagogical plans and strategies; (3) a student module that contains information about the student currently using the ITS; (4) a device simulation that provides an instructional environment; and (5) an interface through which student and ITS interact.

Two of the more critical modules in an ITS are the instructional and student modules. The instructional module is responsible for constructing a pedagogical plan, or

strategy, that will present lesson material so that it is individualized for the particular student who is currently using the ITS. The instructional module, therefore, must interact quite closely with the student module, for it is the student module that is supposed to provide information about the present cognitive states and requirements of a student. Consequently, it is critical that the instructional module be provided with accurate, informative, and useful information about this student. This process of constructing, or inferring, information about a student is called diagnosis because the ITS is attempting to uncover a hidden cognitive state (the student's present knowledge) from his or her observable behavior (VanLehn, 1988).

The research conducted an investigation into the feasibility of using a new technique to diagnose information about a student. There were three primary reasons for this choice. In the first place, present areas of research in ITSs focus not so much on issues of system design or first principles, but rather on the creation of new tools, or techniques, that will help the basic system modules work together more effectively, or that will enhance the operation and efficiency of one of the identified modules. Given the critical role played by the student model in an ITS, I decided to investigate methods for improving its functionality and performance.

Secondly, hands-on experience with existing ITSs and

investigation of the literature on student modules revealed that the present state of this ITS component was seriously underdeveloped in comparison to other system modules. If ITSs are to live up to their true potential, then it is critical that student models present the most revealing and accurate picture of a student as possible. There is no practical advantage to be obtained by having well developed instructional modules, for example, that have little or no significant information about students for whom instruction is intended.

Finally, a study of the student modeling techniques that are presently employed in ITSs showed that one or two general problems seemed to be common to them all. Some current diagnostic techniques require that extremely large databases or lists be constructed that consist of typical errors students make (so called "bugs") or that contain inference or production rules that attempt to account for every possible scenario that may occur during a specific tutoring session (Anderson, Boyle, Yost, 1985; Clancey, 1982; Burton, 1982). As the size of these databases increases, the computations necessary to locate the correct "bug" or rule become more and more demanding. This combinatorial explosion problem is particularly acute for domains where the subject matter is more complex, less structured, and more prone to have elements of uncertainty pervade the reasoning process.

The other general problem faced by current diagnostic techniques pertains to the speed of the diagnostic algorithms. Present algorithms that attempt to provide true interactive and dynamic diagnosis of the student module are too slow to be of practical use in an ITS. There is a natural desire to want to increase diagnostic speed as one does not want the student to become bored with the tutor or lose any willingness to cooperate. This problem is also related to the first problem, for as the size of bug libraries and production rule libraries increases, the complexity of the algorithm needed to search these libraries to locate the correct information also increases.

Because of the presence of these two general problems, and for the other two reasons, the research explored using an alternative method to develop the student model. This alternative is called neurocomputing, and the reasons for investigating its use are explained in the next section.

II. OBJECTIVES.

The main objective of the research effort was to use neurocomputing as an alternative methodology for performing student model diagnosis. Neurocomputing was chosen because it has a number of features that make this computing paradigm highly attractive for performing the tasks required for this procedure, and it can do so in a manner that is not subject to the two general problems noted above.

Neurocomputing departs significantly from the manner in which information is processed in traditional, programmed computing environments. Instead of relying on algorithms coded in software, or on libraries of production rules, the primary information processing structures in neurocomputing, neural networks, develop their information processing capabilities by being exposed to the kind of information processing task one wants the neural network to emulate. Repeated exposure of an information processing task (which is called "training" the neural network) causes the neural network to adjust itself internally such that, after some period of training, it is capable of duplicating this information processing task. This is done without having to write or invent an algorithm that instructs the neural network what to do next, nor does it require large libraries of production rules. This means that neurocomputing can avoid the two bottlenecks that plaque conventional methods of student model diagnosis.

A neural network can be described as an information processing structure that is composed of a number of interconnected processing elements (also called nodes, cells, units, or neurons). Each processing element computes its activity locally based on (a) the activities of the nodes to which it is connected, and (b) the strengths of these connections, or its weights. A specific transfer function then determines the processing element's output

given its input (Rumelhart and McClelland, 1986; Grossberg, 1988; Hecht-Nielsen, 1990).

Since information processing is both parallel and distributed in a neural network, these structures are capable of dealing with large numbers of constraints and variables in a non-monotonic fashion. Thus, they are not subject to the combinatorial problems that plaque conventional algorithmic or rule based programming. Another advantage of neural networks over rule-driven systems is that they are capable of dealing with incomplete, imprecise, or partially correct data. In particular, this feature allows neural networks to continue to process information even in the face of real world ambiguity and uncertainty. Production rule systems, on the other hand, will typically falter if not presented with precise and complete data.

One of the most remarkable things about neural networks is the manner in which they learn how to process information. Two fundamental learning paradigms are presently used in teaching neural networks how to process information. Each has as a goal to get the network to produce some consistent or desired output when exposed to an input vector of information. This is accomplished by having the network adjust its weights, the connection strengths that exist between units, until presentation of the input vector produces the desired, or consistent, output vector.

In one learning paradigm, that of supervised training, the network is exposed to a vector of input stimuli and its output vector (or response) is compared to the desired, or target, output vector. A calculation of the difference, or error, between the network's output and the desired output is made and this error term is fed back through the network so that weights can be adjusted in a manner that minimizes this error. The network is said to be trained when this error term reaches zero or some specified minimum. At this point the network can consistently produce the correct output when exposed to a given input or some partial input vector (Wassermann, 1989).

The other learning paradigm, unsupervised training, does not employ any desired or target responses in training a network. The network is exposed only to input vectors of information and weights are modified so that the network produces consistent output representations of the input. In other words, applying an input vector, or applying some partial representation of an input vector, will cause the network to produce the same output response. In this sense, the network is allowed to construct its own representations of the most significant features of the input patterns (Kohonen, 1988; Grossberg, 1988).

The objective of the research, neurocomputing the student model, was to be pursued using the Foreign Language Instruction Station (FLIS), a computer assisted foreign

language tutor developed by Jim Henry and the principal investigator. FLIS is currently used at Northern Illinois University by forty eight students who are enrolled in Thai and Indonesian classes. Although FLIS should not be considered to be an intelligent tutoring system, it does employ a device simulation, an interface, a small expert domain module, and an intermediate band width student module (VanLehn, 1988). A neural network was to be incorporated into the architecture of the FLIS system so that it could perform the diagnostics on the existing information in the student model. The output of the neural network would be a classification of a student's second language competence level, with this information assisting an instructional planner that could then provide more appropriate language lessons for that type of student.

The research plan called for two approaches to student model diagnosis. One approach was to use a form of supervised training of the neural network to train it to recognize certain (preconstructed) categories or classes of second language competence. This training would take place prior to students using the language tutor. During the course of a lesson, however, the student's actions and responses would provide the input data that the network would use to classify the student as belonging to a particular language competence class. The other approach was to let the neural network develop its own

classifications through an unsupervised learning scheme.

III. SUPERVISED TRAINING.

A substantial amount of time was spent writing the code for the neural network and installing the network into the FLIS system. The network was written in Turbo Pascal so that it would be compatible with the existing software platform of FLIS. This development time included testing the neural network on sample data sets to ensure that it was operating correctly. The learning law employed in the network (or the law that specifies how weight changes during training are to occur) was the back-propagation algorithm described by Rumelhart, Hinton, and Williams (1986). Back-propagation is a supervised learning procedure which attempts to teach a network correct output classifications by minimizing the error between the network's actual output response and the desired, or target, response. In the case under consideration here, the target output responses to be taught to the network were different levels, or stages, of second language competence in students using the FLIS system. In other words, once the network was trained, it would be able to receive input data from the FLIS system as a student was using it, and classify this student into one of five competence categories.

The choice of five categories was not a random choice. The ACTFL Proficiency Guidelines (1988) describe six levels

of second language competence for speaking, listening, reading, and writing. The last level described was not included in this research since it describes the skills possessed by a native speaker (listener) of the language and those who use FLIS are primarily beginning students of the language. The six categories for listening comprehension described are: Level 0: no proficiency; Level 1: elementary proficiency; Level 2: limited working proficiency; Level 3: general professional proficiency; Level 4: advanced professional proficiency; and Level 5: functionally native proficiency.

Throughout the summer of 1990 independent teams of faculty and graduate students were designing new lessons in Thai and Indonesian for FLIS in preparation for the beginning of classes in early September. These teams were operating under the auspices of grants from the U.S. Department of Education and the National Security Agency. Consequently, I had only tangential input into the design of lessons.

The supervised training of the network was dependent on the ability to construct accurate profiles, or descriptive categories, of different stages of second language competence. Although a number of studies (see Krashen, 1982) indicate that language acquirers go through distinct stages in the process of acquiring a new language, including the comprehension theory on which FLIS is based, in practice

it turned out to be extremely difficult to specify the characteristics of these stages to a degree sufficient enough to train a neural network to recognize them.

The efforts to define categories of language competence were complicated primarily by the types of lessons that the lesson authors ultimately ended up designing. The lessons used in the FLIS system permit a wide range of possible actions on the part of students. Typically, an individual frame in a lesson is composed of the following parts: (1) The student hears a prerecorded audio segment, possibly accompanied by the text of the speech on an alternate screen that the student can view with an appropriate combination of keystrokes; (2) Depending on the type of frame, the student may then be presented with a multiple choice question, a true/false question, or be required to type in an answer in English or in the second language; (3) Before students answer a question, they have a number of aids available which they can access to help them solve the problem; these aids include repeating the speech, hearing a slower version of the speech, hearing an English translation of the speech, and accessing menus of clues and glossary items.

Due to the wide variety of available actions and options these lessons permit, it was virtually impossible to specify a set of actions and state with any degree of certitude that this set is indicative of the behavior of, say, a level 1 language learner. Yet this was what was

required if categories of language competence were to be learned by the neural network. Since the network would be performing dynamic classification of a student during the course of a lesson based on that student's actions as represented by a series of keystrokes, the network had to be trained to recognize similar series of actions. There was no way to know in advance, however, what actions a student having trouble with the concept of reverse negation, for example, would take as opposed to one who had mastered that concept. The problem student might look at all the clues before answering because he genuinely needs help. The other student might do the same, not because she needs help, but because she is meticulous and simply wants to see every clue.

Although FLIS was designed to emphasize listening comprehension, there were not any lessons that truly exploited this. Most of the lessons ultimately designed were drill and practice types, or review of vocabulary items. This is not too surprising. Although the importance of listening comprehension is beginning to be noticed by language teachers, relatively little theory has been developed that is specific to second language listening (Richards, 1983; Bernhardt and James, 1987).

In a recent paper, Lund (1990) has suggested a specific methodology for designing listening comprehension lessons that might prove to be more amenable to specifying

categories of language acquirers that could be taught successfully to a neural network. His proposal is to separate listening tasks from listening responses. Listening tasks define how the learner should approach the lesson and suggest what should be derived from the lesson. Six primary listening tasks are specified that range from simple tasks (identifying specific words in a recorded passage) to complex tasks (full comprehension of the main idea plus the details). Listening responses are defined as what the listener does to demonstrate successful listening, and these responses are often related to or suggested by the various listening tasks. Lund suggests nine response categories ranging from simple ones (choosing an item from a list of pictures) to complex ones (carrying on a conversation).

Imposing this task-response matrix on lesson design means that there are potentially 54 cells requiring evaluation, though there may be fewer since every task does not always have to have a specific response associated with it. This would help eliminate much of the variability that exists in the current FLIS environment and permit the development of target responses that could more easily be taught to the neural network. I am currently attempting to persuade the lesson authors to use a strategy similar to this during an upcoming round of lesson design.

III. UNSUPERVISED TRAINING.

Despite the success the back-propagation training procedure has had in a number of other applications, it is not an appropriate training procedure to use when the number and kind of patterns to be exposed to the network is not known in advance. The previous discussion has pointed out that the network input data provided by FLIS (that which results from a student taking a lesson) is far too varied; each student could potentially create a different input vector of information. Under such circumstances, exposing a back-propagation neural network to those input vectors would cause the network to change its weights constantly, never settling down to learn anything.

This drawback associated with back-propagation is called the temporal instability problem and was identified most persuasively by Carpenter and Grossberg (1986). Their argument was that learning should be <u>stable</u> but also <u>adaptive</u>. What a network has already learned should not be destroyed by the presence of new input patterns. Yet the network should have enough plasticity to be able to learn new input patterns, again without this destroying previously learned patterns. Their work on neural networks has generally been classified as a form of competitive learning.

Competitive learning is a form of unsupervised learning in that the desired or target output need not be known or shown to the network in advance in order for it to be able

to learn. Competitive learning always involves some sort of competition process among some or all of a network's processing elements. At the end of the competition, only the winning processing element is allowed to change its weights. Teuvo Kohonen's work on self-organizing feature maps (1988) and Carpenter and Grossberg's work on adaptive resonance theory (1987a, 1987b) are two of the better known instantiations of competitive learning.

A neural network based on adaptive resonance theory (ART) is an input vector classifier. Input vectors of information are presented to the network in a continuous but arbitrary fashion. The network responds in real time by associating each vector with one of some number of previously learned patterns that the new pattern most closely resembles. If the new pattern is unlike any of the previously learned patterns, the network will create a new category. When the network does recognize a new pattern as similar to a stored one, it is essentially matching invariant properties in the input pattern with exemplars in its stored recognition categories. During recognition, the stored category is modified slightly to make it even more like the new input vector.

How closely a new input pattern must match a stored one is controlled by an attentional vigilance parameter. If a match does not occur within the range specified by this parameter, then none of the stored exemplars are modified.

Instead, the network will create a new recognition category. This is how an ART network solves the stability/plasticity problem. New input patterns cause new categories to be created and do not destroy previously learned patterns.

Given the diversity of the input patterns produced by students using the FLIS system, it seemed appropriate to expose a neural network based on the ART paradigm to these input patterns. It was reasonable to assume that the ART network would be able to classify similar input patterns together if there were similarities among the patterns. If the patterns were all so radically different that the ART network was creating a new category for each pattern, this would soon be clear. On the other hand, if the network was able to create a limited number of stored recognition categories, then examining the input patterns that matched a given category might reveal something significant about the behavior or cognitive requirements of those students who grouped together.

The network used in this phase of the research was an ART2 neural network paradigm provided by ExploreNet 3000, a neural network software package from the Hecht-Nielsen corporation. An ART2 network is distinguished from an ART1 network by the fact that ART1 networks can only handle binary inputs. ART2 networks can handle binary inputs as well as continuous range inputs and so was better suited for the continuous range input data produced by the FLIS system.

The architecture of the ART2 network reveals that it has two layers of processing elements (called the F1 and F2 layers), with each layer having multiple "slabs" of processing elements. The F1 layer is divided into seven slabs which represent the short term memory of the network. Two of the more significant slabs in the F1 layer are the input slab and the P slab. The input slab receives the input vector of information and fans out the data entered without performing any calculations on it. Each element in the P slab is connected to every element in the F2 slab and the bottom up and top down weights that govern these connections represent the long term memory of the network.

New input patterns are learned and categorized by changing the bottom up and top down weights (the long term memory traces) that connect the F1 processing elements to the F2 processing elements. The F2 processing elements engage in a "cooperative-competitive" interaction during the recognition process in an attempt to match the input patterns. The winning processing element is the one that most closely matches an input pattern. The attentional vigilance parameter, which determines the degree to which an input pattern must match a stored category, is provided by the top down weights from the F2 layer to the F1 layer.

As the network operates, the vigilance parameter determines the degree of recognition required for a classification response to be issued. If an F2 processing

element matches closely enough, the network is said to have achieved resonance. If there is a mismatch, the F2 processing element is inhibited and the network's processing is repeated until a match is found. A rapid search takes place over the existing learned categories in an attempt to find a matching category. If no existing categories match the input pattern, the network will create a new category via unsupervised learning. If all the existing F2 processing elements have been assigned a category, then it is impossible to create a new one. In this case, all F2 elements are inhibited and the network is said to be in a reject state.

The input vectors presented to the ART2 network consisted of raw data derived from the student model information that the FLIS system collects and stores for each student. At the time of testing the network's performance, each student in the Thai class had completed, on average, 18 lessons, and each student in the Indonesian class had completed an average of 14 lessons. A filter program was written to process each of the lesson data files associated with each student and to prepare the input vectors for the network. Thus, for a student in the Thai class, for example, this program read each of the 18 lesson data files, processed the keystroke history contained in these data files, and produced an input vector to be presented to the ART2 network that consisted of the

following items:

1. Average number of minutes spent on a lesson.

2. Percentage of correct first attempts at answering a question. (Many lesson frames permit a student a second or third chance to answer a question correctly. We were only looking at correct first attempts.)

3. Percentage of incorrect first attempts at answering a question.

4. Average number of times the student accessed any form of help before attempting to answer a question. (Any form of help refers to the student accessing any of the items listed in 5 through 11 below.)

The remaining items listed represent the average number of times the student accessed, looked at, or listened to that item during the course of a lesson.

5. the Clues menu

6. the Glossary menu

7. a Repetition of the second language audio segment

8. a Slower version of the second language audio segment

9. an English audio translation of the audio segment

10. a Text version of the second language audio segment

11. a Repetition of the story associated with the lesson frame (Some lesson frames are preceded by a small story told in the second language. A student can repeat that story during the course of the lesson frame.)

The ART2 network was constructed to have eleven

processing elements in the F1 input layer (one for each of the input vector fields) and five F2 processing elements. This latter number was selected to limit the number of categories the network might create to accord with the ACTFL proficiency guidelines for listening comprehension levels. In addition, during the training procedure, the Thai and Indonesian input vectors were presented separately to the network due to the difference in the number of lessons taken for each class and due to differences in the structure and style of the lessons for the two classes.

The results from the ART2 network were much more promising than anticipated. Given the wide diversity among student lesson data files, it was assumed that the network would rapidly assign F2 processing elements to the first few input vectors it reviewed, causing a quick saturation of the network leading to a reject state. Instead, for the 25 input vectors from the Indonesian class, the ART2 network only created four exemplar categories, and for the 23 input vectors from the Thai class, only three categories were created. These are the results achieved after some experimentation with the vigilance parameter. Setting the vigilance parameter too high meant that subsequent input patterns failed to match the first few patterns learned, thereby leading to a reject condition. If vigilance is too high, the result is poor generalization on the part of the network since minor variations of the same pattern lead to

the creation of separate categories. On the other hand, if the vigilance parameter was set too low, patterns that were clearly different would get grouped together, distorting the stored pattern image until it did not resemble any of the input vectors.

Examination of the input vectors and the categories to which the network assigned them revealed that the network, for the most part, was responding most strongly to item 4 above; i.e., the average number of times a student accessed some form of assistance before answering a question. My initial impression was that poorer students would seek more help than brighter students, thereby supporting the hypothesis that the network would create categories according to the level of one's language competence. Subsequent questionnaires administered to the students and personal interviews with the professors of the classes, however, proved this particular hypothesis not to be the case.

Except for three exceptions in the Thai class and five in the Indonesian class, the amount of help sought before answering a question was more closely tied to motivational and attitudinal factors about the students than it was to language competence. Yet, within each of the categories created, the students in that category shared remarkably similar second language skills. It does appear that the network is categorizing students in terms of their language

competence levels, but this categorization seems to result as a secondary effect from the grouping around motivational factors. Perhaps the next discussion will serve to clarify this.

As stated, the network grouped the Thai students into three categories. For simplicity, we will call them categories 1, 2, and 3. Category 1 students sought help an average of 3.25 times per lesson, category 2 was the highest at 5.92 times per lesson, and category 3 was the lowest at 0.78 times per lesson. Except for one exception in category 1, these students were described by their professor as the best in the class. In terms of motivational factors, he used adjectives such as "extremely hard working", "extraordinarily detailed", "overachiever", and "perfectionist" to describe these students. Although the one exception did not have an A average like the others, the professor did remark that this particular student did have the highest motivation of all his student athletes.

Category 2 students were averaging a B in the class. The professor described these students as "attentive", "hardworking", "complexities trouble him", and "motivated, but has trouble with the spoken language" to name a few. It is perhaps worth noting that for the last student described, 90% of his help count consisted in repeating the spoken audio segment (item 7 in the input vector).

Category 3 students made up the largest segment of the

student population. Their average grade was a C and they were described as having "low motivation" and as being "immature" and "lazy." Two students in this group seemed to be exceptions as they had A averages. One was described as "excellent", the other as "error prone." The excellent student had the lowest help count of anyone in this group, while the error prone student had the highest.

It is interesting that the network produced these categories from input vectors that were summations of students' behavioral actions and choices. No attitudinal factors or personal characteristics of students were included in the network's input. Also note that none of the traditional "objective" measures of intelligence (grade point average, SAT scores, etc.) were included. It is worth pursuing whether the network's categories would differ if it was provided with some of this additional information.

The conclusion of this section is that the unsupervised model of network training does appear to create and identify significant categories of language learners based on input that is a reflection of their observable behavior while engaged with the language tutor. Additional study is required to ascertain which aspects of this observable behavior are more telling than others. At this moment, with the FLIS environment, students' access of aids during the course of a lesson is the primary significant aspect of behavior. Different lesson designs, however, or possibly

even a different group of system users (for example, more advanced speakers of the language) might mean that other aspects of their behavior, or even psychological measurements, would be the most significant features for the network.

IV. RECOMMENDATIONS.

The major recommendation of this report is to continue support for projects that explore the use of neurocomputing in Intelligent Tutoring Systems. The findings of this research project suggest that neurocomputing can be applied successfully to areas in which human and cognitive factors are the primary focus as opposed to most research in neurocomputing which concentrates on wave form and signal analysis. While the work reported on in this report has achieved results that are very promising, it is certainly far from finished. The research conducted here has shed some light on the initial working hypotheses, but it has also generated a host of new hypotheses and important questions that still remain to be tested and answered.

Using neural networks to perform student model diagnosis represents an interesting and viable alternative to the methods currently employed in ITSs. The experiments conducted using the FLIS system have shown that unsupervised, competitive learning is capable of generating categories, or exemplars, of people based solely on their

choices and actions while using the tutor. Moreover, these categories do reflect a division along language competence lines. The next logical step is to experiment with the type of input data the network is allowed to review. Perhaps the inclusion of intelligence assessment data and/or attitudinal and psychological data might result in categories that are more finely adjusted than the ones generated here.

Although this report did not indicate similar levels of success with the back-propagation, supervised paradigm of network training, this should not be taken as a recommendation to terminate further research into this paradigm. As the report notes, the current structure of FLIS lessons does not easily lend itself to developing categories of language learners ahead of time. Improved lesson design, possibly employing a taxonomy along the lines suggested by Lund, will help overcome this problem. An alternative method to the construction of language learner profiles may simply be to let an unsupervised network construct the categories. These categories could then be taught to a back-propagation network using supervised training.

Supervised training, however, may be a more appropriate choice for an ITS in a domain that is more structured and more procedural than that of second language acquisition. Learning a second language is a much more complex task than, say, repairing an electrical panel. ITSs that offer

tutoring sessions with well defined, structured, and procedural tasks certainly should be explored as vehicles in which to test the supervised training of a neural network to perform student model diagnosis.

Finally, this report encourages additional explorations into integrating neural networks into ITSs. There does not appear to be any reasons why neural networks could not play a role in embodying the expert knowledge so crucial to ITSs. The elimination of large data structures for knowledge representation and the avoidance of cumbersome search techniques are only two of the reasons why this suggestion merits further attention.

REFERENCES

ACTFL Proficiency Guidelines (1988) Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

Anderson, J.R., C. Boyle, and G. Yost (1985) The geometry tutor. Proceedings of Ninth International Joint Conference on Artificial Intelligence. Los Angeles, CA: Morgan Kaufmann: 1-7.

Bernhardt, E. and C.J. James (1987) The teaching and testing of comprehension in foreign language learning. In Birckbichler (ed.) Proficiency, Policy, and Professionalism in Foreign Language Education. Lincolnwood, IL: National Textbook Company: 65-81.

Burton, R.B. (1982) DEBUGGY: diagnosis of errors in basic mathematical skills. In Sleeman and Brown (eds.) Intelligent Tutoring Systems. New York, NY: Academic Press: 157-183.

Carpenter, G. and S. Grossberg (1986) Neural dynamics of category learning and recognition: attention, memory consolidation, and amnesia. In Davis, Newburgh, and Wegman (eds.) Brain Structure, Learning and Memory. AAAS Symposium Series.

_____. (1987a) A massively parallel architecture for a self-organizing neural pattern recognition machine. Computer Vision, Graphics, and Image Processing. 37: 54-115.

_____. (1987b) ART 2: self-organization of stable category recognition codes for analog input patterns. Applied Optics. 26(23): 4919-4930.

Clancey, W. J. (1982) Tutoring rules for guiding a case method dialogue. In Sleeman and Brown (eds.) Intelligent Tutoring Systems. New York, NY: Academic Press: 201-225.

Grossberg, S. (ed.) (1988) Neural Networks and Natural Intelligence. Cambridge, MA: The MIT Press.

Hecht-Nielsen, R. (1990) Neurocomputing. Reading, MA: Addison-Wesley Publishing Company.

Kohonen, T. (1988) Self-Organization and Associative Memory. 2nd edition, Berlin: Springer-Verlag.

Krashen S. (1982) Principles and Practice in Second Language Acquisition. New York, NY: Pergamon Press.

Lund, R. (1990) A taxonomy for teaching second language listening. Foreign Language Annals. 23(2): 105-115.

Richards, J. (1983) Listening comprehension: approach, design, procedure. TESOL Quarterly. 17: 219-240.

Rumelhart, D., G. Hinton, and R. Williams (1986) Learning internal representations by error propagation. In Rumelhart and McClelland (eds.) Parallel Distributed Processing: Explorations in the Microstructure of Cognition. 1, Cambridge, MA: The MIT Press: 316-362.

Rumelhart, D. and J. McClelland (eds.) (1986) Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge, MA: The MIT Press.

VanLehn, K. (1988) Student modeling. In Polson and Richardson (eds.) Intelligent Tutoring Systems. Hillsdale, NJ: Lawrence Erlbaum Associates: 55-78.

Wasserman, P. (1989) Neural Computing Theory and Practice. New York, NY: Van Nostrand Reinhold.

AUTOMATIC RADIOFREQUENCY RADIATION MEASUREMENT SYSTEM

Prepared By : Dr. Stewart Maurer, Principal Investigator
Dr. Yoshi Saito
Mr. Gary Drazek

Date: December 26, 1991

NEW YORK INSTITUTE OF TECHNOLOGY
DEPARTMENT OF ELECTRICAL ENGINEERING
1855 BROADWAY
NEW YORK, NEW YORK, 10023

THE RESEARCH FOR THIS REPORT HAS BEEN FUNDED BY THE
US AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
UNDER
AFOSR-UES CONTRACT No. S-210-10MG-081

AUTOMATIC RADIOFREQUENCY RADIATION MEASUREMENT SYSTEM

ABSTRACT

This report describes a prototype automatic radio frequency radiation measurement system to be used for, but not limited to, RF hazard assessment. Its purpose is to be able to gather and store RF measurements in the field where the signal environment is not necessarily known. Stored programs which are user friendly enable the operator to collect data. The program minimizes operator intervention and thus makes for easier repeatability of measurements. This system can provide a more complete way to make RF radiation measurements, where the relative strengths of multiple emitters have frequencies that may extend over a wide band. Unlike the use of broadband isotropic RF radiation monitors, the contribution of individual emitters may be identified. The system is adaptable to a variety of RF environments, limited only by the receiving bandwidth of the spectrum analyzer and the calibration frequency range of the various antennas employed. The hardware and software described herein, constitute a prototype system.

ACKNOWLEDGEMENTS

This report and included software is the result of a team effort. Support from many sources besides those listed on the cover page have made this work possible. They are:

Col. Edwin Maher and Lt. Noel Montgomery of what was AF/OEHL at Brooks AFB, SanAntonio, Texas. Their invaluable help and encouragement during the incubation of this project, while I was at Brooks during the Summer of 1989 and during the project time;

George McPherson and the excellent support staff of Hewlett-Packard for providing invaluable technical and equipment support;

Vice-President Herbert Fox, Professors Billis and Heskiaoff at the New York Institute of Technology, for their support in easing the red tape on the administrative end;

Zennabelle Sewell for her support of office duties;

Ellen Solomon formerly of the New York Institute of Technology for her support in the presentation of the research proposal;

For the students of my Antenna Course who were the guinea pigs for the initial work with Spectrum Analyzer.

INTRODUCTION

The intent of this project is the development of a computer controlled system whereby RF radiation power density measurements across a user-definable band may be recorded, by means of a calibrated antenna and a spectrum analyzer. The frequencies of the RF sources or emitters may either be known or unknown. In particular the system is designed for the Hewlett-Packard series HP859* programmable spectrum analyzers which are equipped to accept programmable memory cards. These memory cards are 32 kilobytes in size and can be used to download programs into the analyzer's on-board computer or to serve as a recorder for data output from the analyzer. The input to the spectrum analyzer is provided by the use of calibrated antennas. In this project the system is illustrated by using a HP8591_A spectrum analyzer and an EMCO 6509 loop antenna.

The spectrum analyzer is controlled by an internally stored program that allows the operator to control the RF sweep bandwidth. Using this program, large frequency bands can be pre-screened by the operator until the areas of interest are determined. The peaks in a specific frequency band can then be recorded in each of three directions - x, y, and z. This can be repeated for many different measurement sites and all the recorded data can be stored on a memory card inserted into the analyzer. The data is then processed by a data-processing program (running on a PC connected to the spectrum analyzer). This program performs the conversions and calculations necessary, and allows the appropriate antenna factors to be entered in order to output the correct power densities.

This system has some advantages over broadband isotropic field strength meters. The operator can define the range over which measurements are taken and the frequency of the emitter is identified along with its peak value. The system also eliminates

measurement errors inherent in the non-linearity of elements in broadband meters and where multiple emitters are present. Added flexibility is possible by the use of different antennas. Various spectrum analyzer functions can provide for discrimination and/or rejection of certain low-level emitters in multi-emitter environments.

PHYSICAL SETUP

The physical setup of the field measurement system consists of the HEWLETT-PACKARD HP8591_A spectrum analyzer connected to an EMCO 6509 loop antenna by a coaxial cable. The loop antenna is mounted on a non-metallic tripod and can be rotated about its three major axes. An HP Think-Jet printer is connected to the HPIB port on the spectrum analyzer via an HPIB cable. (A personal computer (PC) is used to process the stored data but not needed for field measurements.)

The portable system (for taking measurements only) consists of the spectrum analyzer, memory cards (the size of credit cards) for storing the program and data, the printer, the antenna, and connecting cables. When all measurements have been taken, and the data has been stored on cards, the PC is then connected to the spectrum analyzer and the data is down-loaded to the PC for processing. The PC running the data-processing program is NOT needed for taking measurements. Data-processing may done at any convenient location. (See Figure 1 for the set up).

DESCRIPTION OF SWEEP PROGRAM

INTRODUCTION

** NOTE:

The filename for the sweep program is "COLLECT.BAS". This file and others are contained on the software disk provided. Display the "README" file on this disk for a short description of each file, and instructions on how to load "COLLECT.BAS" into the spectrum analyzer's internal memory.

The sweep program consists of three separate functions which are stored in the analyzer's internal memory (as well as on an external memory card). The functions are named NEWSTART, CONTINUE, AND SAVEDATA. They are accessed by pressing the following sequence on the face of the analyzer:

Press	PRESET	key
Press	MEAS/USER	key
Select	USER MENU(S)	

** The above steps must be performed any time one of the following functions are to be selected. **

At this point the functions may be accessed by pressing one of the three user-defined keys which appear with the following labels:

NEWSTART
CONTINUE
SAVEDATA

The purpose of these functions are as follows :

1) NEWSTART:

This runs the sweep program when beginning with a new memory card. This function is used at the FIRST measurement site only (for a particular memory card).

2) CONTINUE:

This runs the sweep program at successive measurement sites AFTER

the first site values have been taken with NEWSTART. This allows the spectrum analyzer to be turned off between measurement sites.

3) SAVEDATA:

This function is used to transfer measurement data, which till this time, has been stored in the internal analyzer memory to an external memory card. If the ARRAY FULL prompt (at the top of the screen) during the sweep program occurs, then this function must be used. This function is also used at the last measurement site when the operator has completed the collection of data which has not completely filled a 32K memory allotment. (The switch on the memory card should NOT be in the "safe" position, otherwise the card cannot be written on).

** To blank or format a memory card :

1) Insert memory card
2) Press CONFIG key
3) Select MORE from screen menu
4) Select CRD CONFIG from screen menu
5) Select BLANK or FORMAT from screen menu

Memory cards must be formatted when new or the battery in the card has to be replaced. A card which has just been formatted should be also a blank card. Going through the blank procedure is a recommended even after formatting.

While running the sweep program, various instructions and prompts will appear near the top of the screen. Numbers are entered using the numeric keys and the ENTER key located on the analyzer front panel. Frequencies are entered with the numeric keys and either the MHz or KHz keys. Amplitude functions are entered with the db key. VALUES MUST BE RE-ENTERED EVERY TIME THE PROMPT APPEARS ON THE SCREEN, even though the correct number appears in the data area.

PROGRAM DESCRIPTION

A "setup" condition begins when either the NEWSTART or CONTINUE functions are selected. This allows the operator to set various parameters on the spectrum analyzer and also to enter information for each measurement site. The available options (which appear sequentially one at a time in the upper left part of the analyzer screen) are:

1) USE PRINTER ?
2) VIDEO AVG OR MAX HOLD
3) SITE ID NUMBER (3 digits)
4) START FREQUENCY
5) STOP FREQUENCY
6) ATTENUATION
7) PEAK EXCURSION

Items 3 through 7 have a re-do option after the last item has been set. This allows the operator to experiment with attenuation and peak excursion for different frequency bands while looking for areas of high emitter density (a screening function). Once decided upon, these options are stored in the data array, and are interpreted with the measurement data by a PC resident data-processing program.

For each measurement site, the antenna is oriented in three different axial directions, and the program initiates a spectral sweep for each axis in turn: x, y, and z. Each sweep is operator controlled (via prompts on screen), and by which the parameters START FREQ, STOP FREQ, AND RESOLUTION BANDWIDTH can be entered. These values can be re-entered as often as necessary until a usable sweep is displayed. The RECORD option will then record the amplitudes and frequencies of all the peaks on the screen. The operator may choose to CONTINUE more sweeps on this axis until satisfied with the data taken or END this axis (and store the data) and then proceed to the next axis.

The y and z axis sweeps are similar to x. When a z axis sweep has been completed, the operator may change to a NEW measurement site (CHG_SITE - ENTER 999 and run CONTINUE function), or re-run all the x, y, and z sweeps for the SAME measurement site by choosing the CONTINUE=0 option. This option allows ALL the x, y, and z data for that site to be updated before proceeding to the next site.

The program continuously stores data in internal analyzer memory each time the NEWSTART or CONTINUE functions are selected. (The program automatically inserts delimiters or markers in the data file between measurement sites.) When the program has stored enough data to fill ONE memory card, the ARRAY FULL prompt is issued and the program terminates (regardless of which sweep is being run). At this point, the operator must insert a blank, formatted memory card into the spectrum analyzer and select the SAVEDATA function in order to save the data collected so far. The program may be resumed (will begin a NEW measurement site) by selecting the NEWSTART function (new memory card). See fig. 2

If an HP Think Jet type or equivalent printer (address 01) is connected to the HPIB port on the spectrum analyzer, and the printer option has been selected, the program will print the amplitudes and frequencies of the recorded peaks for each measurement site taken. If a printer is not available, then on the USE PRINTER ? prompt the NO option must be selected.

SAMPLE RESULTS AND USE OF THE DATA PROCESSING PROGRAM

In order to obtain some sample output results the sweep program was run for the FM, 90 -110 MHz range. This covers the New York FM broadcast band which appears on the spectrum analyzer as a number of peaks of different values. The measurements were taken at 1855 Broadway , New York , 12th floor. This environment serves only to illustrate the program output. (Antenna factors for calibration of the EMCO 6509, passive loop antenna, at the frequencies used were not as available and arbitrary values were chosen for illustration.)

The data-processing program "ANALYZE.BAS", is a separate program which resides on a PC having an HP BASIC LANGUAGE PROCESSOR CARD (HP82300C). This program and the PC are not needed for data collection; their purpose is to read in the stored data from the external memory card, compute the required values, and output the results to a printer.

To run the data-processing program:

1) Connect the HPIB connector on the PC to the spectrum analyzer (the printer can remain connected).
2) Copy software disk provided into the subdirectory in which HP BASIC resides.
3) Type BASIC and press enter.
4) Select "LOAD" function key and type ANALYZE.BAS and press enter.
5) Select "RUN" function key.

** NOTE : The spectrum analyzer will not run with the PC connected. **

The menu driven program for the PC with a HP language card begins by displaying the menu and requesting the operator to enter the number that corresponds to the desired selection. A short description of each menu selection follows:

1) READ DATA FROM ANALYZER:

This option reads measurement data from an external memory card into the analyzer and copies this data into the computer. This option is necessary only when NEW data is to be processed. An external memory card containing measurement data stored by the sweep program must be inserted for this option.
(** Entering A for Abort will return the operator to the menu)

2) ENTER ANTENNA FACTOR:

This option prompts the operator to enter the type of antenna used (loop/H-field or linear/E-field), and the appropriate antenna factor values. Up to 30 different values may be entered with the gain values entered with a leading minus sign. New values must be entered each time a DIFFERENT antenna is used.

3) ENTER SITE DESCRIPTION:

This option allows the operator to add a short description to each measurement site ID number. Using the FORWARD or BACKWARD choices will enable the operator to move sequentially through the measurement sites that have been stored.

4) ENTER GENERAL INFORMATION:

This option allows the operator to enter general information about the measurement survey such as : Date, Operator name, Weather conditions, and Location.

5) START ANALYSIS:

This option will begin the analysis assuming valid data have been read form the spectrum analyzer and antenna factor values have been entered. The following data is output in a tabular format for each emitter detected in a measurement site:

A) The measured frequency of this emitter in MHz

B) The measured powers of this emitter in dBm for the x, y, and z direction. (Assumes far field relations hold.)

C) The antenna factor in dB for this frequency

(** The program obtains intermediate antenna factor values from those values entered in menu option 2 by interpolation **)

D) The calculated total power of this emitter in dBm is computed according to the following formula : (H or E fields)

$$P_T = 10\log(10^{P_x/10} + 10^{P_y/10} + 10^{P_z/10}) \quad (dBm)$$

E) The calculated E-field in dBuV/m <u>or</u> the calculated H-field in dBuA/m depending upon the type of antenna selected. This is computed according to the following formula and does <u>not</u> assume far field relations hold:

E-field : $E_T = P_T + 107 + \text{Antenna Factor}$ (dBuV/m)

H-field : $H_T = P_T + 107 + \text{Antenna Factor}$ (dBuA/m)

F) The calculated power density of this emitter in uW/cm^2 is computed according to the following formula which assumes far field relations hold:

E-field : (uW/cm^2)

$$P_D = (10^{E_T/20})^2 * 10^{-10} / 377 = 10^{(E_T/10 - 10)} / 377$$

H-field : (uW/cm^2)

$$P_D = (10^{H_T/20})^2 * 10^{-10} * 377 = 10^{(H_T/10 - 10)} * 377$$

After each individual emitter is tabulated, the TOTAL POWER DENSITY for this site is sent to the output. A sample of the output from the data-processing program is shown in appendix B and program listings are shown in appendix C.

6) <u>VIEW REPORT</u>:

This option allows the operator to view the above tabulated data on the computer screen.

7) PRINT REPORT:

This option allows the operator to obtain a hardcopy of the above tabulated data from a printer connected to the computer.

0) QUIT:

This option suspends operation of the data processing program and allows the operator to leave the menu and return to the HP BASIC prompt. At this point if desired a return to the DOS prompt is be made by Pressing CTL and F10 simultaneously.

THE EMCO 6509 LOOP ANTENNA AND ESTIMATED ANTENNA FACTOR

To illustrate the prototype measurement system, an EMCO model 6509 8 inch loop antenna was used. This antenna has four frequency bands from 1 KHz to 30 MHz that are controlled by a switch on the antenna. This switch controls taps on a transformer connected to the loop. The sample data used for illustration of the program output existed in the range of 90 - 110 MHz which the antenna no difficulty receiving. For these frequencies, the antenna was switched to the 1-30 MHz position. In this range the transformer is not used and the loop is directly connected to the input/output terminal. The H-field antenna factor (AF) must be estimated since it was not calibrated above 30 MHz. The AF has been estimated based on an assumption of the antenna gain to be decreasing at a rate of -20 dB/decade of frequency from an arbitrary initial value.

The sample data frequencies and antenna factor values are for illustration purposes only. In practice, the operator will enter valid antenna factor values for the particular antenna being used.

CONCLUSIONS AND COMMENTS

This report and the software accompanying it, represent a prototype spectral power and field measurement system. The system may be used independently or in conjunction with broadband isotropic probe devices as produced by Narda and other manufacturers. The spectrum analysis based system shows graphically the relative importance of individual sources. Under certain circumstances isotropic probes can give false readings due to multiple loads or spurious responses due to high non-linear waveforms (such as might be found near VDT's or high voltage power lines).

For many applications the portability of using the spectrum analyzer as described in this report is achieved by use of a number of memory cards used to store measurement data. Since the size and complexity of programs that can be stored in the internal memory of the spectrum analyzer is limited, it would be useful in the future to design a system which incorporates a portable PC for both program and data storage. This system would of necessity be more bulky. A design of this nature would eliminate many limitations on the functional performance of the RF radiation measurement system. The PC is not included as part of the measurement system in order to keep the system as portable as possible. Further, if there is any need for security concerning classified operating frequencies the memory card system would appear to be more secure than one relying on a PC and a hard drive.

However, the system capability can be improved by replacing the PC with a laptop computer. The laptop would now become part of the portable system and would have to accommodate the HP BASIC language processor card. With new programming, this could greatly improve the capabilities of the system by allowing more complex programs to be run directly by the computer. Multiple programs and data files can be stored on computer discs. An added

advantage is the capability of having immediate output of field strengths and power densities at the measurement site.

In the present configuration, sweep program capabilities can be improved by making more spectrum analyzer functions available within the program. Measurement accuracy of peak amplitudes can also be improved. Antenna factors for various antennas could possibly be stored on external memory cards instead of being entered in the data-processing program.

To ascertain the accuracy of the existing system a detailed error needs to be preformed. These evaluations and improvements would be part of a future extension of the work initiated by this study.

GLOSSARY OF TERMS

ATTENUATION - used to set the spectrum analyzer's input attenuator from 0 to 60 dB in 10 dB increments. This attenuator is used to reduce the power level of the input signal to the analyzer. (Allows the measurement of large signals).

MAX HOLD - allows the trace on the spectrum analyzer screen to be updated only by the maximum signal level.

VIDEO AVERAGING - used to perform an averaging or smoothing function on signals and noise displayed on the analyzer screen.

RESOLUTION BANDWIDTH - used to limit the bandwidth of the input signal. Values can be set to decade multiples of 1 KHz and 3 KHz. (This function controls the spectrum analyzer's 3 dB IF bandwidth as opposed to the video bandwidth function which controls the post-detection filter.)

PEAK EXCURSION - used to set the minimum amplitude variation of signals that the spectrum analyzer's marker can identify as a peak. Only peaks that rise and fall more than this value above the threshold line (noise floor of the display) can be considered peaks. Two peaks that are close together will only be differentiated if the valley between them is at least peak excursion deep.

REFERENCES

1. 'Installation,Verification and Operation Manual HP 8591A/8593A Spectrum Analyzer', HP Part No, 5985-7092, Hewlett-Packard, Februrary 1990.

2. 'Installing and Using HP Basic in the MS-DOS Environment', Reorder No 82301-90013, Hewlett-Packard, June 1989.

3. 'Programming with HP Basic', Reorder Number 82031-90015, Hewlett-Packard, June 1989.

4. 'HP 8591A/8593A Spectrum Analzer Quick Reference Guide', Manual Part No. 5958-7090, Hewlett-Packard, Dec. 1989.

5. 'EMCO 6509 Passive Loop Antenna Instruction Manual', Electro-Mechanics Co., PO Box 1546, Austin, Texas 78767.

6. 'Radiofrequency Radiation Survey in the McFarland, California Area, EPA/520/6-89/022, U.S. Environmental Protection Agency, Sept. 1989

7. 'Base Level Management of RF Radiation Protection Program', Report No. 89 - 023RC0111DRA, U.S. AFOEHL, Brooks AF Base, San Antonio Texas, April 1989.

APPENDIX A

SOME THEORETICAL CONCEPTS

A.1 POWER DENSITY AND WAVE IMPEDANCE

A transmitting antenna is designed to radiate efficiently and control the directional propagation of this radiation. For a wave far (many wavelengths) from the source in free space, the electric and magnetic field directions are transverse to the direction of propagation and the power passing through a unit area is given by the product of the electric and magnetic field and is called the power density S in Watts per meter squared. Under these conditions, the ratio of the electric to the magnetic field is called the "free space wave impedance" and has a value of 377 ohms. The measured electric or magnetic fields are converted to the power density which would exist in a free space wave by assuming a wave impedance of 377 ohms. This is called the "equivalent far field power density" or simply "power density". The formulas for power density from the electric or magnetic field are:

$$S\ (W/m^2) = E\ (V/m)^2 \ /\ 377\ ohms$$

$$S\ (W/m^2) = H\ (A/m)^2 \ X\ 377\ ohms$$

For similar reasons, magnetic field measurements in amps/meter are often converted to the electric field which would exist in free space by multiplying the value in amps/meter by 377 ohms to give the "equivalent" electric field in volts/meter.

A.2 ANTENNA FACTORS

Receiving antennas are sensors that detect fields by allowing the field to induce current or voltage in a circuit. Loop receiving antennas generate a voltage proportional to the magnetic field and linear antennas generate a voltage proportional to the electric field. The constant of proportionality which allows the field to be calculated from the voltage is the "antenna factor". These antenna factors are a function of frequency and are

determined by calibration with the antenna connected to a 50 ohm impedance circuit. The antenna factors for the magnetic loop antennas are used to convert from voltage to "equivalent" electric field assuming a 377 ohm wave impedance.

A.3 SCALING

The spectrum analyzer output is normally in units of power in dBm which are defined by P (dBm) = 10 Log P (mW). Fields (electric and magnetic) are often specified in dB with respect to a uV/m or dBuV/m and are defined by E (dBuV/m) = 20 Log E (uV/m) with the magnetic field converted to "equivalent" electric field units. Antenna factor (AF) is generally specified as dB/m or simply dB and may be added to the voltage (V) in dBuV to give the field in dBuV/m. However, since the spectrum analyzer output is power in dBm this can be converted using the 50 ohm impedance to dBuV by adding 107 dBuV/mW.

APPENDIX B

FORMULAS USED IN POWER DENSITY CALCULATIONS

B.1 TOTAL POWER

$$P_x \text{ (dBm)} = 10*\text{Log } (P_x/10^{-3})$$

$$P_x \text{ (watts)} = 10^{-3}*[10^{(P_x(\text{dBm})/10)}]$$

$$P_T \text{ (watts)} = P_{total} = P_x + P_y + P_z$$

$$P_T \text{ (dBm)} = 10*\text{Log } (P_t/10^{-3})$$

B.2 E-Field or H-Field MEASUREMENTS

$$P_T \text{ (dBm)} = 10*\text{Log}(P_T(\text{watts})/10^{-3})$$
$$= 10*\text{Log}(V_T^2/(50*10^{-3})$$
$$= 20*\log V_T + 10*\text{Log}(10^3/50)$$
$$= 20*\text{Log}(10^{-6}*V_T/10^{-6}) + 10*\text{Log}(10^3/50)$$
$$= 20*\text{Log}(V_T/10^{-6})+10*\text{Log}(10^3/50)+20*\text{Log}(10^{-6})$$
$$= V_T \text{ (dBuVolts)} \quad -107$$

Hence

$$\boxed{V_T \text{ (dBuVolts)} = P_T \text{ (dBmW)} +107}$$

E-Field

$$E_T(\text{dBuVolts/m}) = V_T(\text{dBUVolts}) + AF(\text{Efield})$$

$$\boxed{E_T(\text{dBuVolts/m}) = PT(\text{dBmW}) + 107 + AF(\text{Efield})}$$

H-FIELD

$$\boxed{H_T(\text{dBUAmps/m}) = PT(\text{dBmW}) + 107 + AF(\text{Efield})}$$

B.3 POWER DENSITY

E-FIELD

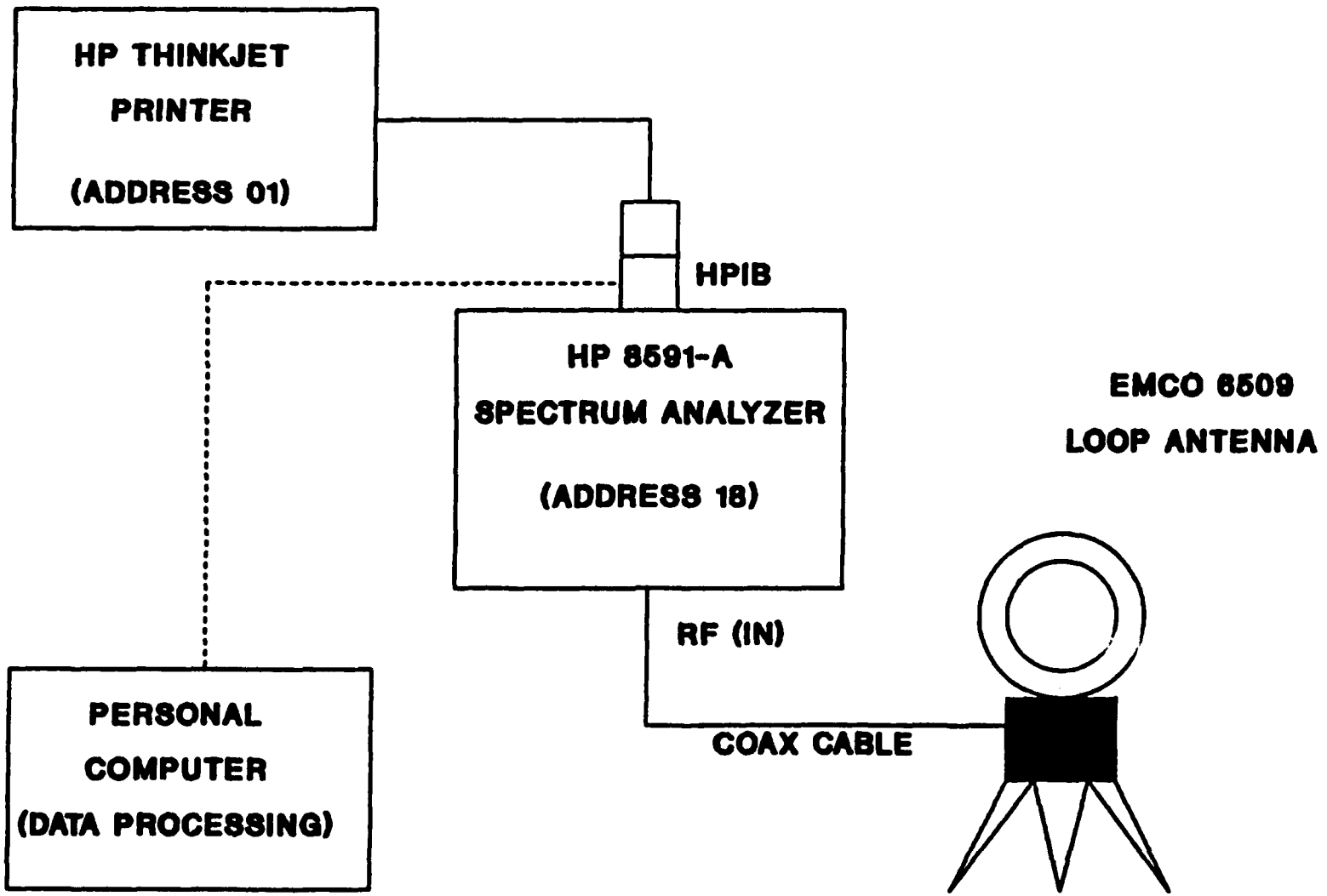
$$P_D(uW/CM^2) = [10^{(ET(dBuV/m)/20)}]^2 * 10^{-10}/377$$

H-FIELD

$$P_D(uW/CM^2) = [10^{(HT(dBuA/m)/20)}]^2 * 377 * 10^{-10}$$

FIGURES

FIG.1, PHYSICAL SETUP OF MEASUREMENT SYSTEM



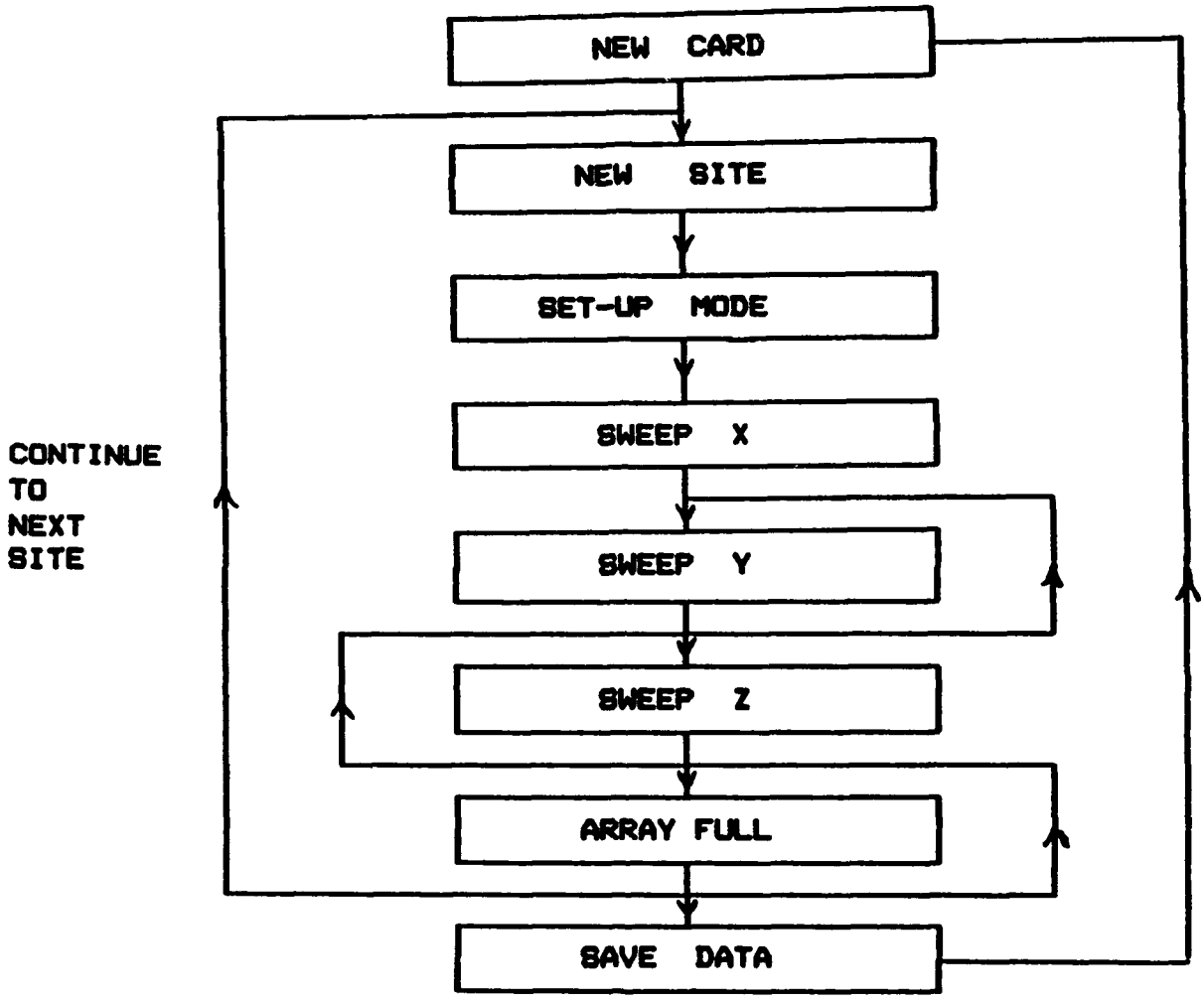


FIGURE 2

FUNTIONAL FLOWCHART

APPENDIX C

SAMPLE OUTPUT OF DATA PROCESSING PROGRAM

S U M M A R Y

Date of Measurement : 10/31/91

Operator : YOSHI SAITO

Weather Condition : PARTLY SUNNY TO CLOUDY TO RAIN

Location : 1855

Total Number of Sites : 2

Antenna Type : Loop / H-field

Number of Transmitters Detected

Site	Number
1	19
2	19

***** S I T E # 1 *****

SITE ID: 201 - 630 WEST END AVE
PEAK DETECT: Video Averaging ATTENUATION: 10 dB PEAK EXCURTION: 6 dB

Freq (MHz)	X-Amp (dBm)	Y-Amp (dBm)	Z-Amp (dBm)	AFact (dB)	T-Pwr (dBm)	H-Fld (dBuA/m)	PwrDsty (uW/cm^2)
92.35	-60.71	-59.30	-60.87	-12.35	-55.46	39.19	.0003126
93.20	-63.69	-63.53	-63.49	-13.20	-58.80	35.00	.0001193
94.80	-63.32	-63.63	-63.21	-14.80	-58.61	33.59	.0000861
95.60	-48.95	-99.00	-99.00	-15.60	-48.95	42.45	.0006628
95.65	-99.00	-47.90	-48.43	-15.65	-45.15	46.20	.0015728
97.25	-55.59	-53.74	-54.91	-17.25	-49.91	39.84	.0003636
98.05	-67.57	-67.16	-66.66	-18.05	-62.34	26.61	.0000173
98.85	-56.53	-56.37	-56.86	-18.85	-51.81	36.34	.0001623
99.65	-67.06	-66.04	-99.00	-19.65	-63.51	23.84	.0000091
99.70	-99.00	-99.00	-65.22	-19.70	-65.22	22.08	.0000061
100.50	-58.29	-99.00	-60.71	-20.50	-56.32	30.18	.0000393
100.55	-99.00	-60.36	-99.00	-20.55	-60.36	26.09	.0000153
101.30	-58.24	-58.58	-57.63	-21.30	-53.36	32.34	.0000646
104.50	-64.50	-64.64	-63.60	-24.50	-59.45	23.05	.0000076
105.30	-61.55	-61.45	-60.24	-25.30	-56.27	25.43	.0000132
106.10	-99.00	-99.00	-54.21	-26.10	-54.21	26.69	.0000176
106.15	-54.39	-54.64	-99.00	-26.15	-51.50	29.35	.0000324
106.95	-58.46	-57.75	-58.72	-26.95	-53.52	26.53	.0000170
107.75	-59.25	-59.29	-58.51	-27.75	-54.23	25.02	.0000120

TOTAL POWER DENSITY = .0035309

***** S I T E # 2 *****

SITE ID: 202 - NEW YORK INSTITUTE OF TECHNOLOGY - OW
PEAK DETECT: Video Averaging ATTENUATION: 10 dB PEAK EXCURTION: 6 dB

Freq (MHz)	X-Amp (dBm)	Y-Amp (dBm)	Z-Amp (dBm)	AFact (dB)	T-Pwr (dBm)	H-Fld (dBuA/m)	PwrDsty (uW/cm^2)
92.35	-60.06	-99.00	-61.57	-12.35	-57.74	36.91	.0001851
92.40	-99.00	-61.26	-99.00	-12.40	-61.26	33.34	.0000814
93.20	-63.29	-64.39	-64.34	-13.20	-59.21	34.59	.0001086
94.80	-62.95	-63.25	-63.63	-14.80	-58.50	33.70	.0000884
95.65	-48.67	-49.51	-49.30	-15.65	-44.37	46.98	.0018791
97.25	-55.81	-56.30	-56.57	-17.25	-51.44	38.31	.0002552
98.05	-68.68	-66.83	-68.55	-18.05	-63.16	25.79	.0000143
98.85	-57.57	-57.78	-56.72	-18.85	-52.56	35.59	.0001365
99.65	-65.49	-64.63	-64.95	-19.65	-60.24	27.11	.0000194
100.50	-59.81	-62.92	-61.71	-20.50	-56.52	29.98	.0000375
101.30	-56.51	-56.21	-55.98	-21.30	-51.46	34.24	.0001002
102.10	-99.00	-68.69	-99.00	-22.10	-68.68	16.22	.0000016
102.90	-66.88	-65.82	-66.64	-22.90	-61.65	22.45	.0000066
104.50	-62.41	-67.90	-68.29	-24.50	-60.53	21.97	.0000059
105.30	-58.67	-58.26	-58.64	-25.30	-53.75	27.95	.0000235
106.10	-54.59	-55.87	-99.00	-26.10	-52.17	28.73	.0000281
106.15	-99.00	-99.00	-56.92	-26.15	-56.92	23.93	.0000093
106.95	-58.51	-60.25	-59.72	-26.95	-54.66	25.39	.0000130
107.75	-59.50	-62.26	-62.[illegible]	-27.75	-56.49	22.76	.0000071

TOTAL POWER DENSITY = .0030011

APPENDIX D

PROGRAM LISTINGS

PROGRAM: ANALYSE.BAS

MENU DRIVEN DATA ANALYZER AND REPORT GENERATOR

PROGRAM:COLLECT.BAS

DATA COLLECTION PROGRAM TO BE LOADED IN SPECTRUM ANALYZER

PROGRAM: ANALYSE.BAS

MENU DRIVEN DATA ANALYZER AND REPORT GENERATOR