UNITED STATES AIR FORCE

SUMMER RESEARCH PROGRAM -- 1995

SUMMER FACULTY RESEARCH PROGRAM FINAL REPORTS

VOLUME 4

ROME LABORATORY

RESEARCH & DEVELOPMENT LABORATORIES

5800 Uplander Way

Culver City, CA 90230-6608

Program Director, RDL
Gary Moore

Program Manager, AFOSR
Major David Hart

Program Manager, RDL
Scott Licoscos

Program Administrator, RDL
Gwendolyn Smith

Program Administrator, RDL
Johnetta Thompson

Submitted to:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Bolling Air Force Base

Washington, D.C.

December 1995

REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-98-

0819

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. F
	December, 1995	Final

4. TITLE AND SUBTITLE
USAF Summer Research Program - 1995 Summer Faculty Research Program Final Reports, Volume 4, Rome Laboratory

5. FUNDING NUMBERS

6. AUTHORS
Gary Moore

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Research and Development Labs, Culver City, CA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NI
4040 Fairfax Dr, Suite 500
Arlington, VA 22203-1613

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Contract Number: F49620-93-C-0063

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The United States Air Force Summer Faculty Research Program (USAF- SFRP) is designed to introduce university, college, and technical institute faculty members to Air Force research. This is accomplished by the faculty members being selected on a nationally advertised competitive basis during the summer intersession period to perform research at Air Force Research Laboratory Technical Directorates and Air Force Air Logistics Centers. Each participant provided a report of their research, and these reports are consolidated into this annual report.

14. SUBJECT TERMS
AIR FORCE RESEARCH, AIR FORCE, ENGINEERING, LABORATORIES, REPORTS, SUMMER, UNIVERSITIES

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
Unclassified	Unclassified	Unclassified	UL

PREFACE

Reports in this volume are numbered consecutively beginning with number 1. Each report is paginated with the report number followed by consecutive page numbers, e.g., 1-1, 1-2, 1-3; 2-1, 2-2, 2-3.

This document is one of a set of 16 volumes describing the 1995 AFOSR Summer Research Program. The following volumes comprise the set:

VOLUME	TITLE
1	Program Management Report
	Summer Faculty Research Program (SFRP) Reports
2A & 2B	Armstrong Laboratory
3A & 3B	Phillips Laboratory
4	Rome Laboratory
5A, 5B, & 5C	Wright Laboratory
6A & 6B	Arnold Engineering Development Center, Wilford Hall Medical Center and Air Logistics Centers
	Graduate Student Research Program (GSRP) Reports
7A & 7B	Armstrong Laboratory
8	Phillips Laboratory
9	Rome Laboratory
10A & 10B	Wright Laboratory
11	Arnold Engineering Development Center, Wilford Hall Medical Center and Air Logistics Centers
	High School Apprenticeship Program (HSAP) Reports
12A & 12B	Armstrong Laboratory
13	Phillips Laboratory
14	Rome Laboratory
15A&15B	Wright Laboratory
16	Arnold Engineering Development Center

ATM DS-3 EXPERIMENTS VIA THE ADVANCED COMMUNICATION TECHNOLOGY SATELLITE

Valentine Aalo and Okechukwu Ugweje
Electrical Engineering Department

Florida Atlantic University
500 North West 20th Street
Boca Raton, FL 33431

Mostafa Chinichian
Electrical Engineering Department

California Polytechnic State University
San Luis Obispo, CA 93407

Final Report for:
Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington DC

and

US Air Force Rome Laboratory

September 1995

ATM DS-3 EXPERIMENTS VIA THE ADVANCED COMMUNICATION TECHNOLOGY SATELLITE

Valentine Aalo
Electrical Engineering Department
Florida Atlantic University

Okechukwu Ugweje
Electrical Engineering Department
Florida Atlantic University

Mostafa Chinichian
Electrical Engineering Department
California Polytechnic State University

Abstract

A series of experiments were conducted via the ACTS DS-3 link carrying ATM traffic in a PLCP frame format. These experiments were performed between Rome Laboratory, Griffiss Air Force Base, New York and Communication Research Centre, Nepean, Canada. These tests were conducted to test the viability of ATM DS-3 bearers via the satellite at a performance level comparable to that of terrestrial fiber optic link or microwave link. In these tests, both single and multiple channels configurations of ATM satellite bearers were tested. The single channel experiments were configure to fully load the channel at 96,000 Cells per second, while the multiple channel experiments were used to assess the ability of ATM protocol to carry information from different sources at different rates. Our main objective of characterizing the DS-3 satellite channel for the transmission of ATM signals was achieved, and the result obtained were presented and analyzed.

ATM DS-3 EXPERIMENTS VIA THE ADVANCED COMMUNICATION TECHNOLOGY SATELLITE

Valentine Aalo
Okechukwu Ugweje
Mostafa Chinichian

1.0 INTRODUCTION

The popularity of Asynchronous Transfer Mode (ATM) has risen to a new height, and all efforts are being made to experimentally verify its potential and usefulness. It is indicated in [1], that in the next decade, most telecommunication traffic will be carried by ATM technology. ATM is now regarded as the communication protocol of the future and will play a significant role in the "information superhighway". The usefulness and adaptability of ATM to current technology is also the subject of lots of research in both Government and industry. ATM technology and its applications are still being developed and tested. Global deployment of ATM technology may take some time as most economical and effective means of implementing this evolving technology is still being studied.

Although ATM was originally intended for fiber optic links transmitting high speed data, it can be used for other links such as the satellite link, which is the subject of this experiment. Satellites will play a vital role in the provision of ATM-based services such as voice, data, video and imaging, on demand, at very high data rates and at any location. Many performance issues regarding the transfer of ATM information bearers via the satellite channel remain unresolved, especially at Ka-band where the performance of satellite communication systems are severely impaired by atmospheric propagation effects, especially rain attenuation. We are interested in studying by means of actual experiment those characteristics of ATM protocol that is suitable and reliable for the transmission of information via satellite Ka-band at DS-3 rates.

The aim of this test was to demonstrate the practicality of transmitting and receiving ATM Cells using the Advanced Communication Technology Satellite (ACTS). This report documents, in a summarized form, the experiment of transmitting ATM signal via the ACTS conducted in the Summer of 1995 at Rome Laboratory (RL), Griffiss Air Force Base (AFB), New York. These experiments are part of the global grid Technical Technology Cooperation Program (TTCP) network, in which RL will communicate via the satellite with the Canadians at the Communication Research Center (CRC) [2]. These experiments were conducted via the satellite between the

CRC, Nepean, Canada and RL, Griffiss Air Force Base (GAFB), New York. This experiment is part of the on-going effort to link the assets of the Air Force at RL, the Army at the Communication Electronic Command (CECOM), and the Navy at the Naval Research and Development (NRaD) [3]. This report is based on the segment of the experiments that were conducted from 27 June 1995 through 24 August 1995. Several tests were performed via a satellite DS-3 link carrying ATM traffic in a Physical Layer Convergence Protocol (PLCP) frame format. In these tests, both single and multiple channels configuration of ATM satellite bearers were tested. The single channel configuration fully loads the channel at 96,000 Cells/s, while the multiple channel configuration was used to assess the ability of ATM protocol to carry information from different sources at different transmission rates.

2 .0 SETUP AND CONFIGURATIONS

The first stage of this experiment involved the characterization of the satellite channel for reliable transmission of ATM signals at DS-3 rates via the ACTS. In the second stage, ATM Cells bearers were transmitted and received via the ACTS satellite at DS-3 rates. In the following sections the resources and experimental procedure used in these tests, particularly at the RL test site, are briefly described.

2.1 Test Sites and Facilities

As indicated earlier, as part of the global grid network, RL, New York, will communicate via the satellite with CRC, Canada. The basic configuration used to connect CRC and RL via the ACTS is shown in Figure 1. This configuration simply shows the terminal equipments and the ATM equipments in block diagram form. Details are not shown. This configuration includes the ATM generator/analyzer equipment and a PC workstation fitted with ATM adapter card. This setup is briefly described below.

2.1.1 Rome Laboratory, Griffiss AFB, New York

At RL, the satellite transmission facilities is located outside in two huts south of building 3. The high data rate antenna is 1.8m with gain of 60.40 dBi for Uplink and 48.53 dBi for Downlink. The antenna is positioned at

Latitude: 43° 13', 13.18'' N

Longitude: 75° 24', 34.00" W

Azimuth 214.1°; Elevation 34.68°; Polarization Tilt 24.43°

In addition to the 1.8m high data rate antenna, other terminal equipments at the RL test site consist of the EF

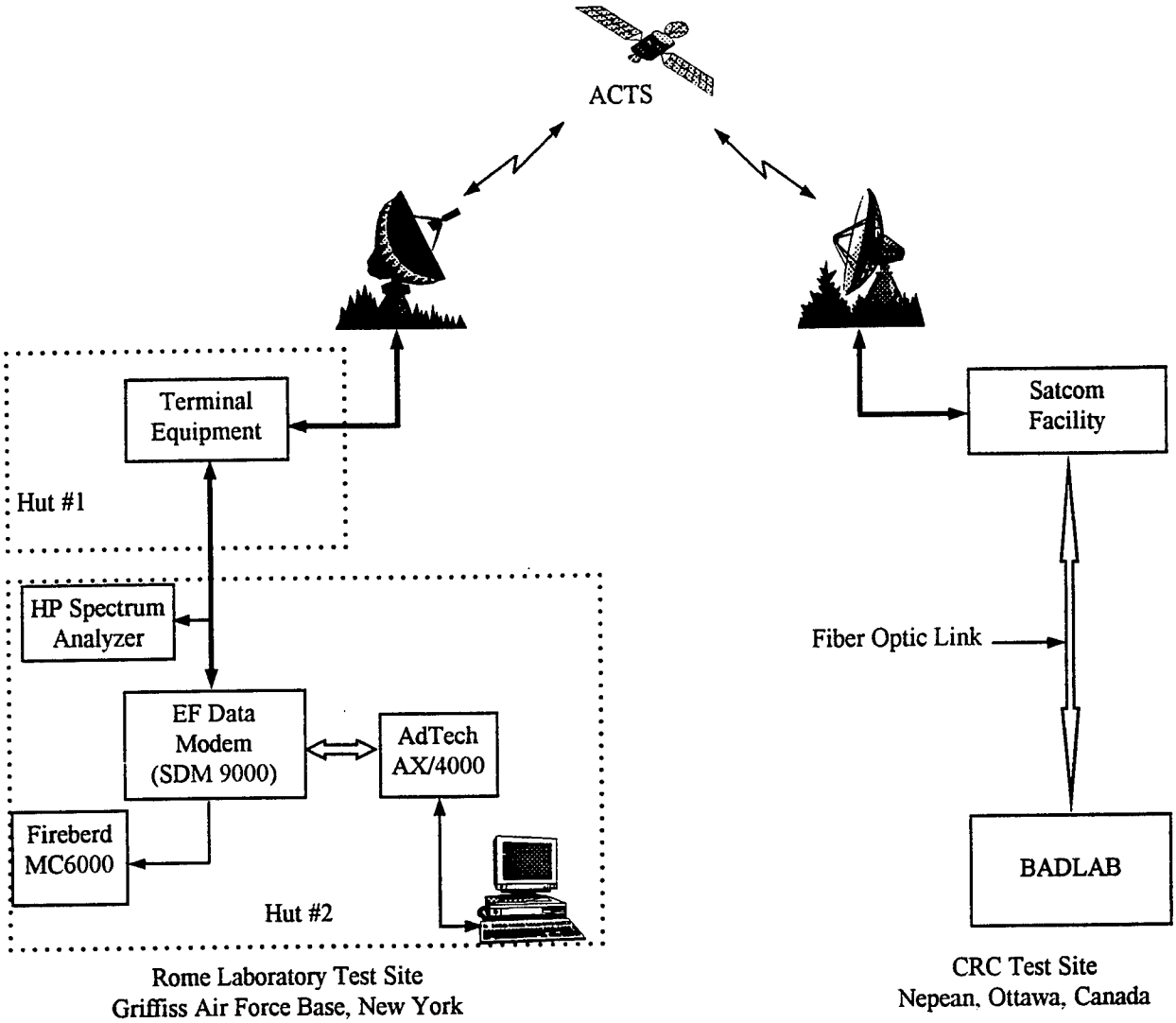


Figure 1: ATM DS-3 Test Configuration

Data modem (model SDM-9000), Ka-band traveling-wave-tube amplifier (TWTA), test loop translator (TLT), converters and amplifiers. The EF Data modem is a high performance, full-duplex, digital-vector modem with data rate capabilities of 6 Mbps to 51.84 Mbps. The modem has built-in scrambler/descrambler, differential encoding/decoding, multi-rate forward error correction capabilities (convolutional encoder and Viterbi decoder) and can be configured to add overhead/framing to the data. For example, a Reed-Solomon (RS) encoding and decoding is provided to work with the built-in Viterbi decoder, in conjunction with additional framing and interleaving, resulting in improved overall performance.

At the RL test site, the ATM generator/analyzer, AX/4000, manufacture by AdTech, Inc. is used as the ATM

test-bed equipment. This equipment provides complete signally, switching, generation and analysis of ATM traffic. It is modular in nature, allowing for custom configuration for a variety of ATM test applications. This unit consists of the mainframe, the ATM generation/analyzer modules and is connected to a window-based workstation. While a variety of port interfaces are possible (for example, SONET OC-3c, E3, TAXI, and DS-3), the interface provided with the AX/4000 is the DS-3 interface using electrical format and BNC connection to the EF Data modem, and IEEE 804.2 interface to the PC workstation. Using the associated AX/4000 Microsoft window-based software, one can select different configurations of single or multiple ports (each port has up to 16 channels or substreams) with capabilities of up to 140 Mbps. However, in this experiment, we are only interested in the DS-3 speed of 45 Mbps, and the AX/4000 is capable of generating a full DS-3 Cell stream.

The AX/4000 ATM equipment is installed in the adjacent hut, about 25 feet from the first hut as shown in Figure 1. The Fireberd Communication Analyzer (model MC6000), HP Spectrum Analyzers (HP 8568A & HP 8563), and the EF Data modem are co-located with the ATM equipment. Note that all connections at the RL test site uses electrical format.

2.1.2 Communication Research Centre, Nepean, Canada

The ATM test-bed at the CRC test site is located in the Broadband Application and Demonstration Lab (BADLAB) in building 2D [4]. The satellite communication (satcom) facility is located in building 46. The satcom facility and the ATM facility are linked by fiber optic line, over off-net extension via the satellite ATM links.

The antenna at CRC test site is 4.2m with gain of 52.34 dBi for Uplink and 56.59 dBi for Downlink. The position of the antenna is at

Latitude: 45° 21', 9.95'' N

Longitude: 75° 54', 9.59" W

Azimuth 220.85°; Elevation 29.16°; Polarization Tilt 48°

The terminal facilities (Modem, TLT, TWTA) at CRC test site is the same as RL test site. There may be minor variations in terms of the location and connectivity of the equipments but the functionality of all the equipments at the two test sites remain the same. It is perhaps important to note that most of the terminal equipments used in this experiment at both test sites were provided by Canada's CRC in corporation with Air Force's Rome Laboratory. Their contribution include the frequency plan, the redesign of the converters and amplifiers to be compatible with ACTS Ka-band frequencies at DS-3 rates, and the actual test procedure used in these experiments.

2.2 Satellite Facilities

The Advanced Communication Technology Satellite (ACTS) was used for this experiment. Access to the ACTS was provided by NASA Lewis, Cleveland, Ohio. The ACTS operations center at NASA Lewis provided all access for the East Scan 4E spot beam used by RL test site as well as the Steerable spot beam used by CRC.

During these tests, RL transmitted horizontally with polarization tilt of 24.4° clockwise into the East Scan 4E beam at 29.125 GHz and received in the orthogonal plane at 19.505 GHz. On the other hand, CRC, transmitted vertically with polarization tilt of 22.2° clockwise into the steerable antenna at 29.225 GHz and received orthogonally at 19.405 GHz [5]. Because of the enormous loss in orthogonal polarization as seen from the satellite EIRP and G/T coverage, self loopbacks are not practical. It should be pointed out, however, that the ACTS operation center at NASA Lewis can provide satellite loopbacks for both the steerable and East Scan 4E spot beams. This implies that NASA Lewis can serve as a loopback node for either RL test site or CRC test site.

3.0 EXPERIMENTAL PROCEDURE

3.1 Channel Characterization

The sequence of tests involving the transmission of ATM Cells bearers via the ACTS started on 27 June 1995. The first set of tests were aimed at characterizing the channel between RL, and CRC, and to familiarize RL personnel with the ground terminal equipment. The objective of the channel characterization experiments is to evaluate the performance of the EF Data modem with the Fireberd tester at DS-3 rates over the ACTS. For the channel characterization experiments, the setup is as shown in Figure 1 without the AdTech AX/4000 equipment connected.

Two types of modulation schemes, namely, the QPSK and 8PSK were emphasized in this experiment. The characterization tests as well as the ATM transmission tests were conducted using both QPSK and 8PSK with and without RS encoding for different modem power levels. During this period of characterization, a number of transmissions involving continuous wave (CW) at the two modulation schemes were performed. This was done to evaluate the EF Data modem and the Fireberd DS-3 test equipment. The channel characterization and modem evaluation are necessary to ensure that at DS-3 rates and in the Ka-band the modem specifications for the different modulations can be verified. During the CW transmissions, the transmit power levels at both ends (CRC and RL) were varied (with the TX power meter reading recorded) while the carrier power (C) and the noise power (N_o) are read from the Spectrum Analyzer at opposite ends.

The bit error rate (BER) is measured using the Fireberd Analyzer. It is also possible to measure the BER using the EF Data modem but we did not use the measurement from the modem. To evaluate the modem and the Fireberd BER tester, the QPSK with RS, QPSK without RS, 8PSK with RS and 8PSK without RS were used as the modulation schemes. QPSK modulation is at the code rate of 3/4 while 8PSK is at the code rate of 2/3. In each case, both the received BER and E_b/N_o were measured for each setting of the transmit power. All transmissions, by default, included the modem's built-in convolutional and Viterbi decoding at the indicated code rates.

At each transmit modem power level, the transmit power meter reading were recorded, as well as the received C and N_o at the other end. From these measurements, the E_b/N_o is given by $E_b/N_o = C/N_o - 10 \log(44.73X10^6)$, where C and N_o are measured with the Spectrum Analyzer

3.2 ATM Transmission Experiments

A series of experiments were conducted via the ACTS DS-3 link carrying ATM traffic in a PLCP frame format. The main aim of the ATM experiments is to characterize the DS-3 satellite link for the transmission of ATM signals. Specifically, some ATM QoS parameters were measured using different levels of modulation (QPSK, 8PSK) with forward error correction and V.35 data scrambling. For the ATM experiments, the setup is as shown in Figure 1, with the AdTech equipment connected. As indicated earlier, the AdTech equipment , AX/4000, is used as the ATM test-bed. The desired configuration is set by means of the window-based software accompanying the AX/4000. There are two types of transmission configurations (single and multiple channels) used in these tests. In the single channel experiment, one port and one substream was selected and configured with the following parameters:

Substream 1

- AAL Type: Test Cells
  PRBS Type: $2^9 - 1$
- GFC Range: 0h - fh
  VPI Range: 00h - ffh
  VCI Range: 0010h
  PT Range: 0 - 3
  CLP Range: 0 - 1
- Cell Rate: 96,000 cells/s

The single channel payload data stream was set to generate 96,000 cells per second. This is the maximum cell generation capacity of the AX/4000. All other ports and substreams (substream 2 to 16) are disabled.

For the multiple channel case, one port and six substreams (substream 1 to 6) are selected and configured as shown in Table 1. Substreams 7 to 16 are disabled. Table 1 shows that substreams 1 & 2, substreams 3 & 4,

Table 1: Multiple Channel Configuration
PORT 1 SETUP

Substream 1	Substream 2	Substream 3	Substream 4	Substream 5	Substream 6
AAL Type: Test Cells PRBS Type: 2^9-1	AAL Type: Test Cells PRBS Type: 2^9-1	AAL Type: Test Cells PRBS Type: 2^9-1	AAL Type: Test Cells PRBS Type: 2^9-1	AAL Type: Test Cells PRBS Type: 2^9-1	AAL Type: Test Cells PRBS Type: 2^9-1
GFC Range: 0h-fh VPI Range: 00h-ffh VCI Range: 0001h PT Range: 0 - 7 CLP Range: 0 - 1	GFC Range: 0h-fh VPI Range: 00h-ffh VCI Range: 0002h PT Range: 0 - 7 CLP Range: 0 - 1	GFC Range: 0h-fh VPI Range: 00h-ffh VCI Range: 0003h PT Range: 0 - 7 CLP Range: 0 - 1	GFC Range: 0h-fh VPI Range: 00h-ffh VCI Range: 0004h PT Range: 0 - 7 CLP Range: 0 - 1	GFC Range: 0h-fh VPI Range: 00h-ffh VCI Range: 0005h PT Range: 0 - 7 CLP Range: 0 - 1	GFC Range: 0h-fh VPI Range: 00h-ffh VCI Range: 0006h PT Range: 0 - 7 CLP Range: 0 - 1
Cell Rate: 150 Cells/s	Cell Rate: 150 Cells/s	Cell Rate: 604 Cells/s	Cell Rate: 604 Cells/s	Cell Rate: 3640 Cells/s	Cell Rate: 3640 Cells/s

substreams 5 & 6, have ATM Cell generation rates of 150 Cells/s, 604 Cells/s, and 3640 Cells/s respectively. The combined Cell Rate should be less or equal to 96,000 Cells/s. In our own case, for multiple channel configuration, the combined Cell Rate is 8788 Cells/s. In either the single substream or multiple substreams configuration, the PRBS errors are triggered by the Cell Sequence Error, Cell Payload Bit Error or the Loss of Payload Pattern Synchronization.

Using the AdTech AX/4000, it is possible to measure a lot of information for each test transmission. These are known as stream and substream statistics. A sample of the measured statistics using the AX/4000 analyzer module is shown in Table 2. This table shows a sample of the statistics for QPSK experiment conducted on 24 July 1995, and the 8PSK experiment conducted on 11 August 1995. Both experiments are for multiple channel (or substream) tests, with only the first substream shown. Similar statistics was obtained for the single channel tests.

4.0 ANALYSIS OF RESULT

The key issues regarding the reliable transmission of ATM signal via satellite is the subject of this test, especially, the error characteristic and propagation delay. This is because in satellite transmission at the Ka-band, the satellite link is susceptible to atmospheric effects, the most important of which is attenuation due to rain. To this effect, the weather condition at RL during each period of an experiment is recorded and is enclosed in the appendix. The reason for this is to observe any relationship between the collected data and the weather condition. We observed that on several conditions, the satellite link between RL and CRC was lost whenever there is a rain-

Table 2: ATM Transmission Stream and Substream Statistics	
QPSK Modulation	**8PSK Modulation**
Stream Statistics: Aggregate Count: 48343 Uncorrected Header Count: 915 Aggregate Cell Rate: 70,113 cells/s Aggregate Cell Transfer Capacity: 29.728 Mb/s Bandwidth Percentage: 73.03 %	Stream Statistics: Aggregate Count: 17114 Uncorrected Header Count: 3 Aggregate Cell Rate: 22,240 cells/s Aggregate Cell Transfer Capacity: 9.430 Mb/s Bandwidth Percentage: 23.17 %
Substream Statistics: Cell Count: 177695 cells Cell Rate: 149.8 cells/s Cell Transfer Capacity: 0.064 cells/s CLP=1 Cell Count: 41 cells CLP=1 Error Ratio: 2.31E-04 Cell Loss Count: 124 cells Cell Loss Ratio: 0.000697495 Misinsertion Cell Count: 40 cells Cell Misinsertion Rate: 0.0337286 cells/s Out-of-Sequence Count: 0 events Errored Cell Count: 85 cells Cell Error Ratio: 0.000478348 PRBS Bit Error Count: 895 bits PRBS Bit Error Rate: 1.65682e-05 errors/bit PRBS Sync Error Count: 0 resyncs	Substream Statistics: Cell Count: 556180 cells Cell Rate: 150.0 cells/s Cell Transfer Capacity: 0.064 cells/s CLP=1 Cell Count: 0 cells CLP=1 Error Ratio: 0.00E+00 Cell Loss Count: 9 cells Cell Loss Ratio: 1.61815e-05 Misinsertion Cell Count: 0 cells Cell Misinsertion Rate: 0 cells/s Out-of-Sequence Count: 0 events Errored Cell Count: 93 cells Cell Error Ratio: 0.000167212 PRBS Bit Error Count: 215 bits PRBS Bit Error Rate: 1.2716e-06 errors/bit PRBS Sync Error Count: 0 resyncs

storm in RL, CRC or in between. This observation is consistent with theoretical analysis which has shown that rain attenuation severely affects transmission at Ka-bands.

To properly assess the viability of ATM via satellite, the cumulative effect of the satellite link characteristics on the ATM parameters must be determined. Performance consideration for ATM are currently based on the assumption that transmission bit errors are randomly distributed. This may be true, to a high degree of accuracy for most terrestrial microwave and fiber optic based transmission systems. However, the validity of this assumption has not been verified for satellite ATM Cell bearers. The bursty nature of transmission errors via the satellite requires careful evaluation.

4.1 Channel Characterization

For satellite transmission channels, ATM performance parameters can only be quantified if the nature of the transmission bit errors is first characterized. The characterization phase of this experiment was aimed at achiev-

ing this objective.

The entire raw data collected during this experiment can be obtained from the satcom division, RL/C3BA, Rome Laboratory, New York. Some of these data is presented here for the purposes of discussion.

Most of the data presented here are collected from the RL test site. The other half collected from CRC can also be obtained from RL/C3BA or from CRC. For brevity, this other half is not presented here.

The graph of E_b/N_o versus BER for the two modulation schemes, with or without RS encoding is shown in Figures 2 and 3. In Figure 4 and 5, the performance of QPSK and 8PSK are compared. The effect of using RS encoding is shown in Figure 6, while the characteristics of both the forward (RL to CRC) and backward (CRC to RL) links are shown in Figure 7. The actual experiment represented by each graph is also indicated on the plot. The following observation can be made from these plots.

- QPSK provides better performance than 8PSK. This result could be attributed to the fact that the EF Data modem is more stable with respect to QPSK than the 8PSK. In other words, it is much simpler for the EF Data modem to distinguish between four phases as opposed to eight phases. Thus, signal detection is more reliable at QPSK modulation than at 8PSK modulation.
- RS encoding/decoding provides considerable system performance improvement. For example in QPSK at BER = 10^{-6}, there is a power saving of as much as 1.5 dB with the use of RS coding. The RS coding gain increases for lower BER values. This is consistent with expected analytical values.
- From Figure 7, it is observed that RL to CRC link and the reverse link have identical transmission link characteristics. The noticeable variation could be attributed to different antenna gain at both test sites, and to the different ACTS transponders (or spot beams) used for both test sites.

4.2 ATM Transmission

The bursty nature of the error statistics and the corresponding large propagation delay involved in a satellite link affect ATM transmission performance [6]. A number of ATM parameters have been identified as being very important in the assessment of the performance of an ATM network. Some of the most important QoS parameters include the transmission channel bit error ratio (BER), cell loss ratio (CLR), errored cell count (ECC) and the PRBS parameters. In the experiments, emphasis was placed on those parameters and statistics that can be measured by the AdTech ATM test equipment. Some of these parameter is discussed bellow.

- *Bit Error Rate (BER):* This is the rate at which the transmitted bits were changed in the ATM physical layer.

- *Cell Loss Ratio (CLR):* The ratio of the number of lost ATM cells sent by a user in specified time interval. Due to the random nature of the ATM Cells, the limited size of the ATM traffic and the limited size of the buffers, it is usually possible that a cell arriving at a switching node may be lost . Thus CLR are caused by buffer overflows and bit error in the cell header that can be detected but not be corrected.
- *Cell Misinsertion Ratio (CMR)*: Defined as the ratio of the cells delivered to a wrong destination to the total number of cells sent. It occurs as a result of an undetected error in the header that causes a change of the cell destination.

4.2.1 Single Channel

Table A1 in the appendix shows a sample of raw data collected for single channel ATM transmission. While many signal channel runs were made, a selected number of them are plotted in Figure 8 to 13 to illustrate the results obtained. The particular experiment and the date performed are indicated on the plots. From these Figures the following observations may be made:

- CLR and CER appear to have similar relations with channel BER. They both decrease as BER decreases.
- CLC and ECC are lowest at BER $=10^{-5}$ for the range of BER $=10^{-6}$ to 10^{-4}.
- For the received ATM alarms, it is observed that a) Framing rate appears to take its highest value at BER = 10^{-5}, and takes its minimum value in the neighborhood of 10^{-6}, and b) FEBE and P-Bit rate appear to increase steadily with increasing BER.
- For the PLCP alarms, the FEBE rate and P-Bit rate appear to have similar relations with channel BER.
- The forward and backward satellite channels between RL and CRC appear to be symmetric with identical measured ATM parameters.

4.2.2 Multiple Channel

A sample of the raw data collected for the multiple channel ATM transmission is shown as Table A2 in the appendix. As many as six ATM substreams of varying cell rates were combined in the multiple channel case with cell generating rate of 8788 Cells/s. As shown in Table A2, two types of measured parameters may be distinguished. One set has to do with he alarms that are generated for the composite stream, mainly, received frame alarm and received PLCP alarm. The other set of are the ATM Cell characteristics of each of the individual substreams as shown in Table 2. Figures 14 to 16 show the multichannel ATM parameters as function of channel BER. The following observations may be made regarding the Figures.

- Substreams with similar configurations exhibit similar characteristic with respect to the ATM QoS pa-

rameters.

- The received ATM alarms, FEBE, P-Bit, and Framing rates exhibit similar functional relationship with the channel BER.
- The effect of RS coding on the ATM transmission was not evident. However, it is conjectured that the use of RS coding will improve performance considerably.

6.0 CONCLUSION

In this experiment, we have successfully transmitted and received ATM Cells between Rome Laboratory, GAFB, New York and CRC, Canada. These ATM tests were performed at Ka-band frequencies over the ACTS. Two transponders (or beams) - the East Scan 4E spot beams and the Steerable spot beam, were used and the signal modulation schemes employed were the QPSK and 8PSK. It is observed in our experiment that the modulation scheme used affects the stability of the received signal. Also, during these experiments, we observed that rain attenuation has a dramatic effect in satellite transmissions at Ka-band frequencies. Ways of mitigating this effect should be the focus of more studies.

Although we were able to transmit and receive ATM Cells at Ka-band, it will be premature to draw a conclusion on the performance of ATM transmission over satellite as a result of one set of experiments. It is strongly recommended that more experiments be conducted after which a more meaningful conclusion on the performance of ATM signals over satellite can be reached. Some, if not all, of the test performed in this experiment should be repeated in other to verify the validity of these test results.

REFERENCES

[1] Arthur Miller, "From here to ATM - (Special Report)", IEEE Spectrum, June 1994.

[2] G. A. Bivens, "Satellite Networking Research in Scaleable Networking Technology, Rome Laboratory Technology Demonstration with NASA ACTS Satellite - Proposal", May 1994.

[3] V. Aalo and O. Ugweje, "A Program Plan for Transmitting High-Data -Rate ATM/SONET Signals over the ACTS, Final Report AFOSR, September 1994.

[4] J. Butterworth and G. Nourry, "TEST PLAN - for an investigation of THE TRANSPORT OF ATM-BASED TRAFFIC BY Ku-BAND SATELLITE BEARERS, Workshop on the "Global Grid" concept", Sponsored by TTCP STP-6, RL/CRC Internal Document, 16th January 1995.

[5] Corey Pike, Rome Lab/CRC internal document

[6] R.O. Onvural, Asynchronous Transfer Mode Networks Performance Issues, Boston; Artech House, 1994.

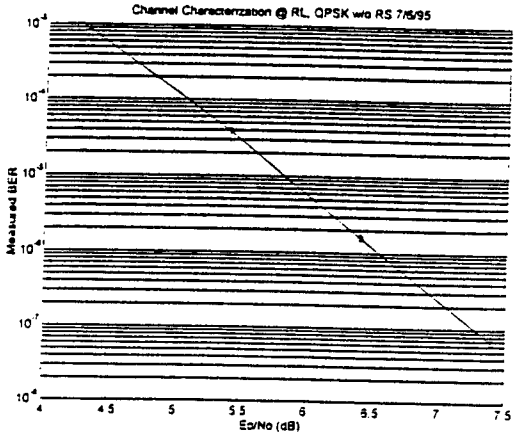


Figure 2. QPSK Channel Characterization without Reed Solomon Encoding

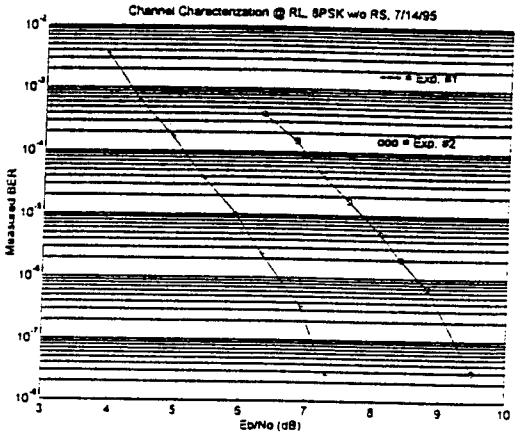


Figure 3. 8PSK Channel Characterization without Reed Solomon Encoding

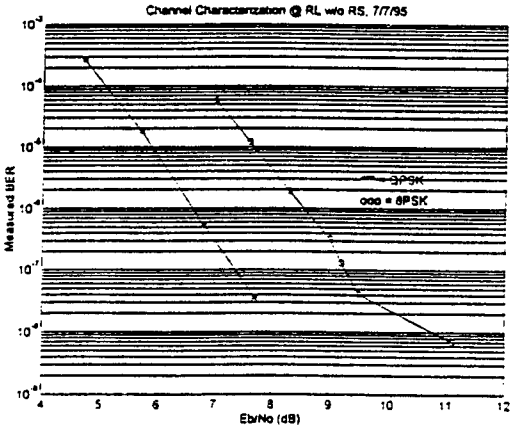


Figure 4. QPSK and 8PSK Channel Characterization without Reed Solomon Encoding

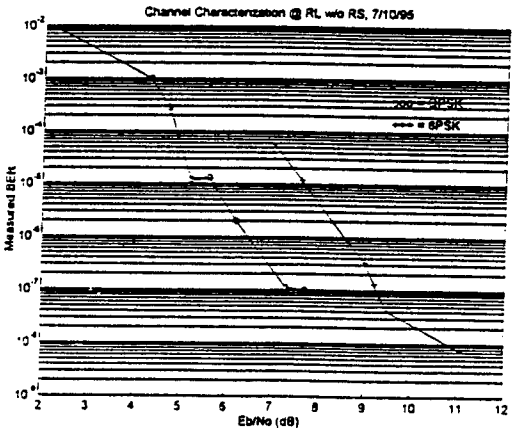


Figure 5. QPSK and 8PSK Channel Characterization without Reed Solomon Encoding

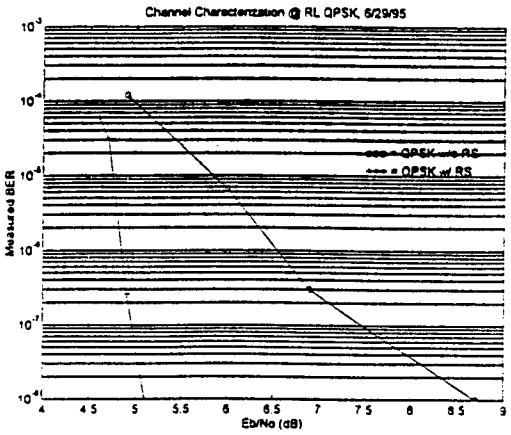


Figure 6. The Effect of Reed Solomon Encoding on Channel Characterization

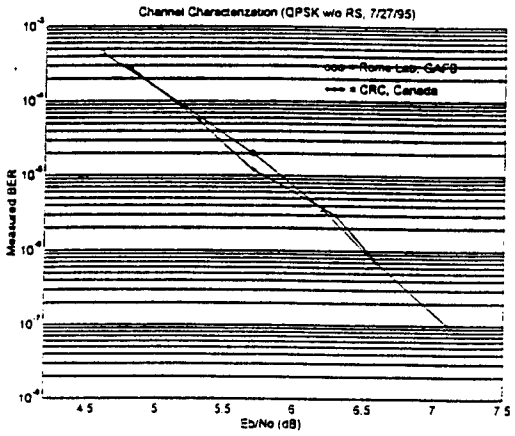


Figure 7. Forward and Backward Link Characteristics w/o Reed Solomon Encoding

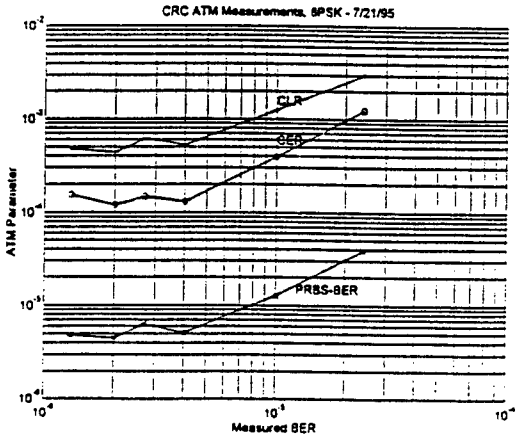


Figure 8. ATM Measurements - 8PSK

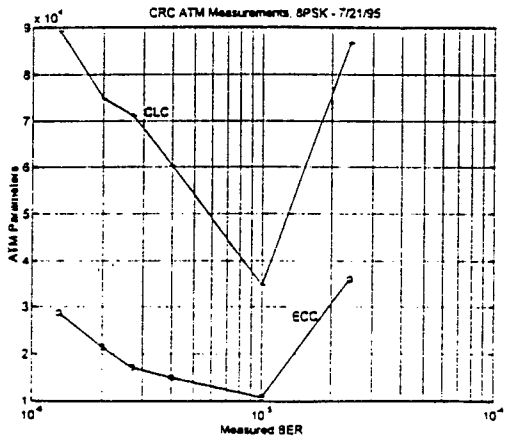


Figure 9. CLC & ECC Measurements - 8PSK

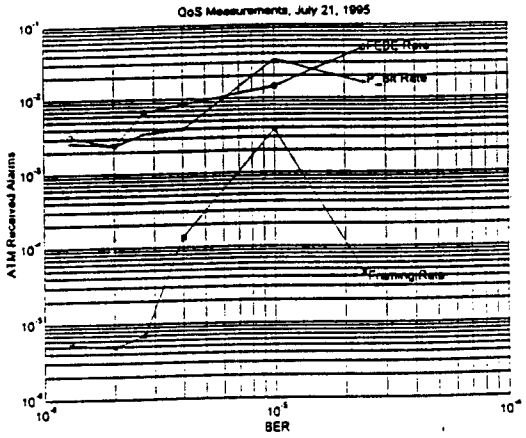


Figure 10. Rx Frame Alarms - 8PSK

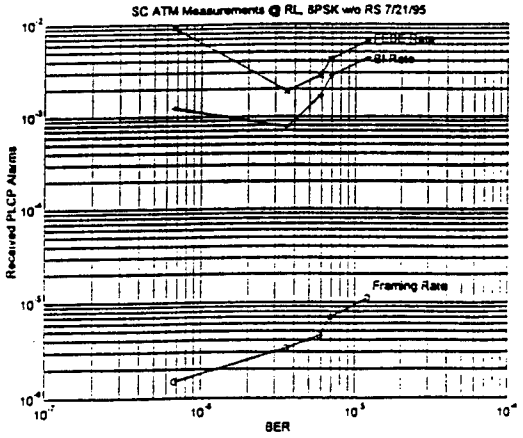


Figure 11. Rx Frame Alarms - QPSK

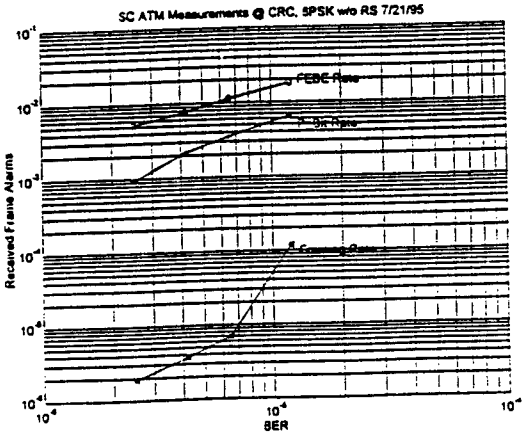


Figure 12. Single Channel w/o RS Measurements @ CRC - 8PSK

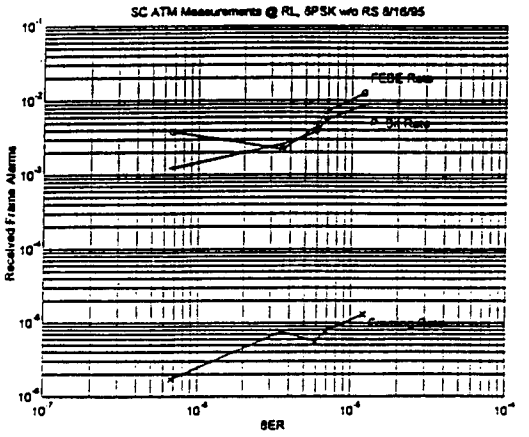


Figure 13. Single Channel w/o RS Measurements @ RL - 8PSK

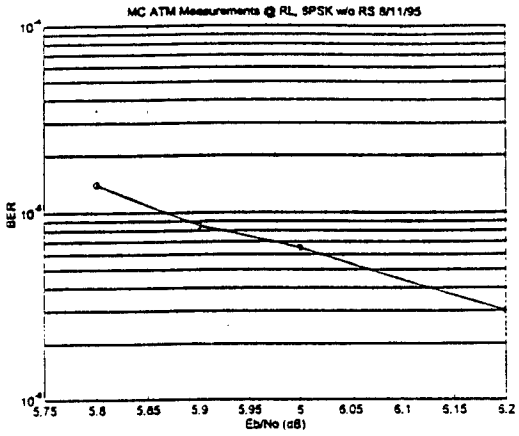


Figure 14. Multiple Channel w/o RS @ RL - 8PSK

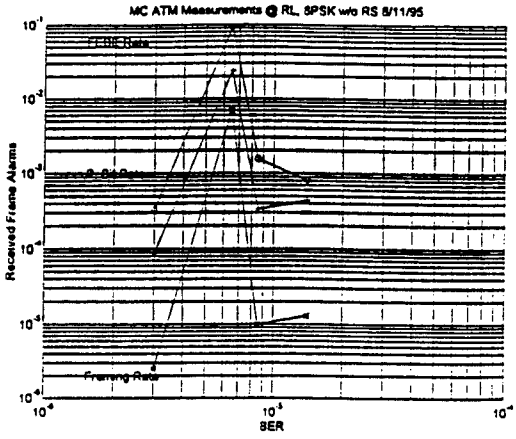


Figure 15. Multiple Channel Alarms w/o RS @ RL - 8PSK

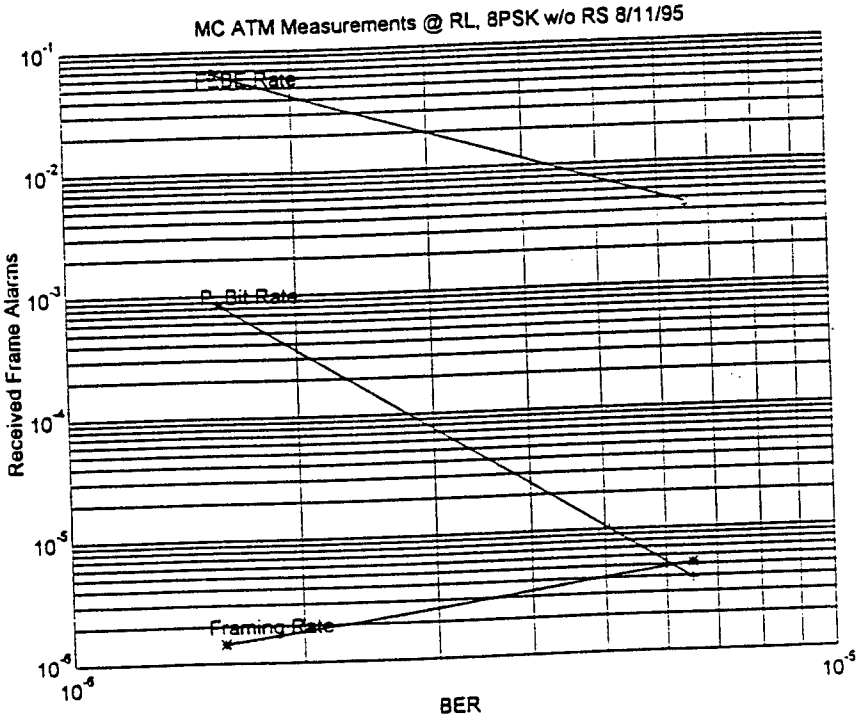


Figure 16. Multiple Channel Alarms w/o RS @ RL - 8PSK

Appendix

A: Weather Conditions and Schedule of Experiments

#	Date	Time	Weather	Comments
1.	27 June 95 (Tuesday)	1815-1930 EDT	Sunny	Terminal Configuration/Setup
2.	29 June 95 (Thursday)	1230-1630 EDT	Sunny	Modem Configuration
3.	5 July 95 (Wednesday)	1200-1600 EDT	Hazy	Modem Configuration
4.	6 July 95 (Thursday)	0800-1100 EDT	Hazy & Hot	C/No Measurement
5.	7 July 95 (Friday)	1300-1545 EDT	Hazy & Hot	BER, Eb/No measurement (QPSK)
6.	10 July 95 (Monday)	1045-1430 EDT	Cloudy	CW measurement
7.	14 July 95 (Friday)	1030-1400 EDT	Clear	8-PSK/ADTECH
8.	17 July 95 (Monday)	1045-1430 EDT	Sunny	ATM (QPSK, No RS)
9.	21 July 95 (Friday)	0730-1230 EDT	Cloudy	ADTECH problem
10.	24 July 95 (Monday)	1045-1430 EDT	Cloudy	ADTECH problem (QPSk, No RS)
11.	27 July 95 (Thursday)	1430-1700 EDT	Cloudy	ADTECH problem (QPSk, No RS)
12.	3 Aug. 95 (Thursday)	1100-1500 EDT	Overcast	Modem problem - no data
13.	4 Aug. 95 (Friday)	1100-1500 EDT	Cloudy	Modem problem - no data
14.	7 Aug. 95 (Monday)	1115-1530 EDT	Sunny	RL loop-back test-Modem problem
15.	11 Aug. 95 (Friday)	0900-1300 EDT	Cloudy	Modem Configuration
16.	14 Aug. 95 (Monday)	0800-1200 EDT	Sunny	ATM (8-PSK, no RS, 6 substreams)
17.	16 Aug. 95(Wednesday)	0800-1100 EDT	Sunny	ATM (8-PSK, no RS, 6 substreams)
18.	20 Aug. 95 (Sunday)	1500-1930 EDT	Sunny	RL loop-back (ATM, QPSK)
19.	22 Aug. 95 (Tuesday)	0915-1300 EDT	Sunny	ATM (QPSK, No RS, 6 Streams)
20.	25 Aug. 95 (Friday)	0800-1100 EDT	Sunny	ATM (QPSK, No RS, 6 Streams)

Table 3. A Sample of the raw data collected for Single Channel ATM Test

Partial Results For Single Channel Test 21 July 95, 8PSK without Reed-Solomon Coding (CRC Blue Sky - Rome Overcast/Rain)												
BER	**Cell Count**	**CLC**	**MCC**	**OOS Count**	**ECC**	**Cell Rate**	**CLR**	**CMR**	**CER**	**PRBS BEC**	**PRBS BER**	**PRBS SEC**
2.4E-5	2.88E+7	35,926	0	1	86,697	95,008	1.24E-3	0	3.01E-3	339,085	3.87E-5	0
1.0E-5	2.74E+7	10,822	0	0	34,678	90,654	3.94E-4	0	1.26E-3	106,500	1.28E-5	0
4.0E-6	1.15E+8	14,987	0	1	60,280	95,763	1.3E-4	0	5.24E-4	177,390	5.08E-6	0

BER: Bit Error Rate; CLC: Cell Loss Count; MCC: Misinsertion Cell Count; OOS: Out of Sequence;
ECC: Errored Cell Count; CLR: Cell Loss Ratio; CMR: Cell Misinsertion Rate; CER: Cell Error Ratio;
PRBS BEC: PRBS Bit Error Count; PRBS BER: PRBS Bit Error Rate; PRBS SEC: PRBS Synchronization Error Count

Rx Frame Alarm										
BER	**Line Code Count**	**Line Code Rate**	**Framing Count**	**Framing Rate**	**P-Bit Count**	**P-Bit Rate**	**CP-Bit Count**	**CP-Bit Rate**	**FEBE Count**	**FEBE Rate**
2.4E-5	0	0	4,025	4.56E-5	91,049	1.6E-2	90,815	1.9E-2	136,481	4.79E-2
1.0E-5	0	0	336,710	3.81E-3	191,829	3.37E-2	191,761	2.24E-2	41,072	1.44E-2
4.0E-6	0	0	48,026	1.37E-4	87,800	3.89E-3	87,676	2.59E-2	95,994	8.51E-3

EF Data Modem Measurements			
BER (Before)	**Tx Power**	**Eb/No**	**BER (After)**
2.4E-5	-10.8 dBm	7.4 dB	3.4E-5
1.0E-5	-10.3 dBm	7.7 dB	1.45E-5
4.0E-6	-10.1 dBm	8.0 dB	6.4E-6

Rx PLCP Alarm						
BER	**Framing Count**	**Framing Rate**	**B1 Count**	**B1 Rate**	**FEBE Count**	**FEBE Rate**
2.4E-5	25,021	3.58E-5	265,534	1.37E-2	362,906	1.87E-2
1.0E-5	7,317	1.05E-5	103,740	5.35E-3	105,783	5.45E-3
4.0E-6	11,616	4.2E-6	179,242	2.33E-3	245,200	3.19E-3

Table 4. A Sample of the raw data collected for Multiple Channel ATM Test

Partial Results For Multiple Channel Test 22August 95, QPSK without Reed-Solomon Coding (CRC clear Sky - Rome Sunny)													
Stream #	BER	Cell Count	CLC	MCC	OOS Count	ECC	Cell Rate	CLR	CMR	CER	PRBS BEC	PRBS BER	PRBS SEC
1	8.53E-6	215,263	101	0	0	74	149	4.69E-4	0	3.44E-4	562	8.59E-6	0
2	8.53E-6	215,260	104	0	0	71	149	4.83E-4	0	3.30E-4	702	1.07E-5	0
3	8.53E-6	865,334	401	0	0	253	598	4.63E-4	0	2.92E-4	2,424	9.21E-6	0
4	8.53E-6	865,341	393	0	0	263	598	4.54E-4	0	3.04E-4	2,053	7.80E-6	0
5	8.53E-6	5,224,675	2,425	0	0	1,611	3,611	4.64E-4	0	3.08E-4	14,483	9.12E-6	0
6	8.53E-6	5,224,663	2,437	0	1	1,644	3,611	4.66E-4	0	3.15E-4	14,230	8.96E-6	0

BER: Bit Error Rate; CLC: Cell Loss Count; MCC: Misinsertion Cell Count; OOS: Out of Sequence;
ECC: Errored Cell Count; CLR: Cell Loss Ratio; CMR: Cell Misinsertion Rate; CER: Cell Error Ratio;
PRBS BEC: PRBS Bit Error Count; PRBS BER: PRBS Bit Error Rate; PRBS SEC: PRBS Synchronization Error Count

Rx Frame Alarm										
BER	Line Code Count	Line Code Rate	Framing Count	Framing Rate	P-Bit Count	P-Bit Rate	CP-Bit Count	CP-Bit Rate	FEBE Count	FEBE Rate
8.53E-6	0	0	4,189	9.94E-6	9,055	3.33E-4	8,794	2.16E-4	20,961	1.54E-3

EF Data Modem Measurements			
BER (Before)	Tx Power	Eb/No	BER (After)
8.53E-6	-14.6 dBm	5.9 dB	1.37E-5

Rx PLCP Alarm						
BER	Framing Count	Framing Rate	B1 Count	B1 Rate	FEBE Count	FEBE Rate
8.53E-6	28,961	8.89E-6	196,563	2.12E-3	319,966	3.46E-3

Ultimate Limits on Spatial Light Modulator Performance for Adaptive Optics in Turbulent Atmospheres

Richard Barakat
Electro Optics

Tufts University
Medford, MA 02155

Final Report for:
Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington DC

and

Rome Laboratory

October 1995

The research conducted during the 1995 summer faculty research program at Hanscom Air Force Base, MA, in the laboratory of Dr. Charles Woods (RL/EROP) consisted of several topics. Each topic will be discussed separately. It should be noted that I do not type, therefore written manuscripts are being typed at Tufts University for publication. Additionally the manuscripts are very mathematical but an attempt will be made to describe the contents and their possible applications. The topics are:

Ultimate Limits on Spatial Light Modulator Performance for Adaptive Optics in Turbulent Atmospheres

Adaptive optics to correct atmospherically induced wavefront aberrations that degrade imaging system performance is still a topic of current interest to the Air Force. The basic ideas (call them collectively, the strategy) behind adaptive optics is now over 25 years old, and it can be said that there is very little disagreement among investigators as to what must be done. Rather the tactics (i.e., the actual carrying out of the necessary mathematics, physics and engineering) are still undergoing change mostly because sensor technology has undergone rapid advances and computers have also been speeded up over the years. These two factors combine to open up new tactics for adaptive optics that were virtually impossible even a few years ago.

To this end, Dr. Charles Woods (my focal point at Rome Labs, Hanscom AFB) has requested that I initiate (in conjunction with him) a research effort to determine the performance of spatial light modulators (SLMs) for possible use as wavefront correctors. It should be noted that I have had considerable experience in various aspects of adaptive optics having been funded several times by Rome Air Development Center in the late 1970's and through the 1980's dealing

mostly with the development of numerical techniques for adaptive optics. see for example

R. Barakat and G. Newsam, "Algorithms for reconstruction of partially known, bandlimited Fourier transform pairs from noisy data", J. Opt. Soc. Am. A2, 2027-2039 (1985). Contains references to much of the previous work of B. and N. appearing in applied mathematics journals.

as well as problems with adaptive optics per se, see for example

P. Nisenson and R. Barakat, "Partial atmospheric correction with adaptive optics", J. Opt. Soc. Am. A4, 2249-2253 (1987).

and

R. Barakat and B. Sandler, "Determination of the wavefront aberration function from measured values of the point spread function: a nonlinear two dimensional phase retrieval problem", J. Opt. Soc. Am. 9A, 1715-1723 (1990).

Previous technology employed various acousto-optical devices, and although results were excellent it is of some importance to employ new (and hopefully) better methods. Although there are a vast number of papers on the theory and application of adaptive optics, for our purposes two relatively old reviews are still of paramount importance for the SLM problem:

J. Hardy, "Adaptive optics: a new technology for the control of light", Proc. IEEE, 66, 651-697 (1978).

J. Pearson, R. Freeman, and H. Reynolds, "Adaptive optical techniques for wavefront correction", in: R. Shannon and J. Wyant (eds.), Applied Optics and Optical Engineering, Vol. 7 (Academic Press, New York, 1979) pp. 246-340.

The second is especially useful because it deals exhaustively with the acousto-optical device issue.

The SLM problem is formulated in the following fashion. Let W(p,q) be the spatially random wavefront (induced by the turbulent atmosphere) in the exit pupil of the optical system. A thin plate mirror made up os contiguous SLMs is governed by a "flexibility" matrix, F, which defines the mirror deflection, D, at points over the mirror due to the array of SLMs, S. Thus

$$D = FS$$

in matrix form. Our problem is to solve for S in terms of a given (measured D) and a given (measured) F, so that in symbolic form

$$S = F^{-1}D$$

The actual inversion of S is quite complicated and involves singular value decomposition and the corresponding pseudo-inverse. The actual details will be written up in the near future.

Now D and W are directly related

$$D = cW$$

where c is a numerical constant. Consequently

$$S = cF^{-1}W$$

This equation should be understood in the sense that since W is a sample realization of the random wavefront, then so is the solution S a sample realization.

The question now arises as to the numerical generation of the two-dimensional, zero-mean, Gaussian random process with prescribed covariance function (a very difficult problem). Fortunately an algorithm to perform this complicated task has already been worked out, see:

R. Barakat and J. Beletic, "Influence of atmospherically induced random wavefronts on diffraction imagery: a computer simulation model for testing image reconstruction algorithms", J. Opt. Soc. Am. 7A, 653-671 (1990).

I have investigated the use of the IBM parallel supercomputer on Maui (AMOS project) to carry

out the needed manipulations. The previous calculations were carried out on the vonNuemann supercomputer at Princeton University (now permanently closed). However I have spent time during the summer in ways of speeding up the algorithm by a factor of our over the original approach.

The only remaining problem we have is to specify the SLM's; in continuing discussions with Dr. Woods we have settled on a few SLMs currently available (or soon to be available) as candidates. Future work will depend upon a Summer Research Extension Program grant for 1996.

Log Intensity Statistics of Optical Wave Propagation in a Turbulent Atmosphere

Several years ago, I developed a theory of weak scattering of laser beams in atmospheric turbulence with direct application to LIDAR, optical imaging issues, and optical communication:

R. Barakat, "Weak scatter generalization of the K-density function with application to laser scattering in atmospheric turbulence", J. Opt. Soc. Am., 3A, 401-409 (1986).

R. Barakat, "Weak scatter generalization of the K-density function. II probability density of total phase", J. Opt. Soc. Am. 4A, 1213-1219 (1987).

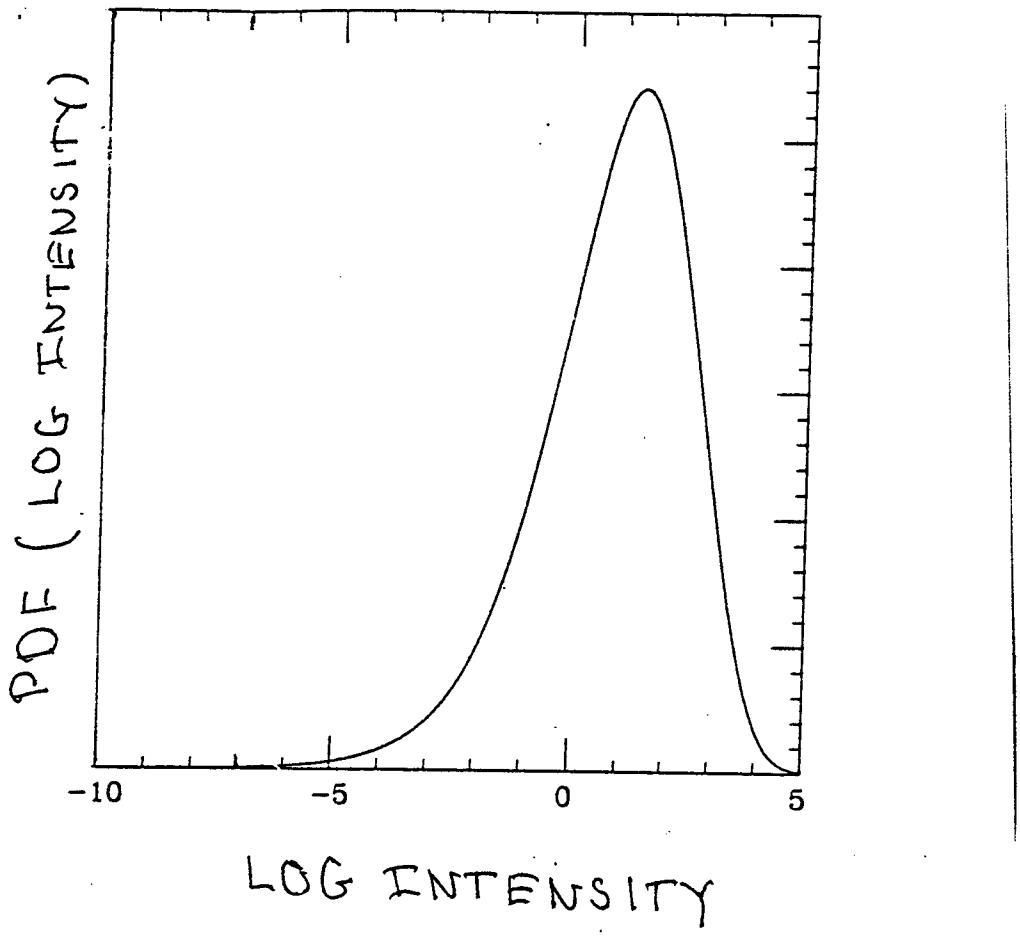
We now wish to extend the theory in terms of the logarithm of the intensity (log-intensity) because many of the field measurements are carried out in terms of the log-intensity and it would be desirable to have expressions for the probability density and moments, particularly the variance, in terms of the log-intensity. Under last summer's research program, I began the analysis and was able to essentially complete it during this summer. Only some additional numerical calculations are needed to complete the manuscript (and they are being completed at

Tufts University). The work has been summarized in a manuscript entitled:

"Log-intensity statistics for scintillation described by generalized K-distributions"

The manuscript is being typed at Tufts University for publication in an archival journal, most likely the Journal of the Optical Society of America.

The results of this work permit a considerable generalization of the theory to cover situations that are inaccessible to the strong scatter theories because the weak scatter theory contains two independent parameters, the strong scatter theories only one independent parameter. Although we cannot go into detail here, a figure is appended which shows the probability density of the log-intensity for a typical weak scatter situation.



Hankel Transform Quadrature Algorithm

This development of numerical algorithms for fast, but accurate evaluation of Hankel transforms occupied some of my attention last summer and I felt that the work should be completed at the end of this summer's research funding. Hankel transforms, i.e.,

$$h(r) = \int A(p) J_n(rp) p \, dp$$

Where J_n is the Bessel function of integral order (n = 0,1,2,...) and A(p) is given, function appears in a wide variety of problems in optics such as signal processing, laser resonators, waveguide propagation, etc. in most applications the integral must be evaluated a large number of times and with reasonable accuracy (say 6 to 8 digits). Thus any numerical quadrature algorithm must be efficient and accurate, if it is to be effective.

The present algorithm employs a change of variables of the independent variable p and the dependent variable r using logarithms to rewrite the Hankel transform as a convolution integral. I have worked the remaining mathematical details needed to maintain accuracy in the evaluation of a key component of the algorithm, the gamma function of complex argument. The function A(p) with its logarithmic change of p is expanded in the Whittaker cardinal function (a type of sampling expansion). It can be shown that the resultant expression for h(r) has coefficients that can be evaluated in terms of the Fourier transform of the gamma function of complex argument by sing the Filon quadrature scheme as developed in

R. Barakat, "Calculation of integrals encountered in optical diffraction theory", in R. Fruden (ed.), *The Computer in Optical Research* (Springer-Verlag, New York, 1980).

The algorithm has two virtues: it is fairly fast and quite accurate (I estimate that 8 digits can be routinely obtained); it is applicable to any real positive integer n. The actual programming of the algorithm will be carried out during the late fall; unfortunately it is very complicated as far as

programming is concerned.

Two possible uses are:

1. semiconductor strip lasers have, to say the least, very unusual mode structures that must be analyzed accurately if their capabilities are to be efficiently utilized. At present, there are various ad hoc and hybrid approximate approaches that have because of the difficulty in evaluating the radial part of the mode structure. I expect that it will be possible to entertain a more exact approach since we can now numerically evaluate the integrals.

2. optical signal processing operations often require the rapid, accurate numerical evaluation of Hankel transforms. Some of these issues (but at an elementary level) are discussed in

V. Das, *Optical Signal Processing* (Springer-Verlag, New York, 1991).

In summary, the algorithm should be of great help in numerical evaluation of Hankel transforms that occur in various optical contexts.

Numerical Modelling Problems in Photorefractive Media

A method of achieving lock-in detection using the photorefractive effect has been developed by scientists at the Rome Laboratories, Hanscom AFB; their work on phase sensitive detection appears to offer advantages over previous attempts:

J. Khoury, C. Woods and M. Cronin-Golomb, "Photorefractive time correlation motion detection", Opt. Comm., 85 5-9 (1991).

and

J. Khoury, V. Ryan, M. Cronin-Golomb and C. Woods, "Photorefractive frequency converter and phase sensitive detection", J.Opt. Soc. Am., 10B, 82-82 (1993).

They describe a technique for extracting a small signal phase modulation embedded in a large noise environment.

The final form of their analysis contains a double integral whose integrand is a rapidly oscillating function, the rapidity of the oscillations increasing with time. They were able to evaluate the double integral in special cases using various analytic techniques, but were not able to study more general situations of practical interest. During the summer, I investigated ways of evaluating the double integral numerically. The problem is very difficult because the integrand is so highly oscillatory in that conventional integration schemes such as Simpson or Gauss quadrature are ineffective.

I have devised an integration scheme which separates the integrand thusly

(integrand) = (slowly varying component) x (rapidly varying exponential)

The slowly varying component is approximated by a quadratic function over one of the 2-D cells, while the rapidly varying exponential is left alone. By various elementary (but extremely tedious) manipulations one can obtain a quadrature formula for the double integral for which the error induced is essentially independent of the time. Although the mathematical manipulations are (as noted) cumbersome, the final expression is quite easy to program!

Dr. Charles Woods plans to program the double integral algorithm sometime this fall and we plan to begin examining numerically further problems of signal processing that are too complicated to be dealt with by analytic methods.

A THEORY OF THE STRUCTURE OF DOMAIN THEORIES

D. Paul Benjamin
Assistant Professor
Computer Science Department

Oklahoma State University
Stillwater, Oklahoma 74078-0599

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

August, 1995

A THEORY OF THE STRUCTURE OF DOMAIN THEORIES

D. Paul Benjamin
Assistant Professor
Computer Science Department
Oklahoma State University

Abstract

Domain theories are used in a wide variety of fields of computer science as a means of representing properties of the domain under consideration. These fields include artificial intelligence, software engineering, VLSI design, cryptography, and distributed computing. In each case, the advantages of using theories include the precision of task specification and the ability to verify results. As the complexity of systems in these fields increases, these advantages become crucial. A great deal of effort has gone into the development of tools to make the use of theories easier. This effort has met with some success. However, a fundamental problem remains: the choice of formulation for a theory. This paper describes fundamental research on the representation of domain theories. The perspective of this work is to view a problem's state space as though it were physical space, and the actions in the state space as though they were physical motions. A domain theory should then state the laws of motion within the space. Following the analogy with physics, a representation is a coordinate system, and theories are transformed by changing coordinates. The mathematical basis for this analogy is given, and illustrated on two simple examples.

A THEORY OF THE STRUCTURE OF DOMAIN THEORIES

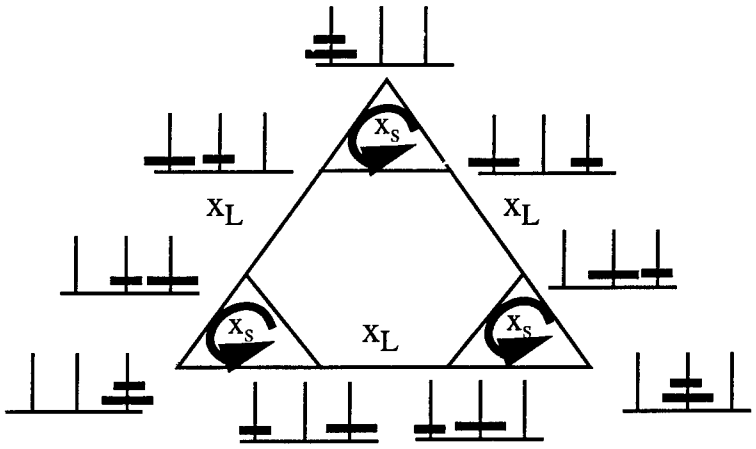
D. Paul Benjamin

Introduction

Domain theories are used in a wide variety of fields of computer science as a means of representing properties of the domain under consideration. These fields include artificial intelligence, software engineering, VLSI design, cryptography, and distributed computing. In each case, the advantages of using theories include the precision of task specification and the ability to verify results. As the complexity of systems in these fields increases, these advantages become crucial. A great deal of effort has gone into the development of tools to make the use of theories easier. This effort has met with some success. However, a fundamental problem remains: the choice of formulation for a theory. This paper describes fundamental research on the representation of domain theories. The perspective of this work is to view a problem's state space as though it were physical space, and the actions in the state space as though they were physical motions. A domain theory should then state the laws of motion within the space. Following the analogy with physics, a representation is a coordinate system, and theories are transformed by changing coordinates. The mathematical basis for this analogy is given, and illustrated on two simple examples.

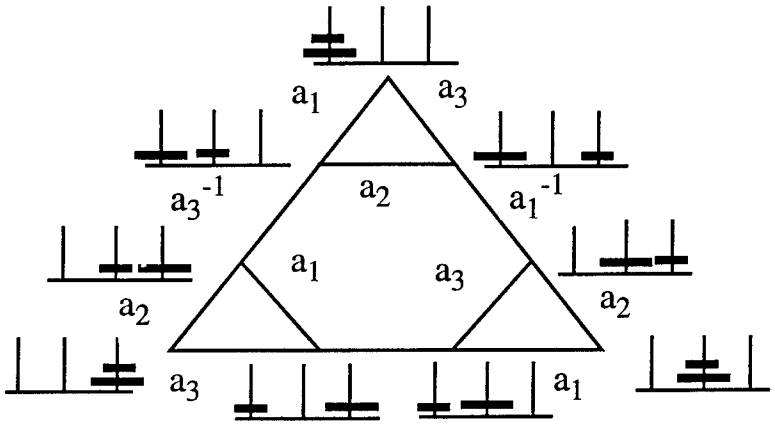
Theories must be organized so that appropriate theories can be efficiently retrieved in new situations. If there are too many specialized theories, appropriate theories will not be found quickly enough, so general theories are needed. But overgeneralization of theories causes the retrieval of theories that are inappropriate. So it is necessary to precisely identify the class of problems on which a given theory can be reused. Removing implementation detail from the representation is necessary to produce as general a theory as possible, which can be reused in many different implementations. Furthermore, even at a given level of generality, a theory can be represented in a large number of different ways that vary in their computational effectiveness. A good choice of notation and a good choice of formulation within that notation are absolutely necessary for effective use of theories. A simple example is given by the follow-

ing two representations for the two-disk Towers of Hanoi.



x = Move disk d one peg right (wrapping around)

x^{-1} = Move disk d one peg left (wrapping around)



a = Move the top disk on peg p one peg right (wrapping around)

a^{-1} = Move the top disk one peg left to peg p (wrapping around)

Even in this simple puzzle there are many possible formulations. The two formulations given above differ in that the first indexes the moves according to the disk moved, whereas the second indexes the moves by the peg moved from (or to, for the inverse moves.) The advantage of the first formulation is clear: it scales up to theories for more disks, because the actions x_S and x_L will be included unchanged in those theories. New moves will be added for the new disks. However, the actions in the second formulation must be redefined as more disks are added, so this theory does not scale up. The result is that analysis and synthesis of the two-disk problem performed using the first formulation can be reused when larger problems are attempted, but analysis and synthesis from the second formulation must be discarded when larger problems are attempted.

Computer science is not the first field to be faced with the problem of properly formulating theories. Throughout the history of science, it has always been desirable to formulate theories in as general a way as possible, so that important regularities are identified and separated from details particular to individual situations. In particular, physics has had a great deal of success in formulating theories of wide generality yet high predictive accuracy. In this paper, we will see that many of the mathematical structures employed in the statement of physical theories can be usefully generalized to the statement of abstract

theories.

We will begin with a brief discussion of the properties of physical theories in the next two sections. The remainder of the paper discusses the appropriate mathematics for analyzing these properties, and gives three simple examples of the analysis and reformulation of theories.

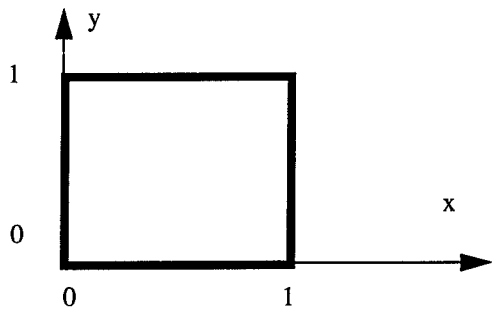
Invariants of Laws

The ability to formulate any law of nature depends on the fact that the predictions given by the law, together with certain initial conditions, will be the same no matter when or where the results of the predictions are observed. In physical theories, the fact that absolute time and location are never relevant is essential for the statement of laws; without this fact, general laws could not be stated, and the complexity of the world would eliminate the possibility of intelligent comprehension of the environment. This irrelevance is stated in terms of the invariance of laws under translation in time and space. Such invariance is so self-evident that it was not even stated clearly until less than a century ago. It was Einstein who recognized the importance of invariance in the formulation of physical law, and brought it to the forefront of physics. Before Einstein, it was natural to first formulate physical law and then derive the laws of invariance. Now, the reverse is true. As the eminent physicist Wigner states, "It is now natural for us to try to derive the laws of nature and to test their validity by means of the laws of invariance, ..."(Wigner, 1967, p.5). This is especially clear in the development of quantum mechanics.

Invariance is important not only in physics. As Dijkstra states, "Since the earliest days of proving the correctness of programs, predicates on the programs's state space have played a central role. This role became essential when non-deterministic systems were considered.... I know of only one satisfactory way of reasoning about such systems: to prove that none of the atomic actions falsifies a special predicate, the so-called'global invariant'." (Dijkstra, 1985.) In other words, as the system moves in its state space, the global invariant is a law of motion. Dijkstra goes on to point out the central difficulty in the use of invariants: "That solves the problem in principle; in each particular case, however, we have to choose how to write down the global invariant. The choice of notation influences the ease with which we can show that, indeed, none of the atomic actions falsifies the global invariant." We need a mathematics of invariance to help us formulate invariants.

Symmetry

A symmetry is a mapping of an object or system to itself such that the result of the mapping is indistinguishable from the original. For example, the human body (idealized) has a left-right symmetry. The following square has a number of symmetries, including flipping it onto itself about the lines $x = 1/2$ or $y = 1/2$ or $x = y$, and rotating it ninety degrees either clockwise or counterclockwise:



Symmetries can exist in physical space or in state space. For each invariance, there is a corresponding set of symmetries, each of which maintains the invariant. For example, the invariance of physical law under translation in space corresponds to the symmetries of space under all translations. Also, the global invariant of a non-deterministic program corresponds to all permutations of the atomic actions; each permutation maintains the invariant. This correspondence holds in reverse, also. For each set of symmetries, there is a corresponding invariant.

For example, the square above can be represented by the following theory:

$$x = 0, \quad 0 \le y \le 1$$
$$x = 1, \quad 0 \le y \le 1$$
$$y = 0, \quad 0 \le x \le 1$$
$$y = 1, \quad 0 \le x \le 1$$

This theory has syntactic symmetries corresponding to the symmetries of the square, e.g. interchanging x and y gives the flip about the line x = y, and interchanging the first and second lines of the theory flips the square about the line x = 1/2.

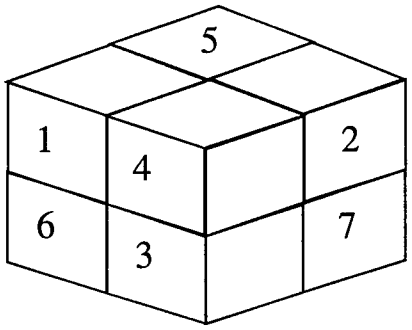
Many of the important symmetries in physics are geometric. In other words, they are symmetries of the space in which motion takes place. By viewing a program as "moving" in its state space, we can take the same approach as physicists: formulate geometric symmetries of the space, and use them to derive invariants, thereby obtaining laws governing the use of the program.

The Mathematics of Symmetry: Groups Theory

Given a global invariant, the corresponding set of transformations is closed under composition, and for every transformation, its inverse is also a transformation that preserves the invariant. The identity transformation is also always in the set. Thus, the set of transformations form a *group*. Group theory is the language of symmetries, and has assumed a central role in modern physics. Group theory can also be used to analyze the symmetries of a task and derive invariants, which are then used to synthesize a program. The following example is given in Benjamin (1994). (This paper will not provide any background

in group theory. The reader is referred to any standard text.)

Let us begin by examining a simple task, the 2x2x2 Rubik's Cube with 180-degree twists (we use such a small example for brevity of presentation, but the techniques are generally applicable, as will be shown). Let the 8 cubicles (the fixed positions) in the 2x2x2 Cube be numbered in the following way (8 is the number of the hidden cubicle):



The goal configuration for the 2x2x2 Rubik's Cube with 180-degree twists.

Number the cubies (the movable, colored cubes) similarly, and let the goal be to get each cubie in the cubicle with the same number. For brevity of presentation, we will consider only 180° twists of the cube. Let f, r, and t denote 180° clockwise turns of the front, right, and top, respectively (cubie 8 is held fixed; Dorst (1989) shows that this is equivalent to factoring by the Euclidean group in three dimensions). Note that this cubie numbering is just a shorthand for labeling each cubie by its unique coloring. This holds true for the Cube with only 180° twists, as position determines orientation.

The actions for the Cube can be represented as a group, which is generated by the actions f, r, and t. We use group representation theory to represent f, r, and t as matrices:

$$f = \begin{bmatrix} 0&0&1&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 1&0&0&0&0&0&0 \\ 0&0&0&0&0&1&0 \\ 0&0&0&0&1&0&0 \\ 0&0&0&1&0&0&0 \\ 0&0&0&0&0&0&1 \end{bmatrix} \qquad r = \begin{bmatrix} 1&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&1 \\ 0&0&0&0&1&0&0 \\ 0&0&0&0&0&1&0 \\ 0&0&0&1&0&0&0 \end{bmatrix} \qquad t = \begin{bmatrix} 0&1&0&0&0&0&0 \\ 1&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0 \\ 0&0&0&0&1&0&0 \\ 0&0&0&1&0&0&0 \\ 0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&1 \end{bmatrix}$$

These matrices are 7-dimensional, corresponding to the 7 unsolved cubies. We find eigenvectors of eigenvalue 1; these are the invariants. Any invariant of all the actions is irrelevant and can be projected away. To do this, we first change the coordinate system so that the invariant eigenvectors are axes, and then project to the noninvariant subspace, removing all irrelevant information at once. In this case, the eigenvectors are:

$$r: \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \text{ for } \lambda = 1, \text{ and } \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} \text{ for } \lambda = -1 \qquad f: \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \text{ for } \lambda = 1, \text{ and } \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \text{ for } \lambda = -1$$

t: $\begin{bmatrix}1\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix}$, for $\lambda = 1$, and $\begin{bmatrix}-1\\1\\0\end{bmatrix}$ for $\lambda = -1$ and the common invariant eigenvector is: $\begin{bmatrix}1\\1\\1\end{bmatrix}$.

Note that we have shortened these eigenvectors to save space; they are actually 7-vectors, with additional zeroes. We then change the basis. The appropriate matrix is:

$$P = \begin{bmatrix} \frac{1}{\sqrt{3}} & ""0 & ""\frac{1}{\sqrt{2}} & ""\frac{1}{\sqrt{6}} & ""0 & ""0 & ""0 \\ \frac{1}{\sqrt{3}} & ""0 & ""\frac{-1}{\sqrt{2}} & ""\frac{1}{\sqrt{6}} & ""0 & ""0 & ""0 \\ \frac{1}{\sqrt{3}} & ""0 & ""0 & \frac{-2}{\sqrt{6}} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{-1}{2} & \frac{-1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{-1}{2} & \frac{1}{2} & \frac{-1}{2} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & \frac{-1}{2} & \frac{-1}{2} \end{bmatrix}$$

yielding the new representations for r, f, and t:

$$r = \begin{bmatrix} 1&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 0&0&\frac{1}{2}&\frac{\sqrt{3}}{2}&0&0&0 \\ 0&0&\frac{\sqrt{3}}{2}&-\frac{1}{2}&0&0&0 \\ 0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&-1 \\ 0&0&0&0&0&-1&0 \end{bmatrix} \quad f = \begin{bmatrix} 1&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 0&0&\frac{1}{2}&-\frac{\sqrt{3}}{2}&0&0&0 \\ 0&0&-\frac{\sqrt{3}}{2}&-\frac{1}{2}&0&0&0 \\ 0&0&0&0&0&0&-1 \\ 0&0&0&0&0&1&0 \\ 0&0&0&0&-1&0&0 \end{bmatrix} \quad t = \begin{bmatrix} 1&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 0&0&-1&0&0&0&0 \\ 0&0&0&1&0&0&0 \\ 0&0&0&0&0&-1&0 \\ 0&0&0&0&-1&0&0 \\ 0&0&0&0&0&0&1 \end{bmatrix}$$

This procedure computes the irreducible invariants of a group. The irreducible factors of dimension 1, 1, 2, and 3 are found along the diagonals of the matrices. Projecting to these subspaces yields two subproblems:

$$r = \begin{bmatrix} \frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{bmatrix} \quad f = \begin{bmatrix} \frac{1}{2} & -\frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{bmatrix} \quad t = \begin{bmatrix} -1&0 \\ 0&1 \end{bmatrix}$$

On cubelets 1, 2, and 3, the subgroup generated is
{i. r. f. t. rt. tr}

$$r = \begin{bmatrix} 1&0&0 \\ 0&0&-1 \\ 0&-1&0 \end{bmatrix} \quad f = \begin{bmatrix} 0&0&-1 \\ 0&1&0 \\ -1&0&0 \end{bmatrix} \quad t = \begin{bmatrix} 0&-1&0 \\ -1&0&0 \\ 0&0&1 \end{bmatrix}$$

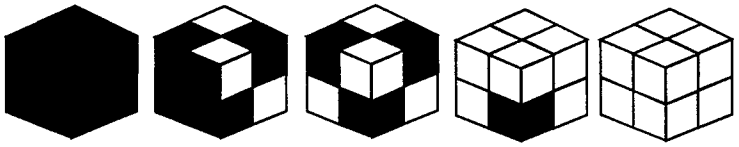
On cubelets 4, 5, 6, and 7, the subgroup generated is
{i, r, f, t, rf, rt, fr, ft, tr, tf, rfr, rft, rtr, rtf, frt,ftr, ftf, trf, tfr, rfrt, rftr, rftf, rtrf, rtfr}

Using each set of matrices as generators, we get two subgroups of actions, the second of which is a faithful representation of the whole group. The first subgroup moves cubies 1, 2, and 3, while holding 4, 5, 6, and 7 in position. The second subgroup moves cubies 4, 5, 6, and 7 while holding 1, 2, and 3 in their positions. We then repeat this procedure on the first set of actions to obtain a full set of prime factors of the group.

These factors can be assembled in different ways to form serial algorithms. There is more than one way to decompose this group, analogous to the different ways of listing the prime factors of a number. Five serial algorithms are obtained in this way. We examine two of them.

Serial Algorithm 1:

$$R = \{i, frtr\} \circ \{i, rtft\} \circ \{i, t\} \circ \{i, r, f\}$$



One of { i,r,f } brings cubelet 3 into cubicle 3.

One of { i,t } brings cubelet 1 and 2 into their places.

One of { i, rtft } brings (4,6) and (5,7) in the proper planes (`the front face looks right').

One of { i, frtr } finishes the Cube.

The above figure is read right-to-left; shaded cubicles have been solved. "i" denotes the identity (null) action. The average number of moves required to solve the Cube in this way is 5.17.

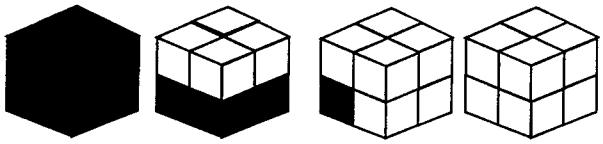
We see that there are two "subspaces" of cubies in this Cube: one consisting of cubicles 1,2,3 and 8, and the other of cubicles 4,5,6, and 7. For each subspace, there is a subprogram that interchanges the cubies in its cubicles while holding the cubies in the other subspace invariant. We are thus justified in thinking of these two sets of cubicles as *independent* subspaces. Recognizing the symmetries that characterize such subspaces is essential for synthesizing such algorithms.

Each step in the algorithm brings a subset of cubies to its goal value. Subsequent steps hold that pattern of cubies invariant. In this way, a divide-and-conquer algorithm is synthesized. For example, the first step solves cubicle 3. Knowing the colors of the solved cubicle 8, we know the colors of cubicle 3 - it has the same color as the bottom of cubicle 8, and two new colors. There is only one such cubie, and it must be in one of three locations: in its goal position, or in cubicle 1 or 2. The system need only examine those 3 values to determine what action to take. Once cubicle 3 is solved, it need not be examined again. The system next solves cubicles 1 and 2; it need only examine either position to see if the proper cubie is there; if so, it does nothing, otherwise, it twists the top. Finally, the system uses the appropriate sequence of actions to solve the remaining four cubicles, by first examining the front face to see if it is a uniform color, and then examining the top or right face to see if it is of uniform color.

We have transformed the global 7-dimensional description of the Cube into a composition of local descriptions, each characterized by a set of symmetries. This decomposition has average cost of 5.17, whereas an optimal solution is of average length 2.46. There are better decompositions. We now examine the best decomposition.:

Serial Algorithm 2:

$$R = \{i, t\} \circ \{i, r, rt\} \circ \{i, f, ft, fr\}$$



One of { i,f,fr,ft } brings cubelet 6 into cubicle 6.

One of { i,r,rt } brings 3,7 in place (bottom layer correct).

One of { i, t} finishes the Cube.

The average number of moves to solve the Cube using this decomposition is 2.75.

Each decomposition can be thought of as a coordinate system whose origin is the goal state. For example, the second serial algorithm can be thought of as a 3-dimensional coordinate system (a,b,c) where a is in {i,f,ft,fr}, b is in {i,r,rt}, and c is in {i,t}.

From the Cube example, we see that we can view a task representation as a coordinate system whose axes are the components of the task. Using group representation theory, we represent the actions as matrices. Changing the basis so that invariant eigenvectors are axes re-expresses the task space so that subtasks are coordinate subspaces, thereby identifying a good decomposition. The divide-and-conquer algorithm transforms the group of the problem into a product of smaller groups (a composition series). This is done by factoring the group at each step by an irreducible normal subgroup, yielding a quotient group. This procedure is similar to that used in Galois theory, in which an equation is solvable by extraction of roots iff its group of symmetries is solvable. However, the situation with domain theories is more general, in that a divide-and-conquer algorithm can be generated even if the transformation group is not solvable, because all that we require is that the decomposition yield components that are small enough that global search can be applied efficiently to each of them.

The Mathematics of Local Symmetries: Inverse Semigroups

The previous section illustrates how the analysis of global symmetries can be useful in synthesis. However, global symmetries, and in general global properties, are not sufficiently general. We must also examine *local symmetries*, which are symmetries between parts of a system, rather than of the whole system. Local symmetries correspond to local invariants, which are predicates that hold for some part of a system, but not necessarily for the whole system. An example is a loop invariant in a program, which

holds during the execution of the loop, but not necessarily elsewhere. This is the sort of invariant more often encountered in computing.

To handle local symmetries, we use a more general theory than group theory: the theory of inverse semigroups. A semigroup is a closed, associative system, and an inverse semigroup has the additional property that every action is invertible. The difference between a group and an inverse semigroup is that the transformations in the group must be globally defined (they correspond to global symmetries) whereas those in the inverse semigroup can be defined on only part of the space, and are thus partial functions on the space. Inverse semigroups are thus more appropriate for reasoning about programming constructs, e.g., rules, which can be defined only on some variable bindings.

In physical theories, space-time is represented in terms of a coordinate system. Invariance under translation in time or space then becomes invariance under coordinate change. Inverse semigroups possess a similar notion of coordinate system, and we use this in the same way.

We consider a task theory $M = (Q, A, \delta)$ consisting of a set A of actions defined as partial functions on a state space Q, together with a mapping $\delta: Q \times A \rightarrow Q$ that defines the state transitions. We are not concerned with the syntactic details of the encoding of the actions A, but rather with which actions should be labeled the same and should therefore be considered instances of a common abstraction. In other words, we are concerned with the algebraic structure of A. A task is specified by a pair (i,g), where $i: 1 \rightarrow Q$ maps a one-element set into Q identifying an initial state, and $g: 1 \rightarrow Q$ identifies a desired state. Without loss of generality, we can restrict our attention to semigroups of partial 1-1 functions on the state space Q (Howie, 1976). This formalism is extremely general. It encompasses nondeterministic systems and concurrent systems.

The description of a system for the synthesis of a plan (or control) for M differs from the description of M in at least one essential way: the process of planning an action is reversible whether or not the action itself is reversible (assuming the synthesizing system can backtrack.) Thus, the process of synthesizing plans can be described by a theory whose actions form an appropriate inverse semigroup containing the original actions together with newly added inverses corresponding to backtracking. As this paper is primarily concerned with system synthesis, we will analyze only inverse semigroups in the remainder of this paper.

To analyze the structure of such a semigroup of transformations, a usual step is to examine Green's relations (Lallement, 1979). Green's equivalence relations are defined as follows for any semigroup S:

$a \mathrel{R} b$ iff $aS^1 = bS^1$ $\quad a \mathrel{L} b$ iff $S^1a = S^1b$ $\quad a \mathrel{J} b$ iff $S^1aS^1 = S^1bS^1$

$H = R \cap L$ $\quad D = RL$

where S^1 denotes the monoid corresponding to S (S with an identity element 1 adjoined). Intuitively,

we can think of these relations in the following way: aRb iff for any plan that begins with "a", there exists a plan beginning with "b" that yields the same behavior; aLb iff for any plan that ends with "a", there exists a plan ending with "b" that yields the same behavior; aHb indicates functional equivalence, in the sense that for any plan containing an "a" there is a plan containing "b" that yields the same behavior; two elements in different D-classes are functionally dissimilar, in that no plan containing either can exhibit the same behavior as any plan containing the other.

Green's relations organize the actions of a transformation semigroup according to their functional properties, and organize the states according to the behaviors that can be exhibited from them. This allows us to define a basic local neighborhood of a semigroup:

Definition. Given a transformation semigroup $S = (Q, A)$ and a point $q \in Q$, a D-class of S containing an action whose domain includes q is called a *basic local neighborhood* of q.

Each basic local neighborhood in state space consists of behaviorally similar states. If the agent's perceptual capabilities are not sufficiently precise to distinguish different neighborhoods, then it cannot plan and move effectively in the space. A state q may be in more than one basic local neighborhood. Clearly, every element of Q is in at least one basic local neighborhood. The set of all basic local neighborhoods forms a neighborhood base for a topology on Q; however, this direction will not be pursued in this paper, so we will often just refer to basic local neighborhoods as neighborhoods.

Utilizing Green's relations, Lallement defines a local coordinate system for a neighborhood:

Definition. Let D be a D-class of a semigroup, and let $H_{\lambda\rho}$ ($\lambda \in \Lambda, \rho \in P$) be the set of H-classes contained in D (indexed by their L-class and R-class). A coordinate system for D is a selection of a particular H-class H_0 contained in D, and of elements $q_{\lambda\rho}$, $q'_{\lambda\rho}$, $r_{\lambda\rho}$, $r'_{\lambda\rho} \in S^1$ with $\lambda \in \Lambda$, $\rho \in P$ such that the mappings $x \rightarrow q_{\lambda\rho} x r_{\lambda\rho}$ and $y \rightarrow q'_{\lambda\rho} y r'_{\lambda\rho}$ are bijections from H_0 to $H_{\lambda\rho}$ and from $H_{\lambda\rho}$ to H_0, respectively. A coordinate system for D is denoted by $[H_0;\{(q_{\lambda\rho}, q'_{\lambda\rho}, r_{\lambda\rho}, r'_{\lambda\rho}): \lambda \in \Lambda, \rho \in P\}]$.

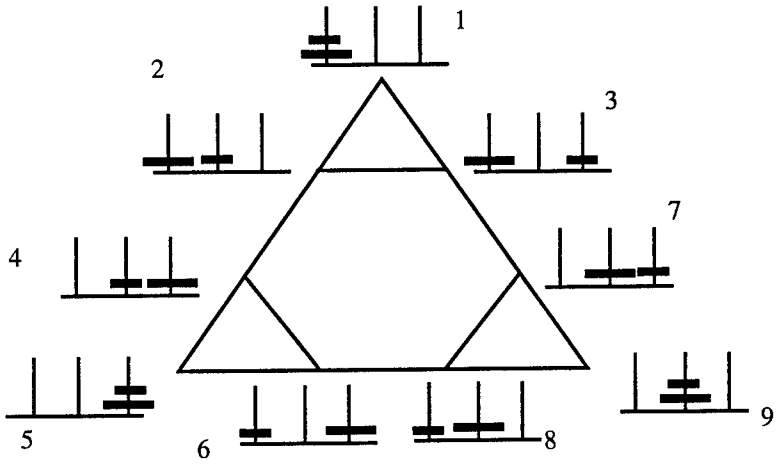
There may be more than one local coordinate system for a D-class. Each coordinate system gives a matrix representation in much the same way that a coordinate system in a vector space gives a matrix representation, permitting us to change coordinates within a local neighborhood by performing a similarity transformation (inner automorphism) in the usual way (the reader is referred to Lallement for details).

Each local coordinate system within the semigroup expresses a distinct syntactic labeling of a subsystem, as we can choose a point in the neighborhood to be the "origin", and label points in the neighborhood according to the actions that map the origin to them. In addition, each coordinate system can be used to create a matrix representation for the semigroup, in much the same way that a coordinate system

for a vector space yields matrices for the linear transformations of the vector space. The reader is referred to Lallement (1974) for details.

Example: The Towers of Hanoi

Let us number the nine states of the 2-disk Towers of Hanoi as follows:



Let the two possible actions be denoted by "x" and "y".

x = 1 2 3 4 5 6 7 8 9 y = 1 2 3 4 5 6 7 8 9
2 3 1 5 6 4 8 9 7 4 8 3

"x" moves the small disk right one peg (wrapping around from peg 3 to peg 1), and "y" moves the large disk one peg to the left (wrapping around from peg 1 to peg 3). In the figure, "x" is shown by narrow, counterclockwise arrows, and "y" is shown by thick, counterclockwise arrows. These two actions generate a semigroup of 31 distinct partial functions on the states. Green's relations for this semigroup are:

D0 | **0**

D1 | x, xx, **xxx**

D2

xyx xyxxyx **xyxxyxxyx**	xy xyxxy xyxxyxxy	xyxx xyxxyxx xyxxyxxyxx
xxyx xxyxxyx xxyxxyxxyx	xxy xxyxxy **xxyxxyxxy**	xxyxx xxyxxyxx xxyxxyxxyxx
yx yxxyx yxxyxxyx	y yxxy yxxyxxy	yxx yxxyxx **yxxyxxyxx**

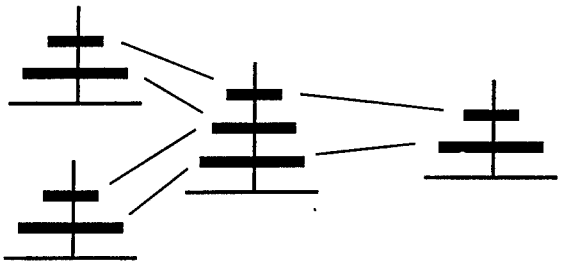
There are three D classes, shown as the three separate large boxes. In each D class, the R classes are rows and the L classes are columns, and they intersect in the small boxes, which are H classes. Note that D0 and D1 consist of only one R class and one L class, and hence one H class. The idempotents are in bold type.

There are no nontrivial inner automorphisms of D0 and D1. The group of inner automorphisms of D2 is a cyclic group of order three. These coordinate transformations are global within D2, but are local in the semigroup. These inner automorphisms are calculated by the matrix techniques explained by Lallement (1979). A generator for this group is the automorphism that maps xyx to xxy, xxy to yxx, and yxx to xyx. Factoring D2 by this map gives:

Define z = case { little disk left of large disk: xxy
little disk on large disk: xyx
little disk right of large disk: yxx }
where z is a new symbol.

x is as before, and z moves both disks left one peg. z is a macro-action, which is implemented as a disjunction of sequences of actions. x and z are independent controls; x solves the position of the small disk, and z solves the big disk. x does not change the position of the big disk, and z does not change the relative positions of the disks. The disks can be solved in either order, as the new representation is an abelian group. This new representation captures the important property of the Towers of Hanoi task: the disks can be solved in any order. In the original theory, this property was obscured by details of the implementations of the disk moves.

This decomposition applies not only to the two-disk problem, but to all Towers of Hanoi problems. This is seen by forming free products of the semigroup with itself, amalgamating the coordinate axes in all possible ways. There are three free products of this semigroup with itself that amalgamate coordinates: D2 can be identified with itself, D1 with itself, and D2 with D1 (by mapping xxy, xyx, and yxx to x). These correspond to identifying the moves of the larger disk of one copy of the semigroup with the moves of the larger disk in another copy, identifying the moves of the smaller disk with the moves of the smaller disk, and identifying the moves of the larger disk with the moves of the smaller disk, respectively. The result of this construction is the three-disk Towers of Hanoi, which is viewed as three copies of the two-disk task running concurrently.



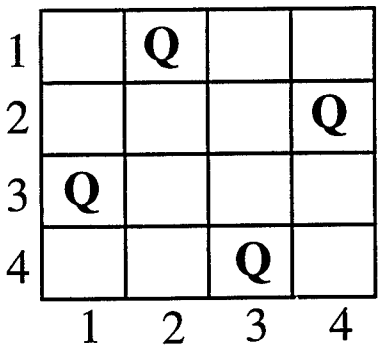
From the logical perspective, this can be viewed as composing three copies of a theory for the two-disk Towers of Hanoi, to yield a valid theory for the three-disk problem. Copies of the theory are joined by unifying variables between theories. Two copies joined at the middle disk will not suffice, as then the largest disk could be placed on the smallest disk. A third copy of the theory prohibits this. The purpose of computing a coordinate system is that the coordinate axes determine what to unify.

The interaction of these three smaller tasks yields a set of relations. For example, consider moving the middle disk one peg to the right. When considering this disk as the larger disk in the task consisting of the upper two disks, this move is xyxxy (in the state when all disks are on the left peg), yxxyx (when the smallest disk is to the right of the middle disk), or xxyxxyxx. On the other hand, when considering this disk as the smaller disk in the task consisting of the lower two disk, this move is x (in the other copy of the semigroup). This means that we add the relation xyxxy = yxxyx = xxyxxyxx to the presentation of the semigroup.

This method of composing larger tasks from smaller ones guarantees that this decomposition generalizes to any Towers of Hanoi problem with n disks, by considering a coordinate system for the three-disk task, and forming free products with amalgamation in the same manner as before. These free products with amalgamation will always reduce to free products with amalgamation of coordinate systems of the two-disk task, so that by induction the properties of the two-disk task determine the properties of all Towers of Hanoi tasks.

Example: n-Queens

The n-Queens task is to place n queens on an nxn chessboard so that no two queens are on the same row, column, or diagonal. We are posed the task of enumerating all solutions for a given value of n.



A direct way to construct a theory of this domain that specifies the legal solutions is to specify a solution as satisfying a predicate that is the conjunction of the goal conditions: no-two-queens-on-same-row & no-two-queens-on-same-column & no-two-queens-on-same-up-diagonal & no-two-queens-on-same-down-diagonal. In Smith (1990), it is shown how this domain theory can be combined with a the-

ory of global search to derive a correct high-level program. However, even with efficient heuristics, the search space grows exponentially with the size of the problem. For example, in the 4-queens problem shown above, there is no solution if we place the first queen in the upper left-hand corner or lower-left hand corner, but the system cannot know this until it has searched all ways of placing the remaining queens. This exponential growth in the number of possibilities is the well-known Achilles' heel of problem solving, as it precludes the efficient solution of very large problems. Global search cannot be used to enumerate the solutions to the million-queens problem.

Effective search heuristics are known for finding a single solution to a problem of that magnitude, but they cannot efficiently enumerate all solutions. This prevents consideration of the space of solutions, and in particular prevents optimal problem solving. More seriously, the heuristic approach requires engineering a new heuristic for each problem class. This does not even partially solve the designer's problem of finding an efficient algorithm; it merely transforms it into the problem of finding an efficient heuristic.

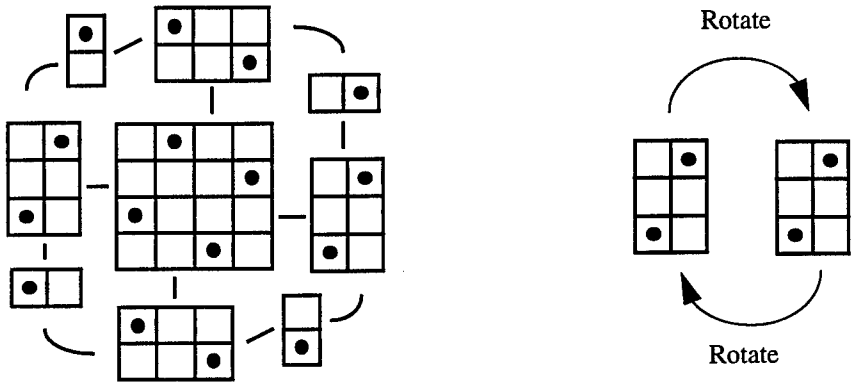
What the designer needs is a method for finding good decompositions of the problem. A good decomposition of the problem has at least the following two properties: it splits the problem into components such that solutions within the components compose to form solutions to the whole problem, and it applies to all problems in the class.

The two forms of decomposition usually implemented in systems such as KIDS (Smith, 1990) are first-rest decomposition and half-half decomposition. In first-rest decomposition, the first part of the problem description is solved and its solution is combined with the solution of the rest (this constituting a recursive call to first-rest decomposition.) In half-half decomposition, the problem description is split roughly in half, and solutions of each half are combined (again leading to recursive calls.) Mergesort is a typical example of half-half decomposition, while bubble sort is an example of first-rest decomposition.

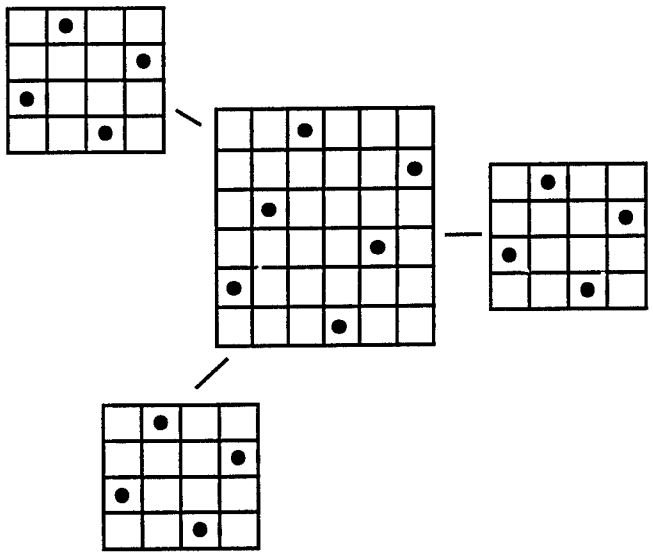
These forms of decomposition are not generally applicable. For example, neither of these forms of decomposition applies to the n-queens problem. In first-rest decomposition, we would place one queen, then place the remaining n-1 queens. But as we saw above, if we place the first queen in the wrong place (a corner), the 1-queen solution will not compose with any solution to the n-1 queens to give a valid n-queen solution. Similarly, half-half decomposition would find any solution for placing 2 queens, and compose it with another 2-queen solution (which could be a copy of itself.) This need not be a solution to the 4-queen problem.

The next figure shows the hierarchy of local symmetries for the solution to the 4-queens problem. The top-level local symmetry is that of the 3x4 subproblems, but the four 3x4 subproblems in the 4x4 are mapped to each other only by the cyclic global symmetry of order 4. The first new local symmetries arise

from considering the 2x3 subproblem that contains 2 queens. The four instances of this subproblem are of course also mapped to each other by the cyclic global symmetry, but also map to themselves by a local symmetry of order 2. The next figure shows these four instances, together with the 1x2 subproblems that arise from intersecting the 2x3 subproblems:



As in the other problems considered, we construct larger instances of the n-queens problem from copies of the 4x4 solution, amalgamating the local symmetries. For example, the 6-queens solution can be constructed from four copies of the 4-queens solution by amalgamating 2x3 subproblems:



The 6-queens solution can be viewed as three 4-queens solutions in this way.

Stated another way, the construction of the 6-queens solution from multiple 4-queens solutions can be thought of as follows: take 2 copies of the 4-queens solution and amalgamate a 2x3 subproblem in one copy with a 2x3 in the other. There are 6 distinct queens. The four queens in the first copy are guaranteed to be on their own rows, columns, and diagonals, and the four queens in the second copy are similarly guaranteed to satisfy the solution conditions. But we are not guaranteed that the two non-amalgamated queens in one copy are on distinct rows, columns, or diagonals from the two non-amalgam-

ated queens in the other copy. It is necessary that those four queens also satisfy the solution conditions, i.e., we need them to be a 4-queens solution by themselves. Therefore, composing three 4-queens solutions in this way is both necessary and sufficient to give a 6-queens solution, which implies that all 6-queens solutions are constructed in this way.

Thus, we can construct n-queens solutions from smaller solutions. It is not necessary to search a space containing non-solutions. Instead, analysis of the structure of the base case reveals both the decompositions of the base case and the inductive definition of the class of problems with the same structure. Ideally, a system for software design should have the capability to perform this analysis, to synthesize programs that construct the solution space directly in this fashion whenever possible.

Summary

A good formulation of a domain theory omits irrelevant detail and identifies the essential structure of the domain. In particular, it identifies the subgoal decompositions that lead to efficient solutions. This paper has described a research project exploring the mathematical foundations of theory formulations. An analogy is constructed between state spaces and physical space. In this analogy, a domain theory gives the laws of motion in the state space, and a coordinate system gives a formulation of a theory. The formulation of theories is then governed by the established principles of symmetry and invariance. The symmetries of state spaces of base cases can be used to formulate invariants for recursive programs that solve an entire class of tasks. The goal of this research is a comprehensive theory of the symmetries and invariants of programs, and an implemented system for program synthesis based on the theory.

REFERENCES

Benjamin, D. Paul, (1994), Formulating Patterns in Problem Solving, *Annals of Mathematics and AI*, **10**, pp.1-23.

Benjamin, D. Paul, (1992a). Towards an Effective Theory of Reformulation, in *Proceedings of the Workshop on Change of Representation and Problem Reformulation*, Michael R. Lowry (ed.), NASA Ames Research Center Technical Report FIA-92-06, pp.13-27, April, 1992.

Benjamin, D. Paul, (1992). Reformulating Path Planning Problems by Task-preserving Abstraction, *Journal of Robotics and Autonomous Systems*, **9**, pp. 1-9.

Benjamin, D. Paul, Leo Dorst, Indur Mandhyan, and Madeleine Rosar, (1990). An Introduction to the Decomposition of Task Representations in Autonomous Systems, in *Change of Representation and Inductive Bias*, D. Paul Benjamin (ed.), Kluwer Academic Publishers.

Benjamin, D. Paul, (1995). Analyzing Languages of Action for the Purpose of Synthesis, in *Proceedings of the AAAI Spring Symposium on Extending Theories of Action: Formal Theory and Practical Applications*, Stanford University, March, 1995.

Bobrow, Leonard S., and Arbib, Michael A., (1974). Discrete Mathematics, Saunders, 1974.

Dijkstra, E. W., (1985). Invariance and non-determinacy, in Mathematical Logic and Programming Languages, Hoare and Shepherdson, eds., Prentice-Hall, pp.157-165.

Dorst, Leo, (1989). Representations and Algorithms for the 2x2x2 Rubik's Cube, Philips Technical Report TR-89-041.

Howie, J. M., (1976). An Introduction to Semigroup Theory, Academic Press.

Lallement, Gerard (1979). Semigroups and Combinatorial Applications, Wiley & Sons.

Petrich, Mario, (1984). Inverse Semigroups, John Wiley & Sons, Inc., New York.

Smith, Douglas R., (1990). KIDS: A Semi-Automatic Program Development System, *IEEE Transactions on Software Engineering*, Vol. 16, No. 9, Special Issue on Formal Methods, pp.1024-1043, September, 1990.

Wigner, P. (1967) Symmetries and Reflections, Indiana University Press.

An Analysis of Adaptive DPCA

Rick S. Blum

Assistant Professor

Department of Electrical Engineering and Computer Science

Lehigh University

Bethlehem, PA 18015

Final Report for:

Summer Faculty Research Program

Rome Laboratory

Sponsored by:

Air Force Office of Scientific Research and

Rome Laboratory

August 1995

An Analysis of Adaptive DPCA

Rick S. Blum

Assistant Professor

Department of Electrical Engineering and Computer Science

Lehigh University

Abstract

A low complexity space-time adaptive processing (STAP) scheme, called the adaptive displaced antenna (ADPCA) technique is analyzed. Conditions under which this scheme is optimum are given for the case of observations with a known covariance matrix. These conditions can be satisfied if exactly two pulses are used in the ADPCA scheme, but they can not be satisfied for ADPCA schemes which use more than two pulses. An interesting interpretation of the ADPCA scheme is provided which may explain its good performance for some airborne surveillance radar problems. For some cases with noise and ground clutter only, the ADPCA scheme approximates a pulse canceling scheme applied to optimally processed pairs of pulses. The approximation is exact for cases where the clutter is produced by ground scatterers with zero azimuth angle. In other cases considered the approximation error was found to be quite small for cases where the clutter is produced by ground scatterers with any azimuth angle. The detection performance was analyzed for a case where the training data does not include the effects of a scatterer that produces clutter in the range cell under test. Here the ADPCA scheme outperforms scheme which is optimum for the case of a known covariance matrix.

An Analysis of Adaptive DPCA

Rick S. Blum

4-1 Introduction

The adaptive displaced phase centered antenna (ADPCA) technique is a low complexity alternative to joint-domain optimum space-time adaptive processing (STAP) [1]. It can provide benefits for cases where the training data is missing some details of the interference environment. ADPCA gets its name from the displaced phase centered antenna (DPCA) technique, which has been extremely useful in reducing the deleterious effects of platform motion in radar systems for a number of years [2, 3], but there is not much similarity in these two techniques. One similarity is that both techniques were developed to reduce the ground clutter seen by airborne radar. The basic DPCA technique [4, 5] involves either physically or electronically displacing the receive antenna to compensate for platform motion. Here we analyze the ADPCA technique which is part of the Rome Laboratory Space-Time Adaptive Processing Algorithm Development Tool (RLSTAP/ADT) [1].

In the ADPCA technique we compare the test statistic

$$T(X) = |V(S)^H \hat{R}^{-1} V(X)| \tag{1}$$

to a threshold to make a decision where $V(X)$ is the vector of complex data samples observed, $V(S)^H$ is the conjugate transpose of the steering vector $V(S)$, and $\hat{R}$ is the estimated noise-plus-clutter covariance matrix. If $T(X)$ is larger than the threshold then a decision is made that a target is present. The threshold that $T(X)$ is compared to can be fixed or it can be made adaptive to provide a constant false alarm rate (CFAR). We do not discuss a CFAR scheme here, but we refer the reader to [7] which discusses an approach for achieving CFAR when using a test statistic which is similar to $T(X)$. Assume that samples are taken from from M different pulse returns received at N antenna array elements. Each return

[1] see http:www.rl.af.mil:8001/Technology/Demos/STAP for more information on RLSTAP.

is assumed to contain a possible signal in additive clutter-plus-noise. The possible signal corresponds to a return from a transmitted pulse and this signal indicates the presence of a target at a particular range. For simplicity, we assume periodically transmitted pulses. Denote the observation corresponding to the j^{th} pulse at the k^{th} antenna element as x_{kj}. Each observation is a complex number corresponding to the in-phase and quadrature components of the complex envelope of the matched filtered (matched to the pulse waveform) received signal. The observations are ordered as

$$V(X) = (x_{11}, x_{21}, \ldots, x_{N1}, x_{12}, \ldots x_{NM})^T \tag{2}$$

where a^T denotes the transpose of the vector a. Define $\tilde{x}_{kj} = x_{kj} - E[x_{kj}], k = 1, \ldots, N, j = 1, \ldots, M$. The estimated noise-plus-clutter covariance matrix $\hat{R}$, whose $(k-j)$th entry we denote by r_{kj}, is an estimate of

$$E[(V(X) - E(V(X)))(V(X) - E(V(X))^H] = \begin{bmatrix} E[|\tilde{x}_{11}|^2] & E[\tilde{x}_{11}\tilde{x}_{21}^*] & \ldots & E[\tilde{x}_{11}\tilde{x}_{NM}^*] \\ E[\tilde{x}_{21}\tilde{x}_{11}^*] & E[|\tilde{x}_{21}|^2] & \ldots & E[\tilde{x}_{21}\tilde{x}_{NM}^*] \\ \vdots & \vdots & \vdots & \vdots \\ E[\tilde{x}_{NM}\tilde{x}_{11}^*] & E[\tilde{x}_{NM}\tilde{x}_{21}^*] & \ldots & E[|\tilde{x}_{NM}|^2] \end{bmatrix} \tag{3}$$

which is the true covariance matrix. The estimate is based on a set of reference data, typically chosen from the surrounding range cells.

Finally, if the array is looking broadside then the steering vector for the ADPCA technique is composed of the binomial coefficients, with each coefficient repeated for the number of antennas and where every other coefficient is altered in sign (start with positive). As a particular example, we have

$$V(S) = (1, 1, 1, 1, -2, -2, -2, -2, 1, 1, 1, 1)^T \tag{4}$$

for a three pulse, four antenna case. If $\hat{R}^{-1}$ is an identity matrix then application of this steering vector (as in (1)) has a simple interpretation. At each antenna, subtract the amplitude of neighboring pulses. Next subtract neighboring results and repeat this process until

only a single output is obtained. Finally the outputs from each antenna are summed. The processing at a single antenna is illustrated in Figure 1 for a three pulse case. This is clearly a type of multiple pulse canceling at each antenna.

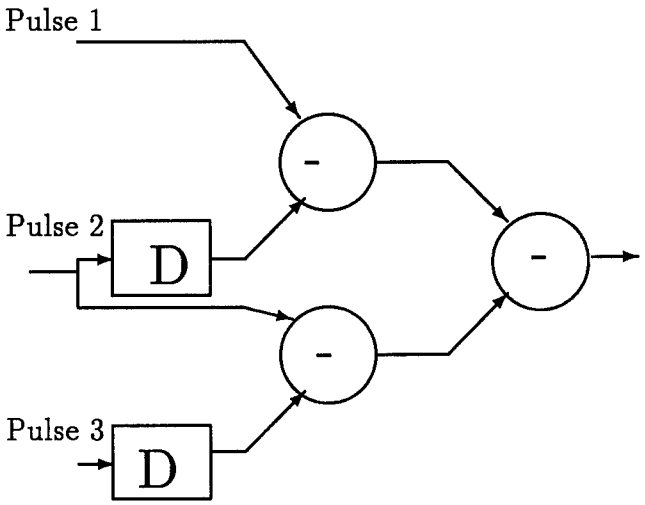


Figure 1: Block diagram of pulse differencing at one antenna.

More generally, the elements of the steering vector in (4) are multipled by the appropriate member of an exponential progression to steer the beam in the direction of the target. To be more explicit, for the rest of this paper we consider a uniformly spaced linear array antenna as described in [5]. We assume the array is oriented along the x-axis and consider a target at azimuth angle ϕ and elevation θ. To steer the beam in the target direction [5], we use

$$\begin{aligned} V(S) = (&p_\nu^0, p_\nu^1, \ldots, p_\nu^{N-1} \\ &-2p_\nu^0, -2p_\nu^1, \ldots, -2p_\nu^{N-1} \\ &p_\nu^0, p_\nu^1, \ldots, p_\nu^{N-1})^T \end{aligned} \tag{5}$$

for a three pulse, N antenna case where

$$p_\nu = \exp(j2\pi\nu_t), \tag{6}$$

$$\nu_t = \frac{d\cos\theta\sin\phi}{\lambda} \tag{7}$$

In (6) and (7) d is the distance between the equally spaced elements in the line array and λ is the radar wavelength.

4-2 Optimality of ADPCA for Two Pulse Case

The Neyman-Pearson optimum STAP scheme (assuming the estimate of $\hat{R}$ is exactly equal to the true covariance matrix) [6], under the typical assumptions of a known signal with unknown amplitude and uniform phase angle which is observed in additive Gaussian noise-plus-clutter, uses a test statistic which is of the form of (1) with $V(S)$ a scalar multiple of the complex envelope of the received signal to be detected. Thus $V(S)$ should be a scalar multiple of [6]

$$\begin{aligned} V(S) = (&p_\nu^0 p_\omega^0, p_\nu^1 p_\omega^0, \ldots, p_\nu^{N-1} p_\omega^0, \\ &p_\nu^0 p_\omega^1, p_\nu^1 p_\omega^1, \ldots, p_\nu^{N-1} p_\omega^1, \\ &\cdots, \\ &p_\nu^0 p_\omega^{M-1}, p_\nu^1 p_\omega^{M-1}, \ldots, p_\nu^{N-1} p_\omega^{M-1})^T \end{aligned} \tag{8}$$

for the N element, M pulse case, where

$$p_\omega = \exp(j2\pi\hat{\omega}_t), \tag{9}$$

$$\hat{\omega}_t = \frac{2vT}{\lambda}, \tag{10}$$

v is the velocity of the antenna array and T is the pulse repetition interval. Using the test statistic in (1) with (8) for the case where $\hat{R}$ is a maximum likelihood estimate of the covariance matrix is sometimes called the sample matrix inversion scheme. In general, we see from (5) and (8) that the ADPCA technique is generally different from the optimum STAP scheme. Now let us show one case where they are the same.

Now consider a case where only two pulses are used with N antenna elements. Further, assume that the data received by the array is preprocessed to simulate controlled platform

motion so that it appears that $v = \lambda/(4T)$ in (10). Under these conditions we find that $\hat{\omega}_t = \frac{1}{2}$ and thus the optimum steering vector for this case (8), which is

$$V(S) = (p_\nu^0, p_\nu^1, \ldots, p_\nu^{N-1}, \\ -p_\nu^0, -p_\nu^1, \ldots, -p_\nu^{N-1})^T \tag{11}$$

is exactly equal to the ADPCA steering vector for this case.

4-3 Optimality of ADPCA in General Case

One might ask "for what other cases is the ADPCA scheme described here is identical to the optimum STAP scheme in (8)". From (1) this requires that

$$V_D(S)^H \hat{R}^{-1} = C V_O(S)^H \hat{R}^{-1} \tag{12}$$

where $V_D(S)$ is the ADPCA steering vector (as in (5)) and $V_O(S)$ is the steering vector for the optimum STAP scheme (as in (8)) and C is some constant multiplier which can take on any finite complex value. Assuming $\hat{R}^{-1}$ must have an inverse $\hat{R}$ we can multiply both sides of (12) on the right by $\hat{R}$ to obtain

$$V_D(S) = C V_O(S) \tag{13}$$

as an equivalent statement to (12). Thus the only way that the ADPCA scheme is optimum is if it uses the same steering vector, to within a complex constant, as the optimum STAP scheme. This is true regardless of $\hat{R}^{-1}$, assuming it has an inverse. This indicates that the ADPCA scheme can only be optimum for the two pulse case described above. In any case with more that two pulses, the binomial coefficients have magnitudes which vary and are not constant as required by (13).

4-4 Interpretation of ADPCA

We have already shown that the ADPCA approach we study here is not optimum for any case with more than two pulses. On the other hand the multipulse canceling scheme that the

ADPCA technique uses has been used successfully in the past in a number of applications. This technique operates on pairs of pulses. Thus, one might ask if the ADPCA technique described here can be shown to be equivalent to our multipulse canceling scheme applied to the optimum processing of pairs of pulses, using (1) and (8). Here we investigate this hypothesis.

4-4.1 Single Antenna

To simplify matters, we initially consider a three-pulse case with a single antenna as illustrated in Figure 2. We want to know if

$$\begin{aligned}
&\begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}^T \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{12}^* & r_{22} & r_{23} \\ r_{13}^* & r_{23}^* & r_{33} \end{pmatrix}^{-1} \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \end{pmatrix} \\
&= \begin{pmatrix} 1 \\ -1 \end{pmatrix}^T \begin{pmatrix} r_{11} & r_{12} \\ r_{21}^* & r_{22} \end{pmatrix}^{-1} \begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix} \\
&- \begin{pmatrix} 1 \\ -1 \end{pmatrix}^T \begin{pmatrix} r_{22} & r_{23} \\ r_{23}^* & r_{33} \end{pmatrix}^{-1} \begin{pmatrix} x_{12} \\ x_{13} \end{pmatrix}
\end{aligned} \tag{14}$$

for a given $(r_{11}, r_{12}, r_{13}, r_{22}, r_{23}, r_{33})$. In (14) we have again used the notation that the $(k-j)$th entry of $\hat{R}$ is denoted by r_{kj}. Note that the conjugate symmetry in the matrices in (14) is a requirement of any covariance matrix (assuming an exact estimate). After simplification we find that a solution to (14) requires

$$\begin{aligned}
B\frac{r_{22}+r_{21}}{r_{11}r_{22}-|r_{12}|^2} &= r_{22}r_{33} - |r_{23}|^2 + 2r_{21}r_{33} \\
&\quad -2r_{23}r_{31} + r_{21}r_{23}^* - r_{22}r_{13}^*
\end{aligned} \tag{15}$$

$$\begin{aligned}
B\left(-\frac{r_{12}+r_{11}}{r_{11}r_{22}-|r_{12}|^2} - \frac{r_{33}+r_{23}^*}{r_{22}r_{33}-|r_{23}|^2}\right) &= -r_{12}r_{33} \\
-r_{13}r_{23}^* + 2r_{11}r_{33} - 2|r_{13}|^2 + r_{11}r_{23}^* - r_{12}r_13^* &
\end{aligned} \tag{16}$$

$$B\frac{r_{23}+r_{22}}{r_{22}r_{33}-|r_{23}|^2} = r_{12}r_{23} - r_{13}r_{22} + 2r_{11}r_{23} \\ -2r_{13}r_{21} + r_{11}r_{22} - r_{12}^* \tag{17}$$

where

$$B = r_{11}r_{22}r_{33} - r_{11}|r_{23}|^2 - |r_{12}|^2 r_{33} \\ +r_{12}^* r_{13}r_{32} + r_{13}^* r_{12}r_{23} - |r_{13}|^2 r_{22} \tag{18}$$

Since (15) through (17) are a set of nonlinear equations it is difficult to say how many solutions exist. One can verify that provided we choose

$$r_{11} = \frac{r_{13}r_{12}^* + r_{12}r_{23} - r_{13}r_{22}}{r_{23}} \tag{19}$$

and

$$r_{33} = \frac{r_{12}^* r_{23}^* + r_{23}r_{13}^* - r_{22}r_{13}^*}{r_{21}} \tag{20}$$

then we obtain a solution to (15) through (17) for any $r_{22}, r_{23}, r_{31}, r_{32}$. While we can not say that all solutions must be of the form given in (19) and (20), the set of solutions given by all possible choices of $r_{22}, r_{23}, r_{31}, r_{32}$ and (19) and (20) is quite large and contains some interesting cases.

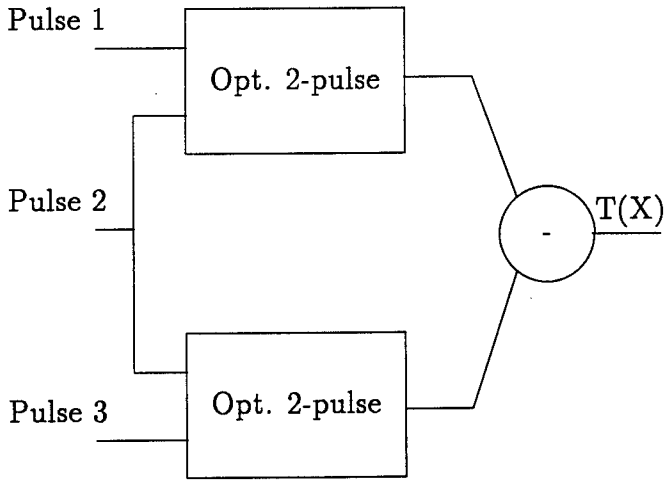


Figure 2: Block diagram showing canceling following optimum two-pulse processing.

4-4.2 Ground Clutter

Consider for example the return from a single discrete ground clutter source at an unambiguous range and at azimuth angle ϕ_c and elevation θ_c. The normalized Doppler frequency of the echo from this source of clutter is

$$\hat{\omega}_c = \frac{2vT \cos\theta_c \sin\phi_c}{\lambda} \tag{21}$$

and we define

$$p_{\omega,c} = \exp(j2\pi\hat{\omega}_c) \tag{22}$$

and

$$V_c = (p^0_{\omega,c}, p^1_{\omega,c}, p^2_{\omega})^T. \tag{23}$$

The clutter covariance matrix for this single patch of ground clutter is calculated as $\hat{R}_c = \xi_c V_c V_c^T$ [5] where ξ_c is related to the expected value of the square of the amplitude of the clutter return which can be calculated from the system parameters for a given clutter model [5]. A direct calculation shows that this covariance matrix $\hat{R}_c$ satisfies (19) and (20) for any ξ_c in the limit as $\phi_c \to 0$ (this is also true as $\phi_c \to \pm 90^o$). This means that the conditions in (19) and (20) are approximately satisfied for main beam clutter.

Since the conditions in (19) and (20) are satisfied for any ξ_c as $\phi_c \to 0$ (they are invariant to a multiplication of $\hat{R}$ by a scale factor), they are also satisfied, as $\phi_c \to 0$, by the autocorrelation matrix which results from considering a sum of any number of clutter patches provided the clutter amplitudes from different patches are uncorrelated. The uncorrelated amplitude assumption is frequently made in practice [5]. Further, if additive noise is considered, the resulting covariance matrix still satisfies the conditions in (19) and (20). A key result that enables the above limiting statements to be easily verified is that (19) and (20) are satisfied by any covariance matrix of the form

$$\hat{R}_c = \begin{pmatrix} C+N & C & C \\ C & C+N & C \\ C & C & C+N \end{pmatrix} \tag{24}$$

for any C and N. If the clutter arrives from an angle with ϕ_c close to 0 then the conditions in (19) and (20) are nearly satisfied. In some practical cases the conditions are approximately satisfied for all angles. This is discussed further in the next section, where we consider a case with two antennas.

4-4.3 Antenna Arrays

The above conclusions extend directly to cases with two antennas as we now demonstrate. For the three pulse and two antenna case we want to show that

$$\begin{aligned}
&\begin{pmatrix} 1 \\ 1 \\ -2 \\ -2 \\ 1 \\ 1 \end{pmatrix}^T \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{16} \\ r_{12}^* & r_{22} & \cdots & r_{26} \\ \vdots & \vdots & \vdots & \vdots \\ r_{16}^* & r_{26}^* & \cdots & r_{66} \end{pmatrix}^{-1} \begin{pmatrix} x_{11} \\ x_{21} \\ x_{12} \\ x_{22} \\ x_{13} \\ x_{23} \end{pmatrix} \\
&= \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}^T \begin{pmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ r_{12}^* & r_{22} & r_{23} & r_{24} \\ r_{13}^* & r_{23}^* & r_{33} & r_{34} \\ r_{14}^* & r_{24}^* & r_{34}^* & r_{44} \end{pmatrix}^{-1} \begin{pmatrix} x_{11} \\ x_{21} \\ x_{12} \\ x_{22} \end{pmatrix} \\
&- \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}^T \begin{pmatrix} r_{33} & r_{34} & r_{35} & r_{36} \\ r_{34}^* & r_{44} & r_{45} & r_{46} \\ r_{35}^* & r_{45}^* & r_{55} & r_{56} \\ r_{36}^* & r_{46}^* & r_{56}^* & r_{66} \end{pmatrix}^{-1} \begin{pmatrix} x_{12} \\ x_{22} \\ x_{13} \\ x_{23} \end{pmatrix}
\end{aligned} \tag{25}$$

Due to the increase in the number of unknowns the conditions corresponding to (19) and (20) are quite complicated and difficult to interpret here. However similar statements can be made with respect to ground clutter as for the single antenna case. With $p_{\omega,c}$ as defined in (22) and

$$V_c = (p_\nu^0 p_\omega^0, p_\nu^1 p_\omega^0, p_\nu^0 p_\omega^1, p_\nu^1 p_\omega^1, p_\nu^0 p_\omega^2, p_\nu^1 p_\omega^2)^T \tag{26}$$

the clutter covariance matrix for a single patch of ground clutter at an unambiguous range and at azimuth angle ϕ_c and elevation θ_c is again calculated as $\hat{R}_c = \xi_c V_c V_c^T$. Once again, a careful calculation shows that this covariance matrix $\hat{R}_c$ will satisfy (25) for any ξ_c in the limit as $\phi_c \to 0$ provided sufficient noise is added so that the inverses in (26) are well behaved. The conditions which require well-behaved inverses are also present in the single antenna case, but these problems can be avoided by using the conditions in (19) and (20) which allow one to approach limiting cases without numerical instabilities even when the inverses are not well behaved.

One can also show that in some practical cases the conditions on $\hat{R}_c$ given in (25) are approximately satisfied for clutter with small but nonzero ϕ_c. As one example consider a case with noise power normalized to 1 and clutter to noise power ratio of 1×10^8. We obtained plots of error, in satisfying (25). versus the clutter azimuth angle ϕ_c and β where

$$\beta = \frac{2vT}{d} \tag{27}$$

gives the slope of the clutter ridge [5]. For all $90^\circ < \phi_c < 90^\circ$ and $0 < \beta < 10$ we found an error less than 10^{-4}. Similar error results were also obtained for other clutter-to-noise ratios and for cases with more than two antennas. Specifically we obtained errors less than 2.5×10^{-4} for cases with $3, 4, 5, 6, 7, 8$ and 9 antennas.

4-5 Performance Loss Compared to Optimum

If the environment can be exactly characterized ($\hat{R}$ is equal to the true covariance matrix) and the exact position of a target is known then one can easily characterize the loss in performance that results from using ADPCA as opposed to an optimum scheme. In this case the loss is exclusively due to using a non-optimum steering vector, as shown in (13), and this loss is measured by the signal-to-interference-plus-noise ratio (SINR) of the ADPCA technique divided by the SINR of the optimum technique (assuming perfect knowledge of

$\hat{R}$) which is

$$\frac{SINR_D}{SINR_O} = \frac{|V_D(S)^H \hat{R}^{-1} V_O(S)|^2}{V_D(S)^H \hat{R}^{-1} V_D(S) V_O(S)^H \hat{R}^{-1} V_O(S)}. \tag{28}$$

$\frac{SINR_D}{SINR_O}$ ranges between 0 and 1. From the Schwartz inequality, the maximum occurs when $(V_D(S)^H \hat{R}^{-1})^H = V_O(S)$. The minimum occurs when $(V_D(S)^H \hat{R}^{-1})^H$ is orthogonal to $V_O(S)$. In general the performance is somewhere in between these extremes and depends on the exact value of $\hat{R}^{-1}$ and $V_O(S)$. Define the unit vectors

$$U_D(S) = \frac{\hat{R}^{-1/2} V_D(S)}{||\hat{R}^{-1/2} V_D(S)||} \tag{29}$$

and

$$U_O(S) = \frac{\hat{R}^{-1/2} V_O(S)}{||\hat{R}^{-1/2} V_O(S)||} \tag{30}$$

where $||V||$ denotes the magnitude of a vector V. Using (29) and (30) in (28) yields

$$\frac{SINR_D}{SINR_O} = |U_D(S) U_O(S)|^2. \tag{31}$$

From (31) it is clear that the loss here is determined only by the angular distance between the vectors $U_D(S)$ and $U_D(S)$. If $\hat{R}$ is equal to the true covariance matrix then performance is monotonic in SINR and (31) completely characterizes the loss in performance. In the more general case the performance is more difficult to characterize and probability of detection may be more appropriate measure of performance. In these cases the loss in performance, as compared to the optimum case for the case where $\hat{R}$ is equal to the exact covariance matrix, is a random variable whose distribution has been characterized for some cases of interest [8, 9], including those where the reference samples used to estimate $\hat{R}$ are statistically equivalent to the actual clutter-plus-noise present in the range cell under test. If a mismatch is present then the performance of the ADPCA scheme may actually exceed the performance of the test in (8), which is no longer optimum due to the poor estimate of $\hat{R}$. We study this interesting case in the next section.

4-6 Performance Compared to Other STAP Schemes

Consider a case where the noise-plus-clutter from the range cell under test and the noise-plus-clutter observations used to estimate $\hat{R}$ are mismatched so that each has a different statistical description. For simplicity let us assume the clutter-plus-noise observations consist of additive contributions of noise and clutter and that the noise and clutter are statistically independent. Assume that the clutter portion of the reference samples is Gaussian distributed with the two dimensional power spectral density (psd) described in [1]

$$P_c(f_t, f_s) = \sum_{d=1}^{6} \frac{\sigma_{c.d}^2}{2\pi\sigma_{ft,d}\sigma_{fs,d}} \exp\left[-\left(\frac{(f_t - f_{ct,d})^2}{2\sigma_{ft,d}^2} + \frac{(f_s - f_{cs,d})^2}{2\sigma_{fs,d}^2}\right)\right] \tag{32}$$

which is a function of Doppler frequency f_t and spatial frequency f_s. The psd in (32) models the case where the ground clutter is assumed to be dominant over other sources of clutter. This psd allows a simple approximation to the clutter ridge which would be obtained from the ground clutter received by a moving platform. One advantage of this approximation is that (32) can be inverted easily to obtain an expression for the covariance matrix [1]. The approximation in (32) consists of six Gaussian-shaped humps, the dth of which has amplitude $\sigma_{c.d}^2$, is centered at $(f_t, f_s) = (f_{ct,d}, f_{cs,d})$, and has a spread in angle and Doppler controlled by $\sigma_{ft,d}^2$ and $\sigma_{fs,d}^2$. To be specific consider a case where the training data is described by the psd in (32) with the specific parameters shown in Table 1. This psd is shown in Figure 3.

Now assume that the range cell under test has a clutter contribution to its received signal which has the two-dimensional psd shown in Figure 4 which has the same parameters as for the psd shown in Figure 3 except that it has $\sigma_{c6} = 1.155$. The psds in Figure 3 and Figure 4 could model a case where there is a large clutter return from a few discrete ground-based scatterers which are not present in the training data. Such cases have been observed in some recently measured airborne data [10]. Alternatively the difference in the psds in Figure 3 and Figure 4 could be the result of jamming [11]. For both the reference samples and for the data under test we assume the noise contribution to the noise-plus-clutter is also Gaussian distributed and that the observations at different sensors and from different

d	$\sigma_{c,d}$	$\sigma_{ft,d}$	$\sigma_{fs,d}$	$f_{ct,d}$	$f_{cs,d}$
1	0.5774	0.0217	0.0909	-0.1522	-0.3182
2	0.5774	0.0217	0.0909	-0.0652	-0.1364
3	1.0	0.0217	0.0909	0.0217	0.0455
4	0.5774	0.0217	0.0909	0.1087	-0.2273
5	0.5774	0.0217	0.0909	0.1957	0.4091
6	0.5774	0.0217	0.0909	0.2826	0.5909

Table 1: Parameters of assumed psd for training samples.

pulses are statistically independent with constant power of 0.0001 so that the noise power is significantly less than that of the clutter, a case which arises frequently in practice. This leads to a covariance matrix which is diagonal with the constant power along the diagonal. The overall noise-plus-clutter covariance matrix is the sum of the covariance matrix from each acting alone.

Consider a case with two sensors and three pulses and assume that a sufficiently large number of reference samples are used so that a perfect estimate of the covariance matrix is made based on the training samples. Consider the case of a single target at an azimuth angle $\phi = 0$ in (7) and at a normalized Doppler frequency $\omega_t = 0.3$ in (10). Assume that the signal and noise-plus-clutter combine in an additive way so that

$$V(X) = \alpha V(S) + V(C) \tag{33}$$

where $V(X)$ is the observed data vector used in (1), $V(S)$ is steering vector is defined in (8), and $V(C)$ represents the additive noise-plus-clutter vector. In (33) α is an complex number whose amplitude is unknown (thus we don't know if target is present) and whose phase is typically assumed to be uniform or possibly known (possibly from a previous scan). We compared the probability of detection performance of the scheme which uses the test statistic in (1) with the steering vector in (8), which we call fully adaptive STAP here, to that obtained when the ADPCA scheme is used. For the case of mismatched training data outlined here, a Monte Carlo simulation using 100000 runs produced the results in Figure 5

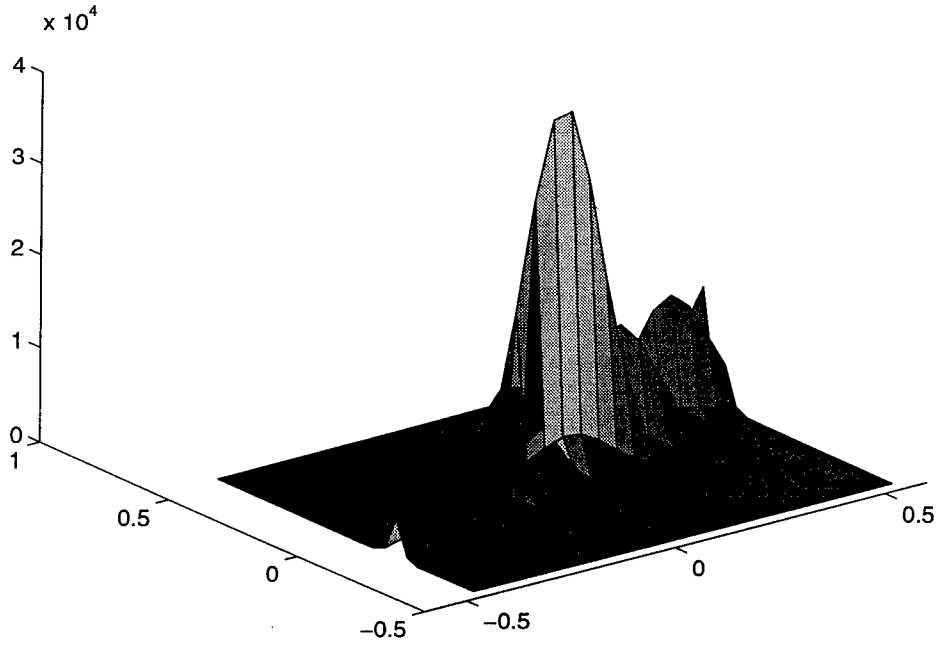


Figure 3: Two dimensional power spectral density of the training data.

for the case where the phase of α is known. Another Monte Carlo simulation produced the results in Figure 6 for the case where α has uniform phase. Larger run lengths were also tested, but produced little change in the results. The ADPCA scheme outperforms the fully adaptive STAP scheme for all signal amplitudes. Similar results were obtained for cases with different values of $\sigma^2_{ft,d}$ and $\sigma^2_{fs,d}$, including small $\sigma^2_{ft,d}$ and $\sigma^2_{fs,d}$. This indicates that the ADPCA scheme outperforms the fully adaptive STAP scheme in these cases, even if the mismatch is due to only a single discrete ground-based scatterer. The reason for the improvement in performance in the ADPCA scheme is apparently based on its pulse canceling structure, which allows this scheme to cancel clutter with high correlation across several pulses even if this correlation is not present in the training data. Recall that for cases where the training data does not exhibit dependence, where $\hat{R}^{-1}$ is an identity matrix, then ADPCA implements a pulse canceling scheme exactly.

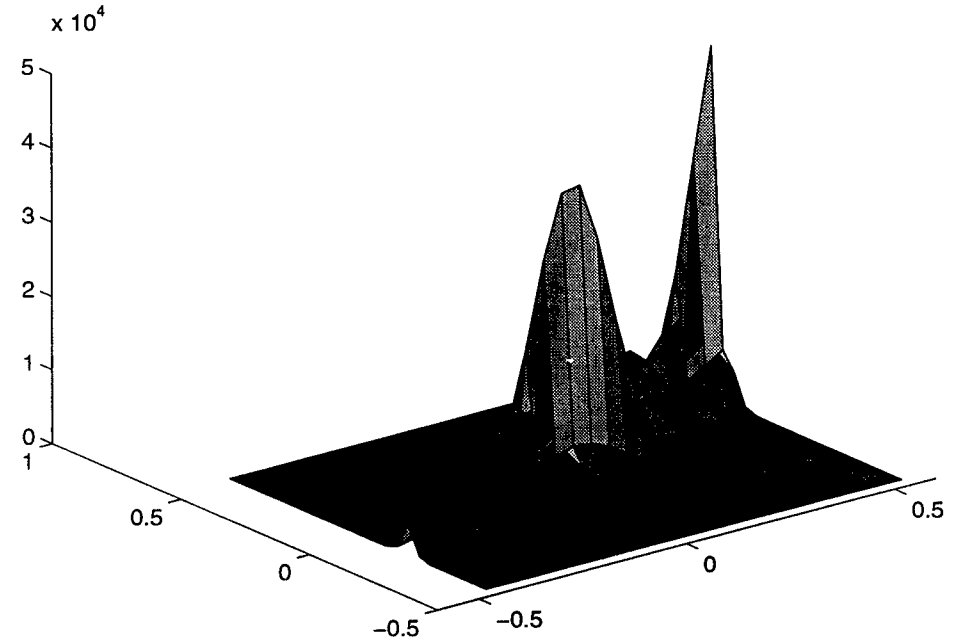


Figure 4: Two dimensional power spectral density of the cell under test.

4-7 Conclusion

An analysis of the ADPCA STAP scheme has been given. Conditions under which this scheme is optimum are given for the case of observations with a known covariance matrix. We demonstrate that these conditions can occur if exactly two pulses are used in the STAP scheme and that the conditions can not be satisfied for ADPCA schemes which use more than two pulses. However, there is a nice interpretation of the ADPCA scheme which may explain its good performance under some conditions. For cases with noise and ground clutter only, the ADPCA scheme sometimes approximates a pulse canceling scheme applied to optimally processed pairs of pulses. The approximation is exact for cases where the clutter arrives with zero azimuth angle, but it is very nearly true for cases with three pulses for ground clutter arriving at any angle. Finally, we analyzed the detection performance for a case where the training data does not characterize all of the clutter and we showed cases where the ADPCA

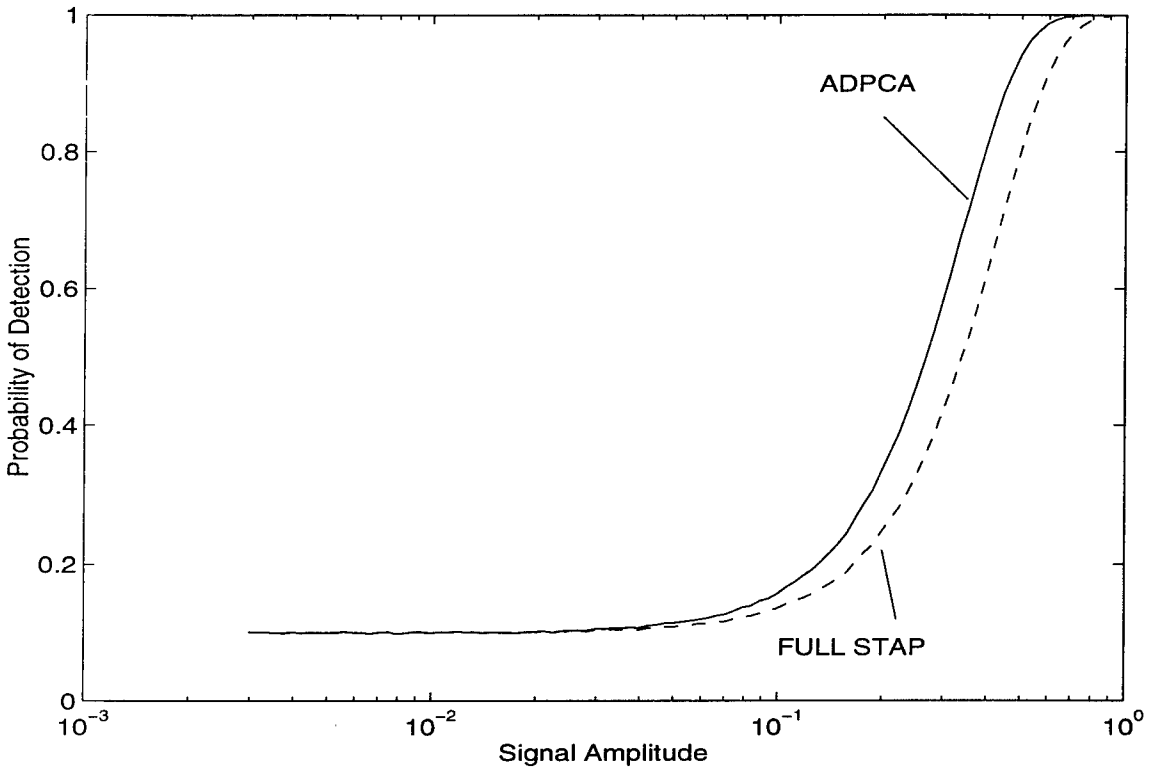


Figure 5: Probability of detection of ADPCA as compared to fully adaptive STAP for a case with additional interference which was not present in the training data, a false alarm probability of 0.1 and known target phase.

scheme outperforms the fully adaptive STAP scheme, which is optimum for the case where $\hat{R}$ is equal to the true covariance matrix.

References

1. H. Wang and L. Cai, "On adaptive spatial-temporal processing for airborne surveillance radar systems," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 30, July 1994, pp. 660-669.

2. F. R. Dickey, Jr., M. Labitt, and F. M. Staudaher, "Development of airborne moving target radar for long range surveillance," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 27, Nov. 1991, pp. 959-971.

3. F. M. Staudaher, "Airborne MTI," Chapter 16 of *Radar Handbook*, editor M. I. Skolnik, NY: McGraw-Hill, 1990.

4. W. C. Morchin, *Airborne Early Warning Radar*, Norwood, MA: Artech, 1990 (esp. pp. 310-314).

5. J. Ward, *Space-Time Adaptive Processing for Airborne Radar*, Technical Report 1015, Lincoln Laboratory, 1995.

6. L. E. Brennan and I. S. Reed, "Theory of adaptive radar", *IEEE Trans. on Aerospace and Electronic Systems*, Vol. AP-24, Sept. 1976, pp. 607-615.

7. H. Wang and L. Cai, "On adaptive multiband signal detection with the SMI algorithm," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 26, Sept. 1990, pp. 768-773.

8. D. M. Boroson, "Sample size considerations in adaptive arrays," *IEEE Transactions on Aerospace and Electronic Systems*, AES-16, July 1980, pp. 446-451.

9. E. J. Kelly, "Performance of an adaptive detection; rejection of unwanted signals," *IEEE Transactions on Aerospace and Electronic Systems*, AES-25, March 1989, pp. 122-133.

10. W. Melvin, H. Wang, and M. Wicks, "Multichannel airborne radar array data analysis," *41st Annual Tri-Service Radar Symposium*, June 1995.

11. H. Wang, Y. Zhang, and Q. Zhang, "A view of current status of space-time processing algorithm research," *IEEE International Radar Conference*, Alexandria, Virginia, pp. 635-640, May 1995.

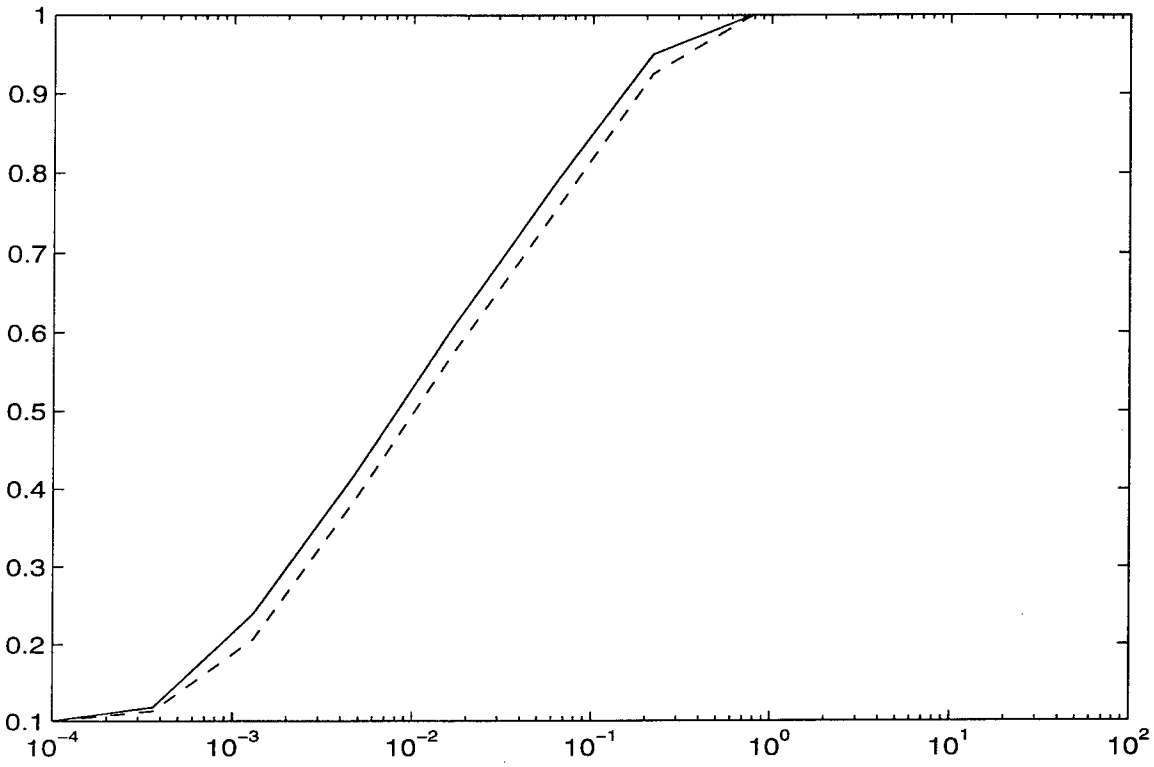


Figure 6: Probability of detection of ADPCA as compared to fully adaptive STAP for a case with additional interference which was not present in the training data, a false alarm probability of 0.1 and uniformly distributed target phase.

Partitioning Procedure in Radar Signal Processing Problems

Pinyuen Chen
Professor
Mathematics Department

Syracuse University
Syracuse, NY

Final Report for:
Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington DC

and

Rome Laboratory

August 1995

PARTITIONING PROCEDURE IN RADAR SIGNAL PROCESSING PROBLEMS

Pinyuen Chen — Syracuse University
William Melvin — Rome Laboratory
Michael Wicks — Rome Laboratory

Abstract

This report considers the problem of partitioning radar signal data according to their covariance structures. A partitioning procedure is developed from the classical ranking and selection approach to include in the selected subset those secondary populations which have the same covariance matrix as the primary population and exclude those populations which have significantly different covariance matrices from the primary population. The formulas for the performance measure, the probability of a correct partition, is derived and the least favorable configuration is found so that the sample size needed to guarantee certain probability requirement can be computed. Guideline for further research is given at the end of this report.

PARTITIONING PROCEDURE IN RADAR SIGNAL PROCESSING PROBLEMS

Pinyuen Chen William Melvin Michael Wicks

1. INTRODUCTION

The problems of statistical inference are basically treated as those of estimation (point and interval) or testing of hypotheses. This formulation does not suit the objectives of an experimenter in situations when one needs to compare among several alternatives. These may be, for example, different varieties of a grain, different competing manufacturing processes for an industrial product, different drugs (treatments) for a specific disease, or clutters with different covariance structures from the signal processing data. Some other examples are comparison of common stocks for investment and traffic fatality rates of several states. In all these problems, we have k (> 1) populations and each population is characterized by the value of a parameter of interest θ, which may be, in the example of signal processing, an appropriate measure of the covariance matrix of a clutter. In the example of comparing common stock, an appropriate measure may be the average ratio of price to earnings.

The classical approach in statistics to the preceding examples is to test the null hypothesis of homogeneity H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$ where $\theta_1, \theta_2, , , \theta_k$ are the values of the parameter θ for these populations. Tests based on Neyman-Pearson theory have been developed for various such situations and these are available in the statistics literature.

If a test of homogeneity is the primary and final goal of an investigator, alternative methods of statistical analysis are not needed. However, there are many practical situations in which other kinds of information or other goals are of interest. The experimenter may be interested in (1) determining which populations differ from which others, and in what direction, or (2) to see which populations can be considered best in some well-defined sense of "best". For the example of signal processing problem, we might be interested in selecting those secondary data (or the reference data) which are more comparable to the primary data (or the test data) in their covariance structures. After the selection process, we shall use the selected data to test the signal by using the conventional testing procedures which were designed under the ideal situation while the secondary data and the primary data have the same covariance structure. The ranking and selection procedures have been designed specifically to resolve such problems.

The formulation of ranking and selection procedures has been generally accomplished under one of two basic approaches, namely, the indifference zone approach (IZ) and the subset selection approach (SS). Most of the literature in ranking and selection theory comes under one of these two approaches and modifications. The main concern of this report, namely, screening procedures for the secondary data which have different covariance structures from the primary data, falls into the category of a modification of the subset selection formulation: partitioning populations (secondary data) with respect to a control (test cell data). Here we shall introduce the basic concepts of the IZ and the SS approaches. Partitioning approach which is more pertinent to radar signal processing data will be defined formally in Section 2.

Consider k independent populations $\pi_1, \pi_2, \ldots, \pi_k$ where π_i has the underlying distribution F_{θ_i}, $i = 1, 2, \ldots, k$. The θ_i are unknown real-valued parameters which represent the value of a quality characteristic of interest for these k populations. We define π_i to be better than π_j if $\theta_i > \theta j$. The ordered θ_i are denoted by $\theta_{[1]} \leq \ldots \leq \theta_{[k]}$. For the basic problem of selecting the best population, the indifference zone formulation of Bechhofer (1954) defines that the selection of the population associated with $\theta_{[k]}$ results in a correct selection (CS). He required that for any procedure R to be valid, R should guarantee a specified minimum probability of a correct selection P(CS), say P* (1/k < P* < 1) whenever the best and the second best populations are apart at least by a specified amount. Let $\delta(\theta_i, \theta_j)$ denote an appropriately defined non-negative measure of the amount of separation between the population associated with θ_i and θj. For any specified $\delta^* > 0$, let Ω_{δ^*} be the subset of the

parameter space $\Omega = \{\tilde{\theta} | \tilde{\theta} = (\theta_1, \ldots, \theta_k)\}$ defined by $\Omega_{\delta^*} = \{\tilde{\theta} | \delta(\theta_{[k]}, \theta_{[k]}) \geq \delta^*\}$. The subset Ω_{δ^*} is called the preference zone. Letting P(CS|R) denote the PCS of a rule R, in order to be valid, it should satisfy P(CS|R) $\geq$ P* for all $\tilde{\theta} \in \Omega_{\delta^*}$. Both δ^* and P* are specified by the experimenter in advance. Suppose R is based on samples of fixed size n from each population. The problem then is to determine the smallest sample size n for which the P* requirement holds. The compliment of the preference zone Ω_{δ^*} is called the indifference zone where no requirement is made on the PCS.

In the subset selection approach of Gupta (1956), the basic problem is to select a non-empty subset of the k given populations so that the best population is included in the selected subset with a guaranteed minimum probability P*. In case of a tie for the best place, we assume that one of the contenders is tagged as the best. Selection of any subset that includes the best results in a correct selection. Any valid rule R should satisfy P(CS|R) $\geq$ P* for all $\tilde{\theta} \in \Omega$. Note that the size S of the selected subset is not decided in advance but is determined by the data.

The procedures should satisfy the probability requirement P*. Any parameter configuration $\tilde{\theta}$ which yields the infimum of the P(CS) over the preference zone Ω_{δ^*} or the entire parameter space Ω, depending on the approach (IZ or SS), is called the least favorable configuration (LFC). Many variations and generalizations of the basic formulation using either of the two approaches have been studied. One important related selection problem is selecting populations better than a standard or a control π_0. In our study in selection procedure based on radar signal processing data, the standard or control can be taken as the primary data. The re-grouped secondary data are those independent populations $\pi_1, \pi_2, \ldots, \pi_k$. We will be interested in selecting those populations from those secondary data $\pi_1, \pi_2, \ldots, \pi_k$ that have the same covariance structure as the control population from the primary data.

Radar signal processing data usually comes in the forms of random vectors and they were analyzed statistically in the framework of multivariate analysis. In Section 2, we will briefly review the selection theory in multivariate analysis. In Section 3, we try to relate the radar signal processing problem and the testing procedure studied in Chen (1994) to the statistical selection concepts introduced in Sections 1 and 2. In Section 4, we formally define the partitioning formulation of selection theory that will be used in our applications in radar data. In Section 4, we will also propose a procedure to solve the partitioning problems for radar signal processing data. We will derive P(CP), the probability of a correct partition, for the proposed procedure in Section 5. Final remarks for future investigations will be given in Section 6.

2. MULTIVARIATE SELECTION PROBLEMS

A multivariate distribution is the joint probability distribution of p variables or components where p > 1. A single observation from a multivariate population is a measurement on each of these components, that is, a pair of measurements if p=2, a triplet of measurements if p=3, and in general, a p-tuple of measurements. Therefore a random sample of size n from a multivariate distribution consists of a number n of p-tuples of measurements, one p-tuple for each of the n sample observations. We sometimes refer to these as vector observations. The main multivariate distribution considered in this report is the multivariate normal distribution (complex- or real- valued) which has as parameters not only the means and variances of each of the p variables, but also covariances between pairs of these components. A multivariate normal distribution with mean vector $\tilde{\mu}$ and covariance matrix Σ is denoted by $N_p(\mu, \Sigma)$.

Let $\pi_1, \pi_2, \ldots, \pi_k$ be k p-variate normal populations, $N_p(\mu_i, \Sigma_i)$, $i = 1, 2, \ldots, k$. We consider several measures θ for selecting the populations such as the generalized variance and the Mahalanobis distance. These measures θ are unknown real-valued parameters which represent the value of a quality characteristic of interest for these k populations. In the real world applications of statistical ranking and

selection, the experimenter usually first decide which measure is to be used to rank the populations. Then he/she will study selection procedures based on a good estimate of θ. The review of multivariate selection problems here is to introduce some familiar measures and the procedures that statisticians use in general practice. Those measures may not be applicable in our study in radar signal processing which, as in any field of specialty, has its own unique properties. However, as it will become clear in Section 3, the general format of our selection goal and probability requirement in radar signal processing problems follows closely to those of existing procedures that are reviewed here.

2.1 Selection in terms of Mahalanobis Distance

Let $\lambda_i = \tilde{\mu}_i' \Sigma_i^{-1} \tilde{\mu}_i$, the Mahalanobis distance from the origin. We consider the selection of the population associated with the largest λ_i.

2.1.1 Subset Selection Approach

Let $X_{ij}, i = 1,\ldots,n$, denote n independent observations from $\pi_i, i = 1,\ldots,k$. Define

(2.1) $$Y_{ij} = X_{ij} \Sigma_i^{-1} X_{ij}, Y_i = \sum_{j=1}^{n} Y_{ij}, Z_i = \bar{X}_i' \Sigma_i^{-1} \bar{X}_i,$$

where $\bar{X}_i$ and

S_i are the sample mean vector and covariance matrices.

Case 1: When the covariance matrices Σ_i are known:

In this case, Gupta (1966) proposed the rule

(2.2) R_G : Select π_I if and only if $Y_i \geq c\, Y_{[k]}$

where $0 < c = c(k,p,n,P^*) < 1$ is the largest number for which the P*-condition is satisfied. It has been shown that the LFC is $\tilde{\lambda} = (\lambda_1,\ldots,\lambda_k) = (0,\ldots 0)$. Thus the constant c in the procedure can be found to be the solution of the following integral equation:

(2.3) $$\int_0^\infty G_v^{k-1}(x/c)\, dG_v(x) = P^*$$

where $G_v(x)$ is the distribution function of a Gamma random variable with probability density function

(2.4) $$g(x) = \frac{x^{v-1} e^{-x}}{\Gamma(v)} \qquad x \geq 0.$$

The values of c to implement the procedure can be found in Gupta (1963). For the analogous procedure for selecting the population associated with the smallest λ_i, the procedure R_G defined in (2.2) can be modified to "select π_I if and only if $Y_i \leq d\, Y_{[1]}$". The appropriate constant d can be found in Gupta and Sobel (1962).

It is natural to use the statistic Z_i instead of the Y_i in the rule R_G. . In this situation however, the infimum of the P(CS) and hence the procedure parameter c do not dependent on n, the sample size of the experiment; this is an unsatisfactory feature for a statistical inference procedure. One can. of course, study a different type of procedure, for example, "select π_I if and only if $Y_i \leq Y_{[k]} - d$" $d > 0$. Such a procedure has not been investigated in the literature.

Case 2: When the covariance matrices Σ_i are not known:

In this case Gupta and Studden (1970) proposed the rule

(2.5) R_{GS} : Select π_I if and only if $T_i \geq e\ T_{[k]}$

where $0 < e = e(k,p,n,P^*) < 1$ is to be chosen suitably to satisfy the P*-condition. As in the previous Case 1, the LFC is $\tilde{\lambda} = (\lambda_1,\ldots,\lambda_k) = (0,\ldots 0)$ and the constant e is given by

(2.6) $$\int_0^\infty F_{p,n-p}^{k-1}(x/e)dF_{p,n-p}(x) = P^*$$

where $F_{p,n-p}(x)$ is the distribution function of a F random variable with probability density function

(2.7) $$f(x) = \frac{(p/n-p)^{\frac{1}{2}p}\, x^{\frac{1}{2}(p-2)}(1+(p/n-p)x)^{-\frac{1}{2}n}}{B(\frac{1}{2}p,\frac{1}{2}(n-p))} \qquad x \geq 0.$$

Values of e to implement the procedure have been tabulated for selected values of k, n, p, and P* by Gupta and Panchapakesan (1969), who also tabulated the needed constant h for the analogous rule "Select π_I if and only if $T_i \leq h\ T_{[1]}$" for selecting the population associated with the smallest λ_i.

2.1.2: Indifference Zone Approach

Alam and Rizvi (1966) have investigated procedures for selecting the populations associated with the t largest λ_i, $1 \leq t \leq k-1$, by taking for the preference zone $\Omega(\delta_1^*,\delta_2^*) = \Omega_{\delta_1^*} \cap \Omega_{\delta_2^*}$ where

(2.8)

$$\Omega_{\delta_1^*} = \{\tilde{\lambda}: \lambda_{[k-t+1]} - \lambda_{[k-t]} \geq \delta_1^*\} \quad \text{and} \quad \Omega_{\delta_2^*} = \{\tilde{\lambda}: \lambda_{[k-t+1]} / \lambda_{[k-t]} \geq \delta_2^*\}$$

for specified $\delta_1^* > 0$ and $\delta_2^* > 1$. They considered the natural selection rule, namely,

(2.9) R_{AR}: Select the populations that yielded the t largest Z_i in the case of known Σ_I 's

and the t largest T_i in the case of unknown Σ_I 's.

The LFC is given by

(2.10) $$\lambda_{[1]} = \ldots \lambda_{[k-t]} = \delta_1^*(\delta_2^*-1)^{-1}, \lambda_{[k-t+1]} = \ldots \lambda_{[k]} = \delta_1^*\delta_2^*(\delta_2^*-1)^{-1}.$$

The smallest value of n required to satisfy the P* condition can be obtained in the case of known Σ_I 's from

(2.11) $$t\int_0^\infty G_p^{k-t}(x,\frac{n\delta_1^*}{\delta_2^*-1})\{1-G_p(x,\frac{n\delta_1^*\delta_2^*}{\delta_2^*-1})\}^{t-1} g_p(x,\frac{n\delta_1^*\delta_2^*}{\delta_2^*-1})dx = P^*$$

where $g_p(x,\theta)$ and $G_p(x,\theta)$ denote the probability density function and the distribution function, respectively, of a non-central chi-square random variable with p degrees of freedom and noncentrality parameter θ. The smallest value of n required to satisfy the P* condition can be obtained in the case of unknown Σ_I 's from

(2.12) $$t\int_0^\infty F_{p,n-p}^{k-t}(x,\frac{n\delta_1^*}{\delta_2^*-1})\{1-F_p(x,\frac{n\delta_1^*\delta_2^*}{\delta_2^*-1})\}^{t-1} f_{p,n-p}(x,\frac{n\delta_1^*\delta_2^*}{\delta_2^*-1})dx = P^*$$

where $f_{p,n-p}(x,\theta)$ and $F_{p,n-p}(x,\theta)$ denote the probability density function and the distribution function, respectively, of a non-central F random variable with p and n-p degrees of freedom and noncentrality parameter θ.

2.2 Selection in terms of the Generalized Variance

The generalized variance $\theta = |\Sigma|$ of a multivariate normal population with covariance matrix Σ serves as an over-all measure of the variability among the components. Gnanadesikan and Gupta (1970) considered selection of the population associated with the smallest θ_i. They proposed a subset selection rule

(2.13) R_{GG} : Select π_i if and only if $W_i \le l\, W_{[1]}$

where $W_i = |S_i|$ is the sample generalized variance and $0 < l = l(k,p,n,P^*) < 1$ is to be chosen to meet the P* condition. They have shown that

$$\inf_{\Omega} P(CS|R_{GG}) = P(Y_1 \le \frac{1}{l} Y_j, j = 2,\ldots,k) \tag{2.14}$$

where the Y_j are i.i.d., each distributed as a product of p independent factors where the rth factor has a Chi-square distribution with (n-r) degrees of freedom. An exact solution for l is obtained for p=2 and is tabulated by Gupta and Sobel (1962). For p > 2, Gnanadesikan and Gupta (1970) have studied approximations using the normal approximation of log χ^2 and Hoel's approximation of the distribution of $Y_i^{1/p}$ by an appropriate gamma distribution.

Eaton (1967) considered a decision-theoretic approach to ranking the k populations according to the values of the θ_i, assuming reasonable properties for the loss function which dependents only on the θ_i. He showed that the natural rule which ranks the populations according to the values of W_i is minimax, admissible, and uniformly the best among rules that are invariant under permutations of $(W_1,\ldots, W_k)$.

As we can see from the rules described in (2.2), (2.5), and (2.12) above, the selection procedures under the SS approach (i.e., when the goal is to select a subset containing the best or to eliminate bad populations) are all of the ratio type whose decision rule depends on the ratio of the statistics of two samples. Another type used as commonly as the ratio type in univariate selection theory is the difference type whose decision rule depends on the difference of the statistics of two samples. The difference type procedures have not been fully explored in multivariate selection theory. The procedure described in (2.9) is neither ratio nor difference type since under the indifference zone approach one is interested in selecting only the best population, not a subset containing the best.

In the multivariate selection problems we have discussed so far, the multivariate normal populations have been ranked according to the values of a scalar function of the parameters which are usually in vector or matrix form. In the indifference zone approach, a measure for the scalar function of the parameters would be used to define the indifference zone. Then a procedure based on the estimators, the so-called selection statistic, of the scalar function of the parameters would be proposed for the selection goal. In the subset selection approach, we usually do not define the indifference zone. Thus the measure for the parameters is not needed. However the selection procedure, again defined by the estimator of the parameters of interest, would depend on some measure of the parameters. For example in Section 2.1, $\lambda_i = \tilde{\mu}_i' \Sigma_i^{-1} \tilde{\mu}_i$, the Mahalanobis distance from the origin, is the scalar function of the parameters of interest. The measure used in the subset selection procedure "R_G : Select π_I if and only if $Y_i \ge c\, Y_{[k]}$ " is ratio. Notice that the procedure also depends on the estimator Y of the parameters. The statistic Y is the selection statistic here. In the indifference zone approach, both the "difference" and the "ratio",

$\Omega_{\delta_1 *} = \{\tilde{\lambda}: \lambda_{[k-t+1]} - \lambda_{[k-t]} \geq \delta_1 *\}$ and $\Omega_{\delta_2 *} = \{\tilde{\lambda}: \lambda_{[k-t+1]} / \lambda_{[k-t]} \geq \delta_2 *\}$ are used to define the indifference zone. Again, the selection procedure depends on the selection statistic. In order to implement a selection procedure, we need to derive the PCS and the LFC. That is where the distribution theory of the selection statistic comes into the picture.

In a recent paper, Bofinger (1992) has considered the problem of multiple comparisons with the best for multivariate normal populations using a "multivariate" approach. Bofinger's results are mainly for bivariate normal populations. She finds that " for comparisons with the "best" of each variate, repeated univariate comparisons appear to be almost as efficient as multivariate comparison, at least for the bivariate case and, under certain circumstances, for higher dimensional cases". These aspects of the selection and related inference problems are worth exploring further.

3. APPLICATIONS OF SELECTION THEORY TO RADAR SIGNAL PROCESSING

Statistical inferences in radar signal processing problem usually involve two of the major areas in statistical research. These two areas are (1) testing of hypotheses and (2) multivariate analysis. The null hypothesis of the mean vector equals to 0 (i. e., $H_0: \tilde{\mu} = 0$) implies that these is no target in the test cell. The alternative hypothesis of the mean vector equals to a given non-zero vector (i.e., $H_0: \tilde{\mu} = \xi$) implies the present of a target in the test cell. Since all the data obtained from radar systems are assumed to be correlated multi-components vectors (usually complex-valued vectors), the testing procedures usually fall into the framework of multivariate analysis. Although some attentions in radar problems have been given to non-normal theory recently, we will focus our study on normal (Gaussian) theory.

Classical radar detection theory was developed under the assumption that target returns follow a homogeneous Gaussian noise. The term "homogeneity" here means the covariance matrices of the reference cells (the secondary data) all have the same structure as the covariance matrix of the test cell (the primary data). Under this assumptions, the likelihood ratio tests and their modifications provide optimum signal processors whose performances have been well studied and documented in literature (see for examples, Kelly (1986), Wang and Cai (1990), Cai and Wang (1990)). As it was well reported in Wicks (1993), the detection of a target in a heterogeneous (non-homogenous) clutter and noise interference environment in modern radar is an important and challenging signal processing problem. The classical detection algorithms, designed for homogeneous situation, may suffer a significant loss in their performances if the true environment is heterogeneous. Statistically, heterogeneous problems can be handled in two manners : (1) One can develop algorithms for the heterogeneous case and compare their performances with those of the algorithms developed for the homogenous case; (2) One can make inferences about the covariance structures of the reference data. If the covariance structure of the reference cells is tested or estimated to have the same structure as that of the test cell, we can apply the existing algorithms which were designed under homogenous environment to the reference data.

Raghavan, Qiu, and McLaughlin (1995) made an attempt to the approach (1). They considered the detection problem assuming the clutter has unknown correlation properties. However, the distribution theory needed for calculating the performance measures (probability of false alarm and probability of detection) of their proposed statistic is far from complete. Therefore, their probability in the null hypothesis (i.e., the probability of false alarm) could be computed only when the signal is a deterministic value which causes no change on the second order statistic (the covariance) structure. Their probability in the alternative hypothesis (i.e., the probability of detection) was not computed for the test statistic $z^H S^{-1} z$ proposed in the paper. In stead of using the sample covariance matrix S in the above test statistic, they used an approximated covariance matrix R_c which is available only under large sample assumption. Strictly speaking, a complete solution to the radar problem parallel to Kelly (1986)'s result should consists of a distribution theory for the test statistic $z^H S^{-1} z$ while z is a normal random vector with zero mean (under the null hypothesis) and non-zero mean (under the alternative hypotheses) and S is a Wishart random matrix, distributed independently of z, with a covariance matrix R_c in a different form from R_s, the covariance matrix of z.

Then one can make comparison between their procedure with Kelly's GLR procedure as they tried to do in the paper.

Chen (1994) made an attempt to the second approach (2). In the report, the author studied a test statistic which took exactly the same form as the test statistic $z^H S^{-1} z$ studied in Raghavan, Qiu, and McLaughlin (1995). However, the goal of Chen (1994) was to test if the covariance structure is the same for the primary data and for the secondary data while the goal of Raghavan, Qiu, and McLaughlin (1995) was to test if a signal is present in the primary data. Chen (1994) derived the distribution of the test statistic $z^H S^{-1} z$ while z is a normal random vector with zero mean (under both the null hypothesis and the alternative hypotheses) and S is a Wishart random matrix, distributed independently of z, with a covariance matrix R_c in a different form from R_s, the covariance matrix of z. Based on the findings in Chen (1994), the research in this report is a ranking and selection aspect of the second approach (2).

As mentioned in Section 1, if a test of homogeneity is the final goal of an investigator, alternative analysis is not needed. However in our radar detection applications, we are not only interested in which secondary data have the same covariance structure as that of the test cell. We are also interested in selecting those clutters whose covariance structure is the same as that of the test cell so that a follow-up testing on the signal by traditional algorithms can be applied. The ranking and selection theory has been designed specifically to solve such problem.

The indifference zone approach of ranking and selection theory does not meet our need in radar signal detection problem because it selects only the best population. The usual subset selection approach is more appropriate since it eliminates the bad populations. In our application of radar detection, we need a procedure which can eliminate those secondary data which have different covariance structures (in the language of ranking and selection the "bad populations") from the primary data. Suppose that the secondary data comes from the independent random vectors $Y_1, Y_2, \ldots, Y_K$ and the primary data comes from the random vector Y_0. Assume that we can divide the vectors $Y_1, Y_2, \ldots, Y_K$ into k subgroups of vectors $\{Y_{a_1}, \ldots, Y_{b_1}\}, \{Y_{a_2}, \ldots, Y_{b_2}\}, \ldots, \{Y_{a_k}, \ldots, Y_{b_k}\}$ where each subgroup serves as a population. This structure can also be seen in a multiband data discussed in Cai and Wang (1990) where they referred k as the number of subbands. Thus our goal is, with high probability of confidence, to eliminate those subgroups whose covariance structures are significantly different from the primary data (or the control population in the language of ranking and selection theory). To achieve the goal, we need to modify the classical subset selection approach so that the idea of comparing with a control is included in the formulation. The new formulation, which will be called partitioning with respect to a control, will be formally defined in the next section. We will also propose a procedure, based on the test statistic $z^H S^{-1} z$ studied in Chen(1994), to select the good populations and to eliminate the bad populations.

4. PARTITIONING WITH RESPECT TO A CONTROL

4.1 General Formulation:

This sub-section is concerned with a general problem of partitioning a set of independent populations into two subsets according to their parameters with respect to a control population. The distribution model for this type of experiment is k populations $f(x;\theta_1), f(x;\theta_2), \ldots, f(x;\theta_k)$, and a single control population $f(x;\theta_0)$. Each of these k+1 distributions is of the same form; they differ only by the parameter θ which measures the goodness of the respective populations. The values of $\theta_1, \theta_2, \ldots, \theta_k$ are all unknown and the value of θ_0 may be either known or unknown. For arbitrary but fixed constants δ_1^*, δ_2^*, we define two disjoint and exhaustive subsets Ω_G, and Ω_B of the set $\Omega = \{\pi_1, \pi_2, \ldots, \pi_k\}$ by

(4.1)

$$\Omega_G = \{\pi_i : \delta_2^* \le d(\theta_i, \theta_0) \le \delta_1^*\}, \Omega_B = \{\pi_i : d(\theta_i, \theta_0) < \delta_2^*\} \cup \{\pi_i : d(\theta_i, \theta_0) > \delta_1^*\}$$

where $d(x,y)$ is a distance measure of the parameters. The most commonly used measures d are the difference and the ratio. After observations have been taken, the set $\Omega = \{\pi_1, \pi_2, \ldots, \pi_k\}$ is partitioned into two disjoint subsets S_G and S_B (the subscripts G stands for "good", B stands for "bad"). The definition of a correct partition is defined as follows:

(4.2) A partition is correct (CP) if $\Omega_G \subset S_G$.

Let P* be an arbitrary but preassigned constant. The statistical problem is to find a procedure R which consists of a sampling procedure and a terminal decision rule such that the appropriate probability requirement stated below is satisfied.

(4.3) $$P(CP|\theta_0, \theta_1, \theta_2, \ldots, \theta_k) \ge P^*.$$

4.2 Multivariate Partitioning According to the Covariance Structure:

Now we turn our direction to the multivariate partitioning problems that we will use in radar detection problems. Let $\pi_1, \pi_2, \ldots, \pi_k$ be k p-variate complex normal populations $CN_p(\mu_i, \Sigma_i)$, i = *1, 2, ..., k,* and let π_0 be the control, a p-variate complex normal population $CN_p(\mu_0, \Sigma_0)$. Thus the f function in Section 4.1 is now taken to be the complex multivariate normal density and the parameter θ is taken to be $(\tilde{\mu}, \Sigma)$. We further assume that $\tilde{\mu} = 0$ since our primary concern is on the covariance structure of the data. A meaningful distance measure between two variances is the ratio function $d(x,y)=x/y$ since the variance is a scale parameter in normal distribution. The covariance matrix of a multivariate normal distribution plays a similar rule as the variance of an univariate normal distribution in many applications, especially in distribution theory. Thus we will also use the ratio of two covariance matrices to define the distance measure in our study. We define the distance function

(4.4) $$d(\Sigma_i, \Sigma_0) = \text{the largest eigenvalue of } \Sigma_i \Sigma_0^{-1}.$$

Now we define the two disjoint and exhaustive subsets Ω_G, and Ω_B of the set $\Omega = \{\pi_1, \pi_2, \ldots, \pi_k\}$ by using the distance function d. They are

(4.5)
$$\Omega_G = \{\pi_i | \delta_1^* \le d(\Sigma_i, \Sigma_0) \le \delta_2^*\}$$
and
$$\Omega_B = \Omega - \Omega_G.$$

where $\delta_1^* < \delta_2^*$ are pre-assigned positive real numbers used to differentiate the good and the bad populations. The values of δ_1^* and δ_2^* should be near 1 since for the special case when dimension p=1 (i.e., the univariate case), the distance measure d has the same meaning as the ratio of two variances. A population is good if the ratio of its variance and the variance of the control population is near 1. This also explain why the function d is chosen to define the "good" and the "bad" populations. Our goal is to partition the populations obtained from the reference data into two disjoint subsets S_G and S_B. The partition is correct (CP) if $S_G \subset \Omega_G$. It means that all the population included in selected subset S_G are good, or they all have similar covariance structure as the control population. It also means that all the bad populations, i.e., the populations with significantly different covariance structure from the control population are eliminated from our detection algorithm We require a procedure R that will satisfy the pre-determined probability requirement P(CP| R) ≥ P*. The proposed procedure R_c is defined as follows:

Procedure R_c : For each population π_i (*i= 1, 2, ..., k*), computer $T_i = x^H S_i^{-1} x$ where x is the test cell random vector and S_i is the sample covariance matrix of the population π_i. Partition the set of the populations $\Omega = \{\pi_1, \pi_2, \ldots, \pi_k\}$ into two subsets S_G and S_B. The subset S_G consists of those populations π_i with $c \le T_i \le d$ where c and d are chosen so that the probability requirement P(CP) ≥ P* is satisfied.

The selection statistic $T_i = x^H S_i^{-1} x$ in procedure R_c is in the same form as the test statistic used in Chen (1994). This seems to be a reasonable choice in estimating any scalar function of the matrix $\Sigma_0 \Sigma_i^{-1}$ when there is only one observation available in the control population π_0. If we consider S_i^{-1} as the reciprocal of the covariance matrix, then the measure used in the selection procedure can be thought of as a ratio in matrix form.

5. THE PERFORMANCE MEASUREMENT OF THE PARTITIONING PROCEDURE R_C --- THE PROBABILITY OF A CORRECT PARTITION

In ranking and selection theory, we use the probability of a correct selection to measure the performance of a selection procedure. In this report, our goal is to partition the populations according to their covariance structures. A natural measurement of the performance of the proposed procedure is the probability of a correct partition (PCP). In the following, we first derive the least favorable configuration (LFC) based on which we will computer P(CP). Then we'll write the minimum of P(CP|LFC) in terms of the multivariate normal density function and the Chi-Square density function. In order to implement the proposed procedure, we need to solve numerically the procedure parameters c and d from the integral equation minP(CP) = P*.

Assume that there are $k_1 + k_2$ bad populations. Among them, k_1 populations $\pi_1, \pi_2, \ldots, \pi_{k1}$ are significantly better than the control and k_2 populations $\pi_{k-k_2+1}, \pi_{k-k_2+2}, \ldots, \pi_k$ are significantly worse than the control. That is, for $i = 1,2,\ldots,k_1$, the largest eigenvalue λ_i of the matrix $\Sigma_0 \Sigma_i^{-1} < \delta_1^*$. For $i = k - k_2 + 1, \ldots, k$, the largest eigenvalue λ_i of the matrix $\Sigma_0 \Sigma_i^{-1} > \delta_2^*$. As a consequence, there are *k- k_1 - k_2* good populations. We also assume that each population has sample size *n*.

Then

(5.1) P(CP) = P(all the bad populations are eliminated)

$$= P(T_i < c, i = 1,2,\ldots,k_1 ; T_j > d,\ j = k - k_2 + 1,\ldots,k)$$

$$= P\{ (\frac{n-p+1}{p}) \frac{T_i}{\lambda_i} < (\frac{n-p+1}{p}) \frac{c}{\lambda_i}, i = 1,2,\ldots,k_1 ;$$

$$(\frac{n-p+1}{p}) \frac{T_j}{\lambda_j} > (\frac{n-p+1}{p}) \frac{d}{\lambda_j}, j = k - k_2 + 1,\ldots,k \}.$$

We obtain from Chen (1994), page 7-8, the random variable $(\frac{n-p+1}{p}) \frac{T_i}{\lambda_i}, i = 1,2,\ldots,k$ follows a F distribution with $2p$ and $2(n-p+1)$ degrees of freedom. Thus, the above expression can be rewritten as

(5.2) $$P(CP) = P\{ f_i < (\frac{n-p+1}{p})\frac{c}{\lambda_i}, i = 1,2,\ldots,k_1;$$

$$f_j > (\frac{n-p+1}{p})\frac{d}{\lambda_j}, j = k-k_2+1,\ldots,k \},$$

where f's are correlated F random variables. It is clear from (5.2) that P(CP) attains its minimum when the parameters λ_i's, $i = 1,2,\ldots,k_1$ reach their maximum, all at δ_1^* and the parameters λ_j's, $j = k-k_2+1,\ldots,k$ reach their minimum, all at δ_2^*. Thus we can write

(5.3) $$P(CP) \geq P\{ f_i < (\frac{n-p+1}{p})\frac{c}{\delta_1^*}, i = 1,2,\ldots,k_1;$$

$$f_j > (\frac{n-p+1}{p})\frac{d}{\delta_2^*}, j = k-k_2+1,\ldots,k \}.$$

It is clear from the above expression that P(CP) is minimum when the number of good populations, $k - k_1 - k_2$, is 0. Thus we have shown the following theorem.

THEOREM 5.1: The least favorable configuration for any parameter vector $(\lambda_1,\ldots,\lambda_k)$ under procedure R_C is given by

(5.4) $$\lambda_1 = \lambda_2 = \ldots = \lambda_m = \delta_1^* < \delta_2^* = \lambda_{m+1} = \lambda_{m+2} = \ldots = \lambda_k$$

where m is an integer between 0 and k that minimizes the probability of a correct partition given at the right hand side of the inequality in (5.3).

From the above theorem, we can write the minimum of P(CP) as

(5.5) *min* $$P(CP) \geq P\{ f_i < (\frac{n-p+1}{p})\frac{c}{\delta_1^*}, i = 1,2,\ldots,m;$$

$$f_j > (\frac{n-p+1}{p})\frac{d}{\delta_2^*}, j = m+1,\ldots,k \},$$

where the minimum is over all the parameter vectors $(\lambda_1,\ldots,\lambda_k)$.

In order to implement the proposed procedure R_C, we need to find the procedure parameters c and d for given n, p, k, δ_1^*, and δ_2^*. Although (5.5) gives us a simple expression of *min*P(CP), it is not easy to computer the probability since it involves correlated F distributions. In the following, we will use conditioning arguments to rewrite *min*P(CP) in terms of multivariate normal density function and Chi-Square density function.

We shall obtain the same minimum P(CP) as expressed in (5.5) if the original control population has covariance matrix $\Sigma_0 = I$, the identity matrix and the population covariance matrix

$$\Sigma_i = \begin{bmatrix} \frac{1}{\delta_1^*} & 0 & . & . & . \\ 0 & e_1 & 0 & . & 0 \\ . & . & e_1 & 0 & 0 \\ . & . & . & . & . \\ 0 & . & . & . & e_1 \end{bmatrix}$$ for $i = 1,2,\ldots,m$, the population covariance matrix

$$\Sigma_j = \begin{bmatrix} \frac{1}{\delta_2^*} & 0 & . & . & . \\ 0 & e_2 & 0 & . & 0 \\ . & . & e_2 & 0 & 0 \\ . & . & . & . & . \\ 0 & . & . & . & e_2 \end{bmatrix}$$ for $i = m+1,m+2,\ldots,k$, where e_1 and e_2 are any numbers larger than $(1/\delta_1^*)$ and $(1/\delta_2^*)$ respectively. Therefore we have, from (5.1) and Theorem 5.1,

(5.6) $$\text{P(CP|LFC)} = \text{P}(x^H S_i^{-1} x < c, i = 1,2,\ldots,m; x^H S_j^{-1} x > d, j = m+1,\ldots,k)$$

where x is a p-variate complex normal random vector *CN(0,I)* and S_i is the sample covariance matrix with covariance matrix Σ_i and S_j is the sample covariance matrix with covariance matrix Σ_j as defined above. Define $x = (x_1, x_2, \ldots, x_p)$ where x_i's are i.i.d. univariate complex *CN(0,1)*. From Rao (1973), the statistic $W_i = L^H \Sigma_i^{-1} L / L^H S_i^{-1} L$ follows a Chi-Square distribution with *2(n-p-1)* degrees of freedom for any fixed vector *L*. Thus the statistic $L^H S_i^{-1} L = (L^H \Sigma_i^{-1} L) / W_i$ follows the reciprocal of a Chi-Square random variable times a constant. Thus from (5.6), we have

(5.7) $$\begin{aligned} \text{P(CP|LFC)} &= \text{P}\{(x^H \Sigma_i^{-1} x / x^H S_i^{-1} x) > (x^H \Sigma_i^{-1} x)/c, i = 1,2,\ldots,m; \\ &\qquad (x^H \Sigma_j^{-1} x / x^H S_j^{-1} x) < (x^H \Sigma_j^{-1} x)/d, j = m+1,m+2,\ldots,k\} \\ &= \int\int\cdots\int \text{P}\{W_i > (\delta_1^*/c)x_1\tilde{x}_1 + (1/ce_1)[x_2\tilde{x}_2 + \ldots + x_p\tilde{x}_p], i = 1,2,\ldots,m; \\ &\quad W_j < (\delta_2^*/d)x_1\tilde{x}_1 + (1/de_1)[x_2\tilde{x}_2 + \ldots + x_p\tilde{x}_p], j = m+1,\ldots,k\}\phi(x_1,x_2,\ldots,x_p)dx_1dx_2\cdots dx_p \end{aligned}$$

where $W_i, i = 1,\ldots,m$ and $W_j, j = m+1,\ldots,k$ are i.i.d. Chi-Square random variables with *2(n-p-1)* degrees of freedom for given *x*, $\tilde{x}_i, i = 1,2,\ldots,p$ are the respective conjugates of $x_1, x_2, \ldots, x_p$, and $\phi(x_1, x_2, \ldots, x_p)$ is the probability density function of a p-varaiate *CN(0,I)*. Consider the eigenvalues e_1 and e_2 for the two covariance matrices Σ_i and Σ_j as parameters subject to change for minimizing P(CP). Then from the last expression in (5.7), the minimum of P(CP|LFC) is obtained when $e_1 = (1/\delta_1^*)$ and $e_2 = (1/\delta_2^*)$. Substitute $e_1 = (1/\delta_1^*)$ and $e_2 = (1/\delta_2^*)$ in (5.7), we obtain

(5.8) $$\textit{min}\text{P(CP|LFC)} = \int\int\cdots\int \text{P}\{W_i > (\delta_1^*/c)[x_1\tilde{x}_1 + x_2\tilde{x}_2 + \ldots + x_p\tilde{x}_p], i = 1,2,\ldots,m;$$

$W_j < (\delta_2^* / d)[x_1\tilde{x}_1 + x_2\tilde{x}_2 + ... + x_p\tilde{x}_p], j = m+1,...,k\}\phi(x_1, x_2, ..., x_p)dx_1dx_2 \cdots dx_p$

The above probability depends on $x_1, x_2, ..., x_p$ only through $x_1\tilde{x}_1 + x_2\tilde{x}_2 + ... + x_p\tilde{x}_p$ which has a Chi-Square distribution with *2p* degrees of freedom. Thus we can write

$$(5.9) \qquad \min P(CP|LFC) = \int_0^\infty (\chi^2(\delta_2^* y / d))^{k-m}(1-\chi^2(\delta_1^* y / c))^m f(y)dy$$

where χ^2 is the distribution function of a Chi-Square random variable with *2(n-p-1)* degrees of freedom and *f(y)* is the probability density function of a Chi-Square random variable with *2p* degrees of freedom.

6. CONCLUDING REMARKS AND DIRECTIONS FOR FUTURE RESEARCH

The limitation on the sample size prohibits us to get a good estimate for the covariance structure of the control (i. e., the test cell in radar application). Since the statistic $T_i = xS_i^{-1}x$ intuitively gives us more information about the deviation of the covariance matrix Σ_i from the covariance matrix Σ_0 than any other statistic, it was used in Chen (1994) to develop the testing procedure for the homogeneity of the covariance matrix of the test cell and the covariance matrix of the reference cells. The underlying distribution for both the test data and the reference data in Chen (1994) is assumed to have 0 mean. As a matter of fact, the second order statistic (therefore, the covariance matrix) is not even estimable if one has only one observation in the test data when the first order statistic is not known.

This report suggests, as those multivariate selection procedures reviewed in Section 2, the same statistic $T_i = xS_i^{-1}x$ used in testing be used in partitioning the secondary data. It shows that, as in any typical selection problem, the distribution theory studied for the purpose of hypothesis testing (in our case, the test statistic T in Chen (1994)) could be very useful here in partitioning theory. The main results we obtain in this paper are (1) a new selection formulation for partitioning the secondary data in radar signal processing; (2) an inferential selection procedure to achieve the goal of partitioning the secondary data; and (3) the formulas for the performance measurement, P(CP) for the proposed procedure. For the future study of this problem, we shall divide the research into the following three phases.

Phase 1: Computing the procedure parameters c and d.
We plan to write a computing algorithm to calculate the probability in formula (5.9) for general *n, p, k*, δ_1^*, and δ_2^*. Then by using the computing algorithm, we shall be able to tabulate the needed procedure parameters c and d for selected combinations of *n, p, k*, δ_1^*, and δ_2^*. A numerical examples should be given to illustrate the proposed procedure by using the table of c and d. The work requires a numerical integration of a small-valued function which is the product of Chi-square distribution functions and probability density function. It also involves searching for the appropriate solutions for different P* requirements.

Phase 2: Simulating P(CP), E(S), E(G), and E(B).
After the table of c and d is computed, simulation work should be conducted to confirm our theoretical results on minP(CP) and the computing results on c and d obtained in Phase 1. At the same time, we need to include in our simulation yet another performance measurement of the proposed procedure, the expected size of the subset that are partitioned as the 'good' reference data. The expected subset size, E(S), is often used in subset selection approach as a criterion to evaluate a selection procedure. The smaller the expected subset size is, the better the selection procedure is. Since our goal is to screen out

those cells with significantly different covariance structures and simultaneously keep those reference cells which have the same covariance structure as the test cell so that a follow-up study on signal detection can be conducted, the expected number of bad cells, E(B) and the expected number of good cells, E(G) in the subset partitioned as the 'good' subset could also be used as possible performance measurements. The above four measures, P(CP), E(S), E(G), and E(B) can be simulated in the following manner.

(1) For given k, and unequal covariance matrices Σ_i ($i = 1, ..., k$), we generate n complex multivariate normal random vectors. We also generate one complex multivariate random vector for the control test cell with covariance matrix Σ_0.

(2) From the table in Phase 1, we obtain the appropriate c and d values for given k and P*. Calculate the statistic T_i for each of the population π_i ($i = 1, 2, ..., k$).

(3) Use Procedure R_c to partition the populations.

(4) Repeat (1) to (3) 10000 times and estimate P(CP), E(S), E(G), and E(B) respectively by
P(CP) = # of times the selected subset contains all the 'good' populations/10000;
E(S) = sum of the subset sizes of the subsets partitioned as the 'good' populations/10000;
E(G) = total # of the true good populations in the subsets partitioned as 'good'/10000;
E(B) = total # of the true bad populations in the subsets partitioned as 'good'/10000.

Phase 3: Evaluating the robustness of the partition procedure in radar applications.

This is the phase where we will be making assessments on the theoretical selection theory that we propose in this report in radar signal processing applications. From the theoretical derivation of P(CP) in Section 5, we can find the procedure parameters c and d so that with high probability P*, we are making a correct partition no matter what the real situation is. If a radar signal detection test designed for the homogenous data is applied to the good populations which have been identified by our partition procedure, it is reasonable to assume that we will be getting at least as good results as the original data. But whether the improvement on the performance measures in radar system is significant for us to develop the partition procedure is an important issue. We first need to identify which standard procedure in radar signal detection problem to be used in evaluating our partition theory. We then need to define the criteria (or the performance measure) that we should use to assess our proposed study.

In radar detection, Kelly (1986) studied a generalized likelihood ratio (GLR) test to test a given signal. His test has been used in many articles (see for examples, Cai and Wang (1990), Raghavan, Qiu, and McLaughlin (1995), and Wang and Cai (1990)) as a standard in comparing new methodologies in signal detection. We suggest that Kelly's procedure being used as the signal detection procedure to evaluate our selection. The general performance measures in radar detection theory are the probability of false alarms P(FA) and the probability of detection P(D). We suggest that these two measures being used as our comparing criteria for the data before selection and after selection. The following simulation study is proposed:

(1) Simulate multivariate complex normal data with k covariance structures for the reference data. Simulate a multivariate complex normal data for the test data.

(2) Apply GLR test for given P(FA) on the data obtained in (1) and record a detection or not detection.

(3) Set the probability requirement P* and choose the appropriate procedure parameters c and d from the table obtained in Phase 1. Apply the proposed partition procedure R_c to the data obtained in (1). We will obtain a set of 'good' populations.

(4) Apply the GLR test for the same P(FA) on the 'good' populations and record a detection or not detection.

(5) Repeat (1) to (5) 10000 times and calculate P(D|without partition procedure) and P(D|with partition procedure) respectively as follows.

P(D|without partition procedure) = total number of detection in (2)/10000;

P(D|with partition procedure) = total number of detection in (4)/10000.

(6) Use the estimate obtained in (5) to assess our partition procedure.

References:

Alam, K. and Rizvi, M. H. (1966). Selection from multivariate populations. *The Annals of the Institute of Statistical Mathematics*, 18, 307-318.

Bechhfer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics*, 25, 16-39.

Bofinger, E. (1992). Multiple comparisons with 'best' for multivariate normal populations, *Communications in Statistics - Theory and Methods*, 21, 915-941.

Cai, L. and Wang, H. (1990). On adaptive filtering with the CFAR feature and its performance sensitivity to non-Gaussian interference, *Proceedings of the 1990 Conference on Information Science and Systems,* 558-563.

Chen, P. (1994). On testing the equality of covariance matrices under singularity, Final Report, 1994 AFOSR Summer Faculty Research Program.

Eaton, M. L. (1967). The generalized variance: testing and ranking problem. *The Annals of Mathematical Statistics,* 38, 941-943.

Gnandesikan, M. and Gupta, S. S. (1970). Selection procedures for multivariate normal distributions in terms of measures of dispersion. *Technometrics*, 12, 103-117.

Gupta, S. S. (1956). On a decision rule for a problem of ranking means. Mimeograph Series No. 150, Institute of Statistics, University of North Carolina, Chapel Hell, NC.

Gupta, S. S. (1963). Probability integrals of the multivariate normal and multivariate t. *The Annals of Mathematical Statistics*, 34, 792-828.

Gupta, S. S. (1966). On some selection and ranking procedures for multivariate normal populations using distance functions. *Multivariate Analysis* (ed. P. R. Krishnaiah), Academic Press, NY, 457-475.

Gupta, S. S. and Panchapakesan, S. (1969). Some selection and ranking procedures for multivariate normal populations. *Multivariate Analysis-II* (ed. P. R. Krishnaiah), Academic Press, NY, 475-505.

Gupta, S. S. and Sobel, M. (1962). On the smallest of several correlated F-statistics, *Biometrika*, 49, 509-523.

Gupta, S. S. and Studden, W. J. (1970). On some selection and ranking procedure with applications to multivariate populations. *Essays in Probability and Statistics* (ed. R. C. Bose et al.), University of North Carolina Press, Chaple Hill, 327-338.

Kelly, E. J. (1986). An adaptive detection algorithm, *IEEE Transactions on Aerospace and Electronic Systems,* Vol. 22, #1, 115-127.

Raghavan, R. S., Qiu, H. E., and McLaughlin, D. J. (1995). CFAR detection in clutter with unknown correlation properties, *IEEE Transactions on Aerospace and Electronic Systems,* Vol. 31, #2, 647-657.

Wang, H. and Cai, L. (1990) On adaptive multiband signal detection with the SMI algorithm, *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 26,. #5, 768-773.

Wicks, M. (1993). Applications of ranking and selection theory to radar signal processing, Technical Report, Rome Laboratory.

AUTOMATIC MOVING TARGETS DETECTION AND ESTIMATION

USING CONTAMINATED DATA

Julian Cheung
Associate Professor

Department of Electrical Engineering
and
Department of Computer Science

New York Institute of Technology
1855 Broadway
New York, NY 10023

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory/IRRE
32 Hangar Road
Griffiss AFB, NY 13441-4114

September 1995

AUTOMATIC MOVING TARGETS DETECTION AND ESTIMATION USING CONTAMINATED DATA

Julian Cheung
Associate Professor

Department of Electrical Engineering
and
Department of Computer Science

ABSTRACT

In this paper the analysis of variance (ANOVA) is extended to the representation of multi-frame images. The methodology of the optimal detection of moving objects embedded in noisy environments is presented, yielding ultimately a closed form solution. The new detector makes use of the information of the spatial and temporal effects, and is uniformly most powerful (UMP) in Gaussian environment with unknown and time-varying noise variance.

The new detector is primarily an optical-flow based detector without, however, the drawbacks of the latter; it uses advantageously a feature-based approach to improve the efficiency of computation. The versatility of the proposed detector is demonstrated by the fact that, with an adjustment of system parameters, it is applicable to multi-spectral detection and stationary-objects detection. It is also shown that all well-known detectors, generally classified as feature-based or optical-flow-based, become subclasses of the new detector if the proposed detector is subjected to more restrictive conditions.

Performance of the proposed detector is fully supported by extensive real image simulations.

AUTOMATIC MOVING TARGETS DETECTION AND ESTIMATION USING CONTAMINATED DATA

Julian Cheung

1. INTRODUCTION

The detection of moving objects using a sequence of time frames has been of major concern throughout the evolution of computer vision, because of its broad spectrum of applications, i.e., biomedicine, tactical and strategic military applications, industrial automation, meteorology, and data compression.

Because of these efforts, a multitude of moving-object detection algorithms have been developed. Most of them can be categorized into two classes. The first class, generally based on feature-based techniques, involves the extraction of a set of relatively sparse (less than 10%), but highly discriminatory tokens such as edges, corners, and sharp changes in curvature characterizing objects. Moving objects detection problem is thus reduced to the correspondence of the tokens between the consecutive frames [1]-[4]. The second class is based on computing the optical flow, i.e., the 2-dimensional (2-D) field of instantaneous velocities of gray levels at every pixel in the image. By the rigidity constraint that all pixels of an object move with the same velocity, the optical flow segments the image scene into distinct regions associated with the moving objects. Their identification is achieved using some similarity measures [5]-[8].

The principal problem with the first approach is that it depends on the token attribute which is highly affected by the modeling of multi-frame images. Also, the observed image tokens may not correspond to the object tokens, probably because of segmentation errors or occlusion caused by components of the background scene. Furthermore, establishing correspondence between sets of tokens between frames is not a trivial matter.

The main drawback of the second method is that since derivative operation accentuates noise, optical flow technique is highly susceptible to noise influence. As is well-known, in many circumstances image flow does not correspond to optical flow, the common practice of computing image flow (from data) and treat it as the optical flow then becomes questionable. Although to a lesser extent, occulatory effects also may play havoc with object detection.

As is valid in almost all computer vision problems, an over-simplistic or, worse of all, improper modeling, of an image when applied to any of the approaches described above may result in that "we often assume that the scene is composed of surfaces each of which has a simple shape ··· in a simple way. ··· As a result, standard vision techniques often do not perform very well when applied to natural scenes" [9]. Also, most methods are primarily developed for cases that involve either low noise or noise-free images. These idealized conditions are increasingly invalid in the present era of wide usage of the information super-highway. This is a handicap in itself due to a rather high number of false alarms that results in applying such techniques in any actual detection problems.

This motivates the present research, which is directed toward the development of a moving object detector that is uniformly most powerful (UMP) in Gaussian noise environments of unknown

and varying variance noise. The paper is organized as follows.

In section II, we model a scene pixel as contributed by the effects of rows, columns, diagonals, and frames, thus permitting all local pixels to have a reasonably large variation in gray levels. These effects are then estimated from data. In section III, the utilization of local contrasts (per frame), as applied to multi-frame images in edge detection is derived. This ultimately leads to a moving-edge detector which, as shown in Section IV, is based on the difference of local contrasts between two frames. In section V, the optimal detection of moving objects, based on the generalized likelihood-ratio test (GLRT), is derived. The duality of the GLRT and the multi-comparison test is subsequently established, and the generality of the proposed detector is explained. An implementation algorithm and computer simulation results based on real noisy images are considered in section VI. Further discussions and suggestions are presented in section VII.

II. THREE-D GRÆCO-LATIN SQUARE DESIGN FOR DYNAMIC SCENES

A Græco-Latin square (GLS) model is a four-way design which contains four treatments groups with I levels each. Consider the layout of the treatments in the first frame of a sequence of dynamic scenes, i.e., $m=1$ of a $I\times I$, $I=5$, GLS design illustrated in Fig.1. In this design, there are four treatment groups: the rows, the columns, the Tau's and the Lambda's. The Tau's (Lambda's) are arranged in a cyclic row (column) pattern representing 45° (135°) diagonal effects. The treatments are orthogonal in the sense that all treatments appear together only once in the entire design [10]. Let $\{\alpha_{1;i}\}_{i=1}^{I}$, $\{\gamma_{1;j}\}_{j=1}^{I}$, $\{\tau_{1;k}\}_{k=1}^{I}$, $\{\lambda_{1;l}\}_{l=1}^{I}$ be the row, column, 45° diagonal, 135° diagonal effects, respectively, in frame 1. Each pixel value is represented by $y_{1;ij} = \mu + \zeta_1 + \alpha_{1;i} + \gamma_{1;j} + \tau_{1;k} + \lambda_{1;l} + \epsilon_{1;ij}$, $i,j,k,l \in D_1$, and D_1 is the set of I^2 observations in frame 1. Here μ denotes the grand mean of all frames, ζ_1 the mean effect of frame 1; $\epsilon_{1;ij}$ is the uncertainty (noise) associated with the ith row, jth column observation that is Gaussian distributed with mean 0, variance σ^2, and is pair-wise statistically independent on the same frame.

A three-dimensional (3-D) GLS design is an outgrowth of the regular GLS design with the additional dimension pertains to the frame number among a sequence of registered scenes. The parametric model of a multi-frame scene becomes

$$\Omega: \quad y_{m;ij} = \mu + \zeta_m + \alpha_{m;i} + \gamma_{m;j} + \tau_{m;k} + \lambda_{m;l} + \epsilon_{m;ij}, \quad i,j,k,l \in D_m;\ m=1,\cdots,M. \tag{2.1}$$

Here D_m is the set of I^2 observations corresponding to mth frame. Define $D = \{D_1,\cdots,D_m\}$, then D is a collection of M frames of observations. Of course, the orthogonality of the treatments extends readily to the collection of frames.

Since the parameters μ, $\{\zeta_m\}$, $\{\alpha_{m;i}\}$, $\{\gamma_{m;j}\}$, $\{\tau_{m;k}\}$, $\{\lambda_{m;l}\}$ are generally unknown, they are estimated from the data. This is attained by forming the sum of squares

$$S = \sum_{m,i,j,k,l \in D} \{y_{m;ij} - \mu - \zeta_m - \alpha_{m;i} - \gamma_{m;j} - \tau_{m;k} - \lambda_{m;l}\}^2 . \tag{2.2}$$

Taking the derivative of S with respect to μ and equating it to zero to obtain

$$MI^2\mu = \sum_{m=1}^{M}\sum_{i=1}^{I}\sum_{j=1}^{I} y_{m;ij} - I^2\sum_{m}\zeta_m - I\sum_{m}\sum_{i}\alpha_{m;i} - I\sum_{m}\sum_{j}\gamma_{m;j} - I\sum_{m}\sum_{k}\tau_{m;k} - I\sum_{m}\sum_{l}\lambda_{m;l}, \text{ with}$$

the resulting least-squares estimate of μ as

$$\hat{\mu} = \mu = y_{\cdot;\cdot\cdot} - \zeta_{\cdot} - \alpha_{\cdot;\cdot} - \gamma_{\cdot;\cdot} - \tau_{\cdot;\cdot} - \lambda_{\cdot;\cdot} \,. \tag{2.3}$$

The convenient dot notation indicates that whenever a subscripted variable is replaced by a dot it means that unweighted averaging operation has been taken place over the affected variable, i.e., $y_{m;i\cdot} = I^{-1}\Sigma_j\, y_{m;ij}$, $y_{m;\cdot\cdot} = I^{-1}\Sigma_i\, y_{m;i\cdot}$, $y_{\cdot;\cdot\cdot} = M^{-1}\Sigma_m\, y_{m;\cdot\cdot}$, $\zeta_{\cdot} = M^{-1}\Sigma_m\,\zeta_m$, $\alpha_{m;\cdot} = I^{-1}\Sigma_i\,\alpha_{m;i}$, $\alpha_{\cdot;\cdot} = M^{-1}\Sigma_m\,\alpha_{m;\cdot}$, etc.

Analogously, taking the derivative of S with respect to ζ_m, $\alpha_{m;i}$, $\gamma_{m;j}$, setting each of them to zero and recognizing that for fixed m and i, $\Sigma_{(j,k,l)\in D_m}\tau_{m;k} = \Sigma_k\,\tau_{m;k}$, $\Sigma_{(j,k,l)\in D_m}\lambda_{m;l} = \Sigma_l\,\lambda_{m;l}$ while for fixed m and j, $\Sigma_{(i,k,l)\in D_m}\tau_{m;k} = \Sigma_k\,\tau_{m;k}$, $\Sigma_{(i,k,l)\in D_m}\lambda_{m;l} = \Sigma_l\,\lambda_{m;l}$ (see Fig.1), yields

$$\hat{\zeta}_m = \zeta_m = y_{m;\cdot\cdot} - \mu - \alpha_{m;\cdot} - \gamma_{m;\cdot} - \tau_{m;\cdot} - \lambda_{m;\cdot} \tag{2.4}$$

$$\hat{\alpha}_{m;i} = \alpha_{m;i} = y_{m;i\cdot} - \mu - \zeta_m - \gamma_{m;\cdot} - \tau_{m;\cdot} - \lambda_{m;\cdot} \tag{2.5}$$

$$\hat{\gamma}_{m;j} = \gamma_{m;j} = y_{m;\cdot j} - \mu - \zeta_m - \alpha_{m;\cdot} - \tau_{m;\cdot} - \lambda_{m;\cdot} \tag{2.6}$$

respectively.

By the same token, upon taking the derivative of S with respect to $\tau_{m;k}$, $\lambda_{m;l}$, setting each of them to zero, and simplifying, yields

$$\hat{\tau}_{m;k} = \tau_{m;k} = I^{-1}\sum_{i=1}^{I} y_{m;i\tilde{j}} - \mu - \zeta_m - \alpha_{m;\cdot} - \gamma_{m;\cdot} - \lambda_{m;\cdot} \tag{2.7}$$

$$\hat{\lambda}_{m;l} = \lambda_{m;l} = I^{-1}\sum_{i=1}^{I} y_{m;i\breve{j}} - \mu - \zeta_m - \alpha_{m;\cdot} - \gamma_{m;\cdot} - \tau_{m;\cdot} \,, \tag{2.8}$$

where $\tilde{j}$ equals $k+1-i$ if $k+1>i$, it equals $k+1-i+I$ otherwise; also $\breve{j}$ equals $i+l-1$ if $i+l-1\le I$, $\breve{j}$ equals $i+l-1-I$ otherwise.

According to (2.1), $\alpha_{m;i}$ being the row treatment effect is independent of other directional effects. This is also valid for its least-squares estimate $\hat{\alpha}_{m;i}$ and has the implication that the last three terms in (2.5) vanish. The same reasoning leads to the elimination of the last three terms in (2.6-8). Furthermore, μ is the grand mean of all the available frames and it is independent of $\{\zeta_{\cdot}\}$, $\{\alpha_{\cdot;\cdot}\}$, $\{\gamma_{\cdot;\cdot}\}$, $\{\tau_{\cdot;\cdot}\}$, $\{\lambda_{\cdot;\cdot}\}$. Therefore, the term $\zeta_{\cdot}$ disappears. In brief, in the 3-D GLS design a set of natural side conditions are

$$\begin{aligned} \zeta_{\cdot} &= M^{-1}\sum_{m=1}^{M}\zeta_m = 0 \\ \alpha_{m;\cdot} &= I^{-1}\sum_{i=1}^{I}\alpha_{m;i} = 0, \quad \gamma_{m;\cdot} = I^{-1}\sum_{j=1}^{I}\gamma_{m;j} = 0 \\ \tau_{m;\cdot} &= I^{-1}\sum_{k=1}^{I}\tau_{m;k} = 0, \quad \lambda_{m;\cdot} = I^{-1}\sum_{l=1}^{I}\lambda_{m;l} = 0, \qquad m=1,\cdots,M. \end{aligned} \tag{2.9}$$

Therefore, the least-squares estimates (2.3)-(2.8) can be simplified as

$$\begin{aligned} \hat{\mu} &= y_{;..}\ , & \hat{\zeta}_m &= y_{m;..} - y_{;..} \\ \hat{\alpha}_{m;i} &= y_{m;i.} - y_{m;..}\ , & \hat{\gamma}_{m;j} &= y_{m;.j} - y_{m;..} \\ \hat{\tau}_{m;k} &= I^{-1}\sum_{i=1}^{I} y_{m;i\tilde{j}} - y_{m;..}\ , & \hat{\lambda}_{m;l} &= I^{-1}\sum_{i=1}^{I} y_{m;i\check{j}} - y_{m;..} \end{aligned} \tag{2.10-15}$$

respectively, and $\tilde{j}$, $\check{j}$ have been defined before.

For subsequent analysis, introduce a $p \times 1$ parameter vector $\boldsymbol{\beta} = [\,\mu\ \boldsymbol{\zeta}\ \boldsymbol{\alpha}\ \boldsymbol{\gamma}\ \boldsymbol{\tau}\ \boldsymbol{\lambda}\,]'$, $p=1+M+4MI$, with the sub-parameter vectors $\boldsymbol{\zeta}^{M\times 1} = [\,\zeta_1 \cdots \zeta_M\,]'$, $\boldsymbol{\alpha}^{(MI)\times 1} = [\,\alpha_{1;1} \cdots \alpha_{1;J} \cdots\cdots \alpha_{M;1} \cdots \alpha_{M;I}\,]'$, $\boldsymbol{\gamma}^{(MI)\times 1} = [\,\gamma_{1;1} \cdots \gamma_{M;I}\,]'$, $\boldsymbol{\tau}^{(MI)\times 1} = [\,\tau_{1;1} \cdots \tau_{M;I}\,]'$, $\boldsymbol{\lambda}^{(MI)\times 1} = [\,\lambda_{1;1} \cdots \lambda_{M;I}\,]'$. This permits a compact representation of the parametric model (2.1) as

$$\Omega: \qquad \boldsymbol{y}^{n\times 1} = \boldsymbol{X}'\boldsymbol{\beta}^{p\times 1} + \boldsymbol{\epsilon}^{n\times 1}, \tag{2.16}$$

where $\boldsymbol{y}^{n\times 1} = [\,y_{1;11} \cdots y_{1;1J} \cdots\cdots y_{1;I1} \cdots y_{1;IJ} \cdots\cdots y_{M;11} \cdots \gamma_{M;IJ}\,]'$ is the observation vector formed by the row stacking of the sequential frames, $n = MI^2$, $J = I$, and $\boldsymbol{\epsilon}^{n\times 1}$ is the corresponding noise vector with the distribution $N(\mathbf{0}, \sigma^2\boldsymbol{I})$. In the $n\times p$ four-way effect matrix $\boldsymbol{X}'$, each element equals 1 or 0 depending on whether the associated level of a particular treatment group is present or not.

For any 3-D GLS design, $\boldsymbol{X}'$ is known *a priori* while $\boldsymbol{\beta}$ may be estimated by the least squares approach. However, in general $\boldsymbol{X}'$ is not of full rank. In fact, the $n\times p$ effect matrix $\boldsymbol{X}'$ has the rank r, $r=p-t$, $p=1+M+4MI$, $t=1+4M$. Consequently, the least-square estimate $\hat{\boldsymbol{\beta}}$ is not unique. This poses a lot of theoretical and implementation problems because, as will be clear later, statistic for moving object detection is associated with the linear transformation of the parametric vector $\boldsymbol{\beta}$ which is supposedly unique. To circumvent this dilemma, we consider

Theorem 1 (Unique parameter-vector estimate) Suppose in the parametric model $\boldsymbol{y}^{n\times 1} = \boldsymbol{X}'\boldsymbol{\beta}^{p\times 1} + \boldsymbol{\epsilon}$ the $n\times p$ effect matrix $\boldsymbol{X}'$ has rank r, $r<p<n$. By imposing the regularity conditions i) a set of $t = p - r$ linear restrictions on $\boldsymbol{\beta}$, i.e., $\boldsymbol{H}'\boldsymbol{\beta} = \mathbf{0}^{t\times 1}$, where $\boldsymbol{H}'$ is $t\times p$ with rank $\boldsymbol{H}' = t$, and ii) no rows of $\boldsymbol{H}'$ is a linear combination of rows of $\boldsymbol{X}'$ except $\mathbf{0}$, then a unique least-squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \boldsymbol{B}\boldsymbol{y}, \quad \boldsymbol{B} = (\boldsymbol{X}\boldsymbol{X}' + \boldsymbol{H}\boldsymbol{H}')^{-1}\boldsymbol{X}. \tag{2.17}$$

Proof: Refer to [12].

Corollary 1.1 Transferring the inverting matrix in (2.17) to the l.h.s. and utilizing the identity $\boldsymbol{H}'\boldsymbol{\beta} = \mathbf{0}$ we get $(\boldsymbol{X}\boldsymbol{X}' + \boldsymbol{H}\boldsymbol{H}')\hat{\boldsymbol{\beta}} = \boldsymbol{X}\boldsymbol{X}'\hat{\boldsymbol{\beta}} = \boldsymbol{X}\boldsymbol{y}$. Replacing $\boldsymbol{y}$ by (2.16) and upon taking the expectation, then $\boldsymbol{X}\boldsymbol{X}'E(\hat{\boldsymbol{\beta}}) = \boldsymbol{X}E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{X}'\boldsymbol{\beta}$, so that $\hat{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}$. Since for normal distribution least squares estimate, mean square (minimum variance) estimate and maximum likelihood estimate coincide [11], $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimate (BLUE) of the unknown $\boldsymbol{\beta}$.

Corollary 1.2 The least-squares sum can be simplified to $S_e = (\boldsymbol{y} - \boldsymbol{X}'\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}'\hat{\boldsymbol{\beta}}) = \boldsymbol{y}(\boldsymbol{y} - \boldsymbol{X}'\hat{\boldsymbol{\beta}})$. Applying Theorem 1, we obtain the least squares sum

$$S_e = y' Q' Q y, \quad Q = I - X'B. \tag{2.18}$$

In reality, the pixel variance σ^2 is unknown. However, the theorem below indicate the relationship between S_e and the estimate $\hat{\sigma}^2$ of σ^2.

Theorem 2 (unbiased estimate of σ^2) The ratio $s^2 = S_e / (n-r)$ is the BLUE of the unknown variance σ^2.

Proof: Refer to [12].

III. MULTI-FRAME SEGMENTATION

Ideally, an edge can be defined as an abrupt transition from one homogeneous region, in terms of gray levels, to another homogeneous region. This step-like distribution of an edge enables its detection using derivative-based local operators, including Sobel, Kirsch, etc. [13]. However, the available data set is usually corrupted by noise which results in blurred edges and false features. Also, an edge is sometimes associated with a line transition which further complicates the edge detection process. As such, it is advantageous to use a statistical technique that would provide both robustness as well as discrimination of lines in the course of detecting possible edges.

For ease of development, we will focus on the detection of vertical edges. Consider a 5×5 rectangular mask that scans the first five rows of frame 1, horizontally and starting with the first 5 columns, then the next 5 columns, and so forth. When it hits the last column it makes a zigzag return and repeats the same procedure on the next 5 rows, etc. Each pixel's intensity is due to a combination of various effects which we call treatments. In a 5×5 region each treatment has only a very limited number of values which we call levels. This suggest the GLS design (2.1) is a suitable model for the representation of gray levels on a local basis. The edges of frame 1 (taken at time t_1) that is within the mask shown in Fig.1 can be expressed using the following contrast functions

$$\begin{aligned}
\xi^{(e)}_{1;1} &= \gamma_{1;1} - \tfrac{1}{4}(\gamma_{1;2} + \gamma_{1;3} + \gamma_{1;4} + \gamma_{1;5}) \\
\xi^{(e)}_{1;2} &= \tfrac{1}{2}(\gamma_{1;1} + \gamma_{1;2}) - \tfrac{1}{3}(\gamma_{1;3} + \gamma_{1;4} + \gamma_{1;5}) \\
\xi^{(e)}_{1;3} &= \tfrac{1}{3}(\gamma_{1;1} + \gamma_{1;2} + \gamma_{1;3}) - \tfrac{1}{2}(\gamma_{1;4} + \gamma_{1;5}) \\
\xi^{(e)}_{1;4} &= \tfrac{1}{4}(\gamma_{1;1} + \gamma_{1;2} + \gamma_{1;3} + \gamma_{1;4}) - \gamma_{1;5} \cdot
\end{aligned} \tag{3.1}$$

Under no noise situations, if there is no edge (indicating homogeneity), each of the contrast functions equals 0. On the other hand, if an edge, for example, exists between columns 2 and 3, then $\xi_{1;2}$ attains the highest amplitude among all contrast functions. Therefore, the contrast functions approach results in enhancing real edges and suppressing false edges. In the presence of noise, the edge detection problem can be expressed as a pair of hypotheses testing $H_0^{(e)}$: $\xi_{1;1} = \xi_{1;2} = \cdots = \xi_{1;4}$ against $H_1^{(e)}$: $!(\xi_{1;1} = \xi_{1;2} = \cdots = \xi_{1;4})$. Rewrite $\xi_{1;i}$ as $\xi_{1;i} = c_{1;i}\, \boldsymbol{\beta}^{p\times 1}$, $c_{1;i} = [\, c_{1;i1} \cdots c_{1;ip}\,]$, to explicitly indicate that every contrast function is a linear combination of the parametric vector $\boldsymbol{\beta}$ with the

corresponding contrast coefficient vector (3.1). For example, we have $c_{1;2,1+M+I+1} = c_{1;2,1+M+I+2} = 1/2$, $c_{1;2,1+M+I+3} = \cdots = c_{1;2,1+M+2I-1} = -1/3$, and $c_{1;2j} = 0$ otherwise. Define the contrast vector $\xi_1 = [\,\xi_{1;1} \cdots \xi_{1;q}\,]'$ and the contrast coefficient matrix $C_1 = [\,c'_{1;1}\ \ c'_{1;2} \cdots\ \ c'_{1;q}\,]'$, $q = 4$, the set of contrast functions in frame 1 is

$$\xi_1 = C_1^{q\times p}\,\beta^{p\times 1}. \tag{3.2}$$

The edge detection problem can now be reformulated as

$$\begin{aligned} H_0^{(e)}&: \quad \xi_1 = 0^{q\times 1} \\ H_1^{(e)}&: \quad \xi_1 \neq 0^{q\times 1}\ . \end{aligned} \tag{3.3}$$

As discussed before, the parameter vector β is unknown, and so ξ_1 is replaced by its estimate $\hat{\xi}_1$. Since $\xi_1 = C_1\,\beta$ is a linear combination of β, by the Gauss-Markov theorem [14, pp.14-15] the BLUE of $\hat{\xi}_1$ is $\hat{\xi}_1 = C_1\,\hat{\beta}$ among all linear estimates, on the condition that $\hat{\beta}$ is the BLUE of β which, we recall, has been established in Theorem1. Consequently, an efficient estimate of the contrast vector is

$$\hat{\xi}_1 = C_1 B y = A_1 y, \tag{3.4}$$

where $A_1 = C_1 B$, and B is defined in Theorem 1.

Since y is normally distributed, its linear transformation $\hat{\xi}_1 = A_1 y$ is also normally distributed, yielding $\hat{\xi}_1 \sim N(\xi_1, \sigma^2 E_1)$. Applying Scheffé's multiple comparison theorem [15], for any ξ_1 in the q-D contrast space spanned by the basis vectors $\{\xi_{1;1}, \cdots, \xi_{1;q}\}$ its BLUE $\hat{\xi}_1$ falls in the confidence ellipsoid

$$(\xi_1 - \hat{\xi}_1)' E_1^{-1} (\xi_1 - \hat{\xi}_1) \le q s^2 F_{\alpha;q,n-r} \tag{3.5}$$

with the probability $1 - \alpha$. Here $q = \text{rank}\,(\xi_1) = 4$, $n = MI^2$, $r = \text{rank}\,(X')$.

Returning to (3.3), the no edge phenomenon is identical to the situation that every contrast function $\{\xi_{1;i}\}_{i=1}^{q}$ is null. Therefore, every contrast function in the q-D contrast space vanishes. Under H_0 the confidence ellipsoid becomes $\hat{\xi}_1' E_1^{-1} \hat{\xi}_1 \le q s^2 F_{\alpha;q,n-r}$. Testing statistic for (3.3), at the α-level of significance, is

$$\hat{\xi}_1' E_1^{-1} \hat{\xi}_1 \underset{H_0^{(e)}}{\overset{H_1^{(e)}}{\gtrless}} q s^2 F_{\alpha;q,n-r}. \tag{3.6}$$

The rejection of $H_0^{(e)}$ indicates there exists an edge (at least one) in the underlying mask. To locate the edge we may choose the $\hat{\xi}_{1;i}$ corresponding to the highest absolute value among all q contrast functions, since it is the most likely candidate responsible for the rejection of H_0.

The edge detector above can only detect and locate vertical edges and is of very limited

practical use. We thus introduce a modified edge detector $\xi_1^{(1)} = \left(\xi_1^{(r)}, \xi_1^{(d)}, \xi_1^{(c)}, \xi_1^{(s)}\right)$, sensitive to four major directions (in frame 1). By using the information in $\xi_1^{(r)}$, $\xi_1^{(d)}$, $\xi_1^{(c)}$, $\xi_1^{(s)}$, it detects edge features oriented in 0°, 45°, 90°, 135°, respectively. Let $\left\{ z_{1;ij}^{(\pi)} \mid i=1,\cdots,I;\ j=1,\cdots,J \right\}$ be elements of the π-directional mask corresponding to $\xi_1^{(\pi)}$, $\pi = r,d,c,s$ and is associated with the physical data $\{y_{1;ij} \mid i=1,\cdots,I;\ j=1,\cdots,J\}$ by

$$z_{1;ij}^{(r)} = y_{1;ji}, \quad z_{1;ij}^{(d)} = y_{1;i,I+j-i}$$
$$z_{1;ij}^{(c)} = y_{1;ij}, \quad z_{1;ij}^{(s)} = y_{1;i,i+j-1}\ . \tag{3.7-10}$$

In short, the theory of edge detection, discussed hitherto, can be applied to π-directional detection, with $\{y_{1;ij}\}$ replaced by $\{z_{1;ij}^{(\pi)}\}$. The foregoing procedure is applied analogously to the detection of edges in frame m (taken at t_m) with, however, the subscript 1 replaced by m, $m=1, \cdots, M$.

IV. DETECTION OF MOVING EDGES

Consider a rigid body moving in the easterly direction at a rate slower than the shuttering speed of a camera mounted on a stationary platform. Without loss of generality, assume the body is larger than the processing mask (typically 5×5). If it is not, one may decrease the mask size and increase the FAR for a satisfactory power of detection. Suppose at time t_1 a region ($\Re$ hereafter) that covers the object including its leftmost boundary and occupies the mask from column 3 onward. From the point of view of the mask, pixels between columns 3 and 5 belong to the object, while the area from columns 1 to 2 belong to the background. The change of gray intensity between columns 2 and 3 constitutes an edge and is a subset of the outermost contour of the object.

Suppose at t_2 the object moves eastward by one pixel. The rigidity of the object mandates that all pixels of the object move by the same amount and in the same direction, implying that $\Re$ now fills columns 4 to 5. In brief, object movement translates the vertical edge from column 3 to column 4 in the time interval $(t_1, t_2]$. On the other hand, if the object is stationary, its vertical edge stays the same from one mask to the next. This suggests the detection of edge movement is crucial to the discrimination of dynamic objects from stationary counterparts. The mechanism is consistent with the operation of the human visual system (HVS), which differentiates a moving object from a stationary object by comparing whether the contour (possibly closed) of the underlying object moves or not relative to a human observer from one frame to the next (t_1 to t_2). The detection of vertical-edge movement may thus be described verbally as $H_0^{(m)}$: *no vertical-edge movement within masks of two successive frames on the same neighborhood* versus $H_1^{(m)}$: *at least one vertical edge moves in the masks.*

Introduce a set of composite contrast functions

$$\psi_i = \xi_{1;i} - \xi_{2;i}, \quad i=1, \cdots, q. \tag{4.1}$$

With $\xi_{m;i}$ defined in (3.1), $m = 1, 2$, then $\psi_i = \boldsymbol{d}_i \boldsymbol{\beta}$ and so every composite contrast function is a linear combination of the parameter vector $\boldsymbol{\beta}$ with the corresponding composite contrast coefficient vector $\boldsymbol{d}_i = \boldsymbol{c}_{1;i} - \boldsymbol{c}_{2;i} = [\, d_{i1} \cdots d_{ip} \,]$, $d_{ij} = c_{1;ij} - c_{2;ij}$, $j = 1, \cdots, p$. Applying (3.1) to get $\sum_{j=1}^{p} d_{ij} = 0$ and, in conjunction with the linear independence of ψ_{i_1}, ψ_{i_2}, $i_1 \neq i_2$, every composite contrast function is a linear contrast function [14]. Thus $\psi_1, \cdots, \psi_q$ are the basis vectors spanning the q-D composite contrast space ω_{cc} and, in particular, the composite contrast vector $\boldsymbol{\psi}$, $\boldsymbol{\psi} = [\, \psi_1 \cdots \psi_q \,]'$, belongs to ω_{cc}. Rewrite $\boldsymbol{\psi}$ as

$$\boldsymbol{\psi} = \boldsymbol{D}\,\boldsymbol{\beta}, \tag{4.2}$$

where $\boldsymbol{D}^{q \times p} = [\, \boldsymbol{d}_1' \cdots \boldsymbol{d}_q' \,]'$ is the corresponding composite contrast coefficient matrix. Applying Theorem 1 to obtain the unbiased estimate $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$, we have

$$\hat{\boldsymbol{\psi}} = \boldsymbol{D}\,\boldsymbol{B}\,\boldsymbol{y}. \tag{4.3}$$

Making use of the Gauss-Markov theorem [14] together with the recognition that $\boldsymbol{\psi}$ is a linear combination of $\boldsymbol{\psi}$, $\hat{\boldsymbol{\psi}}$ is the BLUE of $\boldsymbol{\psi}$, i.e. $\hat{\boldsymbol{\psi}} \sim N(\boldsymbol{\psi}, \sigma^2 \boldsymbol{K})$, $\boldsymbol{K} = (\boldsymbol{D}\,\boldsymbol{B})(\boldsymbol{D}\,\boldsymbol{B})'$.

Returning to (4.1), ψ_i is a difference of two contrasts each of them is the column *contrast* of two adjacent regions in a frame. Clearly, ψ_i equals zero whenever: i) there is no edge at column i in the same neighborhood in the frames, or ii) if an edge exists at column i in the frames (static object). Therefore, when $\boldsymbol{\psi}$ covers the origin, it signifies no shifts in vertical edges (locally). The hypothesis testing for vertical edge shifting can thus be mathematically expressed as

$$\begin{aligned} H_0^{(m)}&: \quad \boldsymbol{\psi} \equiv \boldsymbol{D}^{q \times p}\,\boldsymbol{\beta}^{p \times 1} = \boldsymbol{0}^{q \times 1} \\ H_1^{(m)}&: \quad \boldsymbol{\psi} \neq \boldsymbol{0}^{q \times 1} \,. \end{aligned} \tag{4.4}$$

Because of the q-D composite contrast space as established in the previous paragraph, the implementation of (4.3) is identical to (3.6); it is rewritten here as

$$\hat{\boldsymbol{\psi}}'\,\boldsymbol{K}^{-1}\,\hat{\boldsymbol{\psi}} \underset{H_0^{(m)}}{\overset{H_1^{(m)}}{\gtrless}} q s^2 F_{\alpha;q,n-r}. \tag{4.5}$$

The detection of four-directional edge movement is developed analogously to edge detection.

V. DETECTION OF MOVING OBJECTS

a) Generalized Likelihood Test

Consider in frame 1 a dynamic but rigid object that encloses a maximal rectangular mask of dimension $I \times J$. Suppose the mask is aligned with the frame and the top left corner is at location (x_1, y_1) of the latter. Since the mask may no longer be a square, the diagonal effects exhibited in a GLS design have no physical meaning, (2.1) is modified to represent mask pixels, for $m = 1$, as

$$\Omega: \quad y_{m;ij} = \mu + \zeta_m + \alpha_{m;i} + \gamma_{m;j} + \epsilon_{m;ij}, \quad i = 1, \cdots, I; j = 1, \cdots, J, \tag{5.1}$$

where the same terminology as before is used.

Consider a second mask of the same size aligned with frame 2 in a manner its top left corner is at (x_2, y_2) in frame 2. Model (5.1) is also valid for the second mask ($m = 2$). Pixel gray levels in the aforementioned masks (aligned at different positions on two successive scenes) are described in terms of row, column, frame effects, and can be represented compactly by $\boldsymbol{\beta}^{p\times 1}$, which is defined in (2.16). From the parametric model (5.1) $\boldsymbol{\beta}^{p\times 1} \in \Omega$, where $p = 1$ (grand mean) + 2 (frame effects) + $2I$ (row effects) + $2J$ (column effects). The corresponding effect matrix $\boldsymbol{X}'$ is similar to (2.1) with, however, no diagonal treatment effects. Due to the displacement of the masks, the matching of them, i.e., effects in one mask equal to that in the other, implies the existence of a moving object. This results in a sub-parametric space ω, $\omega = \Omega \cap (\zeta_1=\zeta_2, \alpha_{1;1}=\alpha_{2;1}, \cdots, \alpha_{1;I}=\alpha_{2;I}, \gamma_{1;1}=\gamma_{2;1}, \cdots, \gamma_{1;J}=\gamma_{2;J})$, $\omega \subset \Omega$, with the corresponding parameter vector $\boldsymbol{\beta}_\omega$. Henceforth, all associated parameters tantamount to the same restriction are emphasized so through the subscript ω. Under the parametric space ω, the frame effects ζ_1, ζ_2 are the same and are absorbed into the grand mean μ. Also, since $\alpha_{1;i} = \alpha_{2;i}$, they are denoted by a common variable α_i, $i = 1, \cdots, I$; similar reasoning leads to the new variable β_j, $j = 1, \cdots, J$, yielding a $p_\omega \times 1$ parameter vector $\boldsymbol{\beta}_\omega$, $p_\omega = 1 + I + J$. The corresponding effect matrix $\boldsymbol{X}'_\omega$ is changed accordingly. It then follows that the detection of a dynamic object corresponds to being able to discriminate whether $\boldsymbol{\beta}$ or $\boldsymbol{\beta}_\omega$ better describe the data. The detection of a moving object can thus be formulated as

$$\begin{aligned} H_0&: \quad \boldsymbol{y} = \boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ H_1&: \quad \boldsymbol{y} = \boldsymbol{X}'_\omega \boldsymbol{\beta}_\omega + \boldsymbol{\epsilon}, \end{aligned} \tag{5.2}$$

where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. As explained before, $\boldsymbol{X}'$ is of $n \times p$ with rank $\boldsymbol{X}' = r$, $r = p - t$, $t = 1 + 2 + 2$; while $\boldsymbol{X}'_\omega$ is of $n \times p_\omega$ and rank $\boldsymbol{X}_\omega = r_\omega$, $r_\omega = p_\omega - t_\omega$, $p_\omega = 1 + I + J$, $t_\omega = 2$.

Defining $p_{\boldsymbol{Y}|H_i}(\boldsymbol{y}|H_i)$ as the conditional probability density function (pdf) of a stochastic image $\boldsymbol{Y}^{2IJ\times 1}$ (row scanning of a 3-D mask) with the realization $\boldsymbol{y}^{2IJ\times 1}$ under hypothesis H_i, i=0,1, the generalized likelihood statistic for (5.2) is

$$\lambda(\boldsymbol{y}) = \frac{\max_\omega p_{\boldsymbol{Y}|H_1}(\boldsymbol{y}|H_1)}{\max_\Omega p_{\boldsymbol{Y}|H_0}(\boldsymbol{y}|H_0)} \underset{H_0}{\overset{H_1}{>\!<}} \lambda_0 \tag{5.3}$$

and the testing statistic rejects H_0 if $\lambda(\boldsymbol{y}) > \lambda_0$, where the constant λ_0 is chosen to give the desired significance level α [16].

Under H_0, the multivariate conditional pdf is

$$p_{\boldsymbol{Y}|H_0}(\boldsymbol{y}|H_0) = (2\pi\sigma^2)^{-\frac{1}{2}n} \exp\{-\tfrac{1}{2} S(\boldsymbol{y}, \boldsymbol{\beta})/\sigma^2\}, \tag{5.4}$$

and $S_\Omega(\boldsymbol{y}, \boldsymbol{\beta})$ the sum of squares error

$$S(\boldsymbol{y}, \boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\beta}) = \| \boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\beta} \|^2. \tag{5.5}$$

Since, practically speaking, $\boldsymbol{\beta}$, σ^2 are unknown, $p_{Y|H_0}(y|H_0)$ cannot be evaluated. To circumvent this difficulty, we replace the unknown parameters by their MLE's [16]. Apply (2.18) to obtain

$$S(y, \hat{\boldsymbol{\beta}}) = y' Q' Q y, \tag{5.6}$$

where $Q = I - X'B$, $B = \left(XX' + HH'\right)^{-1} X$. With the aid of Theorem 2, the maximal pdf is

$$\max_{\hat{\boldsymbol{\beta}}} p_{Y|H_0}(y|H_0) = \left(2\pi\,\hat{\sigma}^2\right)^{-\frac{1}{2}n} \exp\left\{-\tfrac{1}{2} S(y, \hat{\boldsymbol{\beta}}) / \hat{\sigma}^2\right\} = \left(2\pi \frac{S(y, \hat{\boldsymbol{\beta}})}{n-r}\right)^{-n/2} \exp\left\{-\tfrac{1}{2}(n-r)\right\}. \tag{5.7}$$

Similarly, under H_1 the least squares error and the corresponding maximal pdf are

$$S_\omega(y, \hat{\boldsymbol{\beta}}) = y' Q_\omega' Q_\omega y, \tag{5.8}$$

$$\max_{\hat{\boldsymbol{\beta}}_\omega} p_{Y|H_0}(y|H_0) = \left(2\pi \frac{S(y, \hat{\boldsymbol{\beta}})}{n-r}\right)^{-n/2} \exp\left\{-\tfrac{1}{2}(n-r) \frac{S_\omega(y, \hat{\boldsymbol{\beta}})}{S(y, \hat{\boldsymbol{\beta}})}\right\}, \tag{5.9}$$

respectively, where $Q_\omega = I - X_\omega' B_\omega$, $B_\omega = \left(X_\omega X_\omega' + H_\omega H_\omega'\right)^{-1} X_\omega$. Substituting (5.7),(5.9) into (5.3) and taking the natural logarithm, the generalized likelihood statistic is

$$\tfrac{1}{2}(n-r) \frac{S(y, \hat{\boldsymbol{\beta}}) - S_\omega(y, \hat{\boldsymbol{\beta}})}{S(y, \hat{\boldsymbol{\beta}})} \underset{H_0}{\overset{H_1}{>\!<}} \lambda_0 . \tag{5.10}$$

Since $\operatorname{rank} X_\omega' = I + J - 1 < 2I + 2J - 2 = \operatorname{rank} X'$, following the vector space concept of Theorem 2, it is obvious that $S(y, \hat{\boldsymbol{\beta}})$ utilizing higher order of basis vectors incurs a smaller error than $S_\omega(y, \hat{\boldsymbol{\beta}})$, implying that λ_0 is a negative value. Dividing both sides by $-\tfrac{1}{2}q$, we have

$$\lambda(y) = \frac{\{S_\omega(y, \hat{\boldsymbol{\beta}}) - S(y, \hat{\boldsymbol{\beta}})\}/q}{S(y, \hat{\boldsymbol{\beta}})/(n-r)} \underset{H_1}{\overset{H_0}{>\!<}} \lambda_0 , \tag{5.11}$$

where $q = r - r_\omega = I + J - 1$. Apparently, $\lambda(y)$ is a sufficient statistic that decides on H_1, i.e., the existence of moving object, whenever $\lambda(y) < \lambda_0$; otherwise it favors H_0, i.e., no moving objects in the 3-D mask. In the above the constant λ_0 is not necessarily the same from one line to the next.

For a given significance level α, λ_0 and α are related by

$$\mathcal{P}\left\{\frac{\{S_\omega(y, \hat{\boldsymbol{\beta}}) - S(y, \hat{\boldsymbol{\beta}})\}/q}{S(y, \hat{\boldsymbol{\beta}})/(n-r)} \text{ (under } H_0) < \lambda_0\right\} = 1 - \alpha . \tag{5.12}$$

Applying the vector space concept of [12] once more, it is not difficult to see that

$$\begin{aligned} S_\omega(y, \hat{\boldsymbol{\beta}}) &= \left(\textstyle\sum_{i=1}^{n} a_i \boldsymbol{\phi}_i - \sum_{i=q+1}^{r} c_i \boldsymbol{\phi}_i\right)' \left(\textstyle\sum_{j=1}^{n} a_j \boldsymbol{\phi}_j - \sum_{j=q+1}^{r} c_j \boldsymbol{\phi}_j\right) \\ &= \textstyle\sum_{i=1}^{q} a_i^2 + \sum_{i=q+1}^{r} (a_i - c_i)^2 + \sum_{i=r+1}^{n} a_i^2 . \end{aligned} \tag{5.13}$$

With the choice $a_i = c_i$, $i = q+1, \cdots, r$, under H_0, then

$$S_\omega(y, \hat{\beta}) / \sigma^2 \sim \chi_q^2 + \chi_{n-r}^2 = \chi_{q+n-r}^2 , \tag{5.14}$$

which simplifies (5.12) to

$$\mathcal{P}\left\{ \frac{\{\chi_{q+n-r}^2 - \chi_{n-r}^2\}/q}{\chi_{n-r}^2/(n-r)} < \lambda_0 \right\} = \mathcal{P}\{ F_{q,n-r} < \lambda_0 \} = 1 - \alpha , \tag{5.15}$$

where $F_{q,n-r} = \dfrac{\chi_q^2/q}{\chi_{n-r}^2/(n-r)}$ is Fisher's F-statistic with $q, n-r$ degrees of freedom (df). The threshold $\lambda_0 = F_{\alpha;\, q,n-r}$ can be found using the tabulated F-tables. The power of detection is evaluated similarly to [17].

b) Multi-Comparison Contrast Test

Let us introduce a contrast vector $\boldsymbol{\theta}^{q\times 1} = [\, \theta_1' \cdots \theta_q' \,]'$, $q = I + J - 1$, and the contrast functions

$$\theta_i = \begin{cases} \zeta_1 - \zeta_2, & i = 1 \\ \eta_{1;u} - \eta_{2;u}, & 1 < i \le I; \quad u = i-1 \\ \xi_{1;v} - \xi_{2;v}, & I < i < I+J; \; v = i-I, \end{cases} \tag{5.16}$$

where $\{\xi_{1;v}\}$ are as defined in (3.1), and $\{\eta_{1;u}\}$ are identical to $\{\xi_{1;v}\}$ upon replacing column effects by row effects. Here θ_1 is the *contrast* of frame effects; $\{\theta_i;\ 1<i\le I\}$ ($\{\theta_i;\ I<i\le I+J\}$) measures the *contrast* of the difference of two adjacent vertical (horizontal) blocks in a mask to that in another mask. Obviously, θ_i equals zero if the masks (located at different positions in the registered scenes) have the same texture. Since $\{\theta_i\}$ is an independent set of contrasts, by the arguments in Section III,IV hypotheses testing (5.2) can be implemented by a multi-comparison test of whether, corresponding to $\boldsymbol{\theta} = \boldsymbol{R}\boldsymbol{\beta}$, its BLUE estimate $\hat{\boldsymbol{\theta}} = \boldsymbol{R}\boldsymbol{B}y$ is within a confidence ellipsoid or not. Applying (3.6) or (4.5), test statistic for moving objects is

$$\hat{\boldsymbol{\theta}}'\, \boldsymbol{L}^{-1}\, \hat{\boldsymbol{\theta}} \underset{H_1}{\overset{H_0}{\gtrless}} q s^2 F_{\alpha;q,n-r} , \tag{5.17}$$

where $\sigma^2 L = \mathrm{var}(\hat{\boldsymbol{\theta}}) = \boldsymbol{R}\boldsymbol{B}\,\mathrm{var}(y)\,(\boldsymbol{R}\boldsymbol{B})'$. Noting that $S(y, \hat{\beta})/(n-r) = s^2$ and $\lambda_0 = F_{\alpha;\, q,n-r}$, the generalized likelihood test (5.11) is equivalent to the multi-comparison test (5.17). Therefore, the detection of edges as well as moving-edges based on the multi-comparison contrast test yield the optimal tests. All remarks pertaining to edge/moving-edge detection are also valid for moving object detection. Due to its simplicity in implementation as well as the physical meaning attached to the contrasts (similarity measure) between two objects, the multi-comparison test is naturally preferred over the generalized likelihood-ratio test.

c) Practical Aspects Consideration

Although the detector (5.11) or (5.17) is optimal theoretically, it encounters some implementation difficulties: i) object displacement from t_1 to t_2 is unknown; ii) object size is also unknown. For

convenience, denote R_1, R_2 as the top left corner of an object's maximal inscribing rectangle in frame 1,2, respectively.

To circumvent i), we suggest an exhaustive search. Let $C = \{c_1, c_2, \cdots, c_D\}$ be a finite set of positive integers and c_l, $l = 1, \cdots, D$ the separation between R_1, R_2. Introduce a Cartesian co-ordinate system with D_x, D_y as the vertical, horizontal axes centered at R_1 and it is aligned with the frames. For any c_l, $l = 1, \cdots, D$ the loci of R_2, of the co-ordinates (D_x, D_y), are equi-distant lattices on the lines

$$\begin{aligned} D_x + D_y &= c_l, \quad -c_l < D_x \le c_l \\ D_x - D_y &= -c_l, \quad -c_l \le D_x < c_l, \end{aligned} \tag{5.18}$$

which represent the intercepts in the first/second, third/fourth quadrants, respectively. Suppose R_2 is 6 pixels ($= c_l$) away from R_1 and the search step Δ equals 2 pixels. Upon substituting $D_x = 0\Delta$ into (5.18), then $D_y = 6, -6$; similarly, substituting $D_x = 1\Delta$ into (5.18), $D_y = 4, -4$; and so forth. As a matter of fact, when $c_l = 6$, we only have to investigate the following points (relative to R_1): (0, 6), (0, -6), (2, 4), (2, -4), (4, 2), (4, -2), (6, 0), (-6, 0). Summing all pixels in C, the number of operations amounts to $\sum_{l=1}^{D} 4\{[c_l/\Delta] - 1\}$, $[x]$ denotes the integer part of x. Recognizing that R_1 is in the neighborhood of a shifting edge, the searching process is repeated for R_1, leading to the total number of searches for a moving object

$$N_{search} = \left\{ \sum_{l=1}^{D} 4 \left(\left[\frac{c_l}{\Delta} \right] - 1 \right) \right\}^2. \tag{5.19}$$

Since the time separation between frames 1,2 is presumably short, c_l is of low value. Clearly, N_{search} is small as compared to a brute-force approach, which is computationally expensive because it looks for all sets of pixels in the image. Briefly, moving edge detection basically reduces the search domain, and hence the computation time in the detection of moving objects.

To bypass ii) we suggest a rubber mask approach [18]. Set the inscribing rectangle to an initial size, for example, 3×3. Apply the exhaustive search and, if H_1 is decided, increment the mask outward and repeat the process. Mask expansion is stopped whenever the former H_1 decision switches to H_0, implying that the mask has grown too big as to include background pixels. Since background pixels (in the context of the object) are different from one region to the next, it is a prudent practice not to include them; otherwise there will be a *difference* even for two identical objects. The dimension of the mask responsible for the current H_1 decision is the maximal size.

d) Generality of the New Class of Detectors

(1) In (5.2), each pixel is modeled as a function of row and column effects. Based on (5.17) the decision of H_1 being true assures the difference of all pixels in one mask over that in another, after decorrelation, is within a confidence ellipsoid. This is a direct consequence that under H_1 and for

$x_1 \neq x_2$, and/or $y_1 \neq y_2$, every pixel at a particular location in one mask equals, within a confidence interval, to the pixel at the same site in another mask. In this sense, correspondence problem [4] naturally disappears.

To gain more insight, let us eliminate both the row, column effects in (5.1) and rename the resulting equation as (5.1'). Clearly, pixel gray level is now solely a function of frame number. The rank of the effect matrix in (5.1') being 2 is far less than that in (5.1). By the vector space concept in Theorem 2, the least squares error suffered by (5.1') utilizing a lower order of basis vectors is naturally worse than that in (5.1). Note that (5.1') is actually a similarity measure predominantly used in the past [4].

(2) No displacement of masks, i.e., $x_1 = x_2, y_1 = y_2$:

i) Setting $I = 1, J = 2$, from (5.16) we readily obtain the contrast functions $\theta_1 = \zeta_1 - \zeta_2$, $\theta_2 = (\gamma_{1;1} - \gamma_{1;2}) - (\gamma_{2;1} - \gamma_{2;2})$. Evidently, θ_2 is a measure of the rate of gray level change at the same location in two sequential frames and, together with the body-rigidity constraint, form the optical flow concept and is a popular technique for object detection [7]. However, as the window size is small, θ_2 performs poorly in noisy environments (refer also to Remark 1 in Section III). This is easily seen by rewriting the realization of θ_2 as $\hat{\theta}_2 = (y_{1;11} - y_{1;12}) - (y_{2;11} - y_{2;12})$ and taking the second moment to get $\text{var}(\hat{\theta}_2) = \Sigma_{m=1}^{2} \Sigma_{j=1}^{2} y_{m;1j} = 4\sigma^2$.

ii) The proposed detector in the form of (5.11) or (5.17) is a stationary object detector. Suppose we further eliminate m, i.e, ignoring frame effects, then the new detector (5.11) or (5.17) corresponds to a stationary object detector that does not take advantage of the available information between frames (deducible from data). This is the same as the reduction of basis vectors by half and, by the vector space approach in Theorem 2, is naturally worse than is achieved by a detector making use of frame effects. This is similar to [19] where, because the mask engulfs the background, a one-to-one random mapping is required.

iii) Suppose the third dimension m represents the spectral instead of temporal data, the resulting detector (5.11) or (5.17) becomes the multispectral detector. In particular, the new detector in independent noise that assumes object size of 1 pixel and $I = J = 1$ has been investigated before [17].

VI. SIMULATION RESULTS

Fig.2 (a1)-(a2) are two consecutive image frames of Route 49 near Griffiss AFB, Rome, NY, taken at ⅓ second apart. Fig.2 (b1)-(b2) are the corresponding images contaminated by Gaussian noise $N(0, 10)$. To detect edges, the detector $\hat{\xi}_1' E_1^{-1} \hat{\xi}_1$, $\hat{\xi}_2' E_2^{-1} \hat{\xi}_2$ (see eq.(3.6)) are applied in four major directions, yielding the edges shown in Fig.3 (a1),(a2), respectively. Concurrently, the detector (4.5) (again four directions) results in the moving-edges field as appeared in Fig.3 (b). Finally, the determined moving objects, based on detector (5.17), are marked by rectangles, as depicted in Fig.3 (c1)-(c2). The corresponding images when noise is of $N(0, 50)$ are shown in Fig.2(b1)-(b2), with

the resulting outputs in Fig.4. In all cases, the Type-1 error is set to 0.01.

The superior performance of the proposed class of detectors speak for themselves.

VII. CONCLUSIONS

We have derived a closed form solution for the optimal detection of moving objects embedded in a Gaussian noise environment. Since the moving edge detectors, which in turn depend on the edge detectors tantamount to the implementation of the proposed detector, they are developed simultaneously. The new detector is a 3-D filter utilizing the spatial (intra-frame) and temporal (inter-frames) data. Pixels in a 3-D neighborhood are modeled in terms of various effects. This allows a large variation of gray levels (heterogeneity) as compared to the traditional model that usually assumes homogeneity (locally). By the vector space approach in Theorem 2, all well-known detectors based on simpler models become subclasses of the proposed detector. Since the new detector is a F-statistic, it is uniformly most powerful (UMP) in Gaussian environment of unknown and varying noise variance [20].

The equivalence of the generalized likelihood test and the multi-comparison test (MCT) is then established. Since the MCT measures the contrast between frames of a local neighborhood (see section 3(d) *i*), it is analogous to the optical flow technique without its drawbacks, in the combat of noise. In this context, the new detector is a combination of the optical-flow and feature-based techniques, but is primarily a generalized optical-flow based detector that takes advantage of the feature-based approach (such as shifting edge), for efficient computation. Because of the associated confidence ellipsoid (see (5.17)), the proposed detector is efficient in the suppression of noise and the enhancement of similarity. A less obvious advantage is that it may tolerate scene frames that are slightly out of registration.

Simulation results based on real noisy images demonstrate the practicality and superiority of the proposed detector.

REFERENCES

[1] J.K. Aggarwal and R.O. Duda, "Computer analysis of moving polygonal images," *IEEE Trans. Comput.*, vol.C-24, pp.966-976, 1975.

[2] J.-Q. Fang and F.S. Huang, "Some experiments on estimating the 3-D motion parameters of a rigid body from two consecutive image frames," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.PAMI-6, pp.545-554, 1984.

[3] V. Salari and I.K. Sethi, "Feature point correspondence in the presence of occlusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.12, pp.87-91, 1990.

[4] J.K. Aggarwal, L. Davis and W.N. Martin, "Correspondence processes in dynamic scene analysis," *Proc. IEEE*, vol.69, pp.562-572, 1981.

[5] D.H. Ballard and O.A. Kimball, "Rigid body motion from depth and optical flow," *Computer vision, Graphics and Image Process.*, vol.22, pp.95-115, 1983.

[6] H.-H. Nagel, "Displacement vectors derived from second-order intensity variations in image sequences," *Computer Vision, Graphics and Image Process.*, vol.21, pp.85-117, 1983.

[7] A. Mitiche, Y.F. Wong, and J.K. Aggarwal, "Experiments in computing optical flow with the

gradient-based multi-constraint method," *Pattern Recognition*, vol.20, no.2, pp.173-179, 1987.

[8] W.B. Thompson, K.M. Mutch, and V.A. Berzins, "Dynamic occlusion analysis in optical flow fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.PAMI-7, no.4, pp.374-383, July 1985.

[9] A. Rosenfeld, "Computer vision: basic principles," *Proc. IEEE*, vol.76, no.8, pp.863-868, Aug. 1988.

[10] W.G. Cochran and G.M. Cox, *Experimental designs*, John Wiley, 1957.

[11] H.L. Van Trees, *Detection, estimation, and modulation theory*, Part I, 1968, John Wiley.

[12] J. Cheung, L. Kurz, G. Nethercott, and F. Rahrig, "A statistical theory for optimal detection of moving objects in variable corruptive noise," *IEEE Trans. Pattern Anal. Mach. Intell.*, (to appear).

[13] W. K. Pratt, *Digital image processing*, second edition, John Wiley, 1991.

[14] H. Scheffé, *The analysis of variance*, John Wiley, 1959.

[15] H. Scheffé, "A method for judging all contrasts in the analysis of variance," *Biometrika*, vol. 40, pp.87-104, 1953.

[16] H.L. Van Trees, *Detection, estimation, and modulation theory*, Part III, 1971, John Wiley.

[17] J. Cheung, D. Ferris and L. Kurz, "On classification of multispectral image data," *IEEE Trans. on Image Process.*, (to appear).

[18] B. Widrow, "The *rubber-mask* technique," *Pattern Recognition* 5, pp.175-211, 1973.

[19] E.S.H. Chang and L. Kurz, "Object detection and experimental designs," *Comput. Vision, Graphics and Image Process.*, vol.40, pp.147-168, 1987.

[20] E.L. Lehmann, *Testing statistical hypotheses*, John Wiley, 1957.

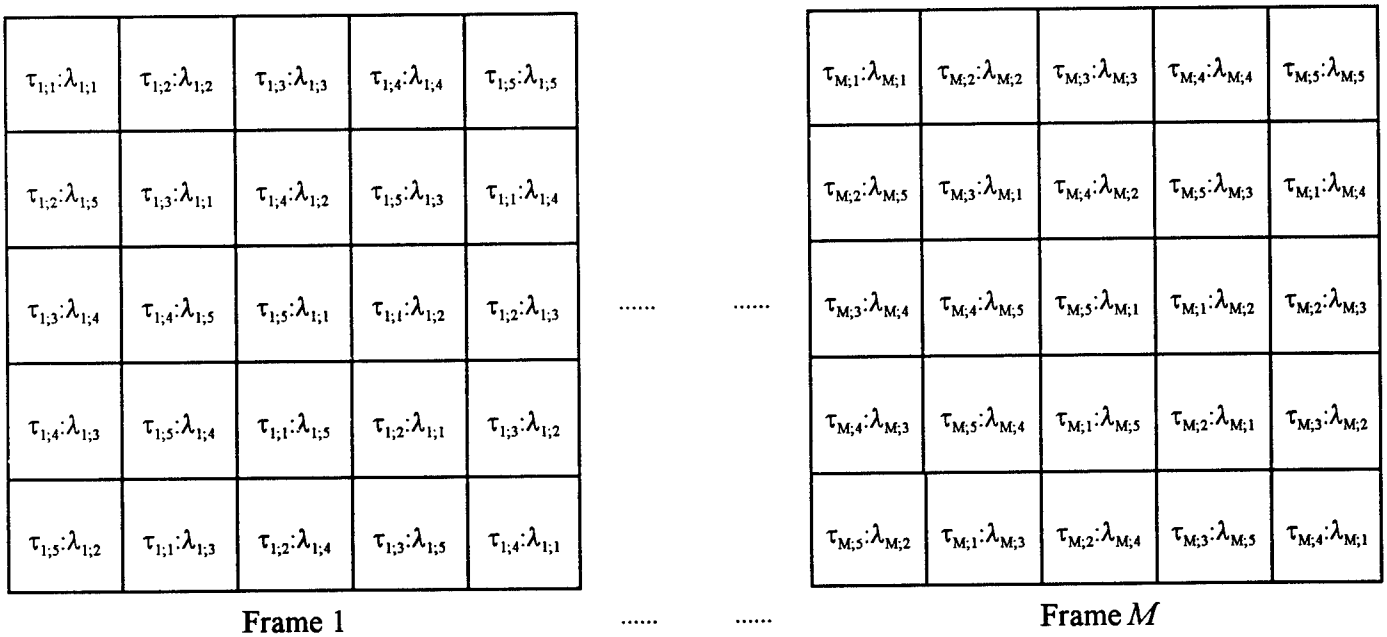


Fig.1 Four-Way Treatment Layout in a $M\times5\times5$ GLS Design

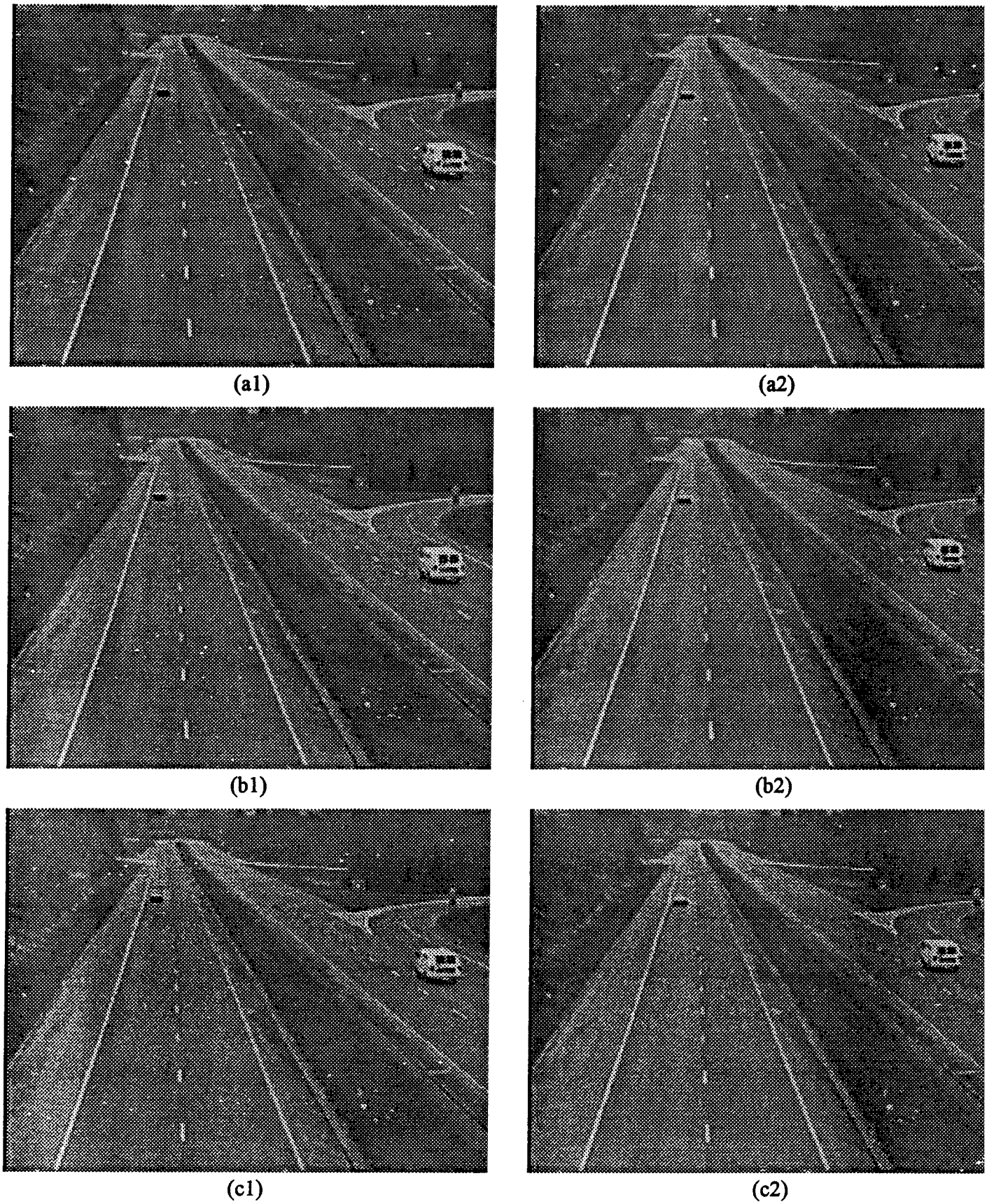
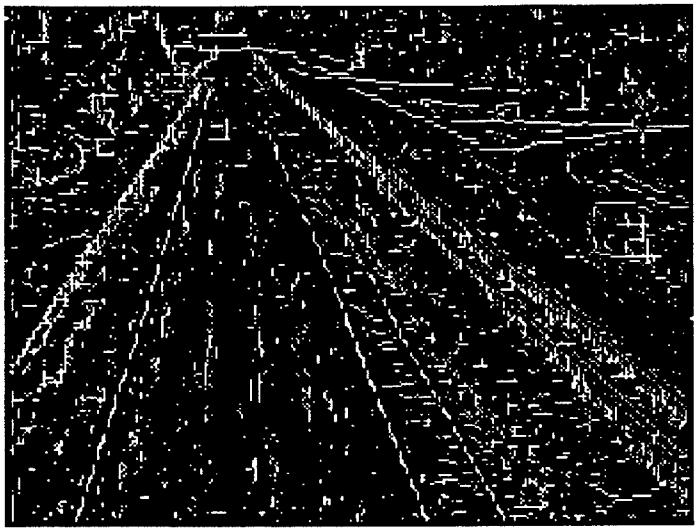
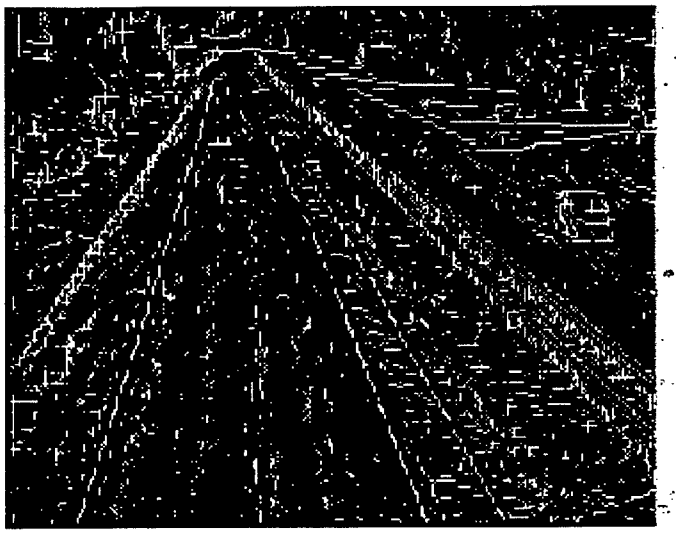


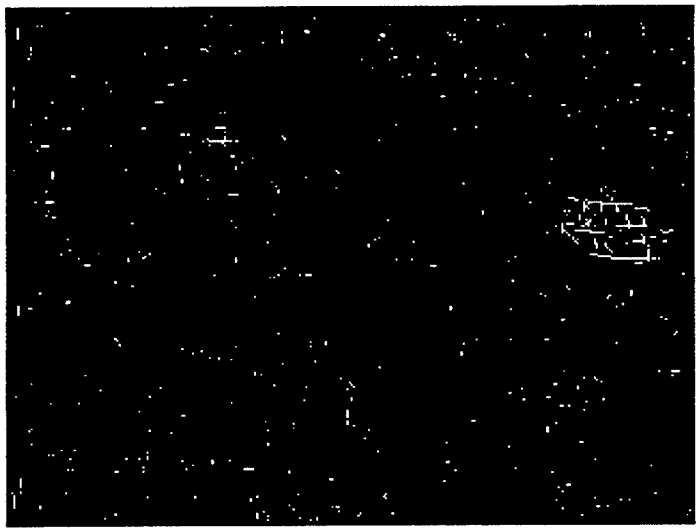
Fig.2
(a1)-(a2): consecutive scene frames of a highway;
(b1)-(b2), (c1)-(c2) are images contaminated by Gaussian noise $N(0, 10)$, $N(0, 50)$, respectively



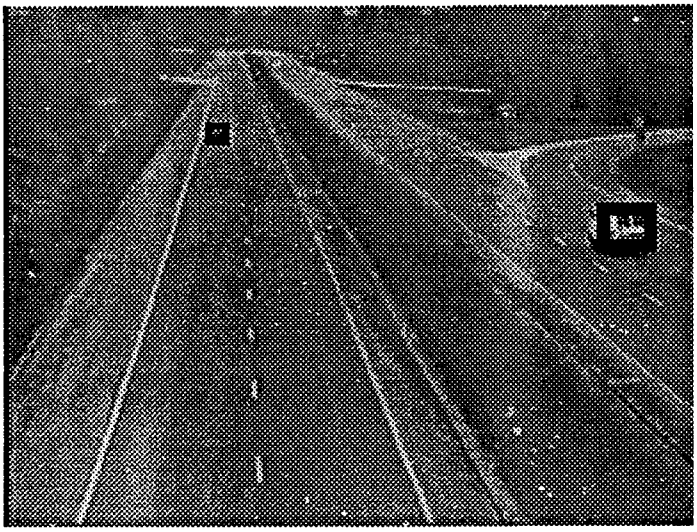
(a1)



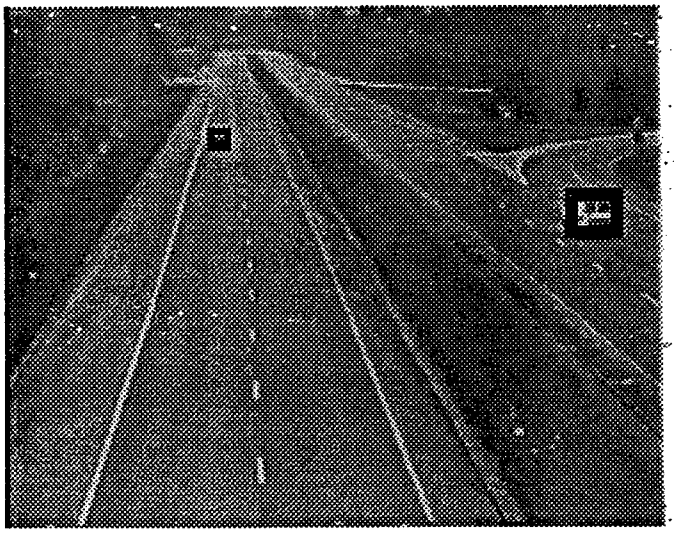
(a2)



(b1)

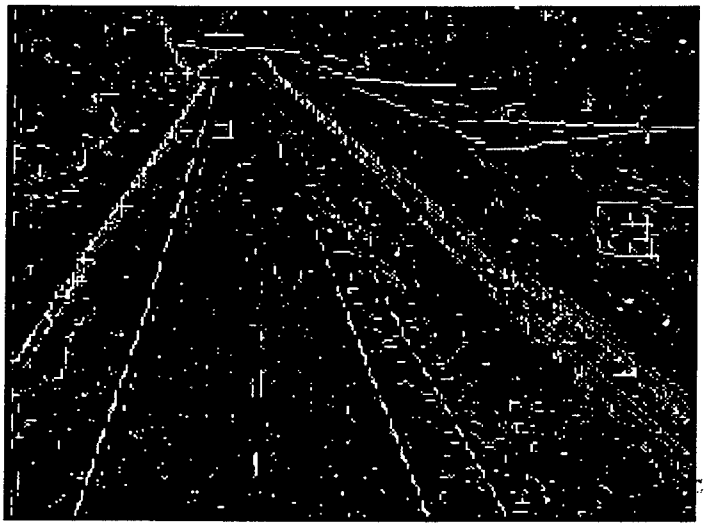


(c1)

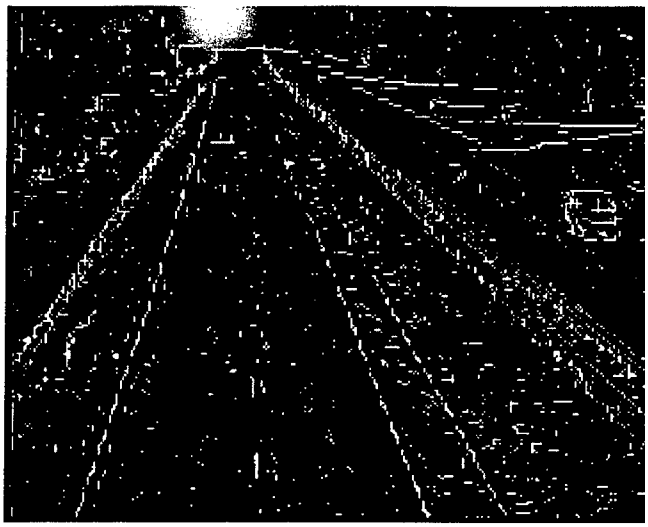


(c2)

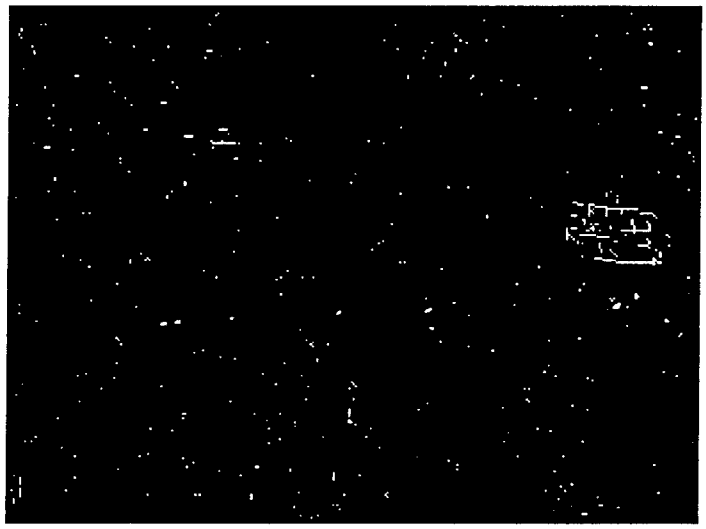
Fig.3
(a1)-(a2): edges; (b): moving-edges; (c1)-(c2): moving-objects
detected using appropriate 3-D detectors on the images in Fig.2 (b1)-(b2), at $\alpha = 0.01$



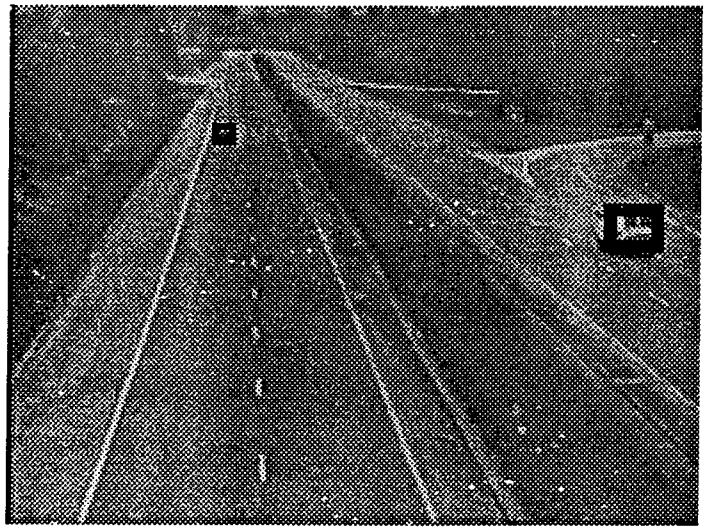
(a1)



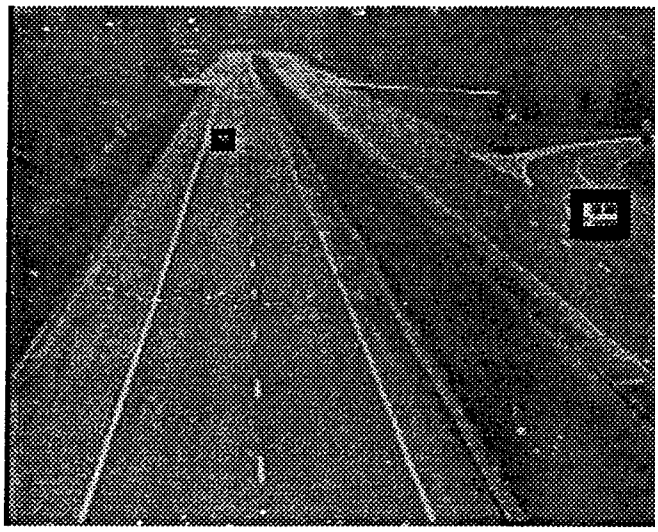
(a2)



(b1)



(c1)



(c2)

Fig.4
(a1)-(a2): edges; (b): moving-edges; (c1)-(c2): moving-objects
detected using appropriate 3-D detectors on the images in Fig.2 (c1)-(c2), at $\alpha = 0.01$

Transport of ATM-Based Traffic Via the Advanced Communication Technology Satellite (ACTS)

Mostafa Chinichian
Professor
Electrical Engineering Department

California Polytechnic State University
San Luis Obispo, CA

Final Report for:
Summer Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington DC

and

Rome Laboratory

August 1995

TRANSPORT OF ATM-BASED TRAFFIC VIA THE ADVANCED COMMUNICATION TECHNOLOGY SATELLITE (ACTS)

Mostafa Chinichian
Electrical Engineering Department
California Polytechnic State University
Grand Avenue
San Luis Obispo, CA 93407

ABSTRACT

These tests will perform as a link test between the two stations, the CRC, Ottawa and Rome Labs, GAFB, New York. Equipment configurations for the ATM tests as well as channel characterization in term of C/No and BER as a function of other ATM parameters in a single channel and multiple channels are studied. Performance analysis are also includes plots for the different QPSK and 8-PSK modulations with or without error correcting coding.

Table of Content

TRANSPORT OF ATM-BASE TRAFFIC VIA THE ADVANCED COMMUNICATION TECHNOLOGY SATELLITE (ACTS)

Mostafa Chinichian

1. INTRODUCTION

The popularity of Asynchronous Transmission Mode (ATM) protocol has risen to a new height, and all efforts are being made to verify its potential and usefulness. It is indicated in [1] that in the next decade, most telecommunication traffic will be carried by ATM technology. The usefulness and adaptability of ATM to current technology is now the subject of lot of research in both Government and industry. Although ATM was originally intended for fiber optic links transmitting high speed data, it can be used for other links such as the satellite link, which is the subject of our experiment.

This report documents the experiment of transmitting ATM signal via the Advanced Communication Technology Satellite (ACTS). Several tests were performed via a Satellite DS-3 link carrying ATM traffic in a Physical Layer Convergence Protocol (PLCP) frame format for a single channel as well as for multiple channels of the ATM via the ACTS. These experiments are part of the Global Grid Technical Technology Cooperation Program (TTCP) network, in which Rome Laboratory (RL) will communicate via the satellite with the Canadians at the Communication Research Center (CRC) [1], [2]. These experiments were conducted between the CRC, Ottawa and Rome Labs, GAFB, New York. This report will be based on the segment of the ATM/ACTS experiments that were conducted from 27 June 1995 through 24 August 1995. The report documents the specific tests that were conducted, presents the collected data and discusses the results obtained.

2 . SETUP AND CONFIGURATIONS

The first stage of the experiment involved the characterization of the satellite channel for reliable transmission of ATM signals at DS-3 rates via the ACTS. The ATM equipment, satellite transmission equipment and the test equipment used during the experiment as well as the characterization procedure are briefly described in this section.

2.1 SATELLITE TRANSMISSION EQUIPMENT

The satellite equipment consist of the EF Data modem (SDM-9000 Satellite Modem, Version 3.1.1), Ka-band Traveling-Wave-Tube Amplifier (TWTA), Test Loop Translator (TLT), power controllers, and the high data rate antenna.

The satellite transmission equipment which included the Up/Down converters, the TLT for ground loop-back, Low Noise Amplifiers (LNA), etc. were located in a hut next to the antenna, where the power meter could be monitored. The EF Data modem and the rest of the baseband equipment (ATM, Firebird tester, Spectrum Analyzer, etc.) were installed in an adjacent hut located about 25 ft from the antenna. An Intermediate Frequency (IF) of 70 MHz was used throughout the test.

The EF Data Satellite modem is a high performance, full-duplex, digital-vector modulator/demodulator with data rate capabilities of 6 Mbps to 51.84 Mbps. The modem has built-in scrambler/descrambers, differential encoding/decoding, multi-rate forward error correction capabilities (convolutional encoder and Viterbi decoder) and can be configured to add over head/framing to the data. For example, a Reed-Solomon codec is provided to work in conjunction with the Viterbi decoder and includes additional framing and interleaving for improved performance. Most of the equipment used was

provided by Canadian on corporation with Rome Labs. Also, the frequency plan used in the design of the TLT is provided by CRC. It should also be noted that at Rome Labs the antenna is 3.7 m (with a gain of 50.80 dBi) while at Ottawa, Canada, the antenna is 4.60 m (with a gain of 53.80 dBi). These antennas were provided by CRC.

2.2 ATM AND TEST EQUIPMENT

The ATM equipment used in this experiment is the ADTECH AX/4000 Series. The AX/4000 Series ATM generator/Analyzer provides complete physical, ATM equipment, switches and networks for proper operation and Quality of Service (QoS) parameter measurements. It is modulator in nature, so it allows custom configuration for a variety of ATM test applications. The main components consist of mainframe, ATM Generator/Analyzer modules, port interfaces and PC controller which is driven with Microsoft Window-based software. While a variety of port interfaces are possible (e.g. SONET OC-3c, E3, TAXI, DS-3, etc.), the DS-3 interface supports 45 MHz using electrical format with BNC connectors was used to interface the ATM equipment with the modem.

3. CHANNEL CHARACTERIZATION

The sequence of tests involving the transmission of ATM cells via the ACTS satellite started on June 27, 1995. The first set of tests were aimed at characterizing the channel and modem and also to familiarize Rome Lab personnel with the ground terminal equipment. To evaluate the EF Data modem (SDM-9000 Satellite Modem), and the Firebird DS-3 test equipment, a number of transmissions involving both CW testing and some modulations were made. The channel characterization and modem evaluation are necessary to ensure that at DS-3 speeds and in the Ka-band, the modem specification for the

different modulations can be verified. Also, through carrier-to-noise ratio measurement the power transfer characterizing of the ACTS transponder can be ascertained in order to determine appropriate operating points for the ATM experiments.

The tests includes obtaining the modem BER characteristics as a function of the Output Power Back Off (OPBO) of the satellite in terms of received Eb/No. For each test, the measurements were taken for three modes:

1. Full-duplex mode (CRC-RL and RL-CRC, both via ACTS).
2. Self Loop-back mode (via- ACTS)
3. Ground Loop-back (TLT)

The ACTS operations center can provide satellite loop-backs for both the East Scan 4E spot beam and the steerable spot-beam. It should be pointed out, however, that during such loop-back there may be considerable loss in the orthogonal polarization and satellite coverage.

3.1 C/No MEASUREMENTS

During the CW transmissions, the transmit power levels at both ends (CRC and RL) were varied (with the Tx power meter reading recorded) while the carrier power (C) and the noise power (No) read from the spectrum analyzer at opposite ends. In this case, the spectrum analyzer received power is calculated from

$$\begin{aligned} Eb/No &= C/No - 10\log(44.73\times10^{6}) \\ &= C/No - 76.5 \text{ dB} \end{aligned}$$

where C and No are read from the spectrum analyzer.

3.2 BER MEASUREMENTS

To evaluate the modem and the Firebird BER tester, the following modulations were transmitted:

1.	QPSK	R = 3/4,	no RS
2.	QPSK	R = 3/4,	with RS
3.	8-QPSK	R = 2/3,	no RS
4.	8-QPSK	R = 2/3,	with RS

In each case, both the received BER and Eb/No were measured for each setting of the transmit power. All transmissions included are convolutional encoding and corresponding Viterbi decoding with indicated rates (R = 3/4 for QPSK and R = 2/3 for 8-QPSK). Also, In our experience convolutional codes with different rates (2/3 and 3/4) are modulated at IF 70 MHz with or without being concatenated to a (208, 192) Reed-Solomon code.

Also, for each power level considered the BER was taken from the Firebird tester as well as from the modem while the reading of Eb/No were made from the modem.

4. ATM TRANSMISSION

The main aim of the ATM experiments is to characterize the DS-3 Satellite link for the transmission of ATM signals. Specifically, some ATM QoS parameters were measured using different levels of modulation (QPSK, 8-QPSK) with forward error correction and V.35 data scrambling. The bursty nature of the

error statistics and the corresponding large propagation delay involved in a Satellite link affect these ATM QoS parameters. Also, at the Ka-band (30/20 GHz) of the ACTS the Satellite link is susceptible to atmospheric effects, the most important of which is attenuation due to rain. To properly assess the viability of ATM via Satellite, the cumulative effect of these Satellite link characteristics on the QoS parameters must be determined.

4.1 ATM QoS PARAMETERS

A number of ATM parameters have been identified as being very important in the assessment of the performance of an ATM network. In the experiments, emphasis was placed on those quality of service parameters and statistics that can be measured by the ADTECH ATM test equipment. Some of these parameter will be discussed in this section.

1. Bit Error Ratio (BER) - this is the Satellite channel bit error rate that determines the rate at which the transmitted bits were changed in the physical layer.

2. Cell Loss Ratio (CLR) - this is the ratio of the number of lost ATM cells sent by a user in specified time interval. Due to the random nature of the ATM traffic and the limited size of the ATM traffic and the limited size of the buffers, it is usually possible that a cell arriving at a switching node may be lost . Thus CLR are caused by buffer overflows and bit error in the cell header that can be detected but not be corrected.

3. Cell Misinsertion Ratio (CMR) - this is defined as the ratio of the cells delivered to a wrong destination to the total number of cells sent. It occurs

as a result of an undetected error in the header that causes a change of the cell destination.

4. Receive Frame Alarm - the following alarm parameters related to the receive frame were measured :

Line Code Count and Rate
Framing Count and Rate
P-Bit Count and Rate
FEBE Count and Rate

5. Receive PLCP Alarm - the following PLCP alarm parameters were measure :

Framing Count and Rate
BI Count and Rate
FEBE Count and Rate

6. Other QoS parameters - In addition to the above, the following QoS parameters were also measured during each run of the ATM tests:

Out of Sequence Count
PRBS Bit Error Count
PRBS Bit Error Rate
PRBS Synchronization Error Count

4.2 SINGLE CHANNEL WITH DIFFERENT MODULATIONS

These tests were performed over the ACTS Satellite link test between the two stations (CRC at Canada and RL at GAFB) using a single payload data

stream which fully loaded the DS-3 bearer (i.e. 96000 cells/sec) with an ADTECH AX/4000 test-set. All tests performed with QPSK and 8-PSK modulations (with convolutional encoders and Viterbi decoders) concatenated with or without Reed-Solomon codes. Also, we used appropriate cell rate to increase observation times.

4.3 MULTIPLE CHANNELS WITH DIFFERENT MODULATIONS

The same tests above as were performed using multiple channels data streams which together fully loaded the DS-3 bearer. We have recorded all the data in order to characterize the DS-3 channel for transmission of ATM signals, in terms of QoS parameters.

5. ANALYSIS OF RESULTS

The following pictures show result of experiments, BER versus Eb/No and other QoS parameters (i.e. QoS/Single Stream and QoS/Multiple Streams). They have been plotted in order to enable reader to compare BER versus Eb/No and other conditions affect ACTS satellite performance parameters (as an example few selected graphs were chosen to demonstrate QoS parameters and Eb/No in terms of BER, see pages 7-13 to 7-20). For more information refer to Final Report for Summer Research Program, September 1995, C3/BA, Rome Labs, GAFB, NY.

6. CONCLUSION

Tests have been conducted between the two stations (the CRC, Ottawa and the Rome Labs, GAFB, New York) to characterize channel and ATM usability and performances. Also, equipment configurations for the ATM tests as well as channel characterization in term of Eb/No and BER as a function of other ATM parameters in a single channel and multiple channels have been depicted for different conditions. Performance analysis have been also included plots for different QPSK and 8-PSK modulations (with convolutional encoders and Viterbi decoders) concatenated with or without Reed-Solomon coding.

7. REFERENCES

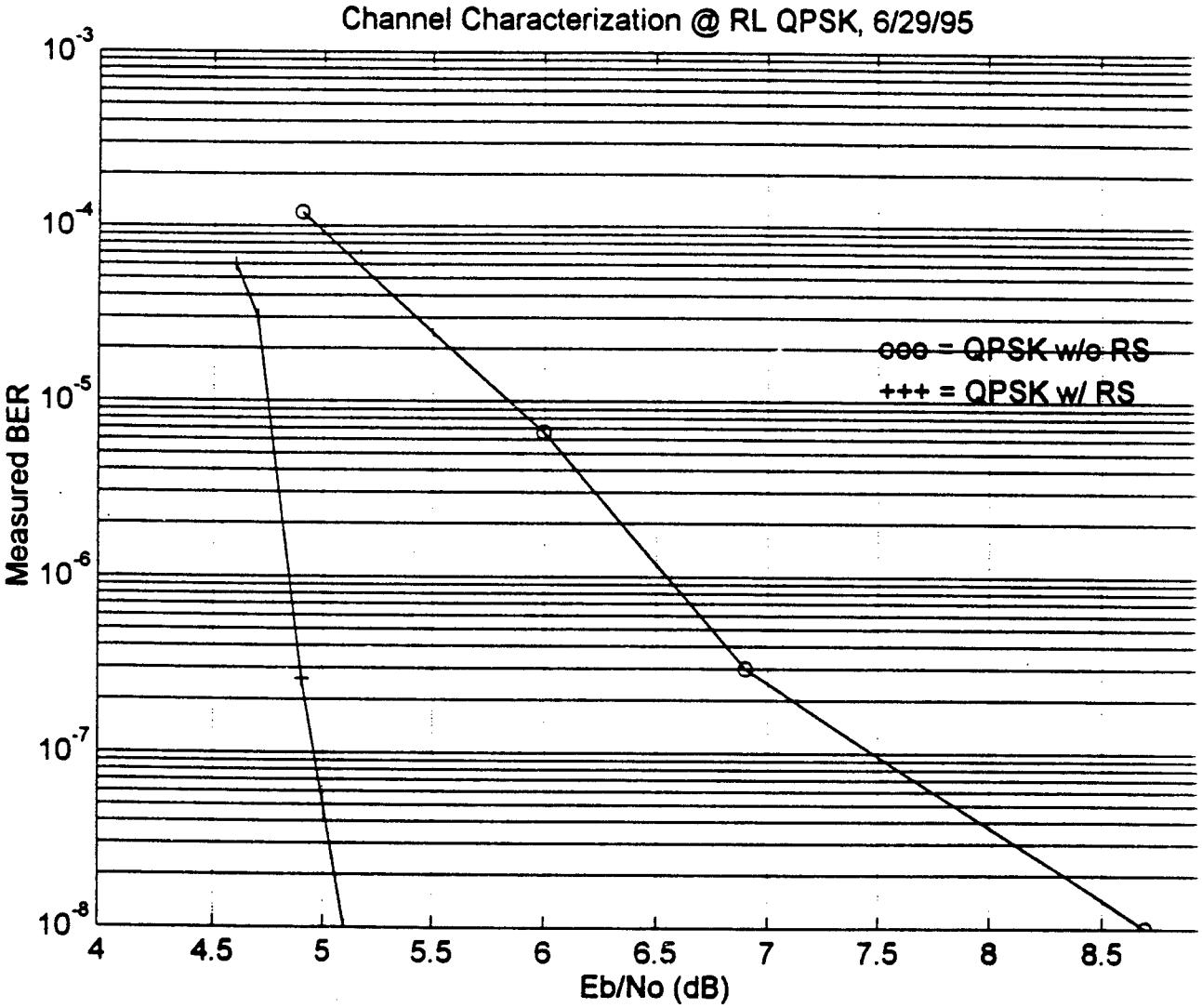
[1] COMSAT, Demonstration of Asynchronous Transfer Mode (ATM) via Commercial Satellite, Technical Report October 25, 1993, COMSAT Technology Services.

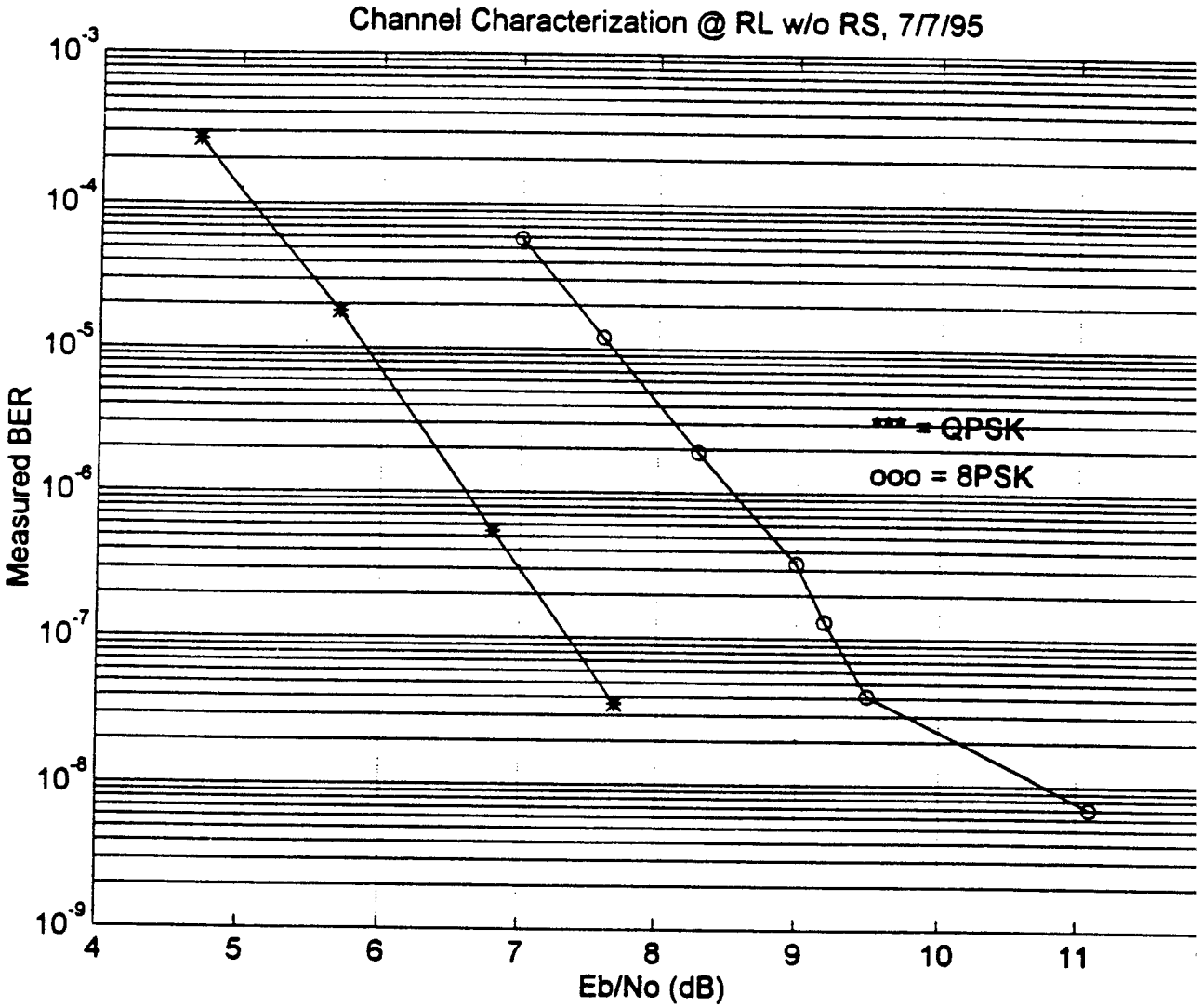
[2] John Butterworth and Gerad Nourry, TEST PLAN for an investigation of THE TRANSPORT OF ATM-BASED TRAFFIC BY Ku-BAND SATELLITE BEARERS, Workshop on the "Global Grid" concept, Sponsored by TTCP STP-6, 16th January 1995.

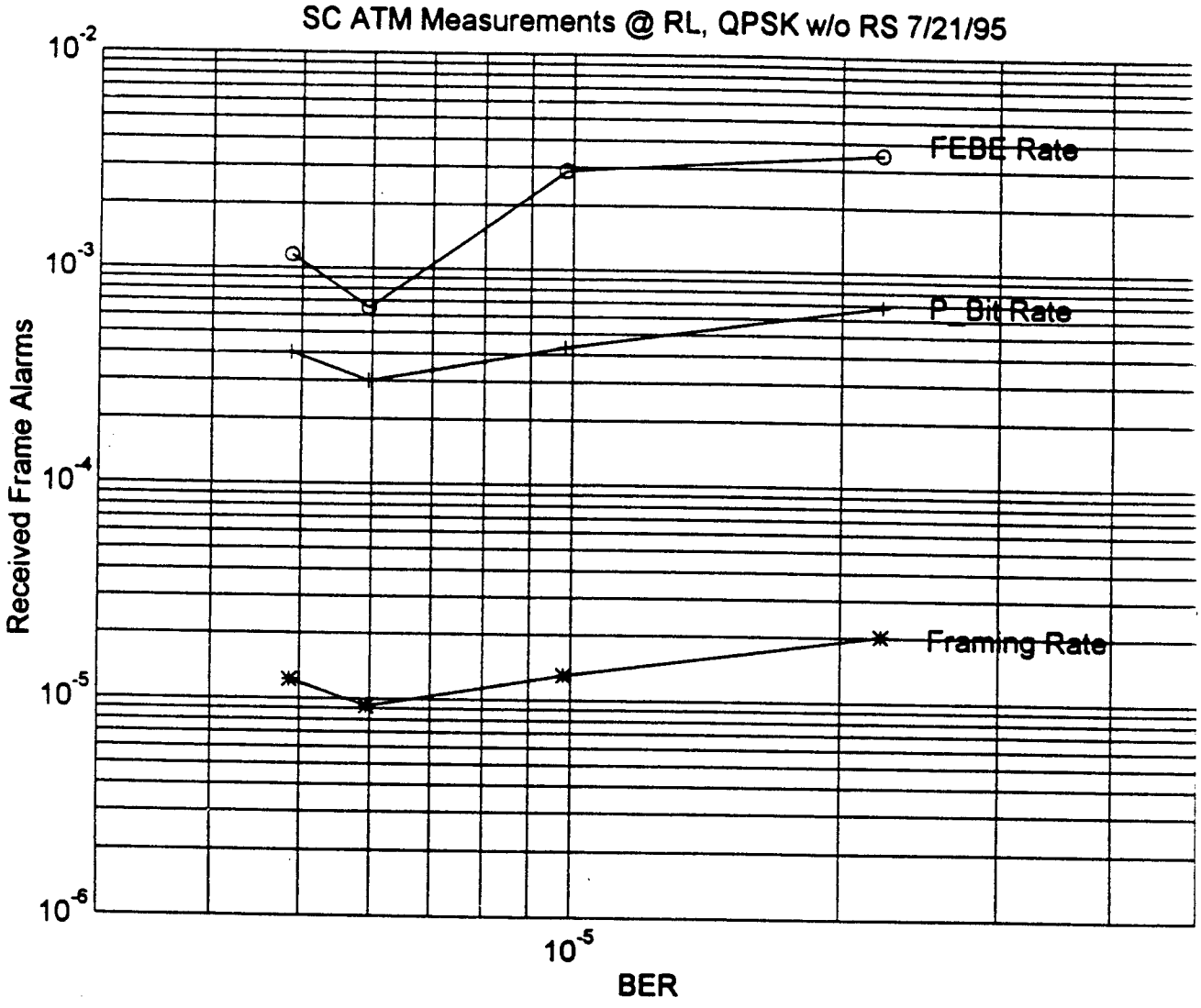
[3] V. Aalo and O. Ugweje, "A Program Plan for Transmitting High-Data - Rate ATM/SONET Signals over the ACTS, Final Report AFOSR, September 1994.

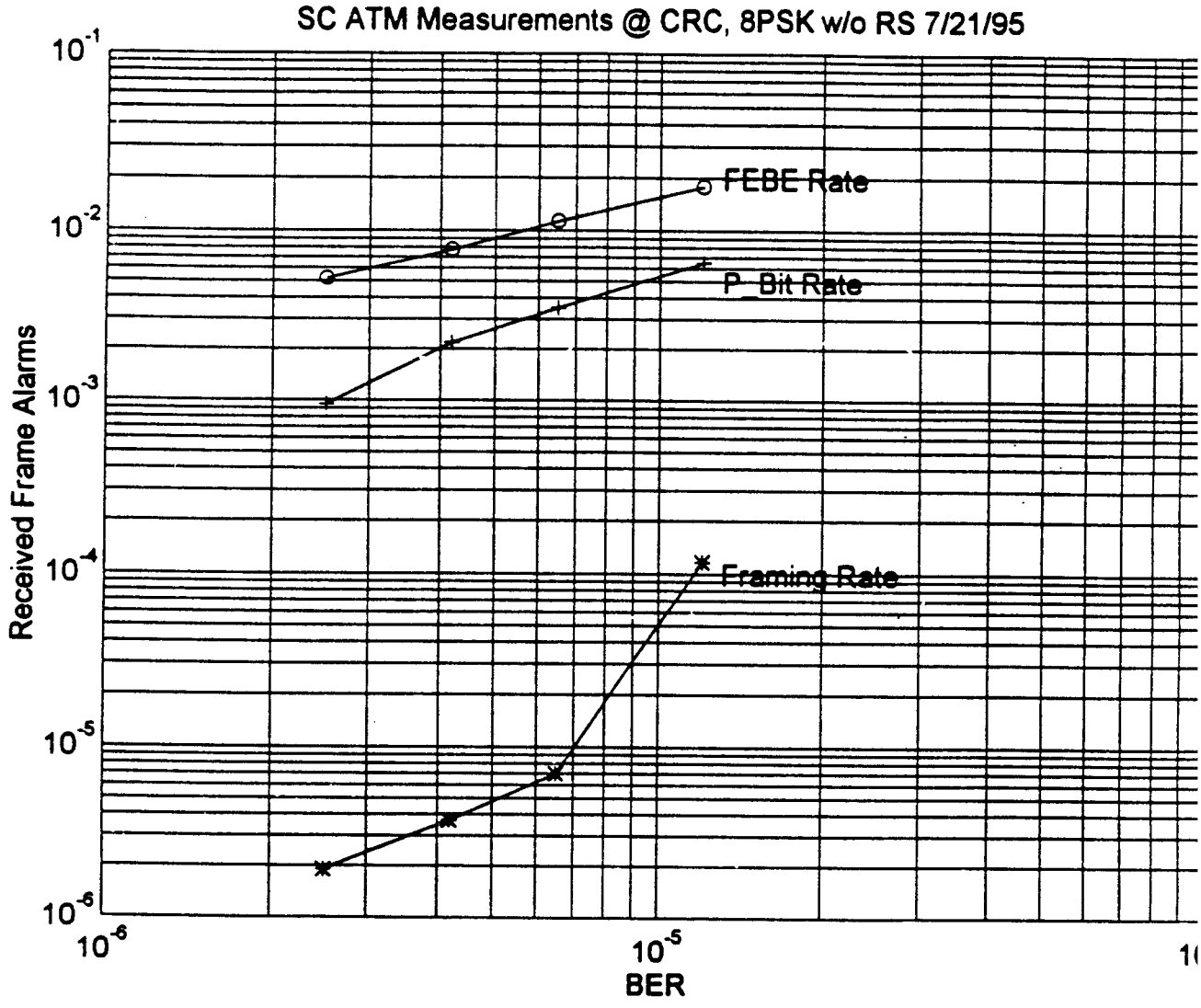
[4] R.O. Onvural, Asynchronous Transfer Mode Networks Performance Issues, Boston; Artech House, 1994.

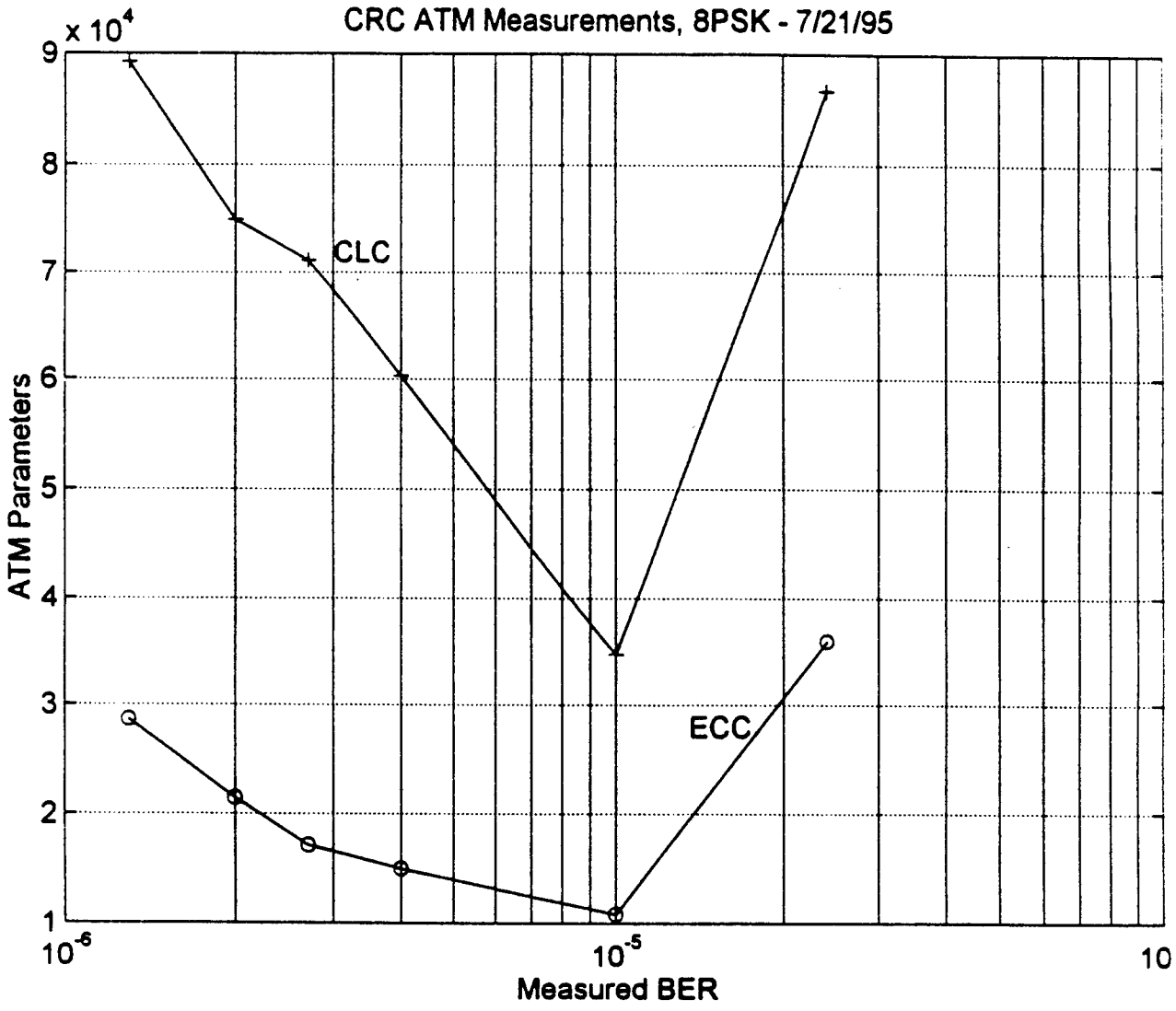
APPENDIX A

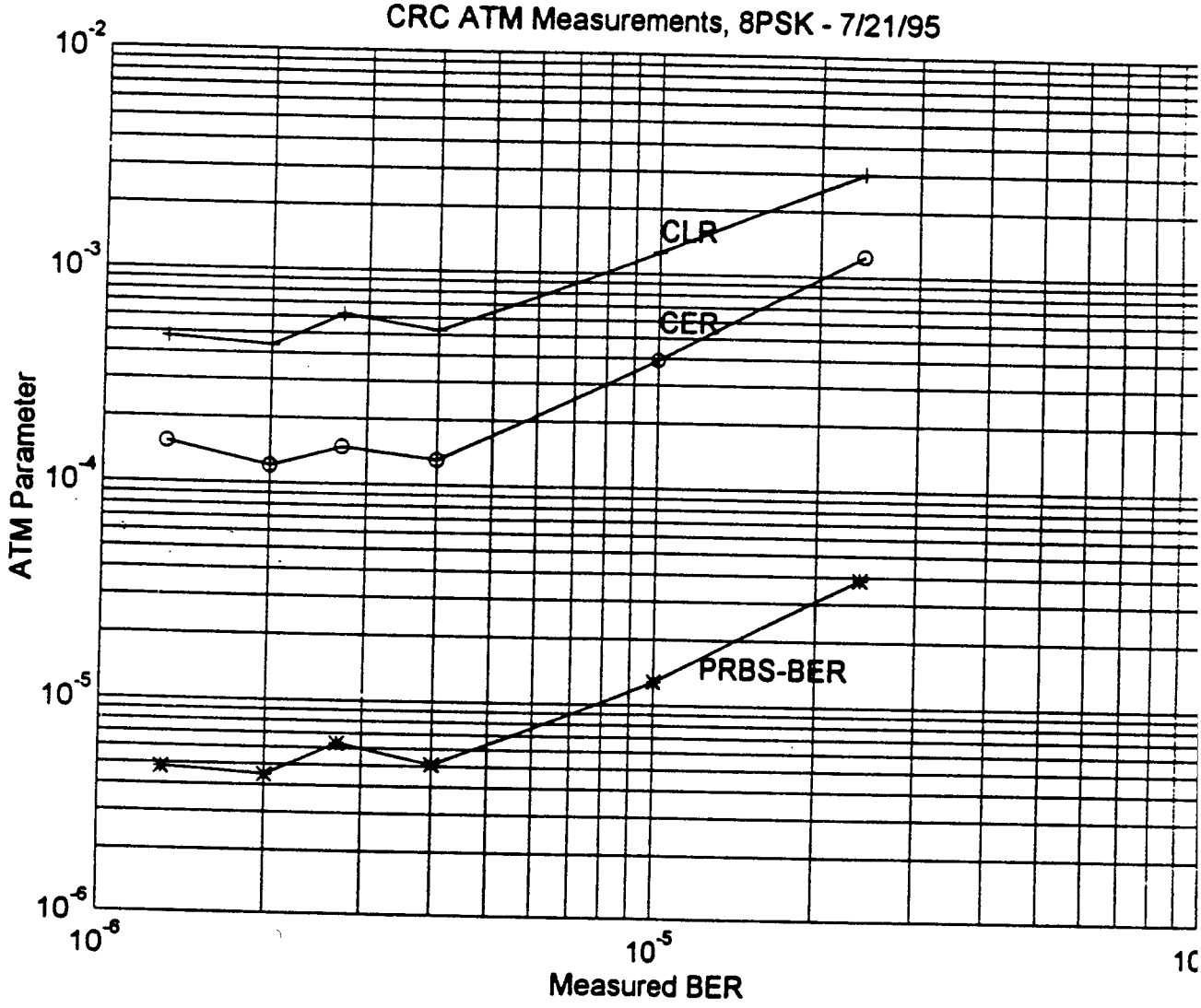

Channel Characterization @ RL QPSK, 6/29/95
Measured BER
10-3
10-4
10-5
10-6
10-7
10-8
ooo = QPSK w/o RS
+++ = QPSK w/ RS
4
4.5
5
5.5
6
6.5
7
7.5
8
8.5
Eb/No (dB)

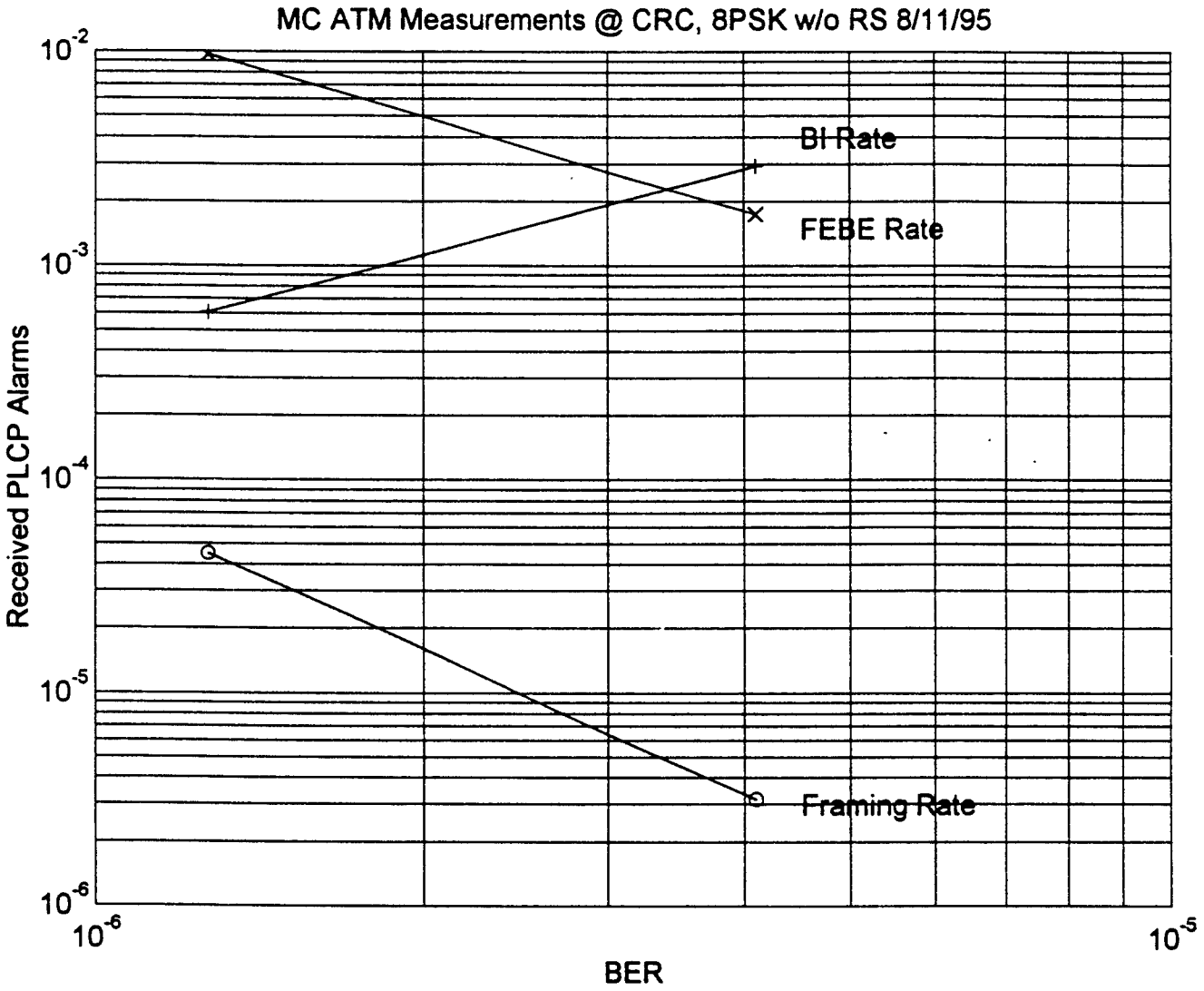

Channel Characterization @ RL w/o RS, 7/7/95
Measured BER
10-3
10-4
10-5
10-6
10-7
10-8
10-9
*** = QPSK
ooo = 8PSK
4
5
6
7
8
9
10
11
Eb/No (dB)

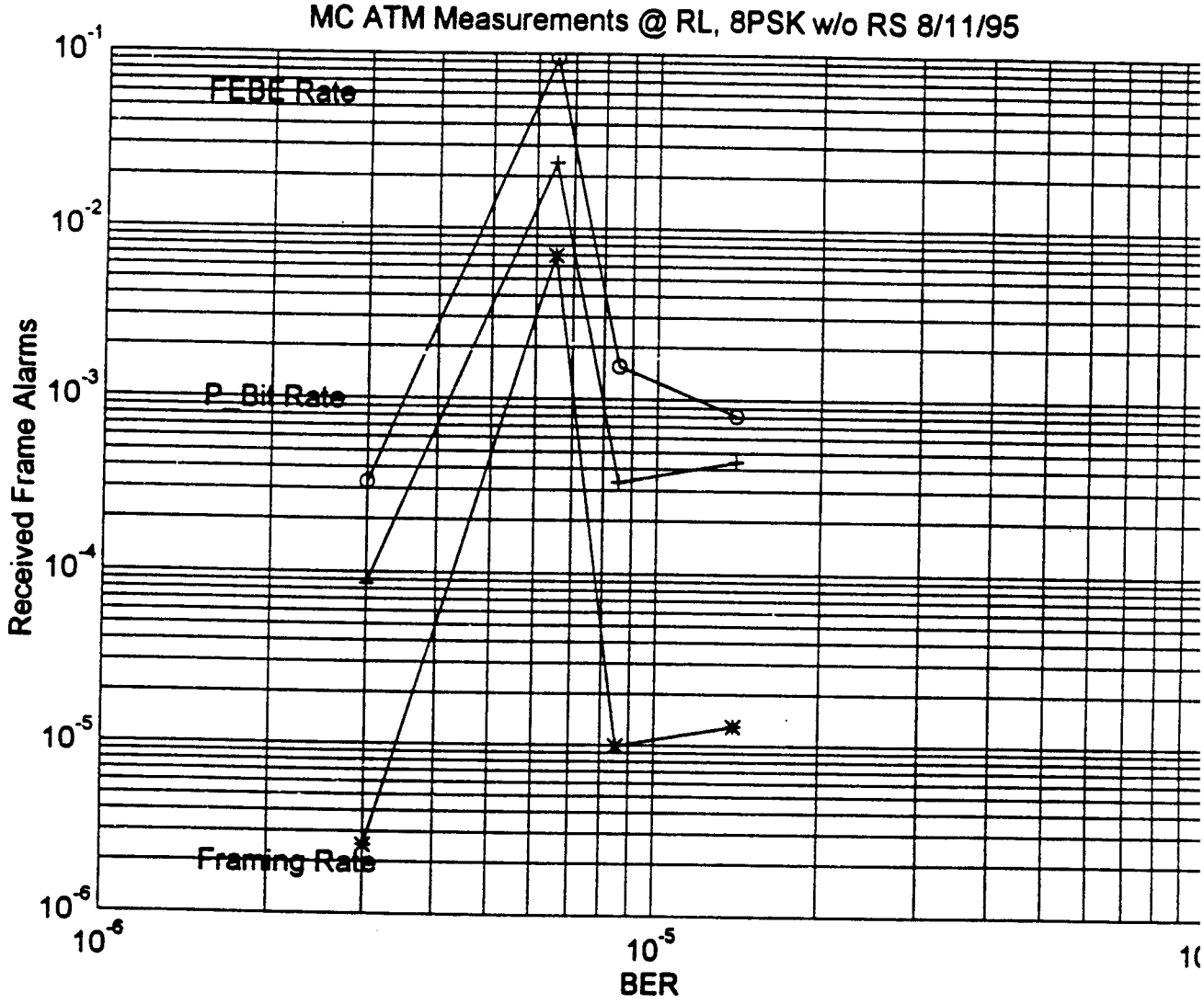

SC ATM Measurements @ RL, QPSK w/o RS 7/21/95
Received Frame Alarms
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
FEBE Rate
P_Bit Rate
Framing Rate
10^{-5}
BER


SC ATM Measurements @ CRC, 8PSK w/o RS 7/21/95
Received Frame Alarms
BER
FEBE Rate
P_Bit Rate
Framing Rate
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-6
10^-5


CRC ATM Measurements, 8PSK - 7/21/95
x 10^4
9
8
7
6
5
4
3
2
1
ATM Parameters
CLC
ECC
10^-6
10^-5
10
Measured BER


CRC ATM Measurements, 8PSK - 7/21/95
ATM Parameter
Measured BER
CLR
CER
PRBS-BER
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-6}
10^{-5}


MC ATM Measurements @ CRC, 8PSK w/o RS 8/11/95
Received PLCP Alarms
BER
BI Rate
FEBE Rate
Framing Rate
10-2
10-3
10-4
10-5
10-6
10-6
10-5


MC ATM Measurements @ RL, 8PSK w/o RS 8/11/95
FEBE Rate
P_Bit Rate
Framing Rate
Received Frame Alarms
10^-1
10^-2
10^-3
10^-4
10^-5
10^-6
10^-6
10^-5
BER

PROBING $V_{In}(PH)_4$ IN InP WITH ELECTRON PARAMAGNETIC RESONANCE

Dennis P. Clougherty
Assistant Professor
Department of Physics

University of Vermont
Cook Physical Science Bldg.
Burlington, VT 05405-0125

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling AFB, Washington DC

and

Rome Laboratory

August 1995

Probing $V_{In}(PH)_4$ in InP with Electron Paramagnetic Resonance

Dennis P. Clougherty

Department of Physics University of Vermont Burlington, VT 05405

Abstract

Electron paramagnetic resonance (EPR) is proposed as a technique to search for the presence of the hydrogen complex $V_{In}(PH)_4$, conjectured to be a donor impurity in InP. A vacancy model is proposed to describe the electronic structure of $V_{In}(PH)_4$. In the absence of a pseudo Jahn-Teller effect, an isotropic g value near the free electron value is predicted. The hyperfine interaction arising from the Fermi contact term on the H sites can be modeled by an effective spin Hamiltonian. The EPR signature of the presence of the $V_{In}(PH)_4$ complex is five transitions in the first-order hyperfine structure which are approximately isotropic. The effect of isotopic substitution ($H \rightarrow D$) is discussed.

Probing $V_{In}(PH)_4$ in InP with Electron Paramagnetic Resonance

Dennis P. Clougherty

I. INTRODUCTION

The abundance and small size of hydrogen insures its presence in semiconductors. Hydrogen can appear interstitially in atomic form, or it can form a variety of complexes either with dopants or with intrinsic defects. The study of hydrogen and its associated complexes in semiconductors is of importance as it is known that its presence can have profound effects on subsequent device performance and reliability.

Electronic properties of hydrogen complexes depend in a sensitive way on their structure in the semiconductor. Infrared (IR) spectroscopy has proven to be a valuable tool, providing in some cases, clues to the structure [1,2]. Experimentally-measured vibrational modes can be assigned to bonds. Vibrational multiplets provide information concerning the symmetry of the local environment. Isotopic substitution of deuterium (D) gives rise to frequency shifts and symmetry-related splittings, providing further information used in the identification of the vibrational modes. While IR spectroscopy provides some insight into structure, it does not unambiguously determine structure. Consequently, complementary techniques will be required to remove the ambiguity.

Recently, it has been proposed by Bliss *et al.* [2] that the presence of a hydrogen complex in InP, consisting of an In vacancy and four tetrahedrally coordinated H atoms, $V_{In}(PH)_4$, can reconcile differences commonly found between impurity concentration measurements and free carrier concentration in high purity samples. Furthermore, their IR studies with deuterated InP find a vibrational triplet, consistent with a three-fold (t_2) vibrational multiplet being split with partial D substitution for H atoms.

The work summarized here is a theoretical investigation of the possibility of using electron paramagnetic resonance (EPR) techniques to confirm the existence of the $V_{In}(PH)_4$ complex in InP. It should be noted that EPR has been used successfully on InP to probe Fe impurities [3]. While the hydrogen complex of interest was not detected in these experiments, one would not expect to see a signal associated with $V_{In}(PH)_4$ as the operating temperature was near room temperature where the probability of occupancy of the defect state is vanishingly small.

II. ELECTRONIC STRUCTURE OF $V_{IN}(PH)_4$

The vacancy model for the electronic structure of $V_{In}(PH)_4$ is proposed below. It is assumed that this complex is isolated in bulk InP. The complex is formed by creating first an indium vacancy. This gives rise to four vacancy states formed from the dangling bonds contributed by the four P atoms surrounding the vacancy. Given the tetrahedral symmetry of the vacancy environment, these dangling bonds can be combined to form a_1 and t_2 orbitals, with the a_1 orbital located inside the valence band and the triply degenerate t_2 orbitals located in the band gap, as illustrated in Fig. 1.

A cluster of H_4 arranged on the vertices of tetrahedron has molecular orbitals of a_1 and t_2 symmetry, both below the top of the valence band. We can approximate the H_4 orbitals as a linear combination of atomic orbitals (LCAO). Labeling the four H atoms as indicated in Fig. 2, the a_1 orbital of the cluster is given by

$$|a_1\rangle = \frac{1}{2}\left(|\phi_a\rangle + |\phi_b\rangle + |\phi_c\rangle + |\phi_d\rangle\right) \tag{1}$$

where $|\phi_i\rangle$ is an atomic s orbital centered on site i. The t_2 states of the cluster are given by

$$|\xi\rangle = \tfrac{1}{2}\left(|\phi_a\rangle - |\phi_b\rangle - |\phi_c\rangle + |\phi_d\rangle\right) \tag{2}$$

$$|\eta\rangle = \tfrac{1}{2}\left(|\phi_a\rangle + |\phi_b\rangle - |\phi_c\rangle - |\phi_d\rangle\right) \tag{3}$$

$$|\zeta\rangle = \tfrac{1}{2}\left(|\phi_a\rangle - |\phi_b\rangle + |\phi_c\rangle - |\phi_d\rangle\right) \tag{4}$$

where $|\xi\rangle$, $|\eta\rangle$, and $|\zeta\rangle$ are orbitals which transform as yz, zx, and xy under the operations of the group T_d.

When introduced into the vacancy, we expect the complex to mix with the vacancy orbitals, forming the following states: (1) a_1 state whose character is bonding between the H_4 cluster and the four P neighbors and whose energy will be deep in the valence band; (2) a bonding t_2 manifold below the valence band edge; (3) an a_1^* state which is H-H bonding and P-H antibonding and is located in the band gap; and (4) a t_2^* manifold which is P-H antibonding. This electronic structure is analogous to that of the Si:Pt^- defect considered by Anderson *et al.* [5].

At suitably low temperature, we expect the purely bonding states to be completely filled, while the a_1^* state contains a single unpaired electron; the t_2^* manifold will be unoccupied. Thus the ground state configuration of the complex is $^2S_{1/2}$. The single electron bound to the complex should have a non-vanishing EPR signal, in analogy with that of the F center in alkali halides [6].

III. EPR SPECTRUM

A. The Effective Hamiltonian

For a doublet ground state in tetrahedral symmetry, an isotropic g factor results. Thus, the Zeeman interaction is modeled by an effective spin Hamiltonian of the form

$$\mathcal{H}_z = g\mu_B \vec{S} \cdot \vec{B} \tag{5}$$

where $S = \frac{1}{2}$. It is anticipated that $g \approx g_e = 2.0023$, as the ground state doublet will not mix appreciably via the weak spin-orbit interaction with the excited sextet. Thus, EPR should indicate a single transition in the fine structure, $M_s = \frac{1}{2} \leftrightarrow -\frac{1}{2}$.

The hyperfine interaction of the doublet with the nuclear spins of the H atoms introduces additional transitions. Motivated by Eq. 1, the hyperfine interaction can be viewed as a superposition of H nuclei– each nucleus i with $I_i = \frac{1}{2}$ – interacting with the atomic 1s states

$$\mathcal{H}_{\rm hf} = A\vec{S} \cdot \sum_i \vec{I}_i \tag{6}$$

Corrections to the LCAO form of the a_1^* wavefunction will result in hyperfine anisotropy.

Within the LCAO approximation, it is anticipated that the first-order hyperfine spectrum has 5 equally spaced transitions, corresponding to the possible values of I_z which can be obtained by summing 4 spin-1/2 nuclei. Enumerating the number of ways of getting the various values of I_z gives the relative intensity of these transitions in the ratio of 1:4:6:4:1. The first-order hyperfine spectrum is given in Fig. 3.

The next nearest shell of nuclei correspond to the phosphorus atoms adjacent to the H atoms of the cluster. The nucleus of phosphorus is spin-1/2. Inclusion of the hyperfine interaction with the phosphorus shell is more involved; since the a_1^* orbital contains substantial phosphorus p-character, the hyperfine coupling to phosphorus is axially anisotropic. It is anticipated, however, that the magnitude of the elements of the hyperfine coupling tensor will be reduced from that of the hydrogen hyperfine coupling, as the amplitude of the wavefunction on the phosphorus nuclei will be reduced.

For dilute isotopic substitution of H $\rightarrow$ D, it is expected that the number of spectral lines will be increased to seven. The three H atoms will contribute four lines, corresponding to the four possible I_z values from $I = \frac{3}{2}$, while the D atom will contribute three additional lines as D has a nuclear spin of 1. The intensity ratios are 1:1:3:1:3:1:1. While for complete isotopic substitution, nine lines are expected with intensity ratios of 1:4:10:16:19:16:10:4:1.

B. Estimation of Parameters

The energy difference between the orbital singlet and triplet states is an important parameter in the analysis. We can approximate this difference by sphericalizing the tetrahedral arrangement of H atoms: we envision the nuclear charge ($Z = 4$) as being smeared out into a spherical shell. This technique of sphericalizing high symmetry structures has recently been used successfully on other clusters [4].

The process of approximating the nuclear potential by its angular average is tantamount to truncating the expansion of the potential in multipole moments after $\ell = 0$. For a cluster with T_d symmetry and one-electron states which transform as a_1 and t_2, it is not necessary to expand the nuclear potential beyond the $\ell = 0$ moment, as the multipole moments $\ell = 1, 2$ vanish.

The radius of the sphere is estimated by using known bond lengths for the In-P bond ($d_{In-P} \approx 2.54$ Å) and the P-H bond ($d_{P-H} \approx 1.60$ Å). The radius of the sphere for the H_4 cluster in the In vacancy is approximately $R \approx d_{In-P} - d_{P-H} \approx 0.94$ Å.

Consider first the eigenenergies of the H_4^{3+} within the spherical approximation. The lone electron experiences a cut-off Coulomb potential given by

$$V_n(\vec{r}) = \begin{cases} Ze/R & 0 < r < R \\ \\ Ze/r & r \geq R \end{cases} \tag{7}$$

The eigenenergy of the 1s state for this potential is $E_b \approx -2.96$ eV, roughly one-fifth as binding as the 1s energy for H.

Once this H_4 cluster is introduced into the In vacancy, the surrounding semiconductor screens the potential, and the binding energy will be further reduced, giving

$$E_b^* \approx \frac{E_b}{\epsilon^2} \tag{8}$$

where ϵ, the dielectric constant of InP, is roughly 10. This yields an estimate of the binding energy of $E_b^* \approx 30$ meV. Thus, any EPR signal for the complex will only appear at temperatures well below room temperature, where there is substantial occupancy of the impurity state.

We estimate the isotropic H hyperfine parameter by considering the magnitude of the Fermi contact term [7] for the hydrogen 1s state, A_{1s}

$$A_{1s} = \tfrac{4}{3} g_p \left(\tfrac{m_e}{M_p}\right) m_e c^2 \alpha^4 \tag{9}$$

$$\approx 1420 \text{ MHz} \tag{10}$$

As the H_4 cluster has its lone a_1^* electron distributed among the four H atoms, we estimate that $A \approx A_{1s}/4 \approx 355$ MHz. This should roughly be the spacing of the hyperfine spectral lines.

IV. DISCUSSION

The analysis summarized does not include several interesting effects. The fine structure was found to be isotropic; however, pseudo Jahn-Teller (PJT) distortions arising from vibronic coupling of the a_1^* and t_2^* states may introduce fine structure anisotropy and substantial shifts in the g-value.

As noted previously, the hyperfine coupling to the phosphorus nuclear spins was not included. This will certainly add transitions, and it may well wash out some of the structure resulting from the H hyperfine transitions; however, this kind of inhomogeneous line broadening can be circumvented experimentally with the use of ENDOR [8], rendering all the hyperfine transitions visible.

Future work would include a study of the linewidths, together with the PJT analysis. In that regard, it should be noted that as the experiments are performed, additional theoretical analysis will certainly be required in refining the theoretical model.

REFERENCES

[1] R. Darwich, B. Pajot, B. Rose, D. Robein, B. Theys, R. Rahbi, C. Prote, and F. Gendron, Phys. Rev. B **48**, 17776 (1993).

[2] D. F. Bliss, G. G. Bryant, D. Gabbe, G. Iseler, E. E. Haller, and F. X. Zach, (preprint).

[3] F. X. Zach, J. Appl. Phys. **75**, 7894 (1994).

[4] D. P. Clougherty and X. Zhu, UVM preprint (1995).

[5] F. G. Anderson, F. Ham, and G. D. Watkins, Phys. Rev. B **45**, (1992).

[6] W. C. Holton and H. Blum, Phys. Rev. **125**, 89 (1962).

[7] H. Bethe and E. Salpeter, Quantum Mechanics of One- and Two-Electron Atoms (Plenum, New York, 1977).

[8] A. Abragam and B. Bleaney, Electron Paramagnetic Resonance of Transition Ions (Dover, New York, 1970).

FIGURES

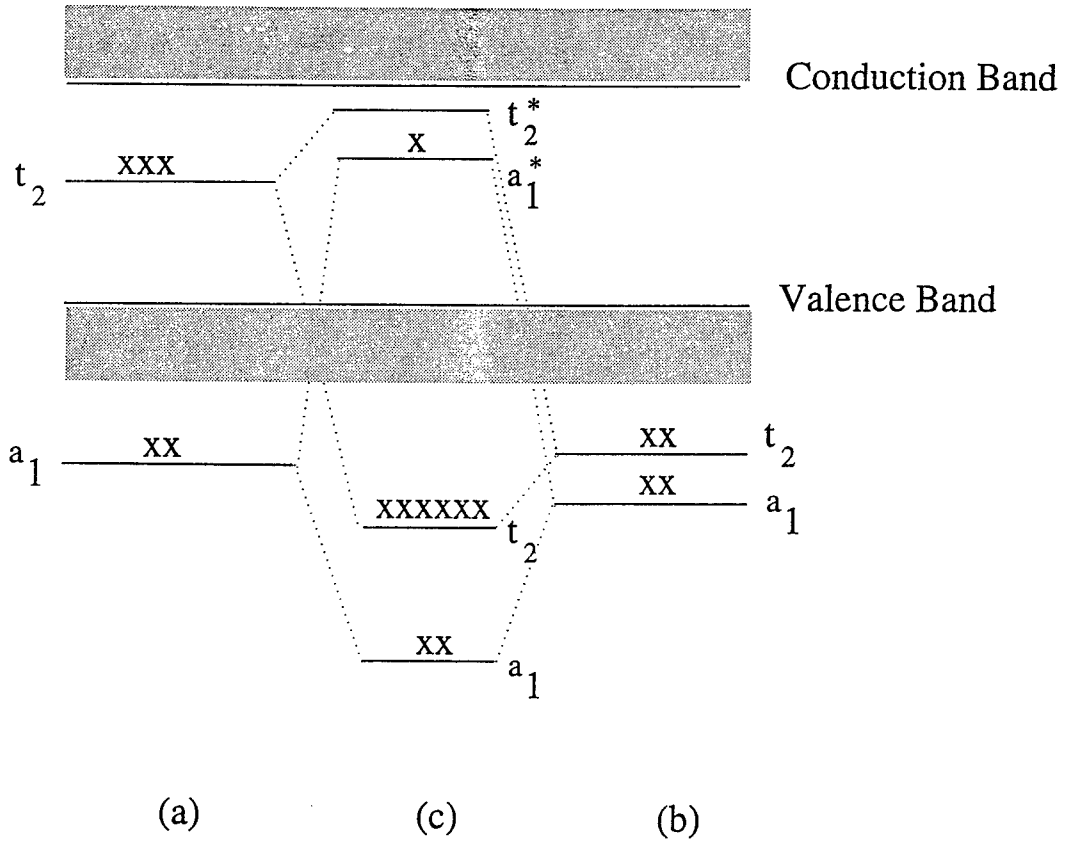


FIG. 1. One-electron states of (a) the In vacancy, (b) the H_4 cluster, and (c) the $V_{In}(PH)_4$ impurity in InP.

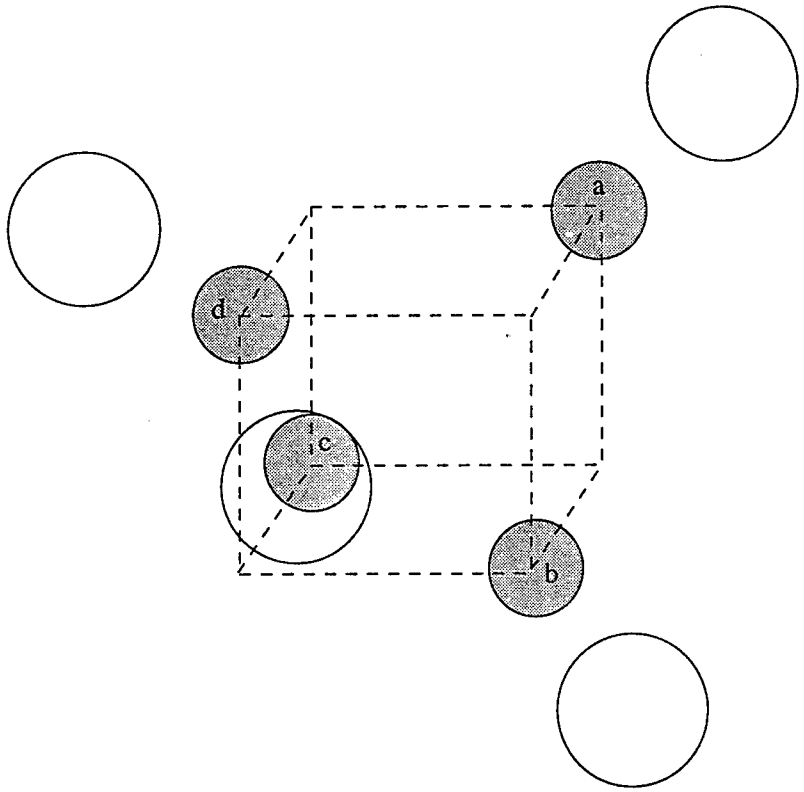


FIG. 2. Geometry of the $V_{In}(PH)_4$ impurity in InP.

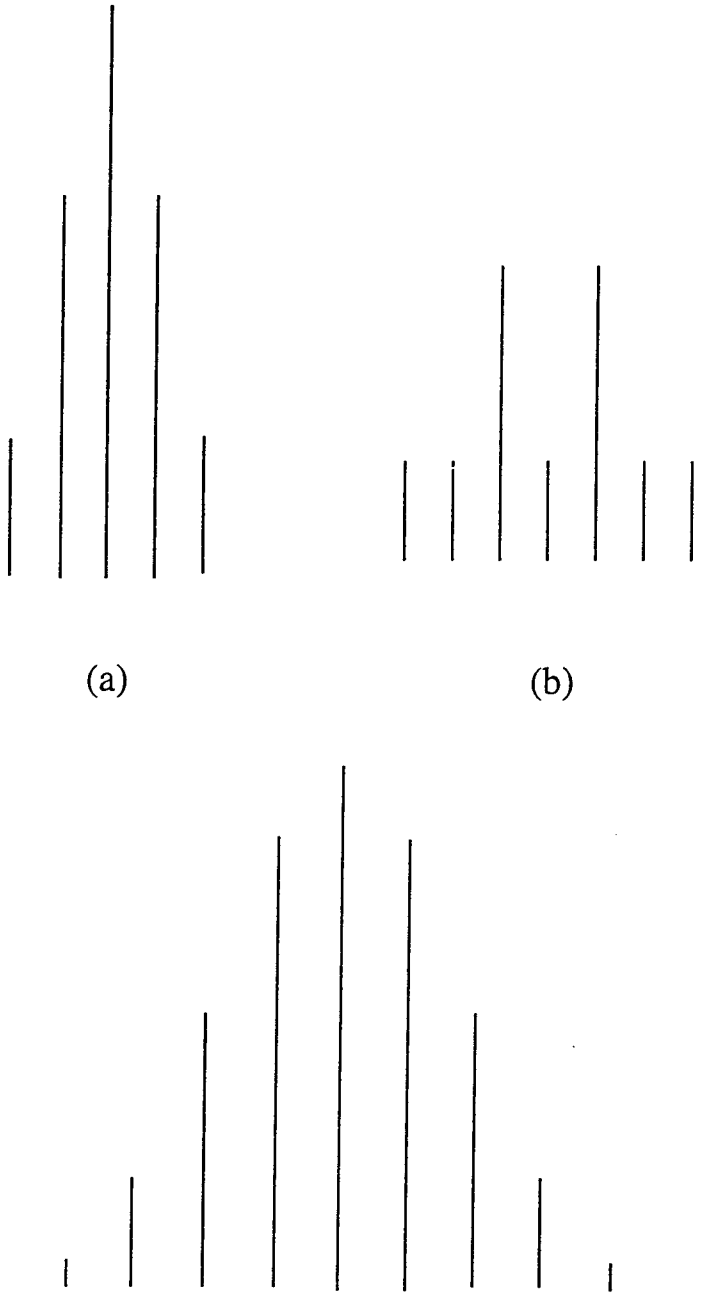


FIG. 3. EPR spectrum (intensity vs. frequency) of (a) $V_{In}(PH)_4$ impurity containing first-order hyperfine structure; (b) $V_{In}(PH)_3(PD)$ impurity resulting from dilute H $\rightarrow$ D substitution; (c) $V_{In}(PD)_4$ impurity resulting from complete H $\rightarrow$ D substitution. Relative intensities within multiplet are approximately to scale, conveying multiplet shape.

A STUDY OF SIMPLE AND EFFICIENT TECHNIQUES FOR LOSSLESS AND NEAR-LOSSLESS COMPRESSION OF DIGITIZED IMAGES

Manohar K. Das
Associate Professor
Department of Electrical and Systems Engineering

Oakland University
Rochester, MI 48309-4401

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington DC

and

Rome Laboratory

August 1995

A STUDY OF SIMPLE AND EFFICIENT TECHNIQUES FOR LOSSLESS AND NEAR-LOSSLESS COMPRESSION OF DIGITIZED IMAGES

Manohar Das
Associate Professor
Department of Electrical and Systems Engineering
Oakland University

Abstract

This report introduces two simple techniques for lossless and near-lossless compression of digitized images and presents a comparative study of the proposed schemes and some of the existing ones. The proposed new techniques represent adaptive and hierarchical versions of the conventional fixed Differential Pulse Code Modulation (DPCM) scheme. These are called Suboptimal Adaptive DPCM (SADPCM) and Hierarchical Block-Adaptive DPCM (HBADPCM) methods, respectively.

The SADPCM lossless coders are mainly useful for archival and/or transmission of still image frames without loss of any data, whereas their near-lossless versions are suitable for applications where loss of some visually imperceptible details may be tolerable without compromising safety. On the other hand, HBADPCM coders provide efficient tools for hierarchical transmission, which allows either accessing the image at progressively improving quality and/or resolution levels, or accessing it gradually according to preferred areas of interest. Such coders are intuitively appealing because they can achieve significant savings in the overall transmission cost by allowing fast browsing of an image database by reducing the volume of unwanted transmission.

The performances of the proposed coders are compared with two of the most popular existing schemes known as the fixed DPCM and Hierarchical Interpolation (HINT) methods. The results of the experimental studies indicate that both SADPCM and HBADPCM coders perform better than the other methods considered. Since the computational complexities of the proposed coders are reasonably low, it is believed that these should find widespread use in applications involving lossless and near-lossless, or hierarchical image compression.

A STUDY OF SIMPLE AND EFFICIENT TECHNIQUES FOR LOSSLESS AND NEAR-LOSSLESS COMPRESSION OF DIGITIZED IMAGES

Manohar Das

1. INTRODUCTION

The sheer explosion of data resulting from the digitization process poses a formidable problem for widespread use of digital techniques for transmission and/or archival of raw images. For instance, a 1024x1024 image, quantized to 8 bits per pixel, requires storage and/or transmission of approximately 1.05 megabytes of picture data. The lossless transmission of such an image over a 19.2 kilobits/second channel requires approximately 7.3 minutes, which is rather long. One of the most promising solution to the data explosion problem is afforded by image coding techniques, which has been a popular research topic for the past two decades. As a result of the concerted efforts of several researchers, a multitude of different image coding techniques have emerged [1]-[4], most of which can be broadly categorized into three classes; namely, lossy, lossless, and near-lossless. A lossless scheme typically achieves a compression ratio of the order of two only, but will allow exact recovery of the original image from the compressed version; a lossy scheme will not allow exact recovery, but can attain much higher compression ratios, e.g., fifty or more. Finally, a near-lossless scheme can deliver higher compression compared to the lossless ones, and at the same, the reconstruction error for each pixel is guaranteed to lie within some predefined limits, such as, ±1, ±2, etc.

Most of the recent advances in image coding have taken place in the area of lossy compression. The currently popular methods include transform coding [2],[3],[5], vector quantization [6],[7], multi-resolution coding [8]-[10], subband coding [11], and adaptive predictive coding (APC) [1],[2],[5],[12]. From the viewpoint of algorithmic complexity, the APC techniques are simpler, but they can only achieve moderately good compression ratio. Thus, for low bit rate lossy coding applications, the other techniques are preferable.

Although lossy compression schemes offer best compression, they are deemed to be unsuitable for medical, analytical, and archival applications, because uncontrolled loss of any information is not tolerable in these situations. Therefore, only lossless or near-lossless schemes are suitable for these applications.

As compared to lossy compression, the lossless case presents a quite different scenario. Although theoretically, any lossy technique, e.g., transform coding or subband coding, can be used for lossless transmission by also transmitting the quantization errors, as Roos et al [13] point out, this results in a rather small compression ratio. Thus, only a handful of different lossless coding techniques and their variants have appeared in the literature till this date. All of them use either a predictive model [13]-[17], or a multi-resolution image model [13],[14],[18]-[20] to reduce the statistical redundancy, and then encode the residuals using an optimal encoder, e.g., Huffman [21], adaptive Huffman [22], Lempel-Ziv [23], or arithmetic coding [24]. For images digitized to 256 gray levels, these techniques typically achieve a lossless compression ratio of 1.5 to 2.5.

There are certain applications involving image analysis, where loss of some quality may be tolerable without compromising safety. Examples range from relatively less quality-constrained applications, such as, preliminary scene analysis, to relatively high quality-critical applications, where some loss in individual pixel values is tolerable only if it is

guaranteed to be within some quantifiable limits set by the user. The latter applications, in particular, need more than just high quality lossy compression, because commonly used lossy coders fall short of providing any guarantee on the loss incurred at each and every pixel; in fact, they can only deliver a reconstructed picture conforming to a desired overall subjective quality level. Near-lossless coders are particularly well suited for the above applications, because they offer significant savings in storage or transmission cost by delivering higher compression compared to the lossless ones, and at the same time, the reconstruction errors for individual pixels are guaranteed to lie within some predefined limits set by the user. Typically, such a technique consists of a combination of a predictive or interpolative lossless coder and a uniform scalar quantizer [25],[26]. The compression gain attainable over a lossless coder depends on the maximum allowable reconstruction error, n. It can be shown [26] that for small values of n, a near-lossless coder achieves a compression gain of approximately $\log_2(2n+1)$ bits less than the underlying lossless coder. For instance, if n = 1, this is a savings of $\log_2 3 = 1.58$ bits per pixel, and if n = 2, this is a savings of approximately 2.3 bits/pixel.

In yet other situations, another kind of image compression methods, known as progressive coders, is required. In progressive coding, first a coarse reproduction of the image is constructed using a small fraction of the pixels, and this coarse reproduction is gradually updated as the remaining pixels arrive. Progressive coding is particularly useful when browsing a large image data base, because in this case the user can quickly determine from the coarse version of the image whether the image is an appropriate one. If not, the user can realize significant cost savings by stopping the transmission of the finer versions of the image.

This study encompasses development of software codes for some existing lossless and near-lossless image coders, evaluation of their performance, and investigation of some improved coding schemes. The specific objectives are summarized below.

2. OBJECTIVES OF THIS STUDY

The three major objectives of this study are:

i. *Development of "C" codes for some lossless and near-lossless coders for incorporation into IE2000 Toolkits*

This involves selection of some existing or new algorithms, and coding them in "C" for incorporation into IE2000 Toolkits.

ii. *Development of new, improved lossless/near-lossless image coders*

This involves investigation of new ideas for improvement of the coding performance of the existing algorithms. Both full-frame and progressive coding schemes are targeted for investigation.

iii. *Comparative performance evaluation of the existing and new coding schemes*

This involves testing the performance of the existing and new coders on a variety of images including standard test pictures, medical images, and IE2000 pictures.

The study of the above objectives is the focus of the following sections.

3. SOFTWARE CODES DEVELOPED FOR INCORPORATION INTO IE2000 TOOLKITS

Although a wide variety of available coding algorithms were tested, the usable "C" codes were developed for three lossless and three near-lossless ones only. These are called space-varying mean (SVM) method, fixed differential pulse code modulation (DPCM) technique, and an adaptive DPCM (ADPCM) method. Among these, the SVM and

DPCM belong to the category of existing methods, whereas the selected ADPCM technique is a new one. The SVM method finds a rounded and equally weighted local mean (alternatively, median) of a pixel from its four nearest causal neighbors [5], subtracts it from the original pixel value to calculate the residual, and transmits the residuals to the receiver. The DPCM technique [5],[13] is similar to SVM, but uses an unequally weighted local mean of only three nearest causal neighbors of a pixel. The ADPCM technique is a new suboptimal technique, the discussion of which is postponed until Section 4.

Each of the above coders consists of two basic modules: a compressor and an expander. The compressor module, in turn, consists of three sub-modules; namely, a residual generator, an integer-to-bit-stream converter, and a shifted Huffman encoder. The residual generator constructs integer residual values using any one of the algorithms mentioned earlier, the integer-to-bit-stream converter converts the residuals first into a string of ASCII characters and then into a bit-stream, and finally, the shifted Huffman encoder encodes the residual bit-stream using either a plain Huffman coder or an adaptive one. Similarly, the expander module consists of three sub-modules; namely, a shifted Huffman decoder, a bit-stream-to-integer converter, and an image reconstructor, the purposes of which are self-explanatory.

There are two program sub-modules that are shared by all the coders; namely, huff.c and bitio.c, which are used for Huffman coding/decoding and input/output of bit-streams, respectively. The remaining modules are:

svmean-c.c (for SVM compression algorithm),

svmean-e.c (for SVM expansion algorithm),

fdpcm-c.c (for fixed DPCM compression algorithm),

fdpcm-e.c (for fixed DPCM expansion algorithm),

badpcm-c.c (for backward ADPCM compression algorithm),

badpcm-e.c (for backward ADPCM expansion algorithm).

The compiled programs are called svmean-c, svmean-e, fdpcm-c, fdpcm-e, badpcm-c, and badpcm-e, respectively. Each of the above programs can be executed as follows:

program-name input output,

which assumes "input" is the file-name for the original image and "output" is the file-name for the compressed image. The programs also ask for the row and column dimensions of the input image, because these are currently meant to be usable with raw images only.

The near-lossless versions of the above coders are organized in an exactly similar fashion, and the executable programs are called svmean-nl-c, svmean-nl-e, fdpcm-nl-c, fdpcm-nl-e, badpcm-nl-c, and badpcm-nl-e, respectively. The usage of these programs is similar to their lossless counterparts, except that in this case the user has to specify the maximum desired error as well.

In terms of relative performance of the three lossless programs, usually badpcm performs slightly better than fdpcm, whereas svmean performs somewhat poorly. However, the current version of badpcm is also the slowest among the three, because it contains a number of redundant arithmetic calculations which should be removed to generate an optimized version of the program. Finally, what was said above about the relative performance of the three lossless

coders, is also true for the near-lossless ones. However, in this case, usually badpcm performs somewhat better than the other two.

4. NEW TECHNIQUES FOR LOSSLESS/NEAR-LOSSLESS IMAGE COMPRESSION

As mentioned before, during the course of this study, two new techniques for lossless and near-lossless image compression were developed. These are called suboptimal adaptive DPCM (SADPCM) and hierarchical block-adaptive DPCM (HBADPCM), respectively. First SADPCM is discussed below, which is followed by HBADPCM.

4.1 A Suboptimal Adaptive DPCM Image Modeling and Estimation Approach

To begin with, we assume that a two-dimensional (2-D) digitized image can be regarded as a nonstationary 2-D signal consisting of pixel intensity values, $\{f(i,j),\ 1 \le i \le L,\ 1 \le j \le L\}$, where i denotes the row index and j stands for the column index, respectively. A conventional 2-D DPCM compression scheme uses an image model of the following form:

$$f(i,j) - a_1 f(i,j-1) - a_2 f(i-1,j) + a_1 a_2 f(i-1,j-1) = w(i,j), \tag{1}$$

where a_1 and a_2 denote the model coefficients, and w(i,j) denotes the modeling error, which is usually regarded as a zero-mean, 2-D white noise sequence. In case of conventional fixed DPCM schemes [13], it is customary to set $a_1 = a_2$ and a commonly used value for both coefficients is 0.95 [13]. Henceforth, this scheme is referred to as the fixed DPCM technique.

In order to develop a suboptimal adaptive DPCM (SADPCM) image model, first rewrite (1) in the form of a multiplicative autoregressive model [14],

$$(1 - a_2 q_1^{-1})(1 - a_1 q_2^{-1})\, f(i,j) = w(i,j), \tag{2}$$

where q_1^{-1} and q_2^{-1} denote the unit backward shift operators in the vertical (i.e., row-wise) and horizontal (i.e., column-wise) directions, respectively. Because of the multiplicative nature of the polynomial operator in the left side of (2), we can further express (2) in the cascade form,

$$(1 - a_1 q_2^{-1})\, f(i,j) = f_1(i,j), \tag{3a}$$

$$(1 - a_2 q_1^{-1})\, f_1(i,j) = w(i,j), \tag{3b}$$

where $f_1(i,j)$ is an intermediate signal regarded as the output of the first stage of the cascade structure.

In order to be useful for adaptive image coding, the coefficients a_1 and a_2, of model (3) must be estimated from the given image data. As discussed in [14], an optimal method of estimating a_1 and a_2 involves either minimization of a nonlinear cost function, or a pseudo-linearization approach. However, both of these approaches are beset with moderately high computational cost, and therefore, they are unsuitable for our purpose. Instead we use a suboptimal estimation scheme, as described below.

4.1.1 A suboptimal scheme for estimation of the model coefficients

The selection of a suboptimal estimation scheme depends on the trade-off between performance and computational complexity. For the simplest coders, presented in this section, we use a fixed value for a_1 and an estimated value for a_2, whereas for the improved coders, presented in a subsequent section, estimated values of a_1 and a_2 are utilized. To find the estimates of both a_1 and a_2, we used the following three-step suboptimal strategy:

Step 1. Estimate a_1 from (3a) ignoring the correlation, if any, between $f_1(i,j)$ and f(i,j);

Step 2. Using the estimated value of a_1, evaluate the residuals of the first stage to form an estimate, $\hat{f}_1(i,j)$, of $f_1(i,j)$ as,

$$\hat{f}_1(i,j) = f(i,j) - \hat{a}_1 f(i,j-1);$$

Step 3. Finally, estimate a_2 from (3b).

Notice that the above procedure is a suboptimal one because the estimate of a_1 is biased due to non-zero correlation between f(i,j) and $f_1(i,j)$. However, it is still useful for predictive coding because of two reasons: i) in a predictive coder, the modeling inadequacies are masked to a large extent by adding the quantized residual errors back to the predicted pixel values, and ii) the error cannot grow provided the estimated model is stable.

Using the above strategy, estimates of a_1 and a_2 are simply obtained as,

$$a_1 = r_f(0,1)/r_f(0,0), \tag{4a}$$

$$a_2 = r_{f1}(1,0)/r_{f1}(0,0), \tag{4b}$$

where $r_f(k,l)$ and $r_{f1}(k,l)$ denote the autocorrelation coefficients with 'lag (k,l)' of f(i,j) and $f_1(i,j)$, respectively. Notice that for raw image data, the value of a_1 almost always lies between 0.95 and 1.0, which represents the typical range of normalized column-wise autocorrelation coefficients for most images. Thus, for the simplest SADPCM coders, which process f(i,j) directly, the value of a_1 can be assumed to be a constant. In our first set of SADPCM experiments, the value of a_1 was kept fixed at 0.99, and only a_2 was estimated using steps (2) and (3) mentioned above. It may be mentioned that the choice of a fixed value for a_1 is not very critical; in fact, only slight differences in the overall performance were noticed by assigning different constant values to a_1 over the range, [0.97, 1.0).

Next, notice that the estimation of a_2 can be carried out either in a forward, or backward fashion. In the forward scheme, the image is first subdivided into smaller, say MxM, blocks, and a separate set of coefficients is estimated for each block. In this case, the blockwise estimates of a_2 need to be transmitted as the side information. On the other hand, in the backward scheme, the estimation of a_2 is carried out over a causal window consisting of a few reconstructed pixel values, and therefore, avoids the necessity of transmitting any side information. The overall coding schemes based on forward and backward estimation techniques are henceforth referred to as forward ADPCM (FADPCM) and backward ADPCM (BADPCM), respectively.

In the forward scheme, two alternative estimation strategies are useful; namely, batch and recursive methods. The batch method involves blockwise estimation of $r_{f1}(1,0)$ and $r_{f1}(0,0)$, followed by evaluation of a_2 from (4b). The recursive method, on the other hand, involves calculation of a_2 using a recursive estimation algorithm, such as, recursive least squares (RLS) or least mean squares (LMS) [27]. Since we need to estimate a single parameter, the use of RLS is preferred because it converges faster and involves only a little extra computation than LMS. The RLS estimation update equations can be summarized as follows:

error update: $$e(i,j) = f_1(i,j) - a_2 f_1(i-1,j), \tag{5a}$$

covariance update: $$p^u = p^o/g, \tag{5b}$$

normalizing gain: $$g = 1 + p^o[f_1(i,j)]^2, \tag{5c}$$

parameter update: $$a_2^u = a_2^o + p^u f_1(i,j)e(i,j), \tag{5d}$$

where the subscripts 'o' and 'u' refer to the old and the updated values of the associated variables.

In the backward scheme, the use of a recursive estimation technique is mandated, and two alternative strategies were found to be useful. The first method uses either RLS or LMS to update the a_2 estimate for each pixel. In the second method, first the values of $r_{f1}(1,0)$ and $r_{f1}(0,0)$ are updated over a sliding causal window, and then pixel-by-pixel estimates of a_2 are calculated from (4b). The window used in our experiments is shown in Fig. 1. Using this window, the estimated values of $r_{f1}(0,0)$ and $r_{f1}(1,0)$ for the $(i,j)^{th}$ pixel are obtained as,

$$r_{f1}(0,0) = (\Sigma\,\Sigma_{m,n\in S}\,[f_1(i-m,j-n)]^2)/7.0, \tag{6a}$$

$$r_{f1}(1,0) = [f_1(i,j-1)f_1(i-1,j-1) + f_1(i-1,j-1)f_1(i-2,j-1) + f_1(i-1,j)f_1(i-2,j) + f_1(i-1,j+1)f_1(i-2,j+1)] / 4.0, \tag{6b}$$

where $S = \{(m,n) \mid m\in[0,2]$, $n\in[-1,1]$, and $n \neq 0$, or 1 when $m=0\}$. It may be mentioned that because of the partial overlap of the windows for two neighboring pixels, the actual computational load required for implementation of (6a) and (6b) is only about five multiplications and four additions per pixel, which can be reduced further by choosing a window of smaller size.

In both forward and backward approaches, the robustness of the overall scheme is greatly improved by adding a stability check for $\hat{a}_2$, the estimated value of a_2, at every step. Notice that the stability of the estimated model is guaranteed if

$$|\hat{a}_2| < 1. \tag{7}$$

If (5) is violated, the estimate is first projected inside the stable zone before continuing with prediction and encoding steps.

Next, lossless and near-lossless image coding schemes based on the above SADPCM modeling approach is presented below.

4.2 Lossless and Near-Lossless Image Coding Using SADPCM Modeling Approach

4.2.1 Lossless coding

As mentioned earlier, two different coding schemes are possible; namely, FADPCM and BADPCM. Since they differ only in the estimation methodology, we summarize the overall coding scheme below for a generic SADPCM model only.

A lossless image coder based on the SADPCM model consists of two main components, namely, a predictor and an entropy coder. The predictor simply uses (1) to calculate a predicted value of f(i,j) as,

$$\tilde{f}(i,j) = 0.99f(i,j-1) + \hat{a}_2 f(i-1,j) - 0.99\hat{a}_2 f(i-1,j-1), \tag{8}$$

where a_1 is chosen to be 0.99 and a_2 denotes the estimated value of a_2. Next, f(i,j) is rounded to generate the integer predicted values, $f_r(i,j)$, i.e.,

$$f_r(i,j) = R[\tilde{f}(i,j)], \tag{9}$$

where R[x] denotes the nearest integer value of x. The residual signals, d(i,j), are then obtained as,

$$d(i,j) = f(i,j) - f_r(i,j). \quad (10)$$

Finally, the residual sequence, $\{d(i,j)\}$, is entropy coded using an optimal encoder, e.g., adaptive Huffman [22], or arithmetic coding [24] and the coded residuals are transmitted to the receiver. At the receiver end, f(i,j) is exactly reconstructed by first synthesizing $f_r(i,j)$ using (8), (9) and then utilizing the decoded d(i,j) to obtain

$$f(i,j) = f_r(i,j) + d(i,j). \quad (11)$$

4.2.2 Near-lossless coding

As mentioned before, the near-lossless coders attempt to improve the compression efficiency of lossless coders by allowing a controlled loss in the reconstructed image. A specific class of such coders, which consists of a lossless coder followed by a uniform scalar quantizer, was introduced in [25],[26]. Since the same strategy is utilized here and details are available in [25],[26], only the main ideas are summarized below.

In the case of near-lossless coders, the predicted value of f(i,j) is obtained as,

$$\tilde{f}(i,j) = 0.99\hat{f}(i,j-1) + \hat{a}_2\hat{f}(i-1,j) - 0.99\hat{a}_2\hat{f}(i-1,j-1), \quad (12)$$

where $\hat{f}(m,n)$ denotes the reconstructed value of f(m,n). Next, the prediction error, d(i,j), and its quantized value, $\hat{d}(i,j)$, are computed as

$$d(i,j) = f(i,j) - R[\tilde{f}(i,j)], \quad (13a)$$

$$\hat{d}(i,j) = Q[d(i,j)], \quad (13b)$$

where Q[.] denotes a center-clipping quantizer whose input-output relationship is dictated by the maximum allowable reconstruction error (MARE). If MARE is ±n, Q is chosen as,

$$Q[k] = m \text{ for integer } k \in [m-n, m+n], \quad (14)$$

where m=0, ±(2n+1), ±(4n+1), etc.

Finally, the reconstructed pixel, $\hat{f}(i,j)$, of f(i,j) is obtained as

$$\hat{f}(i,j) = R[\tilde{f}(i,j)] + \hat{d}(i,j), \quad (15)$$

and at the end, after repeating the above steps in a recursive fashion, the sequence $\{\hat{d}(i,j)\}$ is entropy coded and transmitted. At the receiver, the reconstructed image is obtained recursively by first computing $\tilde{f}(i,j)$ from (12), rounding it to the nearest integer and then adding the same to the decoded value of $\hat{d}(i,j)$, as in (15). This compression scheme is a near-lossless one because by choosing 'n' in equation (15) to be small, a nearly perfect version of the original image can be reconstructed at the receiver end.

4.2.3 Entropy coding of the residuals

As mentioned before, the residuals of both lossless and near-lossless coders can be entropy coded using either Huffman, adaptive Huffman, or arithmetic coders. In the initial stage of our experiments, we performed a comparative study between plain Huffman and its adaptive version, and found that the performance of plain Huffman is only marginally inferior compared to its adaptive counterpart, which achieves an average bit rate that is very close to the first order entropy of the residuals. Therefore, in the subsequent experiments, we compared performance of different coders based on the first order entropies of the residual images only. It may be mentioned that a similar approach has been followed by other researchers [13].

4.3 Experimental Results Using Simple SADPCM Coders

For a comparative performance evaluation of the simple adaptive lossless coders introduced above with the existing schemes, we chose HINT [13] and fixed DPCM [13] because they are known to perform better than other simple lossless coders [13],[14]. For the fixed DPCM method, a_1 and a_2 in (1) are chosen to be 0.95. For the HINT scheme, we implemented the 4x4 version, which is briefly outlined below.

4.3.1 Hierarchical interpolation (HINT) method

The hierarchical interpolation (HINT) [13] is a multi-resolution pyramid coding scheme that begins with a low-resolution version of the original image, P_0, and successively generates the higher resolutions $P_k, 1 \leq k \leq n$, using noncausal interpolations. The lowermost resolution, P_0, is entropy coded and transmitted first. Thereafter, in a hierarchical fashion, the interpolation scheme is used to generate estimates of the unknown pixel values of P_k by calculating the average of its four nearest neighbors obtained from P_{k-1}. The estimates are rounded to their nearest integers and then subtracted from the true pixel values. The difference signals pertaining to each of the higher resolutions, D_k, $1 \leq k \leq n$, are also entropy coded and transmitted. Details of the operations can be found in [13].

As far as encoding of the difference signals is concerned, two kinds of HINT schemes are possible; namely, i) a scheme that uses only one codebook for coding all the residual errors, which is simply referred to as HINT, and ii) a scheme that uses five different codebooks for coding the residual errors pertaining to different resolutions, which is called HINTS. Although HINTS is expected to perform better than HINT, the former requires more computation and additional side information.

4.3.2 Comparison of computational complexity of SADPCM, fixed DPCM, and HINT/HINTS

Notice that the computational complexity of SADPCM is only marginally higher than that of either DPCM or HINT. Consider, for instance, the FADPCM scheme described in Section 2 above. Using the autocorrelation approach for estimation of a_2, the extra computational load for FADPCM compared to fixed DPCM, or HINT, consists of approximately three multiplications and three additions per pixel. The comparison between SADPCM and HINTS becomes more difficult because of the fact that HINTS requires construction and transmission of five codebooks instead of only one. Thus, the extra computational load of HINTS, as compared to HINT, depends upon the choice of the entropy coding algorithm.

4.3.3 Comparative experimental results of FADPCM/BADPCM, HINT, and fixed DPCM

First, the lossless coding results are compared. Eight images, digitized to 256 gray levels, are chosen for this study. Two of these are digitized radiographs, called R1 and R2, respectively, and six others are standard test images available at a number of internet sites. Most of the standard pictures in our experiments are taken from the test-image database at the RPI site. These are referred to as Lena, Pepper, Jet, Flower, Cameraman, and Bridge, respectively. All the pictures selected for this experiment. with the exception of Flower, Cameraman and Bridge, are (512x512) in size. The picture, Flower, is of size (480x512), whereas both Cameraman and Bridge are of size (256x256). In this study, performances of both FADPCM and BADPCM are compared with those of fixed DPCM, HINT, and HINTS. The block size chosen for FADPCM is 32x32, and the estimation technique used include: i) block-by-block autocorrelation

based approach for FADPCM, and ii) the sliding window autocorrelation approach for BADPCM. It may be mentioned that virtually identical results were also obtained using the RLS estimation technique.

As mentioned before, the performances of different methods are compared on the basis of the average bit rates, as measured by the first order entropies of the residuals plus the side information bit rates, if any. Table 1 summarizes these lossless coding results. The results clearly indicate that in terms of lossless compression efficiency, SADPCM generally performs better than the other methods considered.

Next, we compare the performance of near-lossless coders based on FADPCM, BADPCM, fixed DPCM, and HINT. The results obtained for MARE values of ±1, and ±2 are shown in Table 2. Once again, the results clearly indicate that the proposed techniques generally perform better than the fixed DPCM technique.

4.4 Improved SADPCM Coding Schemes

The purpose of this section is to demonstrate how the coding performance of the above SADPCM coders can be improved at a little additional computational cost. The basic idea and the motivation behind it are summarized below.

Notice that most image data are nonstationary in-the-mean, which is usually caused by the combined effect of both reflectance and lighting variations over different parts of an object or scene. Since predictive coders work best on stationary data, it makes sense to attempt to improve the performance of the proposed SADPCM coders by removing the nonstationary attributes of the image data.

The simplest method of eliminating nonstationarity-in-the-mean consists of subtraction of an estimated local mean from each pixel value. Since the SADPCM coders utilize two coefficients associated with the nearest causal neighbors of each pixel, f(i,j), it makes sense to obtain the staionary-in-the-mean image as follows:

$$f_m(i,j) = R[0.5*(f(i,j-1) + f(i-1,j))], \tag{16a}$$

$$f_s(i,j) = f(i,j) - f_m(i,j), \tag{16b}$$

where $f_m(i,j)$ denotes the local mean at location (i,j), R[.] denotes the operation of rounding to the nearest integer value, and $\{f_s(i,j)\}$ denotes the stationary-in-the-mean image data.

Next, the SADPCM scheme is utilized to code $\{f_s(i,j)\}$ using either forward or backward approach, as discussed earlier. However, in this case, both a_1 and a_2 need to be estimated because a fixed value for a_1 cannot be assumed any more. The modified methods are henceforth called improved FADPCM (I-FADPCM) and improved BADPCM (I-BADPCM) schemes, respectively. Next, some lossless compression results using I-FADPCM only are given in the following subsection.

4.4.1 Experimental results using I-FADPCM

Table 3 shows the lossless compression results on the same set of eight images using I-FADPCM. For the sake of comparison, the results using basic FADPCM are also quoted there. As can be readily seen, the improved version of FADPCM performs better, albeit with some extra computational complexity. Similar improvements are also attainable using I-BADPCM rather than the basic BADPCM.

Next, HBADPCM coders are discussed.

4.5 Hierarchical Block-Adaptive DPCM Coders

There are two main reasons that motivated us to pursue the development of hierarchical block-adaptive DPCM (HBADPCM) coders. First, we wanted to adapt SADPCM coders for hierarchical transmission in order to achieve better compression than that afforded by the existing simple and elegant hierarchical coders, such as, HINT. Second, we wanted to develop a scheme that allows the user to access different areas of an image gradually according to his/her preference. As far as the first goal is concerned, since predictive coders work well only on a full image frame or subsections of it, and performs rather poorly on subsampled images, we propose to develop a block-by-block encoding and transmission method. This also seems to be good for meeting our second goal because an user can selectively access specific blocks in any desired sequence.

In order to achieve blockwise hierarchical transmission, first we must subdivide the image into a number of smaller, say MxM, blocks. The next thing needed is adoption of a suitable hierarchical framework for coding and transmitting different blocks. Although a variety of different hierarchical frameworks could have been chosen for this purpose, we picked the framework of HINT in this study mainly because of its excellent performance as a hierarchical lossless coder. Also, as shown in the following subsection, this framework allows a nice way of implementing a backward HBADPCM coder.

Following [13] then, we arrange the blocks to be coded in the hierarchical order illustrated in Fig. 2. In this figure, the blocks are marked according to their allotted level of hierarchy. Thus, the blocks marked "4" denote the ones to be coded and transmitted first, followed by the blocks marked "3", and so on.

Next, the coding and transmission of each individual block is carried out using either FADPCM (which requires blockwise estimation of a_2 only), or its improved version, I-FADPCM (which requires blockwise estimation of both a_1 and a_2). Then the block-by-block residuals pertaining to each hierarchical level are encoded using an entropy coder and the coded residuals are transmitted. It is pointed out that there is no need to construct separate codebooks for different hierarchical levels, because all the residuals generated by FADPCM or I-FADPCM exhibit narrow peaks and can be efficiently coded using a single codebook. The overall coding schemes are henceforth referred to as forward HBADPCM (F-HBADPCM) and improved F-HBADPCM, respectively.

Until now, our only use of the HINT framework has been restricted to the task of laying out a hierarchy of blocks. As expected, fruitful exploitation of the above framework allows further improvement of the HBADPCM coder discussed above. One such improvement, namely, a backward HBADPCM coder, is presented below.

4.5.1 A backward HBADPCM scheme

The main objective of a backward HBADPCM scheme would be to modify the parameter estimation strategy such that transmission of the block-by-block model coefficients becomes unnecessary. In this study, two useful strategies are used to realize the above goal; namely, i) a backward estimation scheme, where the model coefficients pertaining to a block are estimated from the past reconstructed pixel values within that block, and ii) a hierarchical approximation scheme, where the block-by-block model coefficients are first laid out in the same hierarchical pattern as the blocks themselves, and an approximate set of blockwise model coefficients is formed through hierarchical interpolation from

the past reconstructed blocks' coefficients. This set of approximate coefficients is used by both transmitter and receiver avoiding the necessity of transmittal of any side information.

A comparative experimental study of the two backward HBADPCM schemes indicates that the hierarchical approximation scheme performs marginally better than the backward approximation one. Because of this and the fact that the latter attempts a true exploitation of the HINT framework, we choose the HINT approximation approach here. However, it may be pointed out that like in the forward case, there exist two versions of the scheme; namely, one based on blockwise estimates of a_2 only, and the other based on blockwise estimates of both a_1 and a_2. The overall coding schemes are henceforth referred to as the backward HBADPCM (B-HBADPCM), and improved B-HBADPCM, respectively. Finally, it is pointed out that like in the case of SADPCM, the residuals can be entropy coded using either Huffman, or arithmetic coders, both of which achieve an average bit rate, which is very close to the first order entropy of the residulas. Hence the performances of different coders were evaluated on the basis of the first order entropies of the residual images only.

Next, one important thing that needs to be addressed relates to the question of how to interpolate for the missing blocks so that higher level approximations can be generated, if needed, from the reconstructed image blocks pertaining to a lower hierarchical level. This issue is addressed in the next subsection.

4.5.3 A scheme for interpolation for the missing blocks

Notice that the problem of interpolation for the missing blocks, in the above context, bears a striking similarity to the problem of error concealment in case of loss of cells during transmission [28],[29]. Thus, techniques used for error concealment - particularly, the spatial domain ones, are well suited to our task.

In view of above, we propose to use a simpler version of the projective interpolation technique [29], for demonstration of interpolation for the missing blocks. In the projective interpolation scheme, Jung et al suggests using bilinear interpolation adapted to the edge pattern of a missing block. Essentially, their method consists of four steps; namely, i) examination of the boundary pixel values of a missing block for possible edge patterns within the missing one, ii) classification of a missing block into one of six categories depending on the possible edge patterns, iii) determination of the interpolation direction, and iv) interpolating the missing block using a weighted bilinear interpolation.

In this study, we use a simpler version of the above technique, which can be summarized as follows:

- For each pixel within a missing block, examine four pairs of boundary pixel values - located along the horizontal, vertical, 45^0, and 135^0 directions;
- Identify the pair possessing minimum absolute difference among the four, and perform a linear interpolation in the corresponding direction.

Finally, some experimental results are presented in the following section.

4.6 Experimental Results Using HBADPCM

In this section we present some experimental results comparing the performance of the proposed coders with that of HINT [13]. Before presentation of the quantitative results, however, first we compare the computational complexities of the methods considered.

4.6.1 Comparison of computational complexity of HBADPCM and HINT

Notice that HBADPCM requires estimation of model coefficients. Thus, its computational complexity is slightly higher than that of HINT. Consider, for instance, the basic forward HBADPCM coder, F-HBADPCM. Using the autocorrelation approach for estimation of $a_?$, the computational load due to parameter estimation consists of approximately three multiplications and three additions per pixel. This, however, gets approximately doubled for the improved F-HBADPCM, which requires blockwise estimation of both a_1 and a_2.

On the other hand, it may be pointed out again that the proposed coders do not require multiple codebooks for transmission of the residuals pertaining to different hierarchical levels. This constitutes one of the key advantages of the proposed coders over HINT, because the latter does not perform well without separate codebooks. To stress this point further, experimental results are given for both single codebook and five codebook versions of HINT. The five codebook version, which uses separate codebooks for different hierarchical levels, is henceforth referred to as HINTS.

4.6.2 Comparative experimental results

In this comparative study, the test-images used are same as the ones utilized earlier for the SADPCM coders. The block size chosen for all HBADPCM coders is 32x32. For the basic forward/backward HBADPCM coders, the fixed value of a_1 is chosen to be 0.99, and the autocorrelation approach is used for estimation of a_2. For the improved forward/backward HBADPCM coders, both a_1 and a_2 were estimated using the autocorrelation approach. It may be mentioned that virtually identical results were also obtained using the RLS estimation technique.

As mentioned before, the performances of different methods are compared on the basis of the attainable average bit rates, as measured by the first order entropies of the residuals plus the side information bit rates, if any. Table 4 summarizes these lossless coding results. The results clearly indicate that in terms of lossless compression efficiency, HBADPCM coders perform better than HINT.

Finally, Figs. (3a) and (3b) demonstrate the effectiveness of the simple interpolation scheme for the missing blocks, presented in Section 4.5.3 above. Fig. (3a) depicts the missing blocks of Pepper in Level 1, assuming a block size of 8x8, whereas Fig. (3b) shows the interpolated picture. As is apparent from these figures, even a simple interpolation scheme produces reasonably good quality of interpolated pictures. Of course, block sizes have to chosen small for achieving good results.

5. Conclusion

Three objectives are achieved in this study. First, "C" codes for some existing lossless and near-lossless image coding schemes are developed. Second, two new classes of lossless/near-lossless and hierarchical coders are developed. Finally, a comparative performance evaluation of all the methods is undertaken. The results of the comparative experimental studies indicate that the new methods perform better than hierarchical interpolation (HINT) and fixed DPCM methods. Particularly, the improved SADPCM coders and their hierarchical versions perform significantly better than the other methods considered. Since the computational complexity of SADPCM is only slightly higher than fixed DPCM, and the same holds true for HBADPCM as compared to HINT, it is believed that SADPCM and HBADPCM coders should find widespread use in applications involving fast lossless/near-lossless, and hierarchical compression.

Acknowledgement

The author gratefully acknowledges the help and cooperation afforded by all the following personnel affiliated with the Rome Laboratory:

Mr. Shawn Bisgrove, Mr. Ed Bohling, Ms. Audrey Copperwheat, Mr. Chuck Ferrara, Mr. Jerry Nethercott, Mr. Patrick O'Connor, Mr. David Quinn-Jacobs, Mr. Fred Rahrig, Mr. John Reilly, Mr. Mark Rosiek, Mr. Richard Simard, and Mr. Paul Smith. Without their help, this study could never have been completed.

References

[1] A. K. Jain, "Image Data Compression: A Review," Proceedings of the IEEE, Vol. 69, pp. 349-389, March 1981.

[2] M. Rabbani, editor, *Selected Papers on Image Coding and Compression*, SPIE Optical Engineering Press, Bellingham, Washington, 1992.

[3] W. B. Pennebaker, *JPEG Still Image Data Compression Standard*, Van Nostrand Reinhold, New York, 1993.

[4] R. B. Arps and T. K. Truong, "Comparison of International Standards for Lossless Still Image Compression," Proc. IEEE, Vol. 82, No. 6, pp. 889-899, JUne 1994.

[5] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall, Englewood Cliffs, NJ, 1989.

[6] A. Gersho and R. M. Gray, " Image Coding Using Vector Quantization," Proc. IEEE International Conference on Acoustics Speech and Signal Processing, pp. 428-431, April 1982.

[7] H. M. Hang and J. W. Woods, "Predictive Vector Quantization of Images," IEEE Transactions on Communications, Vol. COM-33, pp. 1208-1219, November 1985.

[8] P. J. Burt and E. H. Adelson, "The Laplacian Pyramid as a Compact Image Code," IEEE Transactions on Communications, Vol. COM-31, No. 4, pp. 532-540, April 1983.

[9] M. Todd and R. Wilson, "An Anisotropic Multi-Resolution Image Data Compression Algorithm," Proc. IEEE International Conference on Acoustics Speech and Signal Processing, pp. 1969-1972, April 1989.

[10] S. R. Burgett and M. Das, "Predictive Image Coding Using Multiresolution Multiplicative Autoregressive Models," IEE Proceedings - Part I, Vol. 140, No. 2, 1993.

[11] J. W. Woods and S. D. O'Neil, "Subband Coding of Images," IEEE Transactions on Acoustics, Speech and Signal Processing," Vol. ASSP-34, pp. 1278-1288, October 1986.

[12] S. R. Burgett and M. Das, "Predictive Image Coding Using Two-Dimensional Multiplicative Autoregressive Models," Signal Processing, Vol. 31, No. 2, 1993.

[13] P. Roos et al, "Reversible Intraframe Compression of Medical Images," IEEE Transactions on Medical Imaging, Vol. 7, No. 4, pp. 328-336, 1988.

[14] M. Das and S. Burgett, "Lossless Compression of Medical Images Using Two-Dimensional Multiplicative Autoregressive Models," IEEE Trans. on Medical Imaging, Vol. 12, No. 4, pp. 7221-726, 1993.

[15] M. Das and C. C. Li, "Simple Space-Varying Least Squares Model for Lossless Medical Image Compression," Electronics Letters, Vol. 30, No. 11, pp. 849-850, 1994.

[16] N. Tavakoli, "Lossless Compression of Medical Images," Proceedings of Fourth Annual IEEE Symposium on Computer-Based Medical systems, pp. 201-207, 1991.

[17] V. K. Heer and H-E Reinfelder, "A Comparison of Reversible Methods for Data Compression," SPIE Vol. 1233, Medical Imaging IV: Image Processing, pp. 354-365, 1990.

[18] H. Blume and A. Fand, "Reversible and Irreversible Image Data Compression Using S-Transform and Lempel-Ziv Coding," SPIE Vol. 1091, Medical Imaging III: Image Capture and Display, pp. 2-17, 1989.

[19] L. Wang and M. Goldberg, "Comparative Performance of Pyramid Data Structures for Progressive Transmission of Medical Imagery," SPIE Vol. 1232, Medical Imaging IV: Image Capture and Display, pp. 403-413, 1990.

[20] L. Wang and M. Goldberg, "Reduced-Difference Pyramid: a Data Structure for Progressive Transmission," Optical Engineering, Vol. 28, No. 7, pp. 708-716, 1989.

[21] D. A. Huffman, "A Method for Construction of Minimum Redundancy Codes," Proc. IRE, Vol. 40, pp. 1098-1101, 1952.

[22] R. G. Galleger, "Variations on a Theme by Huffman," IEEE Transactions on Information Theory, Vol. 24, No.6, pp. 668-674, 1987.

[23] A. Lempel and J. Ziv, "Compression of Two-Dimensional Images," *Combinatorial Algorithms on Words*, pp. 1410-154, Springer-Verlag, 1985.

[24] I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic Coding for Data Compression," Communications of the ACM, Vol. 30, No. 6, pp. 520-540, June 1987.

[25] M. Das and D. L. Neuhoff, "Near-Lossless Compression of Digitized Images," Communications and Signal Processing Lab Report #283, University of Michigan at Ann Arbor, June 1993.

[26] M. Das, D. L. Neuhoff, and C. L. Lin, "Near-Lossless Compression of Medical Images," Proc. IEEE International Conf. on Acoustics Speech and Signal Processing, held in Detroit, Michigan, 1995.

[27] L. Ljung and T. Soderstrom, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, 1983.

[28] A. Narula and J. S. Lim, "Error Concealment Techniques for an All-Digital High-Definition Television System," SPIE, Vol. 2094, pp. 304-315, 1993.

[29] K-H Jung, J-H Chung, and C. W. Lee, "Error Concealment Technique Using Projection Data for Block-Based Image Coding," SPIE, Vol. 2308, pp. 1466-1476, 1994.

X	X	X
X	X	X
X	□	

Fig. 1. Causal window for estimation of $r_{fl}(1,0)$ and $r_{fl}(0,0)$ for each pixel. □ denotes current pixel, X denotes window pixels.

4	0	2	0	4	0	2	0	4
0	1	0	1	0	1	0	1	0
2	0	3	0	2	0	3	0	2
0	1	0	1	0	1	0	1	0
4	0	2	0	4	0	2	0	4
0	1	0	1	0	1	0	1	0
2	0	3	0	2	0	3	0	2
0	1	0	1	0	1	0	1	0
4	0	2	0	4	0	2	0	4

Figure 2. Hierarchical ordering of blocks. Blocks "4" are coded first, followed by blocks "3", and so on.

Images	**Average bit rates (in bits/pixel) using different lossless coders**				
	HINT	**HINTS**	**Fixed DPCM**	**FADPCM**	**BADPCM**
R1	3.03	2.68	2.65	2.54	2.55
R2	3.02	2.75	2.52	2.45	2.51
Lena	5.20	4.92	5.11	4.98	5.00
Pepper	5.20	4.93	5.38	5.11	5.06
Jet	4.64	4.31	4.23	4.23	4.25
Flower	4.06	3.71	3.68	3.59	3.60
Cameraman	5.38	5.16	5.14	4.91	4.93
Bridge	6.36	6.07	6.13	5.97	6.03

Table 1. Average bit rates (in bits/pixel) using different lossless coders

Images	Average bit rates (in bits/pixei) for maximum error of ± 1				Average bits rates (in bits/pixel) for maximum error of ± 2			
	HINT	Fixed DPCM	FADPCM	BADPCM	HINT	Fixed DPCM	FADPCM	BADPCM
R1	1.80	1.74	1.52	1.49	1.34	1.44	1.21	1.14
R2	1.68	1.48	1.41	1.41	1.16	1.23	1.15	1.08
Lena	3.64	3.56	3.43	3.45	2.94	2.88	2.75	2.77
Pepper	3.64	3.81	3.56	3.50	2.94	3.11	2.86	2.81
Jet	3.18	2.83	2.79	2.83	2.52	2.27	2.23	2.24
Flower	2.59	2.29	2.20	2.20	1.94	1.81	1.72	1.69
Cameraman	3.85	3.56	3.46	3.47	3.18	2.97	2.85	2.84
Bridge	4.79	4.56	4.40	4.45	4.06	3.85	3.68	3.73

Table 2. Average bit rates (in bits/pixel) using different near-lossless coders

Images	Average bit rates (in bits/pixel) using basic and improved FADPCM coders	
	FADPCM	
	Basic Version	Improved Version
R1	2.54	2.48
R2	2.45	2.44
Lena	4.98	4.81
Pepper	5.11	4.89
Jet	4.23	4.10
Flower	3.59	3.40
Cameraman	4.91	4.85
Bridge	5.97	5.88

Table 3. Average bit rates (in bits/pixel) comparing basic and improved SADPCM coders

Images	Average bit rates (in bits/pixel) using different coders					
	HINTS	**HINT**	**Forward HBADPCM** (1 codebook)		**Backward HBADPCM** (1 codebook)	
	5 codebooks	1 codebook	**Basic version**	**Improved version**	**Basic Version**	**Improved Version**
R1	2.68	3.03	2.58	2.64	2.67	2.69
R2	2.75	3.02	2.49	2.54	2.52	2.55
Lena	4.92	5.20	4.99	4.86	5.01	4.87
Pepper	4.93	5.20	5.12	4.95	5.15	4.95
Jet	4.31	4.64	4.26	4.14	4.27	4.15
Flower	3.71	4.06	3.44	3.31	3.49	3.33
Cameraman	5.14	5.38	4.92	4.88	4.92	4.90
Bridge	6.07	6.36	5.97	5.92	5.97	5.94

Table 4. Average bit rates (in bits/pixel) using different coding schemes

Figure 3a. Missing 8x8 blocks of Pepper in Level 1

Figure 3b. Interpolated Pepper from available blocks of Level 1

Fiber-optic Sources for Communications

J. W. Haus
Physics Dept.
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

Final report for:
Summer Faculty Research Program
Rome Laboratory, Rome, NY

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC
and
Rome Laboratory

August 1995

Fiber-optic Sources for Communications and Control

J. W. Haus
Physics Dept.
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

Abstract

Fiber-optics technology has become a cornerstone for advanced information communications in both the military and the civilian sectors. The availability of mode-locked fiber-optic sources would provide compatibility with existing fiber-optic structures and an inexpensive alternative to sub-picosecond semiconductor sources in this wavelength regime. In this program we have continued to develop a unique capability in the simulation of mode-locked fiber-optic sources. I have endeavored to examined several avenues to creating reliable mode-locked laser sources and improving their behavior.

I have also proposed a novel mode-locking scheme using fiber Bragg gratings that is being examined for implementation in the laboratory. The research is a collaboration with James Theimer at Rome Labs and is designed to support ongoing experiments of Reinhard Erdmann and my student. Walter Kaechele, who working is at Rome Labs.

Two papers were prepared and submitted for publication during the summer program. Conference papers have been accepted at the Optical Society Annual Meeting and further conferences have been identified to disseminate research results. Other aspects of the summer research are being completed after the program's end.

1 Introduction

Information systems have become one of the fastest growing sectors of the economy and its need for improving military capabilities cannot be understated. In a little over one century the capability of undersea cables to communicated has risen by a factor of more than 10^9 (from a few words per minute in trans-oceanic communications of the 1850's[1] to GHz data rates today). The demand for more bandwidth in communications continues to drive the technology toward more significant achievements. The continued improvements will also have significant impact in the arenas of local area networks and wide area networks, which will also place further demands toward improvements in optical communications.

The new era of economic revolutions promises to be information intensive, rather than labor intensive, as was the first industrial revolution. In the first industrial revolution work done by individuals was rendered inefficient and animals were set out to pasture as machines took over their function. Large machines and the expenditure of huge amounts of power are now relegated to a lesser role, while intellect or human creativity and information take on a primary role. In lightwave communications the full potential of the information age can be achieved; the workhorse of the information age, the laser, will bring to its innovators increased productivity, reliability and commercial profit.

There are several enabling technologies needed for this development among them are rapid detectors, short-pulse sources and ultrafast optical switches. Detectors are already available, but require improvements in speed and sensitivity to meet the challenges of terabit data rates. Optical switches, on the other hand, have not yet matured as a usable technology and have remained an adolescent technology; nevertheless, they form a critical element for rapidly getting information ON and OFF the so-called Information Superhighway.

There are several sources that can be employed in these networks and provide the workhorse necessary to drive the information age. Already in use are semiconductor lasers, which have become ubiquitous in our daily life as elements in laser scanners and optical disk readers. However, ultrafast semiconductor lasers with the requisite bandwidth are not available and modulation has been relegated to expensive, ultrafast modulators.

A major breakthrough toward the eventual development of high data-rate optical communications occured in the 1980's with the discovery that erbium-doped fibers exhibit gain in 1.55 μm wavelength regime; this regime is of particular interest for long-distance communications, since fibers exhibit their minimum loss in this regime. Erbium-doped fiber amplifiers promise repeaterless transmission of optical data. This eliminated the need for expensive electronic repeaters in long-distance communications systems. Erbium–doped fibers are also an essential building block of mode-locked fiber lasers.

Laser diodes (MOPA - Master Oscillator-Power Amplifier) that are compact and efficient are being used to pump erbium-doped amplifier systems. This same technology is applied to pump mode-locked fiber lasers. The wide bandwidth available for amplifications (about 30 nm) is sufficient to create pulses approximately as short as 70 fs duration; in a network this would correspond to THz data rates. Erbium-doped fiber lasers are a potential ultrashort pulse source for optical communications technologies[2, 3]. They provide high repetition rates in the wavelength regime corresponding to the minimum loss of silica fibers. Such lasers have distinct advantages, for instance, they can be pumped using diode lasers and they are compatible with fiber technologies.

However, they suffer from stability and reproducibility problems because of their complex dependence on the polarization of the light in the laser cavity [4, 5, 6]. Recently, this problem has been addressed and new cavity designs have been proposed to operate with ultrashort pulses while being insensitive to environmental changes [7, 8, 9, 10, 11, 12, 13, 14]. The new cavity designs incorporate birefringent fiber amplifiers as elements to control the polarization state of the light in the fiber. Polarization mode-locking combines the uses of birefringent fibers with polarization elements to create fast saturable absorber-like action.

Generally, when a soliton pulse, which has a well defined polarization state, propagates through a birefringent fiber, then cross- and self-phase modulation effects act to rotate the polarization state of the pulse in addition to the linear birefringence of the fiber. When that fiber is an amplifier though, the perturbation of the pulse phase and amplitude can be large enough that the polarization varies across the pulse. In this situation, the mode-locking element no longer optimally functions as designed. The result will be a pulse that is distorted from its original shape.

The following Section is a discussion of mode-locked fiber lasers using three examples of passive mode-locking elements. The new one we proposed this past summer is the Bragg grating fiber. Section 3 is a presentation of our results on figure-eight lasers; we have developed a simulation technique that is applicable to examining the steady–state operation of these lasers; we also learned from these studies that a technique we dubbed *dispersion balancing* can be used to engineer a more robust figure–eight laser. The final Section is devoted to a brief summary and some comments about future work.

2 Mode-locked Fiber Lasers

Erbium-doped fiber lasers (EDFL) have many cavity designs and operation regimes. Each is being explored as a source of high repetition rate, energetic, ultrashort pulses, that meets the needs of future communication networks. Ultrashort pulses are generated using mode-locking techniques. Active mode-locking, using for example a Mach-Zehnder modulator in the cavity, has been valuable as a source of regularly spaced, high repetition-rate pulses [15]. Passive mode-locking, using a fast saturable absorber, has produced pulses of sub picosecond duration. The pulses in each case are soliton-like, i.e. hyperbolic secant shaped; solitons have proven to be robust against the presence of losses and amplification in fiber transmission systems; i.e. they don't alter their shape in the presence of small perturbations. The topic of soliton transmission in optical fibers has rapidly evolved from a pure research topic to an emerging technology through a series of important technological breakthroughs (overcoming challenging obstacles) in long-distance, high bit-rate communication systems. In addition, soliton interactions have been proposed for logic and routing devices, which perform important information processing tasks [16].

A short-pulse laser requires saturable absorber-like action to assist in achieving shorter pulses. The saturable absorber transmits more of the higher intensity regime of the pulse and the lower intensity experiences higher losses. There are now several fiber-optic devices that can be applied to achieve the fast saturable absorber-like action and there is already an extensive literature developed for EDFA [2] and EDFL [3]. The optimum configuration of the devices in each laser design requires a detailed analysis. Recently, the combination of active and passive mode-locking has been simulated as a source of stable pulse trains and ultrashort pulse operation [17]; Active mode-

locking is an effective strategy to shorten pulses because the modulator has a sharp transmission function; however, there are physical limitations set by the modulator's response time; perhaps a few picoseconds could be attained, but with the introduction of passive mode-locking ultrashort pulses, i.e. less than 1 picosecond duration, can be achieved. By further modeling and simulations, the pulse shape, height and energy can be optimized.

We are developing reliable, numerical simulations of EDFL using three types of passive mode-locking (PML). As previously mentioned, the action of PML relies on a fast saturable absorber-like action. Three devices that perform this task are: nonlinear loop mirrors, dual-core fibers and fiber Bragg gratings; this list is not exhaustive. In the following we briefly discuss each type.

2.1 Nonlinear Loop Mirrors

The nonlinear optical loop mirror (NOLM) [18] works in a manner similar to a Sagnac interferometer. A length of fiber is formed into a loop and the two ends are connected to a four port directional coupler device that splits the input pulse intensity by unequal amounts (splitting ratios between the two output ports are, say, α and $1-\alpha$; the ratio of the two is denoted: $\alpha/(1-\alpha)$) into two counter-propagating directions around the loop; as the pulses travel, each one picks up a different optical path difference that is related to its intensity. If the intensity of both pulses is low, then the phases are nearly the same and the light recombines in the coupler and is mostly reflected back into the same port as the input pulse; this light is rejected from the laser cavity by a isolator. As the intensity is raised the nonlinear contribution to the polarization with Kerr coefficient $\chi^{(3)}$,

$$P = \frac{3}{4}\chi^{(3)}|E|^2 E, \tag{1}$$

changes the optical path length of each pulse by a different amount; thus, the output intensity increases in the second port. A related device is the nonlinear amplifying loop mirror (NALM) [19], which places an amplifier at one end of the loop. In this case a 50/50 coupler can be used and still the pulse amplified at the beginning of the loop will experience a greater optical path difference than the counter-propagating pulse that is amplified after passing around the loop. The transmission through the second (output) port for an NALM is given for a CW wave by

$$T_2 = G\,(1 - 2\alpha(1-\alpha)\{1+\cos\Delta\phi\})\,; \tag{2}$$

where G is the gain and the phase is given by

$$\Delta\phi = (G(1-\alpha)-\alpha)\,|E_{in}|^2 2\pi n_2 L/\lambda. \tag{3}$$

The new parameters are: L, the length of the loop; λ, the wavelength in vacuum; n_2 the coefficient of the nonlinear index, which is related to the Kerr coefficient.

The transmission of the NOLM ($G = 1$ in Eq. (2) has a transmission maximum for the value of the phase shift $\Delta\phi = \pi$. For a pulse whose intensity is less than the maximum on the transmission curve higher loss is experienced for lower intensities and it acts like a fast saturable absorber. For ultrashort pulses the length of fiber required is only a few meters, but self-frequency shifts, due to stimulated Raman scattering limit the power levels of the PML device. In truth, Eq. (2) does not accurately describe the action of an NOLM or NALM. Detailed numerical analysis needs to be

considered. In this device the birefringence of the fiber plays an important role and polarization controls are used to correct for the distortion of the polarization state.

Two other related PML devices also act as as fast saturable absorbers: the dual-core fiber laser recently proposed In Ref.([20]), whose length has been chosen as a half beat or in other words so that light input in one core exits in the other core, and the birefringent fiber with cross polarizers [21, 22, 23, 24]. The dual core fiber distributes the loss over the fiber length and consists of a core with gain enclosed in a cavity, while the second core is open to reject any unwanted CW or pedestal on the pulse; this is the essence of a PML device. The birefringent fiber with polarizers will not be further discussed here.

2.2 Dual-core fibers

In dual core fibers the field can couple between the two cores; as it propagates; in the linear regime, like a pair of coupled oscillators, the energy periodically passes from one core to the other. For nonlinear materials the situation is altered, as the intensity is raised, the pulse couples less and less, until it remains on a single core; this effect has been studied for use as a nonlinear switch beginning with the work of Jensen [25] on CW operation. In the nonlinear dual-core fiber the soliton is special again, when the coupling is a small perturbation, the soliton retains its shape and the pulse is not cut off in the wings. Recent proposals [26] include gain and loss in the fibers to improve the switching characteristics, i.e. lower power and shorter fiber lengths, or applications as a demultiplexer or narrow bandpass filter [27].

Soliton pairs can even interact between the cores and be used for gating and logic operations [28, 29], much like the interaction between two orthogonally polarized solitons used for dragging one soliton with the aid of a second one[16]. In dual–core fibers Abdullaev reported the existence of an attractive and repulsive interaction between solitons and adjacent fibers; this can be used as the basis for new applications in pushing or pulling solitons within some time window or spectral window. One such application is the logic gate operation proposed in Ref. [29].

The same properties of dual-core fibers that permit logic and switching operations also make them suitable as saturable absorbers in a fiber laser geometry [30, 31]. The fiber length chosen is 1/2 a beat length, i.e. the length chosen in the linear regime, such that the input energy in one core would emerge in the other core, to remove the weak continuous wave component (i.e. removes the pulse pedestal) and the pulses have soliton-like shapes. Due to the switching characteristics mentioned above, the dual-core fiber has the characteristics of a saturable absorber. In their calculations Winful and Walton [20] made use of the PML property of the dual-core fiber. They obtained short bandwidth limited pulses using this scheme. Our investigations incorporating Raman scattering, third-order dispersion and saturation show that the pulse is qualitatively changed [31]. The incorporation of this device into other laser cavity designs will require careful consideration of these limitations, as well.

The combination of both positive and negative dispersion in single core fibers has been applied to increase the power output without sacrificing the ability to shorten the pulse width or using external pulse compression techniques [32]. For single-mode fibers operating at 1.55 μm, the dual-core fiber can be designed to have a half-beat length around 1 m and typical peak power threshold for the saturation near 1 KW, when the pulse width is less than 1 ps. The energy in the pulses will be several hundred picojoules.

2.3 Fiber Bragg Gratings

Fiber-optic Bragg gratings have been fabricated with reflectivities in excess of 99.8 %. They are written into the fiber by exposing a photosensitive core to ultraviolet light. The depth of the index difference in the periodic structure can be as much as 10^{-2} and this is sufficient for a stop band of several nm. In practice the transmission function is strongly dependent on the wavelength and on the short wavelength side, light is lost by scattering to the cladding.

The simplest case to consider is a multilayered structure[33, 34]. It exhibits interesting phenomena, such as, high reflectivity over a frequency range, a so-called forbidden band or *stop band.* The theory has a strong correspondence with the quantum theory of periodic lattices, so we draw heavily on physical concepts developed in solid state physics. The band structure for a one-dimensional structure always possesses a stop band, no matter how small the index difference. This is good news for the development of future applications of these devices because they won't place emphasis on extreme material properties.

The nonlinear dielectric response of the material can be enhanced through the electromagnetic resonances at stop–band edges in the medium. The use of periodic dielectric materials can controllably enhance the local field using interference effects. Periodic structures that incorporate a nonlinear response of the medium exhibit interesting phenomena.

Chen and Mills[35] considered a periodic structure with an intensity–dependent polarizability. When the light frequency is tuned inside the stop gap and has a low intensity, the transmissivity is low, as expected. As the intensity is increased though, the transmissivity begins to increase until the sample is transparent. There is a bistable behavior of the transmissivity very similar to the behavior of a Kerr nonlinearity in a Fabry-Perot cavity[36].

This phenomenon can be physically explained as a shift of the dielectric values that moves the edge of the stop band toward the excitation frequency. As it moves out of the stop band, the transmission dramatically changes.

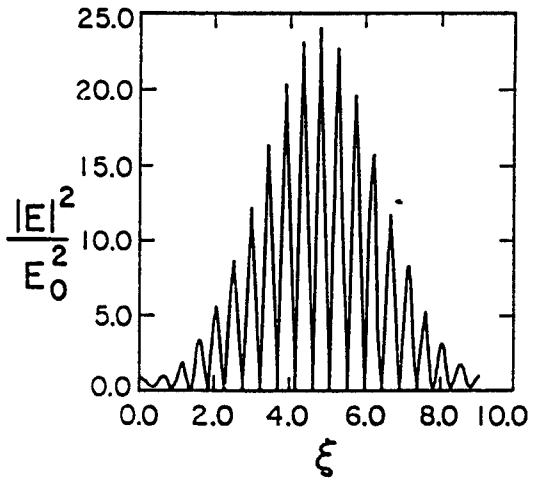


Figure 1: The spatial profile of the field intensity for the input power at which the transmission becomes unity. The oscillations are the Floquet-Bloch function and the envelope is a hyperbolic–secant squared.

Chen and Mills also discovered that the field-envelope inside the structure has a hyperbolic

secant shape, which is a well-known form of soliton solutions, see Fig. 1. Thus, they named the field structure a *gap soliton.* As with many such numerical discoveries, the simplicity of the field envelope suggested that it would be possible to find an appropriate analytical theory to describe the nonlinear phenomena in periodic structures.

Two types of experiments have reported verification of the gap–soliton predictions. Sankey et al.[37] use a corrugated silicon-on-insulator waveguide geometry. The measurements were made using a Nd:YAG laser at a wavelength of 1.06 μm. The pulse duration was around 30 ns and the pulse energies were in the μJ range. The carrier dynamics was complicated, but was also the source of interesting unstable dynamical evolution.

The experiments of Herbert et al.[38] used a dye-doped colloidal crystal as a distributed Bragg reflector. They use a cw dye laser to tune across the stop gap and ramp the intensity while holding the frequency constant. The lattice spacing was around 215 nm and the index modulation is weak; there were about 400 periods of the unit cell. Depending on the tuning position within the gap, the transmissivity showed a monotonic behavior, bistable or multi-stable operation.

We believe that this nonlinear optical behavior will constitute the necessary mode-locking characteristics required for a fiber ring laser. However, the nonlinear response of Bragg gratings in fibers has not been investigated and this mode-locking techniques constitutes and unproven concept. First, the theoretical details need to be flushed out. That task was begun this past summer during my participation in the Photonic Lab summer faculty program. The approach to a feasibility study involves analytic methods[39] that have been developed to study gap solitons[40, 41, 42] and numerical methods[43] that we have built up over the past ten years. The two approaches are designed to provide reliable answers on the operation of a mode-locked fiber laser using nonlinear Bragg gratings as elements. Completing the analysis will require further work and research is continuing on this project. The experimental group of Reinhard Erdmann and Kenneth Teagarten are involved in the project. The project also uses a student, Walter Kaechele, who is developing experiments at the Photonics Laboratory.

3 Figure-eight Laser Model

Figure–eight lasers using erbium–doped fiber amplifiers are of special interest due to their ability to produce nearly transform limited pulses at a very high repetition rate and wavelength operation at the minimum loss region of the optical fiber. Their operation was first demonstrated in 1990 [44, 45] with two mode–locking configurations. One design uses the nonlinear optical loop mirror (NOLM)[44] and the other uses a nonlinear amplifying loop mirror (NALM) by Richardson[46] and by Duling[45, 47]. Bulushev[48] simulated the NOLM based laser by direct integration of the nonlinear Schrödinger equation (NSE) in the NOLM, but no propagation in the amplifier, which was modeled as a homogeneously-broadened, saturable gain medium. Tzelpis et al.[49] reported a similar analysis on the NALM based figure–eight lasers. They can also be analyzed as additive-pulse mode-locking (APM) lasers[50].

It has been experimentally found that periodic perturbations to a soliton-like pulse produce spectral sidebands [52, 53, 54, 55, 56]. The research of Dennis et al.[56] has quantitatively examined this phenomenon. The sidebands are a result of dispersive wave shedding that circulates in the cavity and is amplified along with the pulse when the wavelength is phase matched. The operation

of the figure–eight laser (F8L) is sensitive to the total dispersion in the cavity, which includes the amplifier section of the laser.

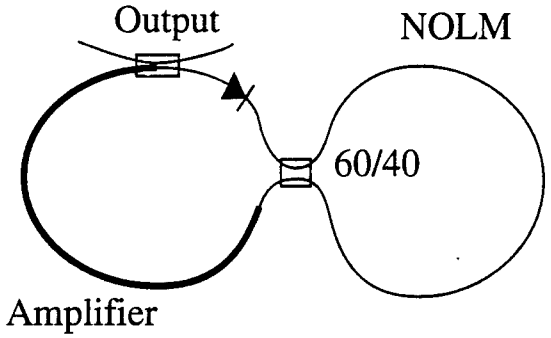


Figure 2: Sketch of the figure–eight laser geometry showing the important elements. The isolator in the amplifier section is depicted as a diode symbol.

Here I discuss simulations of certain operational aspects of the F8L laser with an NOLM that were completed during the summer program. This work has been submitted for publication in the Journal of Lightwave Technology[51]. The emphasis is placed on the role that dispersion and propagation play in the laser operation. Pulse propagation is described by a modified NSE and an initial seed pulse is injected into the cavity. The basic features of our model are shown in the diagram in Figure 2. The NOLM is a loop of fiber, whose length is initially chosen as 4 soliton lengths in scaled units. Two pulses propagate in opposite directions around the loop and by virtue of their different amplitudes, they have different nonlinear phase shifts upon recombining at the 60/40 cross coupler. The amplifier section allows pulse propagation only in the clockwise direction; the counter–clockwise direction is suppressed by the action of the isolator. The output of our model laser is a 10 % coupler placed just before the entrance to the NOLM. Pulse propagation is described by a modified NSE and an initial seed pulse is injected into the cavity. Not shown is the amplifier pump.

The numerical solution is obtained by means of the split-step Fourier transform method[57]. The fiber amplifier was modeled as a two-level system with a parabolic line shape as in Ref.[58]. We consider the case of a 60/40 directional coupler between the two cavity sections in Figure 1 and a 10 % output coupler. Initially, the NOLM loop is four soliton periods long in scaled units. Duling[59] found that this length produces minimum loss and pulse distortion. As the pulse width shortens during successive round trips in the cavity, the soliton period correspondingly shortens. It was assumed that the population inversion in the amplifier was uniform, and for steady–state conditions it totally recovers between passes of the pulse through the amplifier.

$$i\frac{\partial E}{\partial z} + \frac{1}{2}\frac{\partial^2 E}{\partial \tau^2} + E|E|^2 = i\frac{G}{2}E + i\mu\frac{\partial^2 E}{\partial \tau^2}, \tag{4}$$

The equations have been scaled to soliton units throughout [57]. The time is scaled to a value T_0, related to the initial pulse width; the length is scaled by the dispersion length $L_D = T_0^2/|\beta_2|$, where β_2 is the group velocity dispersion parameter; it is negative in the wavelength regime near 1.5

μm; $\beta_2 \approx -20$ ps^2/km. An often used length related to L_D is the *soliton* length whose definition is $Z_0 = \pi L_D/2$. The field amplitude scaling corresponds to a fundamental soliton, $\gamma_2|E_0|^2 = |\beta_2|/T_0^2$, where γ_2 is related to the fiber's Kerr nonlinearity and the fiber's effective core area. The gain parameter G is a variable in our simulations and the gain dispersion parameter $\mu = GT_2^2$ is a product of the gain parameter and the polarization relaxation time we used is $T_2 = 100fs$ in physical units, which is appropriate for erbium–fibers. All our scaled parameters are based on scaling $T_0 = 300fs$.

The left hand side of Eq. (4) has the elements of dispersion and nonlinearity required for pulse propagation in an optical fiber. Our pulse lengths are long enough that effects, such higher-order dispersion, stimulated Raman scattering and self-steepening, are negligible. This portion of the evolution equation is applied to propagation of the two pulses in the NOLM.

The two additional terms on the right hand side of Eq. (4) are included in the simulation when the pulses propagate through the amplifier section of the laser. They describe a gain curve with a maximum at the pulse's center frequency and a parabolic gain profile; the gain profile is much wider than our pulse spectra and this approximation is not a limiting factor.

The saturation energy of the amplifier can be given by $E_s = h\nu_0 a_{eff}/2\sigma$. This expression can be found, for example, in reference [58]. In this expression ν_0 is the central frequency of the pulse, a_{eff} is the effective core area, and σ is the absorption cross section of the laser line. The parameters which effect saturation energy are, therefore either not under our control, or are severely constrained by amplifier design criteria. Typically, saturation energy will be found to be around 10 mJ. This implies that the values of saturation energy used in references [48] and [49] were far too large to be physically realistic.

For realistic values of the saturation energy, the pulse initiation and growth from noise cannot be simulated because it would require thousands of round–trips in the cavity before steady–state operation was achieved. When we start with enough gain to overcome the loss due to the NOLM at low intensities, the amplifier will hardly have saturated at all when the pulse has become intense enough to have reached the first transmission peak of the loop mirror. The pulse will then break up, becoming broad in time and frequency space. This is a condition that our current model does not handle well. For this reason, we have limited ourselves to studying conditions for steady–state pulse operation.

3.1 Results for the Figure-eight Laser

Consider first the effect of the fiber amplifier's length. Altering the length changes the amount of dispersion in the cavity, which also limits the minimum pulse length in the cavity. We found that stable pulses could not be produced if the amplifier was longer than 0.7-0.8 dispersion lengths; we attribute this effect to reshaping the pulse in the laser amplifier, since, during this portion of the evolution, the pulse has a higher energy than a fundamental soliton energy and the pulse tends to shed energy while it evolves toward a fundamental soliton (this will be discussed further below). Since the NOLM loop is 4 soliton periods long, this implies a total loop length of about 7.0 dispersion lengths. For amplifiers shorter than 0.8 dispersion lengths, we could only achieve stable, single pulse operation for a narrow band of values for the gain coefficient. If the gain was either too great, or too small, the pulse would decay away. These findings are summarized Figure 3.

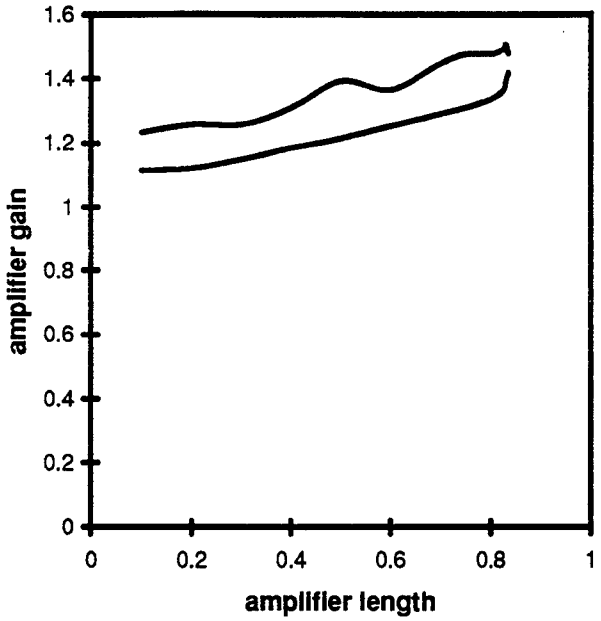


Figure 3: Stable operation regime for a single pulse in the figure–eight laser.

To generate the results in Figure 3, simulations were run using a single pulse that is launched in the cavity. For each amplifier length a range of gain parameter values was used and steady–state operation was sought. The entire range of stable pulse operation was determined by altering the the gain and amplifier length parameters. In Figure 3 we plot the gain, which is related to the previously defined gain parameter by

$$\Gamma = e^{GL_a}. \tag{5}$$

The parameter L_a is the amplifier length. The parameter Γ is a measure of how much the intensity increases after one trip through the amplifier. The curves represent the lower and upper bounds for stable single–pulse operation. For a very short amplifier, the gain only needs to overcome the loss due to the output coupler. Length dependent loss has been discussed in references [56] and [60]. An amplifier length around 0.75 dispersion lengths is sufficient to cause an instability in the pulse shape and its subsequent disappearance due to a loss of transmission through the NOLM.

We also found a maximum amplification, beyond which a single pulse in the loop will not be stable. For higher gains the cavity would adjust to include two pulses in the cavity. The effect of a second pulse in the loop will be to cut the gain coefficient in half, which would reduce the gain to the point where stable pulses could again be formed. This implies a pump dependent transition from a regime with one stable pulse in the loop, to one with multiple pulses. Such a transition is observed in NALM based lasers.

Pulse energy evolution depends upon the gain parameter. One gain is near the near minimum amplification boundary; the pulse amplitude is quite steady and varies little in each successive round trip. For the gain near the maximum amplification boundary, the energy in the pulse decays and recovers; the pulse takes a much greater number of round trips to settle to a steady state. In this latter case, the pulse shape undergoes considerable change during its evolution.

Physical insight into these phenomena can be found by examining a transmission curve for the NOLM, given in Figure 4. The curve is generated using hyperbolic-secant shaped input pulses; input pulse amplitudes near 2 closely correspond to fundamental solitons in the NOLM, where the transmission curve has a strongly peaked maximum. The minimum gain limit corresponds

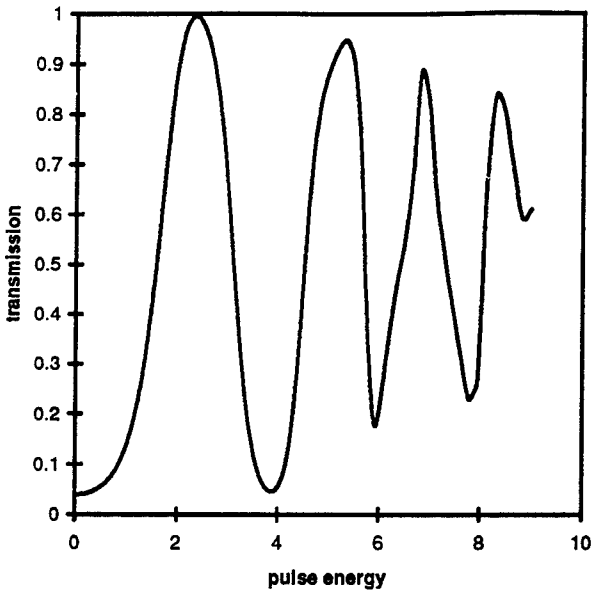


Figure 4: The hyper-secant–pulse transmission curve for a nonlinear optical loop mirror. The splitter has a 60/40 splitting ratio.

to the amplification required to overcome losses in the laser, so greater gain represents a surplus energy that alters the amplitude of the NOLM input pulse. As the pulse intensity exceeds that for maximum transmission, the pulse will experience greater loss in each round trip. If this loss is not too great, it can recover. However, when the gain is so great that the transmission falls below the recovery threshold, then the pulse decays. We note that the transmission has a very deep minimum near a pulse energy of 4. The difference between the minimum and maximum stable gains implies that the pulse with an amplitude such that it is near the transmission maximum, need only an additional amplification by about 10 % (to recover the loss due to the output coupler). When the gain is below the minimum, there is not enough replenishment to sustain the pulse in the cavity and when the gain is above the maximum the pulse amplitude from the amplifier places the pulse it in a very steep part of the NOLM transmission curve. The pulse experiences too great a loss through the NOLM and will not be able to recover in the next pass through the amplifier.

The directional coupler for the NOLM loop produces two pulses of roughly half the input intensity, traveling in opposite directions. This would make them approximately unit solitons, and as such subject to minimum distortion during propagation. However, when these pulses are recombined and propagate in the amplifier, they tend to be reformed into a soliton shape suitable to their amplitude, and to shed some light to dispersive wave radiation. The degree of transmission in the NOLM loop is highly dependent on pulse shape, and the reshaped pulse will experience greater loss than the original sech shape. If this loss becomes too great, the laser will simply cease to function.

The steady-state pulse shape for an amplifier length of 0.1 dispersion lengths is well fit in our numerical simulations by a hyperbolic secant shaped envelope; in the NOLM this pulse is split into two pulses that are good approximations to fundamental solitons. The amplifier is short, so that there is a small amount of additional dispersion in the cavity. The fit of the pulse intensity to a fundamental soliton shape for this case is much narrower than the actual pulse width.

The spectrum of this pulse a central section that is also a hyperbolic–secant shape, but there is additional structure in the wings that is due to dispersive–wave shedding. The shoulders are

consistent with the placement of the sideband peaks given by the formula[47]

$$\omega_m = \sqrt{8mZ_0/L - 1}; \tag{6}$$

where L is the length of the fiber and Z_0 is the soliton length corresponding to the steady–state pulse width.

The amplifier, whose length is 0.7 dispersion lengths, has a central region that is well fit by a soliton envelope, but the wings, having shed radiation are much larger. Here we emphasize again that the soliton envelope is constrained by both its amplitude and width. Comparison between the short and long amplifier cases shows that pulse propagation over a longer distance in the amplifier is better approximated by a fundamental soliton shape. There is a much more pronounced sideband structure that is amplified in the cavity. The sideband peak positions are well approximated by the expression in Eq. (6).

Contrasting the two cases of a short and a long amplifier length, we conclude that the longer amplifier reshapes the pulse toward a fundamental soliton. Therefore, as the pulse is split in two by insertion into the NOLM it undergoes reshaping again toward a much different fundamental soliton amplitude. In other words, the NOLM and the amplifier shaping mechanisms compete against one another. This interpretation of our results suggests a a strategy to improve the mode-locking stability of the F8L. We proposed to balance the dispersion in the NOLM and the amplifier sections in order to minimize the pulse reshaping in each cavity[62].

3.2 Dispersion Balanced Figure-eight Laser

In our previous Subsection, we modeled the behavior of a fiber laser mode-locked by a nonlinear optical loop mirror (NOLM). A schematic of the laser is shown in Figure 2. The loop mirror has a length of four soliton periods. A 60/40 directional coupler was used in the center, and 10% of the pulse energy was coupled out of the cavity. We found that the pulse experienced a loss which was dependent on the length of the amplifier and we couldn't stabilize the pulse output if the length of the amplifier was longer than 0.83 dispersion lengths. This result led us to propose a new strategy to improve the operational stability of the F8L.

Our analysis showed that this instability is due to the competing tendencies of the amplifier and the NOLM to reshape the pulses toward a fundamental soliton shape. When two pulses in the NOLM favor a fundamental soliton shape, then maximum transmission is produced for a pulse with a field envelope with a shape 2.2sech(τ), where the intensity and τ are in soliton units. For this situation the pulse will be transmitted with only a small amount of distortion. The two pulses at the output of the NOLM are recombined and are not a fundamental soliton in the amplifier. When the fiber amplifier is extended, the pulse is reshaped toward a fundamental soliton and for a long amplifier, it approaches the fundamental soliton shape. We attribute the length dependent loss to the fact that this reshaping process resulted in a pulse shape which was not correct for complete transmission through the NOLM.

This understanding of the length dependent loss suggests a possible alternate solution. We should design the amplifier fiber so that the pulse is a fundamental soliton in that fiber, as well as in the NOLM; this balances the two sections of the laser so that the pulses are always close to a fundamental soliton shape and that perturbations of that shape are kept to a minimum. For a

fundamental soliton

$$1 = \frac{\gamma P_0 T_0^2}{|\beta_2|}, \tag{7}$$

where γ is the nonlinear coefficient, P_0 is the peak power, T_0 is the pulse width and β_2 is the dispersion. The parameter γ contains both the intrinsic material parameters and the fiber geometry; we assume that it is constant in our simulations. Since, we desire that T_0 be the same in both sections, then the dispersion must be increased by a factor of around 2.2 to compensate for the increased peak intensity at the output of the NOLM.

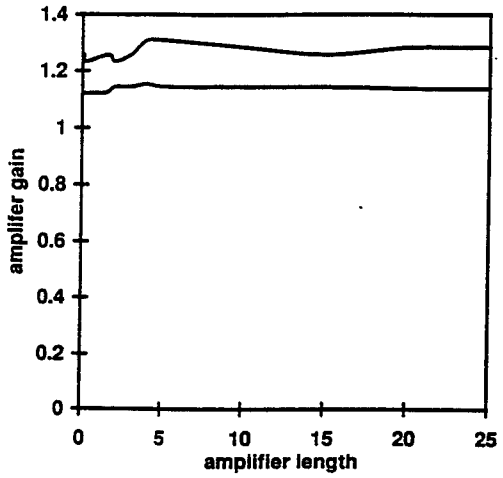


Figure 5: Stability regime for stable single pulse operation is a band in the amplifier gain between about 1.1 and 1.3.

This design feature was incorporated into our laser model. Figure 5 shows the regions in which stable pulses could be formed. This figure is in sharp contrast to the stability regime of the ordinary F8L configuration shown in Fig. 3. It should be noted that in this case the limit of the abscissa of the graph extends to 25 dispersion lengths; the pulse is very well approximated by a fundamental soliton and for further amplifier lengths, no change is expected. It should be further noted that the gain required for stable operation is quite constant, representing minimal gain dependent loss and a result of the soliton pulse shape. When the amplifier is longer than about 5 dispersion lengths, pulse reshaping is minimal, and the maximum and minimum stable gains are nearly constant. If the amplifier is shorter than this, the pulses deviate from a soliton shape and the output from the NOLM is sensitive to its input pulse shape.

To obtain insight into the functioning of this laser, the pulse shape was examined at the point at which it left the output coupler, and before it entered the directional coupler; it is shown in Fig. 6. The case that was examined was for a 25 dispersion length amplifier, with a gain parameter, $G = 0.0056$; This corresponds to an amplifier gain of $e^{GL} = 1.15$. The output pulse has a nearly uniform phase. The pulse intensity corresponds almost exactly to that of a fundamental soliton, with a pulse intensity of about 2.7. In comparing the relationship of the pulse width to pulse height it must be remembered that the ratio has been altered by a factor of 2.2.

Figure 7 shows the pulse in frequency space. Sidebands have clearly developed. The separation of the first order sideband from the center is 0.05 in normalized frequency units, which is consistent with dispersive-wave shedding[53]. In the case of a F8L, the soliton period varies as the pulse

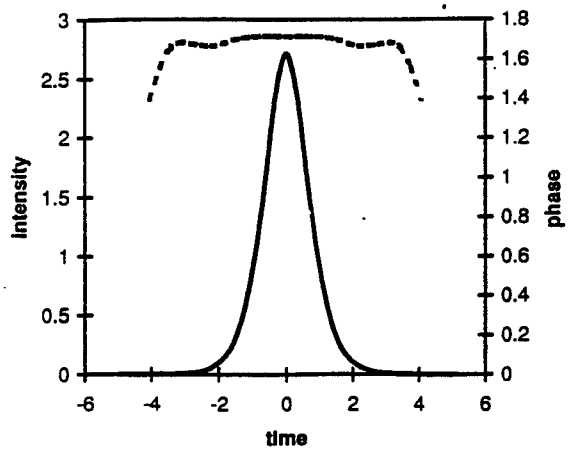


Figure 6: Pulse shape and the phase of the pulse after propagating through a 25 dispersion length amplifier. The minimum amplifier gain of 1.1 was used and the pulse is seen after the output coupler.

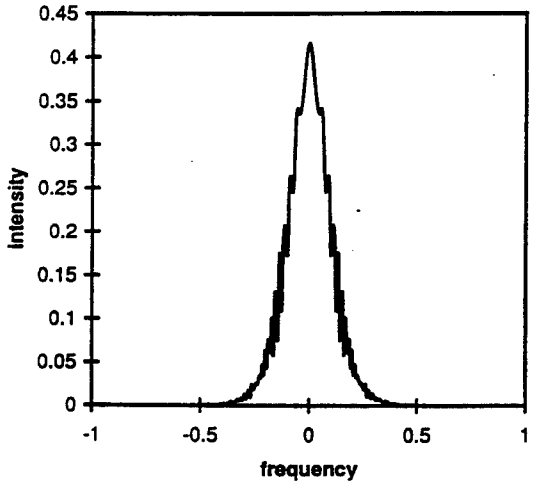


Figure 7: The pulse spectrum of the pulse in Fig. 3. The side bands are reduced in this balanced laser; their position is consistent with soliton perturbation theory, see Refs. 4,6.

travels through the laser. Since the length of the amplifier is 25 dispersion lengths, this section alone has a length far longer than eight soliton periods usually found to limit operation of the F8L; a restriction that also applied to our previous F8L simulation. This is significant as the length appears to represent a length limit for ring lasers and we attribute this to the improved rejection of the dispersive wave in our NOLM because the pulses are very nearly fundamental solitons. We also find that the dispersive wave component in the spectrum is increased when the cross-coupler splitting ratio is closer to 50/50 and the length of the NOLM is correspondingly increased.

The existence of a maximum amplifier length sets a limit of the shortest pulse that can be produced by a given laser. If the physical length of the amplifier is given by κ, and the maximum amplifier length in dispersion lengths is L_{max} the limit would be expressed as

$$\kappa = L_{max} L_D .. \tag{8}$$

Since the dispersion length can be defined as $L_d = T_0^2/|\beta_2|$, Eq. (8) can be transformed to

$$\left(\frac{\kappa|\beta_2|}{L_{max}}\right)^{1/2} = T_0. \tag{9}$$

The limit on L_{max} is related to the dispersive wave soliton resonance of 8 soliton periods [60], but since this laser has reduced side band amplitudes, this limit no longer applies. Our results indicate, however, that it is possible to decrease the shortest possible pulse, while increasing the average dispersion in the laser, which is a possibility which has not been explored before. Previous work has centered on optimizing the performance of the laser by using fiber with normal dispersion [60, 32]. It should be noted that our method for optimizing the laser could be implemented by decreasing the dispersion of the fiber in the NOLM loop.

3.3 Conclusions

The final pulse widths depend on several factors including, the amplifier gain and length, and the length of the NOLM. In Figures 5 and 6, the full width at half maximum corresponds to about 2.5 and 1.5 in scaled units, resp.. In physical units the pulses have widths of about 750 fs and 450 fs. The higher gain parameter has a shorter width due to stronger nonlinear shaping mechanisms. We can also work toward shorter pulses by adjusting the cavity length. Naturally, as the pulses shorten the propagation loss effects mentioned in Section 3 must incorporated into the analysis and the amplifier model should incorporate a more accurate gain profile shape and the dynamical evolution of the active medium.

In conclusion, we have developed a simulation that predicts steady–state properties of the NOLM fiber laser. We have found in our results a length dependent minimum pulse width, a length dependent loss, which has been observed by experimental groups, but has not been reported in previous simulations. All prior F8L simulations do not incorporate propagation in the amplifier. We have also found the existence of a maximum application for stable single–pulse operation and observed how the final pulse shape is dependent on the length of the laser cavity, including propagation in the gain medium. Finally, we have observed dispersive wave shedding in the cavity, especially for larger gains. The side bands in the power spectrum are consistent with the results discussed by Dennis et al. [56]. The maximum gain instability occurs for an effective cavity length around 7-8 soliton lengths.

The final pulse widths depend on several factors including, the amplifier gain and length, and the length of the NOLM. In physical units the pulses have widths of about 750 fs for the shorter amplifier and about 450 fs for the longer amplifier. The higher gain parameter has a shorter width due to stronger nonlinear shaping mechanisms. We can also work toward shorter pulses by adjusting the cavity length. Naturally, as the pulses shorten the propagation loss effects, such as stimulated Raman scattering, must incorporated into the analysis and the amplifier model should incorporate a more accurate gain profile shape and the dynamical evolution of the active medium.

4 Summary of Research

The dispersion-balanced F8L is an excellent candidate for future experimental work. Based on our numerical simulation, when the fiber dispersion parameters are matched, a greatly improved

operational stability of the lasers is found. The pulse shape is more closely approximated by a fundamental soliton shape, i.e. there is a better correspondence of the pulse height and width in both time and frequency domains to that of a fundamental soliton.

During the Summer Faculty Program two papers were prepared for publication and submitted to technical journals. These papers cover steady–state operation of F8Ls. The novel feature we propose for future F8Ls is to apply our design feature called dispersion balancing, which simply places a fiber with approximately double the dispersion of the NOLM in the amplifier section of the laser. This prevents break-up of the pulses and stable operation because both sections propagate pulses that are nearly fundamental soliton shapes.

Research was also begun on a new passive mode-locking element called a fiber Bragg reflector. This element takes advantage of the nonlinear operation of a fiber grating, when a pulse is tuned inside the stop band. The stop band has low transmission when operating in the linear regime, but at sufficiently high intensity, the pulse is transmitted. An expanded program of research on this new mode-locking element is currently being pursued. There is also experimental support for the concept at the Photonics Laboratory in Rome, NY and the details will be worked out during the coming year.

Acknowledgments

The support of the Photonics Laboratory and my colleagues there is gratefully acknowledged. Without their help this research could not have been completed.

References

[1] The first successful trans-oceanic cable was laid in 1858 between Newfoundland and Ireland without repeaters. The signal was so feeble and the time constants so large that special optical detectors had to be improvised, i.e. a mirror attached to a galvanometer rotated and the deviation was amplified by bouncing light off its surface. The cable ceased to function after one month in service; only 400 messages were sent. See: B. Dibner, *The Atlantic Cable*, Burndy Library Inc., Norwalk, CN, 1959.

[2] E. Desurvire, *Erbium-doped Fiber Amplifiers: Principles and Applications*, Wiley, NY (1993).

[3] M. J. F. DIGONNET ed., *Rare Earth Doped Fiber Lasers and Amplifiers*, Marcel Dekker, New York (1993).

[4] A. B. Grudinen, D. J. Richardson and D. N. Payne, " Optical pulse propagation in doped fiber amplifiers," Phys. Rev. A**44**, 7493 - 7501 (1991).

[5] W. Hodel, J. Schön and H. P. Weber, " Intensity discrimination of optical pulses with birefringent fibers," Opt. Commun. **88**, 173-179 (1992).

[6] M. Hofer, M. E. Fermann, F. Haberl, M. H. Ober and A. J. Schmidt, " Mode locking with cross-phase and self-phase modulation," Opt. Lett. **7**, . 502-504 (1991).

[7] I. N. Duling III and R. D. Esman, " Single-polarization fiber amplifier," Electron. Lett. **28**, 1126 - 1128 (1992).

[8] M. E. Fermann, M. J. Andrejco, Y. Silberberg and A. M. Weiner, " Generation of pulses shorter than 200 fs from a passively mode-locked fiber laser," Phys. Rev. A**48**, 7493-7501 (1993).

[9] M. E. Fermann, M. J. Andrejco, Y. Silberberg and M. L. Stock, " Passive mode locking by using nonlinear polarization evolution in a polarization-maintaining erbium-doped fiber," Opt. Lett. **18**, 894 - 896 (1993).

[10] M. E. Fermann, L.-M. Yang, M. L. Stock and M. J. Andrejco, " Environmentally stable Kerr type mode-locked erbium fiber laser producing 360 fs pulses," Opt. Lett. **19**, 43-45 (1994).

[11] V. J. Matsas, T. P. Newson and M. N. Zervas, " Self-starting passively mode-locked fibre ring laser exploiting nonlinear polarisation switching," Opt. Commun. **92**, 61 - 66 (1992).

[12] V. J. Matsas, D. J. Richardson, T. P. Newson and D. N. Payne, " Characterization of a self-starting, passively mode-locked fiber ring laser that exploits nonlinear polarization evolution," Opt. Lett. **18**, 358-360 (1993).

[13] K. Tamura, H. A. Haus and E. P. Ippen, "Self-starting additive pulse mode-locked erbium fibre ring laser," Electron. Lett. **28**, 2226 - 2228 (1992).

[14] M. Y. Frankel, R. D. Esman and J. F. Weller, "Additive-pulse modelocking in fiber lasers," IEEE J. Quant. Electron. QE-**30**, 200 - 208 (1994).

[15] G. T. Harvey and L. F. Mollenauer, "Harmonically modelocked fiber ring laser with an internal Fabry-Perot stabilizer for soliton transmission," Opt. Lett. **18**, 107-109 (1993).

[16] M. ISLAM, *Ultrafast Fiber Switching Devices and Systems*, Cambridge University Press, Cambridge, (1992).

[17] R. L. FORK, K. SINGH, J. W. HAUS, R. K. ERDMANN AND S. T. JOHNS, "Harmonically mode-locked laser and applications ," *SPIE Conference Proceedings* **2216**, 148-159 (1994).

[18] N. J. Doran and D. Wood, "Nonlinear-optical loop mirror," Opt. Lett. **13**, 56-58 (1988).

[19] M. E. Fermann, F. Haberl, M. Hofer and H. Hochreiter, " Nonlinear amplifying loop mirror," Opt. Lett. **15**, 752-754 (1990).

[20] H. G. Winful and D. T. Walton, "passive mode locking through nonlinear coupling in a dual-core fiber," Opt. Lett. **17**, 1688-1690 (1992).

[21] M. N. Islam, C. E. Soccolich, J. P. Gordon and U. C. Paek, "Soliton intensity-dependent polarization rotation," Opt. Lett. **15**, 21-23 (1990).

[22] R. H. Stolen, J. Botineau and A. Ashkin, "Intensity discrimination of optical pulses with birefringent fibers," Opt. Lett. **7**, 512-514 (1982).

[23] M. Hofer, M. E. Fermann, F. Haberl, M. H. Ober and A. J. Schmidt, "Modelocking with cross and self-phase modulation," Opt. Lett. **16**, 502504 (1991).

[24] C.-J. Chen, P. K. A. Wai and C. R. Menyuk, " Stability of passively mode-locked fiber lasers with fast saturable absorption," Opt. Lett. **17**, 417 (1992).

[25] S. M. Jensen, "The nonlinear coherent coupler," IEEE J. Quantum Electron. **18**, 1580-1583 (1982).

[26] Y. Chen, A. W. Snyder and D. N. Payne, "Twin core nonlinear couplers with gain and loss," IEEE J. Quantum Electron. **28**, 239-245 (1992).

[27] B. Wu and P. L. Chu, "Narrow-bandpass filter with gain by use of twin-core rare-earth-doped fiber," Opt. Lett. **18**, 1913-1915 (1993).

[28] S. L. Doty, J. W. Haus, Y. J. Oh and R. L. Fork, " Soliton interactions on dual core fibers," Phys. Rev. E**51**, 709-717 (1995).

[29] Y.J. Oh, J. W. Haus and R. L. Fork, "Soliton Repulsion Logic Gate," Opt. Lett., submitted (1995).

[30] D. T. Walton and H. G. Winful, "Passive mode locking with an active nonlinear directional coupler: positive group-velocity dispersion," Opt. Lett. **18**, 720722 (1993).

[31] Y. Oh, S. L. Doty, J. W. Haus and R. L. Fork, "Robust Operation of a Dual–core Fiber Ring Laser," J. Opt. Soc. Am. B, in press (1995).

[32] K. Tamura, C. R. Doerr, L. E. Nelson, H. A. Haus and E. P. Ippen, "Technique for obtaining high-energy ultrashort pulses from an additive-pulse mode-locked erbium-doped fiber ring laser," Opt. Lett. **19**, 46-48 (1994).

[33] L. Brillouin, *Wave Propagation in Periodic Structures* (Wiley, NY, 1946).

[34] A. Yariv and P. Yeh, *Optical Waves in Layered Media* (Wiley, NY, 1988).

[35] W. Chen and D. L. Mills, "Gap solitons and the nonlinear response of superlattices," Phys. Rev. Lett. **58**, 160-163 (1987).

[36] H. Gibbs, *Optical Bistability* (Academic Press, New York, 1985).

[37] S. D. Sankey, D. F. Prelewitz and T. G. Brown, "All-optical switching in a nonlinear periodic structure," Appl. Phys. Lett. **60**, 1427-1429 (1992); S. D. Sankey, D. F. Prelewitz, T. G. Brown and R. C. Tiberio, "Optical switching dynamics of the nonlinear Bragg reflector: comparison of theory and experiment," J. Appl. Phys. **73**, 7111-7119 (1993).

[38] C. J. Herbert, W. S. Capinski and M. S. Malcuit, "Optical power limiting with nonlinear periodic structures," Opt. Lett. **15**, 1037-1039 (1992). C. J. Herbert and M. S. Malcuit, "Optical bistability in nonlinear periodic structures," Opt. Lett. **18**, 1783-1785 (1993).

[39] J. Kevorkian and J. D. Cole, *Perturbation Methods in Applied Mathematics,* (Springer Verlag, New York, 1981).

[40] J. E. Sipe and H. G. Winful, "Nonlinear Schrödinger solitons in a periodic structure," Opt. Lett. **13**, 132-133 (1988).

[41] C. M. deSterke and J. E. Sipe, "Envelope function approach for the electrodynamics of nonlinear periodic structures," Phys. Rev. A **38**, 5149-5165 (1988).

[42] C. M. deSterke and J. E. Sipe, "Extensions and generalizations of an envelope function approach for the electrodynamics of nonlinear periodic structures," Phys. Rev. A **39**, 5163-5178 (1989).

[43] M. Scalora, J. D. Dowling, A. S. Manka, C. M. Bowden and J. W. Haus, " Pulse Propagation near Highly Reflecting Surfaces: Applications to Photonic Bandgap Structures and the Question of Superluminal Tunneling Times," Phys. Rev. A**52**, 726-734 (1995).

[44] H. Avramopoulos et. al., "Passive modelocking of an erbium-doped fiber laser, Optical Amplifiers and Their Applications," Technical Digest Series., vol. 19, PDP 8, 1990.

[45] I. N. Duling, " All-fiber modelocked figure-eight laser," OSA Annual Meeting 1990, PDP-4.

[46] D. J. Richardson et. al., "Selfstarting, passively modelocked erbium fiber ring laser based on the amplifying Sagnac switch," Electron. Lett. **27**, 542-544 (1991).

[47] I. N. Duling III, "Subpicosecond all-fiber erbium laser," Electron Lett. **27**, 544-545 (1991).

[48] A. G. Bulushev, E. M. Dianov, and O. G. Okhonikov, "Self-starting mode-locked laser with a nonlinear ring resonator," Opt. Lett. **16**, 88-90 (1991).

[49] V. Tzelpis, S. Markatos, S. Kalpogiannis, Th. Schicopoulos, and C. Caroubalos, "Analysis of a passively mode-locked self-starting all-fiber soliton laser," J. Lightwave Technol. **11**, 1729-1736 (1993).

[50] H. A. Haus, E. P. Ippen and K. Tamura, "Additive-pulse modelocking in fiber lasers," IEEE J. Quantum Electron. **30**, 200-208 (1994).

[51] J. Theimer and J. W. Haus, "Figure-eight laser stable operating regimes," J. Lightwave Tech., submitted (1995).

[52] N. Pandit, D. U. Noske, S. M. J. Kelly, and J. R. Taylor, "Characteristic instability of fibre loop soliton lasers," Electron. Lett. **28**, 455-457 (1992).

[53] S. M. J. Kelly, "Characteristic sideband instability of periodically amplified average soliton," Electron. Lett. **28**, 806-807 (1992).

[54] N. J. Smith, K. J. Blow and I. Andonovic, "Sideband generation through perturbations to the average soliton," J. Lightwave Technol. **10**, 1329-1333 (1992).

Metal Strip Polarizing Fibers

Philipp Kornreich
Professor
Department of Electrical and Computer Engineering

Syracuse University
Syracuse, NY 13244

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base
Washington, D.C.

and

Rome Laboratory

September 1995

METAL STRIP POLARIZING FIBERS

Philipp Kornreich
Professor
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
Syracuse University

ABSTRACT.

We have fabricated and measured the polarization dependence of the transmission spectrum of Metal Strip Polarizing Fibers. The measurements were performed at the Photonics Center of Rome Laboratories. These devices were fabricated by Syracuse University. We have shown that the metals survive the fabrication process. We where able to measure the plasma resonance absorption of the metal strips.

METAL STRIP POLARIZING FIBERS

Philipp Kornreich

INTRODUCTION

We have tested the transmission spectra and polarizing properties of Metal Strip Polarizing Fibers (MSPFs) that were fabricated at Syracuse University. The MSPFs consist of a glass core flanked on two sides by thin by a thin semiconductor cylinder. That is the metal strips are located at the core cladding interface, see Fig. 1. Light polarized parallel to the metal strips is absorbed while light polarized perpendicular to the strips is not absorbed. We have used a variety of materials to fabricate these strips. We have used silver, copper and CdTe. A number of eacn type of fiber were fabricated by Syracuse University and tested at the Photonics Center of Rome Laboratories. Spectral measurements at the Photonics Center of Rome Laboratories showed that these materials survived the fiber fabricating process. The fibers had reproducible characteristics.

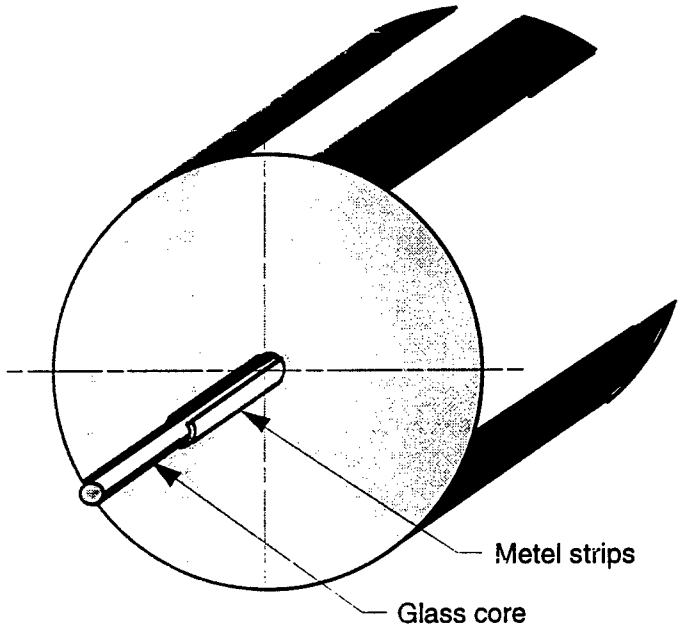


Fig. 1. Typical MSPF construction. The device consists of a glass core with two thin metal strips at the core cladding interface.

Several type of fiber polarizers have been constructed by other researcher[1,2]. A particular successful type was constructed by grinding away some of the cladding of in glass are a prede-

cessor of this technology. SMs have a large number of surface states at their interfaces with the host glass. A metal block was placed in the space where the cladding was ground away, see Fig. 2. Extinction ratios of 70 dB have been reported for these devices.

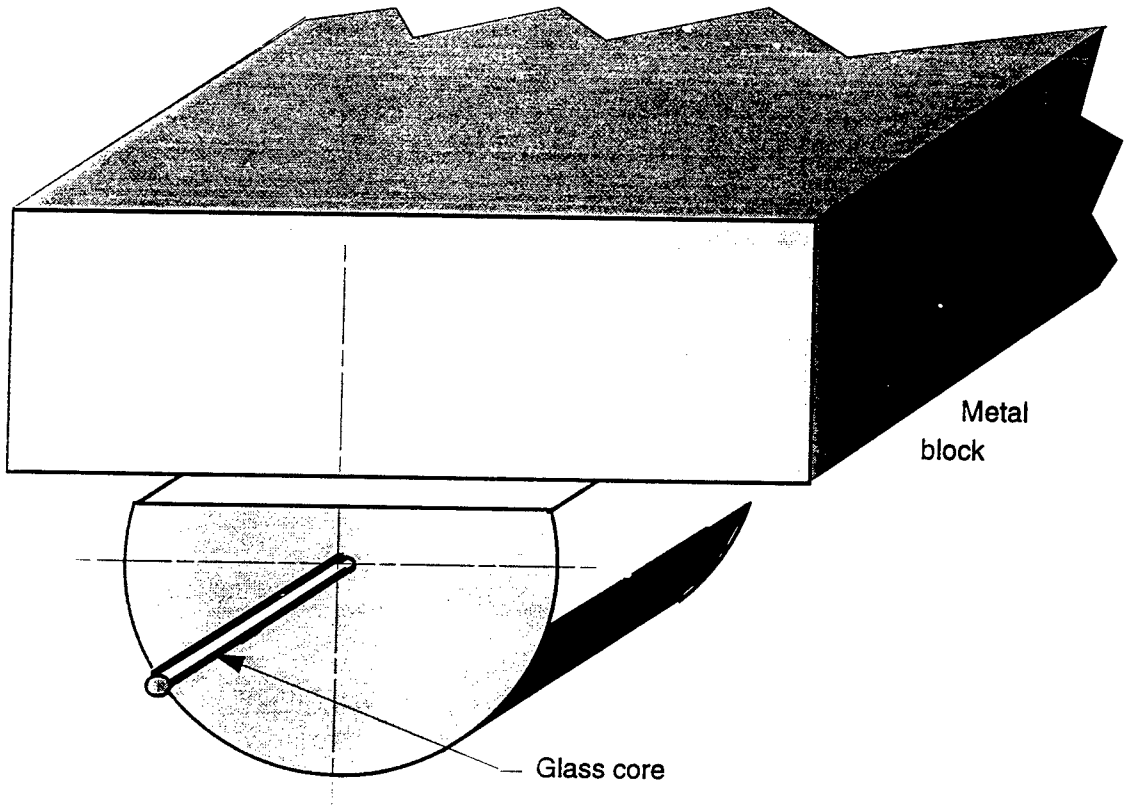


Fig. 2. D shaped polarizing fiber with metal block.

Based on our successful experience of fabricating Semiconductor Cylinder Fibers (SCFs) which we have reported last summer we attempted to use the same techniques for fabricating MSPFs.

The 3M[3,4] company sells commercial polarizing fibers. However each of these is about 40 m long while the MSPF are only e few mm long.

MSPF FABRICATION

The MSPFs were fabricated by, first, vacuum depositing metal film strips on two sides of a Pyrex glass rod that will form the core of the fiber in a diffusion pumped vacuum system. The

glass rod is clamped in a mask as shown in Fig. 3 for this deposition process. A Pyrex glass was selected that has a softening temperature of 720 ^{o}C. This is lower than the melting points of the metals used.

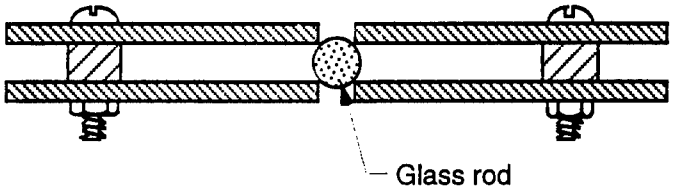


Fig. 3. Mask for vacuum depositing metal strips on core glass rod.

Next, the metal coated glass rod was inserted into a Pyrex glass tube that has been closed at one end. This structure is evacuated and collapsed. N_2 is used as the residual gas in the evacuation process. Finally, a fiber is pulled from the resulting preform.

EXPERIMENTAL RESULTS

In order for this device to be practical it is necessary that only a single mode propagate in the fiber. This requires that the core have a sufficiently small diameter and that the core have a slightly higher index of refraction than the cladding. Unfortunately we were not able to secure glass rod that have a sufficiently small diameter to yield sufficiently small fiber cores, and worth, the core and cladding glasses had the same index of refraction. Recall, the Semiconductor Cylinder Fibers (SCF) that we tested last year had a similar problem. However, since the semiconductors completely enclosed the core it facilitated some guiding. This is not the case with the metal strips.

Several metal strip fibers were fabricated and tested. The metal films deposited on the glass rods where approximately 3000 Å thick. Since all components shrink proportionally in the fiber fabricating process by a factor of about 94 the metal films should have a thickness of about 32 Å. Not even a SEM could resolve these films. Therefore, we are having these films analyzed by a Scanning Force Microscope at Rome Laboratories. To date we are still waiting for these tests to be performed.

We here present the test results of Ag strip fibers. These fibers had approximately 35 μm diameter Pyrex glass cores and 80 μm diameter Pyrex glass claddings. Fibers of various length ranging from 120 mm to 12 mm length were tested. The 12 mm long devices worked best. The results presented here were for the 12 mm long fibers. As stated above material availability restricted us to use glass with the same index of refraction for both the core and cladding.

The following test arrangement was used: Light from a white light source was focused by a microscope objective into the fiber. Another microscope objective was used to retrieve the light from the fiber. An iris was used to block the light from the cladding. These fibers had exceedingly poor guiding. The light passed, next, through a Glen Thompson Polarizing Prism. Finally the light was focused into a ANDO type AQ 1425 optical spectrum analyzer connected to a computer. "Lab. Window" software was used to control the spectrum analyzer and record the data.

The data taking procedure was as follows: It is important to note that the spectrum analyzed is somewhat polarization dependent. A piece of MSP fiber was positioned in the fiber holder. Transmission date using the spectrum analyzer was recorded for consecutive angular positions of the Glen Thompson Polarizer. The data was taken in 10^{o} steps. The Glen Thompson Polarizer was rotated a total of 180^{o}. This experiment for exactly the same Glen Thompson Polarizer positions was repeated for a piece of standard multimode fiber which was used as a reference. The reason for this procedure is that the fiber could never be repositioned with the same angular position in the fiber holder. Of course, it is not possible to rotate the fiber and maintain optical alignment. The data from the multi mode fibers was used to normalize the date from the MSPFs.

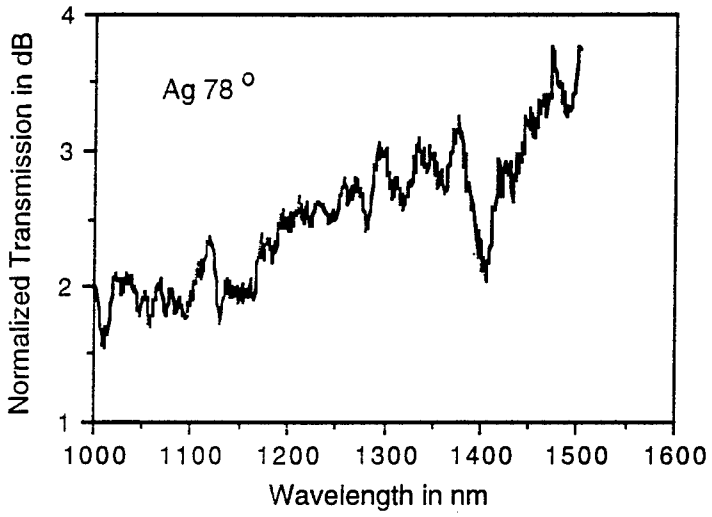


Fig. 4 Normalized transmission spectrum of MSP fiber for Glen Thompson polarizer position of 78^{o}. This corresponds to the maximum transmission through the fiber. The "bulge" at about 1200 nm is the most significant part of the data. Note the OH absorption at about 1400 nm which is larger in the Pyrex fiber than in the commercial fiber.

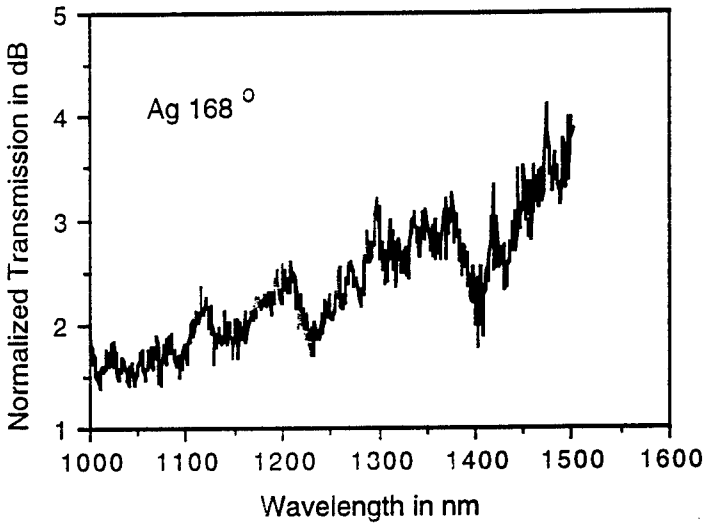


Fig. 5. Normalized transmission spectrum of MSP fiber for Glen Thompson polarizer position of 168°. This corresponds to the minimum transmission through the fiber. Again,note the missing of the "bulge" at about 1200 nm. Note the OH absorption at about 1400 nm which is larger in the Pyrex fiber than in the commercial fiber.

fibers. This limited the effect of the polarization dependence of the optical spectrum analyzer. The polarization dependence of transmission spectra of Ag strip as well as of CdTe strip fiber sections were obtained.

The normalized transmission spectrum of an Ag MSP fiber at Glen Thompson Polarizer positions of 78° and 168°, corresponding to minima and maxima of the transmission through the MSP fiber are shown in Fig's 4 and 5. Of course, we have plotted data of these fibers in 10° polarization steps.

The logarithm of the ratio of maximum and minimum normalized transmission data, the data presented in Fig's. 4 and 5 is shown in Fig. 6. Note the "large" peak at about 1230 nm. The normalization with respect to the multimode fiber should eliminate effects of the polarization dependence of the optical spectrum analyzer. Note from Fig. 6 that the OH absorption was eliminated in this process. In order to make sure that the resonance peak in Fig. 6 is not due to the polarization dependence of the spectrum analyzer we analyzed the ratio of other normalized transmission spectra. For example, the logarithm of the ratio of the normalized transmission spectrum of date with the Glen Thompson Polarizer positioned at 118° and 208° is shown in Fig. 7. No resonance peak appears in this data. Of course, data with the Glen Thompson Polarizer adjusted to angles close to 168° and 78° exhibit smaller resonance peaks.

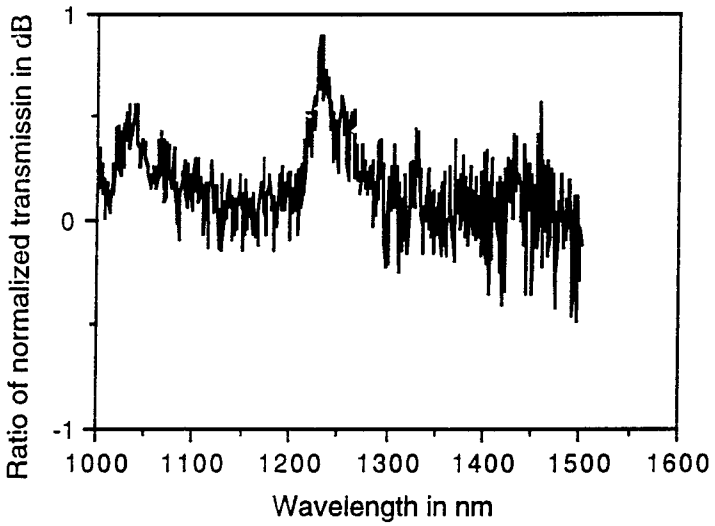


Fig. 6. Ratio of normalized transmission spectrum of MSP fiber for Glen Thompson polarizer positions of 168° and 78°. Note the plasma resonance peak at 1230 nm.

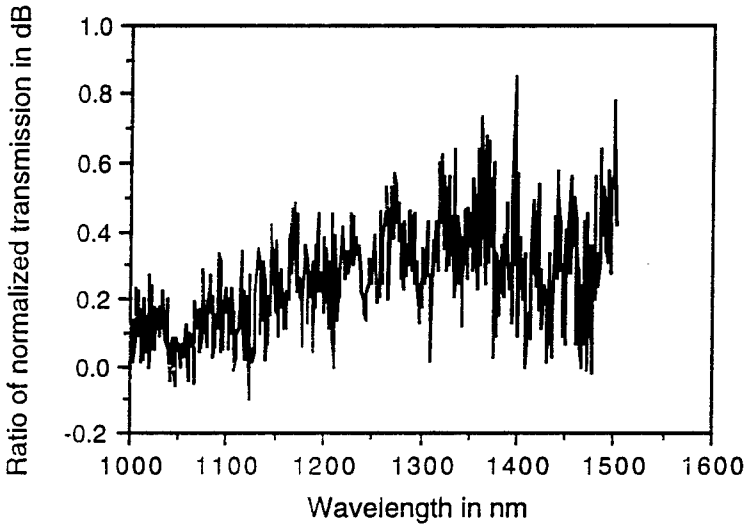


Fig. 7. Ratio of normalized transmission spectrum of MSP fiber for Glen Thompson polarizer positions of 208° and 118°. Note the absence of the plasma resonance peak at 1230 nm.

Metals exhibit plasma resonances at optical frequencies. We believe that the peak observed in Fig. 6 is a plasma resonance of the Ag film.

Of course, the present fibers with their large non guiding cores are not practically useful. Nevertheless, we were able to obtain some interesting data for the Ag strip fibers.

Syracuse University has glass on order that will allow the fabrication of single mode MSP fibers. To make these fiber polarizers commercially viable it is necessary to find metals that have plasma resonances near wavelength of 1.3 nm and 1.55 nm.

REFERENCES

1. "In-Line Fiber-Optic Polarizer" by W. Eickhoff. Electron, Vol. 16, pp 762-763, 1980

2. "Theoretical Study of Metal-Clad Optical Waveguide Polarizer" by Yu Tong and Wu Yizun, IEEE Journal of Quantum Electronics, Vol. 25, No. 6 June 1989 pp1209-1213.

3. "A Broad-Band Single Polarization Optical fiber" by Michael J. Messerly, James R. Oonstott, and Raymond C. Mikkelson, Journal of Lightwave Technology, Vol, 9, No. 7 July 1991.

4. Polarizing Single-Mode Fiber (PZ Series) Type FS-PZ-4611 for 820 nm, Type FS-PZ-5651 for 1060 nm, Type FS-PZ-6621 for 1320 nm. 3M Company specification sheet

James Masi Report not available at time of publication.

A Framework for Visualization for Imagery Exploitation

Timothy S. Newman
Assistant Professor
Department of Computer Science

University of Alabama
Huntsvillel,l AL 35899

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base
Washington, D.C.

and

Rome Laboratory

August 1995

Abstract

This report discusses a framework for visualization in imagery exploitation applications. We evaluate a recently-developed tool (called VIA (Visualization for Image Analysts)) whose goal was to aid image analysts through the use of visualization techniques. The strengths and shortcomings of VIA are discussed and suggestions for improvements and new directions for this tool are made. More generically, we also propose a re-casting of the usage of visualization techniques within the imagery exploitation arena, especially for battle damage assessment (BDA), mission rehearsal, and training regimens for image analysts. We also present our new framework for an improved graphical user interface (GUI) for the VIA tool. The new framework conforms to current directions and preferred practices in graphical user interface design.

1 Introduction

Visualization involves the use of computer-synthesized images or pictures to discover or highlight data geometry and topology (that is, the relationships between the data) [2, 9]. Visualization also often involves the hi-resolution animation of time-varying data; in fact, the original understanding and definition of visualization (from a 1987 National Science Foundation Report) considered this to be the hallmark of visualization [8]. Today, visualization also often involves the use of cross-disciplinary techniques, specifically the use of sophisticated scientific tools from other disciplines [9]. Topics of current investigation within the visualization communities include extraction and display of useful features from large datasets; using imaging representations and interactive graphics in manipulation and rendering, including volumetric modeling and rendering, such as viewing the interior of volumes; studying human perception of information; multi-media presentation of information, the portrayal of error and uncertainty in data; and "augmented reality" (also called virtual reality).

In this final report of our 1995 Summer Faculty Research Program experience, we present our work on the development of a framework for the usage of visualization techniques in imagery exploitation applications, particularly in image analysis. In Section 2 of the report, background material on the usage of visualization techniques for cartographic and image exploitation purposes is presented. In Section 3, we evaluate the (Grumman-developed) Visualization for Image Analysts (VIA) tool's aims and goals and the capability of the tool to meet those goals. The fourth section of the report discusses the framework that we are developing to improve the user interface (UI) of the VIA tool and presents our progress to date on this effort. We conclude the report with a discussion of potential future directions for visualization in the imagery exploitation arena.

An addendum to the report briefly discusses a second area of investigation for our summer study, the dual-use Statistical Multiple Object Detection and Location System (S-MODALS) artificial neural network (ANN) technology. In the addendum, we discuss our attempts to integrate the technology into the IE200C environment.

2 Visualization for Image Exploitation

The presentation of cartographic data has been studied extensively over many years and has produce a number of heuristic guidelines for mapping. For example, it is understood that color must be use carefully and appropriately to highlight relationships and patterns within map data. Many of the mappin guidelines are not formal rules but rather rely on the experience of the mapmaker. Currently, there is interes within cartography in the promotion of human-map interaction through computer-aided visualization o map data. Some of the topics being explored in cartographic visualization include the depiction of chang and movement, the use of multimedia, and the presentation of multiple views of data to aid in human understanding [5]. Many of these investigations currently focus on how the power of interactive computin can be used to aid user exploration of cartographic data.

Some of the experiences and techniques from cartographic visualization may be applicable in task within image exploitation that would benefit from visualization capabilities. Techniques for presentin cartographic information on a computer monitor may, however, differ from many of the techniques useful i visualization of cartographic or image data on paper. For example, the size of a computer monitor is limite whereas paper maps large enough to cover an entire wall can be created. Furthermore, the resolution c most computer monitors is generally limited to no larger than approximately 1000×1000 screen element (i.e., no more than 100-125 pixels per inch). Thus, information must be carefully presented on compute monitors.

The purpose of the project that produced the VIA tool was to explore the feasibility of using visualizatio techniques in an image exploitation task. Specifically, the goal was to aid image analysts. In imag exploitation, two of the interests are (1) searching to find targets and other features of interest and (2 detecting subtle changes to identify clandestine operations. Targets usually occupy very small region in typical imagery products and furthermore are sometimes oriented at a bad angle or camouflaged wit netting. There also tends to be minimal changes between images in a time series, making detection o change very difficult. The presence of shadows in an image can also complicate an image analyst's task By providing visualization cues to the analyst, his or her attention could be focused into areas of the imag more likely to contain targets.

VIA incorporated three visualization techniques including perspective viewing of aerial or satellit imagery (such as Landsat, SPOT, ADRI, etc.), overlaying feature information on images, and displayin 3D target keys. Grumman hoped that the perspective viewing might also help a pilot in navigation o target acquisition, in addition to helping image analysts in the creation of targetting products. Research i perception reveals that perspective and shadowing are very useful in human determination of position an orientation [3].

3 An Evaluation of VIA

In this section, we evaluate the VIA tool and propose a number of potential directions for future enhancements. One of the strengths of VIA is that it provides a set of visual cues and helps for image analysis tasks. The package's visualization capabilities are helpful, although they are also somewhat limited and tend to be somewhat cumbersome to use, particularly for novice users. The package could be made more useful with a number of extensions, including a more intuitive graphical user interface.

3.1 VIA Strengths

VIA has a number of beneficial features. One strength is that it supports the overlay of surface material features. The overlay capability is useful in locating certain features in a sparse environment (for example, roads in desert areas). The overlay feature also assists in the display data that will aid understanding and hasten user decision-making, for example by focusing analyst attention on areas likely to contain target(s) (through the display of the surface material features).

VIA also supports perspective viewing and zooming from any viewpoint. Viewing from different and, occassionally, unusual vantage points may assist users in finding things which might not have otherwise been found. Some have observed that users tend to take too much time to make a decision about image content when using the tool, however. This behavior may be caused by the availability of interactive exploration capabilities, such as perspective viewing and zooming. These features give the user more freedom and capabilities which he or she may wish to exploit fully before making a decision on image content.

VIA also allows the user to synthesize images from CAD models (created with the Army's Ballistic Research Laboratory CAD package (BRL-CAD). The illumination parameters for these synthetic target images can be adjusted for sun position. Shadows cast by these objects can also be rendered in the image, although sun position does not affect the rendering of the (non-synthetic) imagery. The synthetic images can also include opaque or semi-translucent camouflaging through simulation of the appearance of netting. The user is able to observe these synthetic images and, presumably, compare them with the image that is to be analyzed.

3.2 VIA Limitations

Products that include graphical user interfaces generally must follow one of two philosophies in their human-computer interaction. Specifically, the interaction must be either very similar to the interaction model of other products which provide generally similar functionality or the interaction must be *completely* different from other products that provide generally similar functionality [13]. Products with dissimilar interfaces tend to be more difficult for new users to grasp, thus it is generally held that it is advantageous

for graphical user interfaces to be similar, that is, to follow a standard [12, 13]. The VIA product conta several features that do not conform to generally accepted Motif standards and also has several deficienc in its GUI.

One of VIA's weaknesses is that it is not strongly compliant with the Motif Style Guide. Its usage "File" and "Quit" buttons on its main image canvas is not strictly compliant with Motif standards a also differs from general practice in graphical user interfaces. In other GUIs, these commands are m typically invoked from a menu bar at the top of the window or canvas that contains the image (or docum or drawing, depending on the application). Further, in VIA these buttons are physically grouped with otl commands that have different logical purposes. The other buttons that File and Quit are grouped with a have different behaviors. In GUIs, physically adjacent buttons should follow similar steps in accomplishi their tasks and should perform topically similar tasks[1]. The other buttons next to "File" and "Quit" all ca pop-up windows to appear on the screen. The pop-up windows control the illumination, viewpoint, overl parameters, etc., of the main display window. Since the "File" and "Quit" buttons behave differently (t do not bring up any windows or menus or cause any dialog box to be produced, and their behavior somewhat transparent to the user[2]), they should be physically separated (i.e., distant) from the viewpo and illumination control buttons. The "File" and "Quit" buttons also would be improved if they caus some type of dialog box to appear, even if the dialog box contained merely a description of what action t button was causing or about to cause. The "File" button actually causes the on-screen image to be saved a file while the "Quit" button causes VIA to exit. A superior usage of dialog boxes would provide the u with a cancel and confirm option before executing these actions.

The current interface allows the viewpoint to be changed by moving six sliders which govern the positi and orientation of the plane onto which the image is projected for perspective viewing. This interacti is somewhat cumbersome as it is difficult to position the view plane at a desired pose. It is also diffic to precisely position the sliders: if the user clicks on the slider just one or two pixels to the left or rig of where he should click to make a fine adjustment, the slider will move a large amount. This could improved by using sliders where the "micro" and "macro" adjustment areas are larger and more clea demarked. An even more powerful interface modality is that of the direct manipulation model, perha the most powerful mode of human-computer interaction [13]. For example, if the user could click t mouse on the view plane and drag the plane to new positions and orientations, a more direct manipulati mode could be achieved. For newer users and for users desiring to interactively explore imagery, a dir manipulation interactivity mode may be quite useful. A more natural point-and-click direct manipulati interface for zooming would also be useful.

[1]For example, in an editing or word processing application, all "clipboard" operations - (cut, copy, paste, etc.) are group physically close together. These operations have similar logical purposes and behave similarly (i.e., the user uses similar steps executing these operations and the machine appears to carry out execution of the operations in a similar and consistent manner).

[2]The "Quit" button does cause the application to exit, which could be considered to be non-transparent.

In VIA's pop-up windows (triggered by the "View", "Scene", and "Options" buttons), there are buttons labelled "Go" and "Quit." The "Go" button causes the settings in the pop-up window to be applied as new viewing or lighting parameters. In many GUIs, this functionality might be achieved with a button labelled "Apply." The "Quit" button actually causes the pop-up window to be dismissed; it does not cause the VIA application to exit. In many GUIs, this type of functionality is provided by buttons labelled "Dismiss" or "Cancel."

VIA gives the user some control over illumination of the synthetically generated target images through sliders that govern the elevation and azimuth of the sun in the sky. This allows precise control over all possible sun positions[3], however this interface may not seem as natural to some users as an interface that allowed the user to choose a time of day and a time of year and to see how the sun's position at that time affected target appearance. If the interface offered two methods of controlling the sun's position, users could set illumination parameters using whichever technique is most straightforward for them.

The current VIA tool allows only a small number of targets to be synthesized. It also does not allow the user to directly load new target models. It would be useful if a capability to load new models was added to the VIA tool. When images of target models are synthesized, the VIA tool's prior zooming and projection plane parameters also seem to change. Usually, the target is synthesized at a much higher resolution than the current image setting. This makes it somewhat difficult to compare the synthetic target image with regions of the real image that may contain the target. Finally, the targets can only seem to be synthesized in a small number of orientations. It would be more useful if the targets could be synthesized in any position and orientation and in different environments (e.g., a target partially obscured by trees or brush).

Visualization products that allow the user to navigate through an image require innovative ways to show current local position in a lower-resolution global map. Without features like this, users tend to become "lost" in an image. VIA's current display of the viewing plane overlaid on the global image helps the user somewhat in navigating through the image, but the projected image's direction that corresponds to the global image's "up" direction is not always apparent. Moreover, as the viewing plane is "zoomed" into the image, the viewing plane grows smaller and more faint in the image; it is difficult to distinguish the boundary of smaller viewing planes from image structures. This problem might be alleviated by displaying the viewing plane in a different color, such as red, or by displaying the boundary with wider, bolder lines. (Currently the view plane is an off-white or beige color that blends in with the image.)

Data should also be presented in a multiple-window, hierarchical fashion to aid user understanding of the images. It is important in cartographic and image visualization products to avoid a "single map mentality," which tends to be a limiting mindset for understanding the data [6]. One way to organize and present the data is through the use of a multi-resolution pyramid. This may allow quicker presentation of data

[3]VIA also allows the user to illuminate the targets using some sun positions that are impossible, such as a sun position on the northern horizon.

at a high level, helping the user locate his/her position in image while supporting exploration of the data at a low level. Such an approach also might use screen real estate more effectively. As has been noted previously, on-screen images require a new philosophy for data presentation; one cannot merely utilize a brute force conversion of a map into pixels, but rather information must be presented carefully on screen. It is important to not overwhelm the user and also to not under-utilize the screen or try to force too much information onto it. It is necessary for the presentation to be less map-oriented and more video-oriented.

VIA also does not allow the user to change image contrast settings during execution. Image contrast, the choice of whether to overlay the image on a terrain elevation map, and the choice of which image file to load all must be chosen when VIA is invoked and these choices cannot be changed without quitting and then restarting the application. These options should be changeable during program execution.

In VIA, the overlay of images on terrain currently does not make use of color. If VIA contained an option to display either the terrain elevation or the image using color, some users might find it easier to determine elevation or to see new features in the data overlaid on the terrain. It is quite difficult to see most terrain features in the current implementation of VIA. It might also be useful if, whenever the image was being overlaid on the terrain, the user could toggle display of the image on and off. If the user could see the terrain and the image separately and also be able to switch rapidly between one image and the other, the user might be able to better visualize the image in the context of the terrain.

It would be beneficial if visualization functionality was built into the IE2000 toolkit. This could be accomplished by adding support within VIA or other visualization tools to allow the easy input of many image types into the visualization tool. Currently, VIA can only accept images that are in MIVS format, which limits its applicability. Visualization tools will be much more powerful and useful if they are able to share images with other tools in the IE2000 environment. The IP Toolkit image processing utilities are one example of functionality that are not accessible by VIA images at present.

Currently, when images in VIA are zoomed by a very large factor, aliasing artifacts begin to appear. If anti-aliasing filters were included in VIA, these artifacts could be minimized. In order to satisfy any concerns about the filtering altering the fundamental image, it would probably be worthwhile to show both the non-anti-aliased image to the analyst as well as the anti-aliased image to the analyst for comparison purposes.

VIA could also be improved by the following enhancements. (1) More complete lighting models could be used, including the modeling of clouds and other atmospheric effects. (2) The user might be given the capability to add in their own surface material features. (3) A capability could be added for the looped playback of a series of user-selected image views, such as a sequence of views in a fly-through. A capability to pause or play a portion or all of the views backwards could also be added. (4) It may also be useful to provide 3D stereo views of imagery to increase realism and user understanding. (5) VIA needs to be able to render its images faster.

4 A Framework for a Graphical User Interface for VIA

In this section, our framework for an improved GUI for VIA is presented. Our framework is also generically applicable to other visualization tools for image exploitation. We also present a summary of the current status of our GUI development work.

We have begun preliminary development of an extension to the VIA GUI. The extended user interface is being developed using the Tcl/Tk system, a free, easily obtained, and widely available library that operates over X Windows developed by John Ousterhout [10]. Tcl is an acronym for tool command language. It is a scripting language for controlling and extending applications. Tk is a toolkit for X windows which extends the core Tcl with commands for building user interfaces. Tcl/Tk is an open software solution, making the resultant GUI highly portable and extendable. Tcl/Tk allows fairly rapid prototyping of GUIs, although there is somewhat of a learning curve associated with mastering the system. One of the difficult concepts in Tcl/Tk involve local variables being inactivated whenever user button-press, mouse-click, or keystroke input are processed. Another hurdle to mastering Tcl/Tk is its approach to evaluating variables and expressions, which often seems unusual to higher-level language programers and which may seem almost counter-intuitive for some programmers steeped in languages like C and Fortran.

One of the difficult parts of graphical user interface design is that interfaces are best constructed using an iterative design philosophy [13]. This involves a series of design steps, followed by user tests, followed by re-design and re-test. Although iterative design may cause lengthier processes to create a final product it also causes the user to be a more intimate part of the development process and tends to produce products that are very well-suited to the end users. We have tried to seek user feedback whenever possible from VIA users in the IE2000 work area and from those involved in image analyst activities. A commercial-grade end product would need to involve the user even more intimately than we were able to do so and would need to use more rigorously designed tests of user acceptability.

Our accomplishments in developing an improved user interface in the limited amount of time available are described in this section. None of our additions were done on the original program files, rather we made changes to a copy of the files named with a "-new" suffix. A number of data and source files for the original VIA product were found to be duplicated in multiple directories, and we did reorganize these files to ease the maintenance task and to save on storage space. We accomplished the reorganization by keeping one master copy of each duplicated file and creating links from where each other file instance had been.

One of our changes was to re-name certain button names to reflect more standard GUI term usage. The name changes also better reflect the behavior of the buttons. In the pop-up windows that controlled perspective viewing and illumination parameters, the "GO" button was renamed to "APPLY" and the "QUIT" button was renamed to "DISMISS." These changes were made in the Motif code that drives VIA rather than in the Tcl/Tk front-end to VIA where most of the rest of our changes occurred.

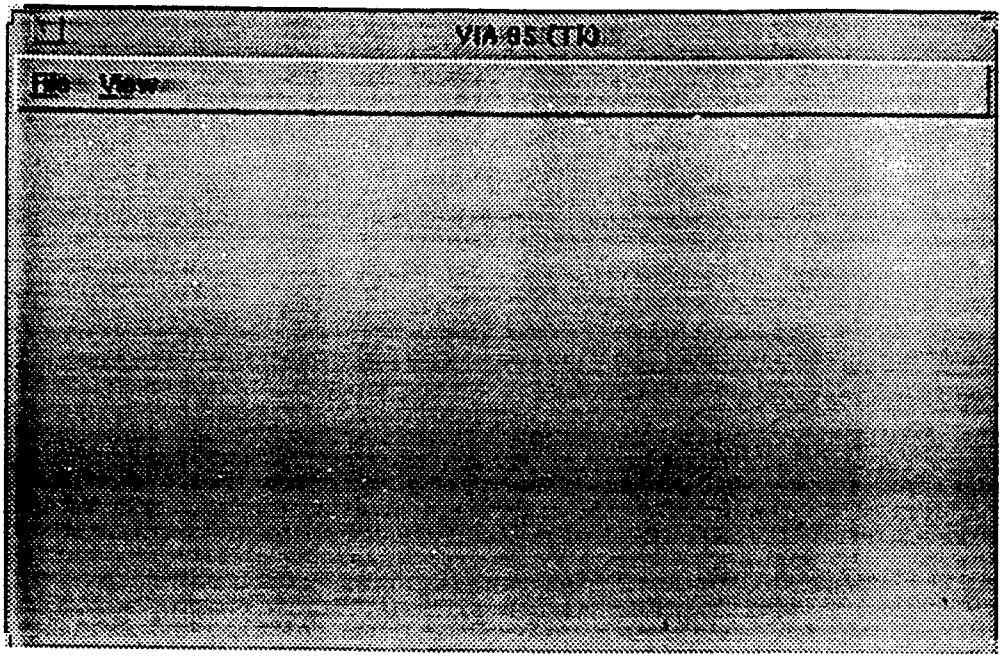


Figure 1: Top Menu

The Tcl/Tk front-end we built is a Motif-style screen and is strongly Motif-compliant. This shoul enable users, particularly new users, to master VIA quicker. Modern visualization and GUI desig practices both recommend that the user interface be consistent, easy to use, and have a clean organizationa structure [4, 13]. In our Tcl/Tk front-end, we re-organized the entire VIA GUI layout, making File an View be the only two top-level menu options. These terms are customary and usual in the X-Window application world. Our top-level screen is shown in Figure 1. Moreover, our goal was to develop a shor well-defined, logical structure to aid in decision-making and in the easy location of commands. These goal are strongly consistent with the choice of a menu-style of interface a good choice [4]. In Motif application which follow the Motif style guide, these options usually are selected from a menubar at the top of th screen. Tear-off menus may exist underneath these or any other top-level options although tear-offs at th topmost level are not generally accepted in the Motif application community. The tear-off capability i demonstrated in Figure 5, which shows the opened View window.

Underneath the File option, we added standard options for loading files, saving files, and quitting. Thes extensions also give VIA the ability to save to any file that the user chooses and to change to a new loa dataset without exiting the program (through a pop-up window opened with the Load selection of the Fil menu). The pull-down File menu is shown in Figure 2. Our re-organization of the layout also group logically and behaviorally similar items closely together on the screens. Logically and/or behaviorall distinct menus or objects are spatially separated on the screen. We also supported keyboard selection o menu choices. Accelerator keys could be easily added to our Tk front-end. Within the File menu, we hav

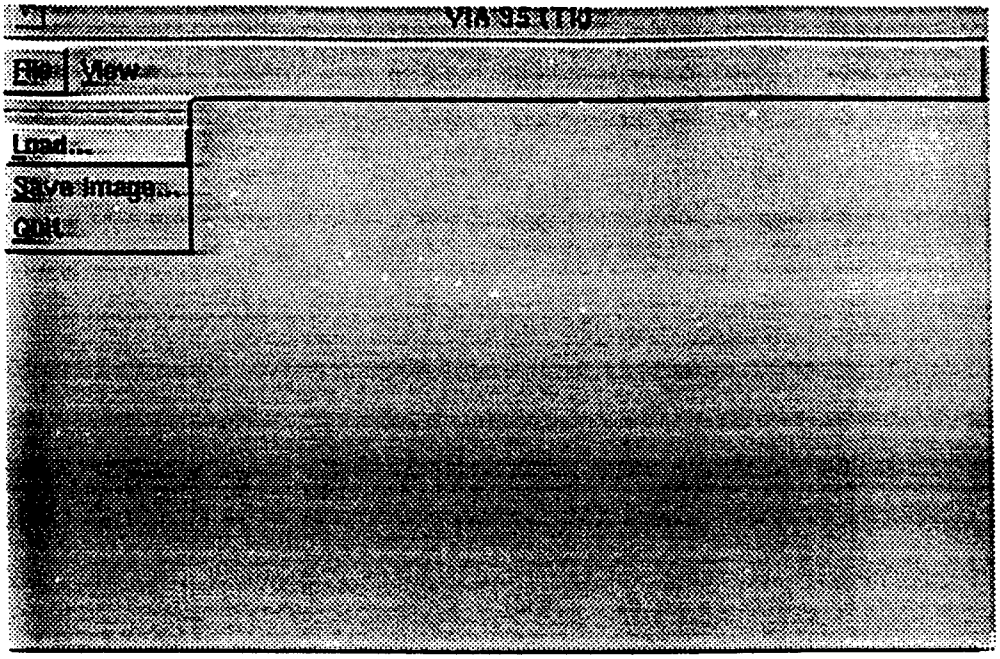


Figure 2: File Pull-down Menu

also built in the capability for the user to enter a save filename through the Save option. This is implemented using a pop-up window, shown in Figure 3.

In the pop-up Load window, we have added a capability for the user to enter an image filename with a browser (currently, input image names are hard-wired into VIA in the start-up script and there is no way to change files in the middle of a session). We also allow the user to choose from menu if VIA will use terrain overlay, elevation scaling, and feature overlay. In addition, the user can choose which contrast option to use. These features add much flexibility and makes for a clean, powerful interface with user. The pop-up Load window is shown in Figure 4.

Additionally, we allow the user to cancel their selections for the options in the pop-up menus. Any settings changed after screen invocation but prior to the cancel will be rolled back to their settings before screen invocation.

A number of items would have been tackled if we had more time. These include:

- Make the file browser click-sensitive for loading subdirectories.
- Actually use the filename that the user entered in the Load and Save menus (right now, these are only demonstration features due to the shortness of time for this project).
- Make the Tcl/Tk quit button destroy the (currently separate) X/Motif- driven VIA process.
- Remove the LOD button from the pop-up X/Motif menu. This button does nothing in the Grumman-

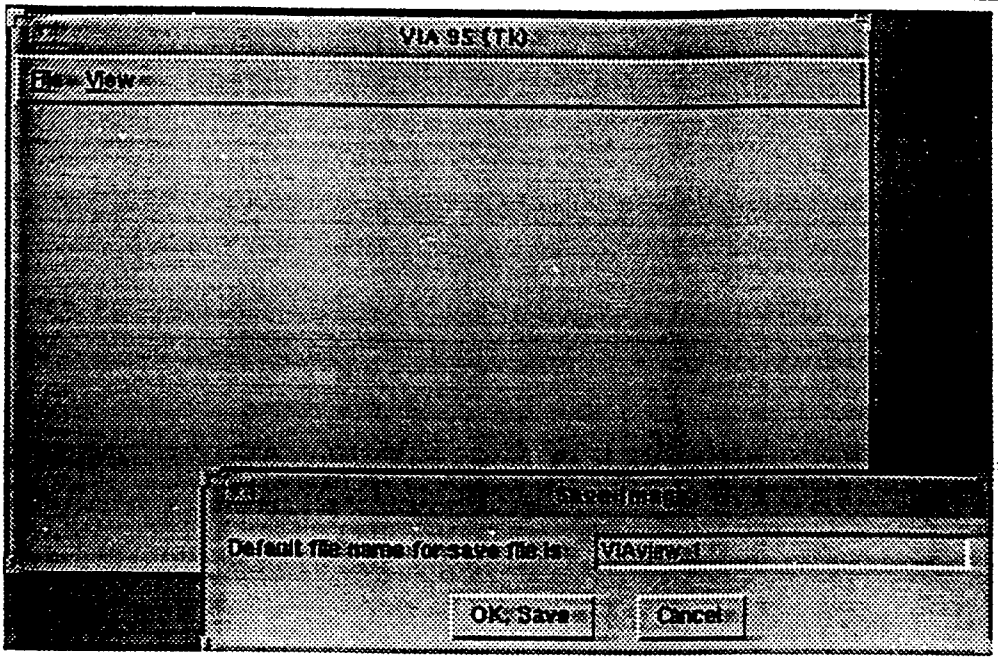


Figure 3: Save menu

supplied VIA program.

- Add a capability to click on locations in the image and find the corresponding position in geographic coordinates.
- Remove the viewing and illumination buttons from the VIA Motif window, replacing them with the menu options in the View menu (shown in Figure 5).
- Have the new VIA application write its main image output onto the Tcl/Tk window rather than opening a separate window. (This could be achieved by allowing the Tcl/Tk front-end to communicate with the X/Motif application. This is possible, although it involves mastering some deeper technical details of Tcl/Tk that we did not have time to resolve.) The main window of Grumman's VIA tool is shown in Figure 6. Figure 7 shows the pop-up window for selecting VIEW parameters.

5 Future Uses of Visualization in Image Exploitation

Visualization techniques have a strong potential to benefit image exploitation applications. Some of the visualization features which could be added to future visualization tools for image exploitation are discussed in this section. Some of the techniques VIA tool provide a starting point from which several applications could be spun off, for both image analysis and training applications. Three potential training applications

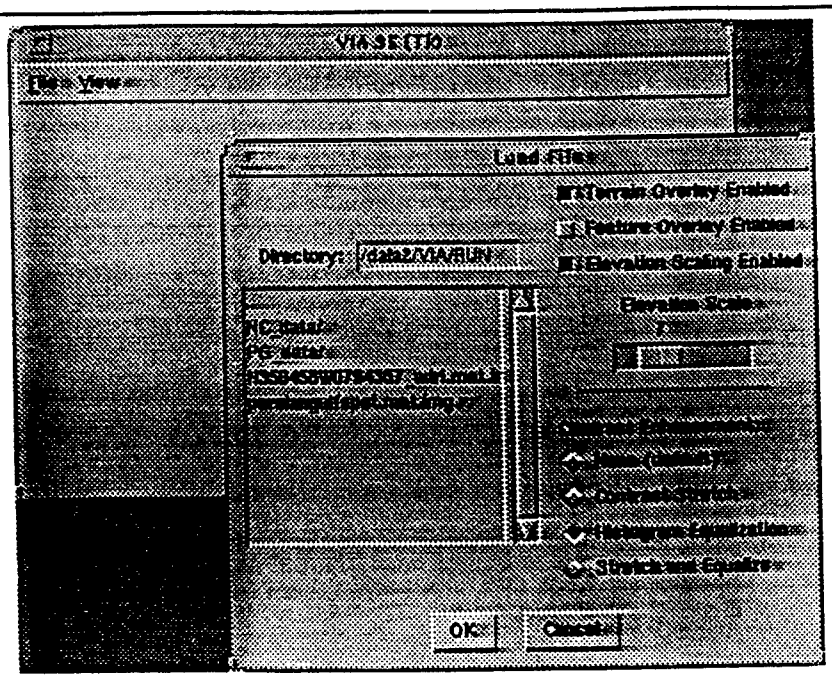


Figure 4: Opened Load

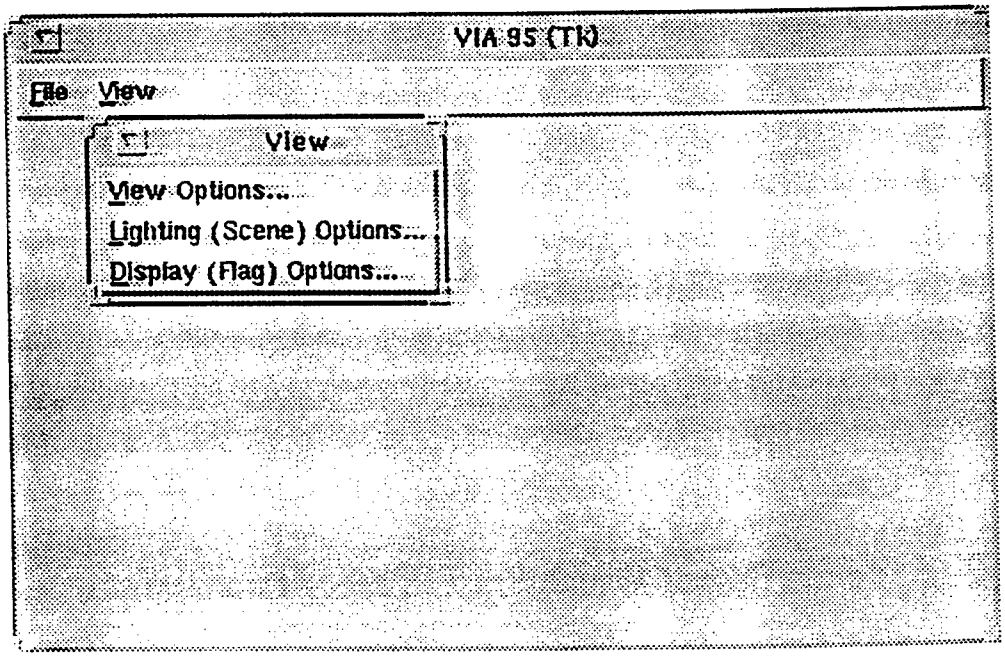


Figure 5: View menu

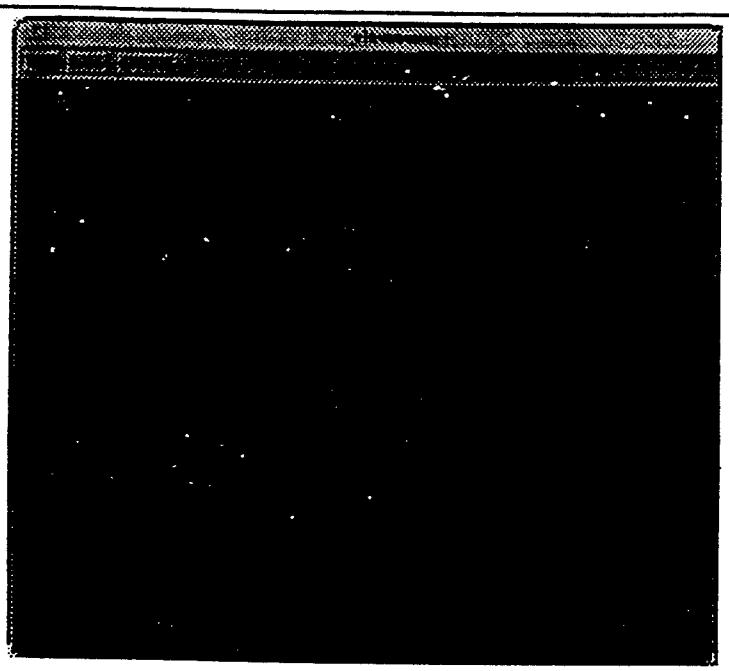


Figure 6: Image associated with selected view

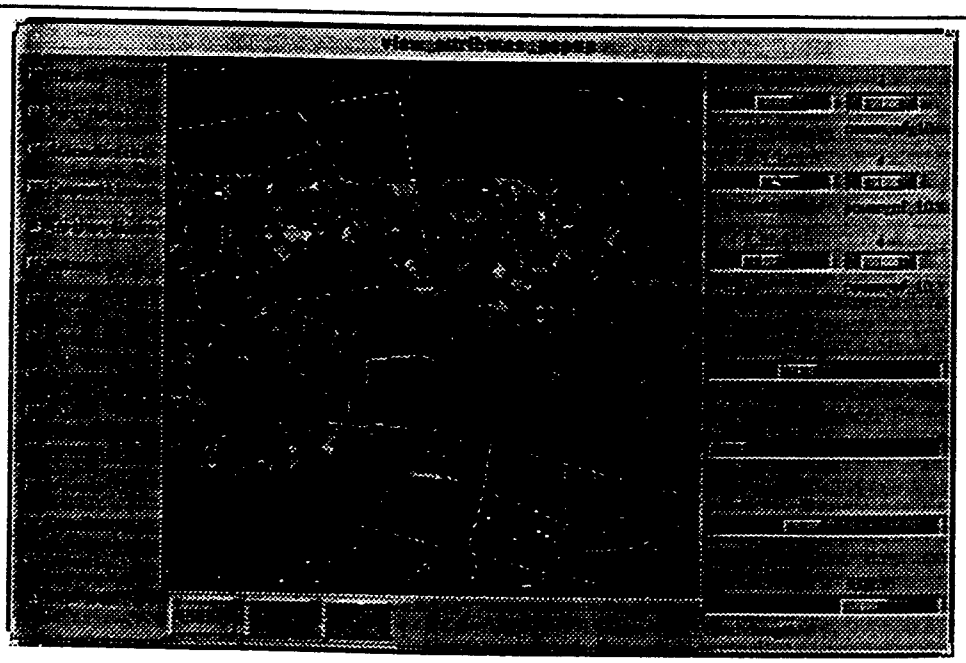


Figure 7: Selection of view

for visualization include training of image analysts, mission rehearsal, and training for battle damage assessment. Actual battle damage assessment activities would probably also benefit from visualization capabilities.

Interactive, exploratory visualization tools may prove useful in image analysis to provide a capability to more fully explore image data. One example of exploratory visualization capability in the VIA tool is the overlay of DTED and DFAD features. These help provide scene context, easing the task of locating some features. For example, some roads were often not readily apparent in the image yet when DFAD feature overlays were added, analysts reported that it was easier to locate road features. However, the current VIA tool only allows feature overlays on its most gross display (that is, it's lowest resolution, most "global" view); the zoomed-in views and the perspective views do not allow any feature overlays. If feature overlays were supported for zoomed-in views, it would probably be necessary to address the mismatch of feature data resolution to terrain and image resolution. Feature data is generally stored at much lower resolution than other image data. A number of techniques have been presented in the literature that could address the poor registration of feature to image data, including the image morphing and warping of computer graphics as well as the physical-based deformation models from computer vision. The mismatch between features and image data would look quite disconcerting in the highest resolution (maximally zoomed) images and is already apparent in some regions of VIA's "global" image display window. An alternate approach to the registration problem would be to give the user the capability to move or adjust the overlays in an interactive fashion. This capability would have to allow for both global and local adjustments because the mis-registration is not merely a simple global rotation and translation; the mis-registration is different in different parts of the image. Additional capabilities that aid in image exploration should be part of future visualization efforts for image exploitation applications.

Visualization tools could also be used to aid image analysis by providing techniques that assist in target identification. For example, the shadow cast by an object can be used to derive 3D object shape. Automated algorithms that reconstruct potential object profiles from shadows in the image could assist image analysts in determining object identity. This could be implemented by allowing the analyst user to interactively segment the rough boundary of a shadow from an image and then use a semi-automatic machine algorithm to complete segmentation and to begin object shape detection and 3D synthesis of potential objects. It might be possible to recognize the object by machine as well, although automatic object recognition is more difficult and perhaps less reliable than an interactive process that keeps the human analyst in the loop. BRL-CAD or other CAD models, DTED, and DFAD could be used to aid 3D shape detection and then used to synthesize the object in the image.

For future visualization tools in the image exploitation arena, capabilities to allow interactive determination of position and image statistical features might also be beneficial. For example, if the user could click on an area of the image with the mouse and be able to see a side window pop up with coordinat

information (in both geographic and UTM coordinates), the ability to navigate through the image and to determine precise positioning within the image would be aided. Furthermore, if a similar point-and-click capability was enabled for DFAD feature data, the user's ability to gather useful information from the image would also be aided. The functionality termed geographic brushing in the literature [7] might also be useful in future products. Geographic brushing means the display of image region features (such as local region histograms), often statistical in nature, for moveable, selectable regions on the map or image. Scatter plot outputs might also be supported to display the nature of the local image features.

It should be possible to correct images for shadows caused by sun position[4]. Through normalization of shadowed images, two sun-normalized images of a pre-stored 3D target model (probably generated from a CAD model) and an image could be presented for comparison to the analyst. This may assist the analyst in the identification of image objects.

Image analysts may also benefit if future software tools allowed the loading and display of images of targets in other images from prior scenarios. This would give the analyst an encyclopedic ability to compare the current image with previous images and subimages containing similar suspected targets/threats. Pattern analysis or template-matching techniques would be useful to screen archived image data and in the selection of images for viewing.

It would also probably be useful to attempt to display data that can't normally be visually seen, such as the visibility window of a ground-based radar site or radar signatures from different orientations. Other non-visual data, such as radioactivity and radio activity also might be useful. These could be displayed as colored translucent "domes" of activity overlaid on the image plane or by other useful visual means. These displays might be useful both for image analysis and for mission rehearsal.

Visualization tools in the image exploitation arena need to be able to support data from multiple sources and to include capability to overlay those data types on top of each other. As has been mentioned earlier, the issues in registration of data are not trivial, indeed a large number of researchers are actively working in this area. In image exploitation, the capability to overlay multiple sensor data, including photographic data, infrared data, elevation data, and radar data (such as synthetic apperture radar) are necessary.

5.1 Detection of Change

Another capability that would be useful in the image exploitation environment, for both battle damage assessment and for detection of clandestine activity, is that of displaying enhanced differences in images collected over time. One way that this capability might be provided is by flickering between "old" images and "new" images, perhaps allowing an analyst to detect clandestine activity. Research in human perception and cognition seems to indicate that humans have an enhanced ability to detect image differences through

[4]Both time of day and time of year of image capture will affect shadowing.

this method [11]. It is not clear if VIA or any other visualization tools have actually been applied to any extent in this area. Other techniques for displaying change in images collected over time might involve the use of statistical pattern recognition to detect image regions that differ and then using color, height maps, or overlaid (perhaps translucent) symbols to highlight the changed regions. Another possibility to detect subtle change over time might involve the use of local region histograms.

Visualization of future battle scenarios might also be useful for battle damage assessment and for war-gaming purposes. If possible damages from munitions could be simulated, battle planning might be aided as well as battle damage assessors trained prior to actual battle situations. Thus, visualization could be useful in a training role for battle damage assessment.

It may be useful to couple visualization functions with the S-MODALS to be able to recognize image regions that have suffered battle damage. S-MODALS may also be useful in recognizing targets in an image or to recognize regions of change in images over time, such as helping determine if SAM sites have been built or camouflaged tanks moved into a region.

5.2 Fly-throughs

The realistic rendering of landscape elements as overlays on images or maps may also be useful in image exploitation. For example, pilots often use waterways for navigation, thus it may be useful to render waterways in a photo-realistic manner, especially for mission rehearsal applications. A capability that provided actual texture mapping of an image product onto a terrain elevation model would probably also be a useful adjunct to the photo-realistic display of waterways. Other landscape elements such as trees, forests, cities, etc., could be rendered using fractal techniques. If Vertical Flight Obstruction Data was available, power lines, towers, chimneys, etc., could also be synthesized for mission rehearsal fly-throughs. Display of minimum flight elevation (MFE) data as a translucent "cloud" over the image might too be useful for this applications. The interpolation of terrain elevation might also add to the feeling of photo-realism. For real-world use, scenes would have to be constructed quickly from the 3D terrain model. There appears to be a significant level of interest from many commanders to train for missions using computer-generated views and fly-throughs. Contour mapping of terrain elevation data may also be useful as a display option for some applications.

For fly-throughs or drive-throughs (i.e., navigation-style applications), it would be useful to support the display of multiple scene views simultaneously on the screen while also taking care not to clutter the screen. This may assist the user in determining their current position and also help maintain a global view of the data and a memory of what the user has previously visited. Techniques similar to this have been recommended by some in the cartographic visualization community [6]. One approach might be to let the user save a series of viewpoints by clicking on the image display and then display the scenes from those viewpoints, maintaining the same projection geometry for all views. This would roughly simulate the user

moving from point to point along flight paths. It may also be useful to interpolate views between points. I a few rendering short-cuts can be used in "navigating" through the different scene viewpoint, near-realtim navigation will become more likely. Rendering "shortcuts" that involve image interpolation may not b suitable for image analysis applications, though, due to the possiblity of missing some scene features. Fo effective navigation, a multiple resolution approach to data representation is likely also required. This wil allow the user to achieve both global navigation and local navigation [3]. Further, use of text-based aid and navigation buttons seems to help avoid information overload [1].

One concern about the viability of fly-throughs or drive-throughs might be the speed at which scene can be rendered. One typical approach to rendering landscape images is to map images or photos ont a polygonal approximation of the terrain elevations. Often, this can produce near-realtime animations However, one open issue in the overlay of photos or other texture maps is how to quickly overlay *multipl* texture maps onto the terrain polygons in near-realtime. To achieve near-realtime operation, it may b necessary to port a fly-through to a parallel machine. If there are few 3D objects to render and little 3l maneuvering, however, a fast serial machine may be adequate.

5.3 Image Analyst Training

Perhaps one of the most useful short-term applications for visualization for image analysis is in the trainin of image analysts. By providing access to powerful visualization tools that allow analysts to see a larg collection of existing images of various targets and by also giving analysts the capability to see synthesize images of targets and threats in new scenarios, it should be possible to provide a large and useful set o training imagery for the beginning analyst. Analysts could thus be exposed to many more situations tha are actually readily available in real imagery. Visualization can play an important role in image analys training.

5.4 Finale

No report on visualization would be complete without mentioning that before final adaptation of visualization product in a non-experimental setting, it would be necessary to run several experiments wit many users and images and to then verify that visualization techniques indeed do improve productivity Carefully constructed visualization tools should be able to meet the challenge of improving productivity b helping end users discover trends in data that were not apparent before.

The focus for future military applications will be the fast integration of multiple sources of data. It wil be necessary to put data together quickly and effectively from multiple sources. Once the data is integrated it will be necessary to quickly plan using the data. It may also be necessary to quickly train using th data, and then to respond quickly to emerging threats. Visualization can assist in providing the requisit

interactivity, connectivity, and integration required in future military enterprises. Visualization has the potential of assisting both the operators in the field and the planners at headquarters.

Addendum: Extension of S-MODALS ANN

The S-MODALS technology has been used for target recognition, for medical diagnostic applications, and for justice applications. This technology has been found to be especially useful in the detection and classification of pulmonary tumors.

One weakness of the ANN software is that it lacks a graphical user interface (GUI). Recently, a TIM40 board with three C40 processors has become available to Rome Laboratory, offering the opportunity to explore parallel implementations of the artificial neural net technology.

For our summer work, we originally proposed to investigate parallel realizations of the artificial neural network on the TIM40 hardware. Furthermore, we proposed to develop a graphical user interface for the ANN software. The addition of a GUI to the S-MODALS software would have made it a very useful part of the IE 2000 environment.

A.1 Parallel Realizations of the ANN S-MODALS System

Our first milestone was to have been a port of the neural network algorithm to the TIM40 environment. We were next planning to investigate the effects of task granularity and memory contention on the parallel algorithm performance. Our goal was to optimize the performance of the algorithm on the TIM40. In addition, we were to attempt to improve the algorithm so that its performance was more robust. An increased usage of knowledge about the problem environment might have delivered this performance improvement.

Complications in ANN Research

Some of the difficulties we encountered in our efforts included:

- A contractor controlled the source code and would not release it to us. We were unaware of this when the project began.
- The contractor would not provide timely assistance in resolving configuration problems in porting the software executable to Rome Lab.
- We eventually discovered that the software would not run on the hardware configuration at Rome Lab. The contractor would not supply the changes necessary to allow the software to run. Due to the contractor controlling the source code, we were unable to make the necessary software configuration changes that would have allowed the software to run on our hardware.

- We did not have a cross-compiler to allow us to generate source code of our own that exercised t TIM-40 boards at Rome Lab.

- We switched from the ANN project in the seventh week of ten at Rome Lab. We had spent a great amount of time and effort to discover and resolve the configuration problems and really had only two weeks to focus on our final project in visualization.

References

[1] H. Asche and C. M. Herrmann, "Designing Interactive Maps for Planning and Education," in [5], pp 215-242, 1994.

[2] A. Kaufman, "Trends in Volume Visualization and Volume Graphics," in *Scientific Visualization: Advances and Challenges*, ed. by L. Rosenblum et al., Academic Press; San Diego, 1994.

[3] M.-J. Kraak, "Interactive Modelling Environment for Three-Dimensional Maps: Functionality and Interface Issues," in [5], pp. 269-285, 1994.

[4] M. Lindholm and T. Sarjakoski, "Designing a Visualization User Interface," in [5], pp. 167-184, 1994.

[5] A. M. MacEachren and D. R. F. Taylor, *Visualization in Modern Cartography*, Pergamon/Elsevier Tarrytown, NY, 1994.

[6] C. McGuinness, "Expert/Novice Use of Visualization Tools," in [5], pp. 185-199, 1994.

[7] M. Monmonier, "Graphic Narratives for Analyzing Environmental Risks," in [5], pp. 201-213, 1994.

[8] Report of the NSF Advisory Panel on Graphics, Image Processing, and Workstations, 1987.

[9] G. M. Nielson, "Visualization Takes its Place in the Scientific Community," *IEEE Transactions o Visualization and Computer Graphics*, Vol 1, No. 2, 1995, pp. 97-98.

[10] J. Ousterhout, *Tcl and the Tk Toolkit*, Addison-Wesley, 1994.

[11] M. P. Peterson, "Cognitive Issues in Cartographic Visualization," in [5], pp. 27-43, 1994.

[12] J. Preece, Y. Rogers, H. Sharp, D. Benyon, S. Holland, and T. Carey, *Human-Computer Interactio* Addison-Wesley: New York, 1994.

[13] H. Thimbleby, *User Interface Design*, ACM Press: New York, 1990.

Calibration of Infrared Thermograms of Electromagnetic Fields

John D. Norgard
Professor
Department of Electrical and Computer Engineering

University of Colorado
Colorado Springs, CO 80917-7150

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base
Washington, D.C.

and

Rome Laboratory

August 1995

CALIBRATION OF INFRARED THERMOGRAMS OF ELECTROMAGNETIC FIELDS

John D. Norgard
Professor/ECE
Department of Electrical & Computer Engineering
University of Colorado

Abstract

The infrared (IR) imaging technique for measuring electromagnetic (EM) fields was further developed this summer to map two-dimensional EM field distributions near a radiator or a scattering body. Both electric and magnetic field intensities have been measured with this technique. Initial tests to prove the validity, accuracy and sensitivity of this technique were performed this summer in the anechoic chamber at the Electromagnetic Vulnerability Analysis Facility (EMVAF) at Rome Laboratory (RL). Empirical calibration tests were performed to determine the absolute intensity level of the magnitude of the electric field. Magnetic field intensity levels were not calibrated this summer.

Experimental tests were performed on a conducting cylinder irradiated by an incident plane wave. E- and H- field patterns of the scattered/diffracted energy from the cylinder where measured. Tests were conducted at various microwave frequencies relative to the resonant frequencies associated with the cylinder and at numerous angles of incident (eg. end-on, broad-side, oblique-incidence) and for several different polarizations of the incident field relative to the axis of the cylinder (eg. horizontal, vertical, and skewed). Each tests was performed in the near and far fields of the cylinder. IR thermograms of the scattered fields were taken. The equi-temperature contour levels in the IR thermograms are being compared to numerical predictions of the scattered electric field intensity. The numerical code was developed at the Jet Propulsion Laboratory (JPL). These experimental results will be used to validate the accuracy of the JPL numerical code.

The imaging technique was also used to map the scattered electric field around a model of an aircraft. A plastic F16 scale model was constructed and sprayed with several coats of silver paint to made it conductive. IR thermograms of the magnitude of the scattered electric field intensity were taken in the horizontal and vertical longitudinal planes through the fuselage of the aircraft. The equi-temperature contour levels in the IR thermograms are being compared to numerical predictions of the scattered electric field intensity. The numerical predictions are being performed with the GEMACS code at Rome Laboratory. These experimental results will be used to validate the accuracy of the GEMACS numerical code.

The lossy IR detector screen was irradiated by a normally incident EM wave in the near and far fields of the antenna. The level of the incident electric field was measured using D-dot probes and was

also predicted theoretically using the Friis Transmission/Reception Formula. The known field radiated from the standard gain horn antenna was used to calibrate the measured temperature levels in the IR detector screen. A correlation of the measured temperature level (a selected color on the IR thermogram) to the measured or predicted magnitude of the incident electric field was used to calibrate the IR technique. A table of color level vs. incident field level was made. This table is the calibration curve for the IR detector screen used in the calibration test. Different IR detector screens with different sensitivities can all be calibrated using this same technique.

Four papers were presented at international IR and EM conferences this summer; two seminars were also presented on the technique. One paper was published in an IR journal; three new papers have been submitted for presentation at several IR conferences next year.

CALIBRATION OF INFRARED THERMOGRAMS OF ELECTROMAGNETIC FIELDS

John D. Norgard

Table of Contents

CALIBRATION OF INFRARED THERMOGRAMS OF ELECTROMAGNETIC FIELDS

John D. Norgard

An infrared (IR) measurement technique is being developed to measure electromagnetic (EM) fields. This technique uses a minimally perturbing, thin, planar IR detection screen to produce a thermal map of the intensity of the EM energy over a two-dimensional region. EM fields near radiating microwave sources and scattering bodies can be measured with this technique. This technique also can be used to correlate theoretical data with experimental observations and to experimentally validate complicated numerical codes which predict electric field distributions inside waveguide cavities (E-Fields) and surface current distributions on metallic surfaces (H-Fields).

1. Introduction

A non-destructive, minimally perturbing IR measurement technique is being developed to observe EM fields. Metallic surface currents and charges also can be measured with this technique.

This IR measurement technique produces a two-dimensional IR thermogram of the electric or magnetic field being measured, i.e. a two-dimensional isothermal contour map or a gray scale of the intensity of the EM field.

Electric and magnetic fields can be measured separately. For example, electric field patterns radiated from microwave horn antennas, electric field intensities coupled through apertures in shielded enclosures, diffraction patterns of electric fields scattered from complicated metallic objects, and electric field modal distributions induced inside cylindrical waveguide cavities can be measured. Also, magnetic field distributions near conductive surfaces and induced surface currents on metallic surfaces can be determined.

The advantages and disadvantages of this new IR measurement technique are discussed later.

2. IR Measurement Technique (Overview)

The IR measurement technique is based on the Joule heating that occurs in a lossy material as an EM wave passes through the material. A thin, planar sheet of lossy carbon loaded paper can be used to map electric fields; a thin, planar sheet of lossy ferrite loaded epoxy can be used to map magnetic fields. In either situation, the absorbed heat energy is converted into conducted and convected heat energy and into re-radiated EM energy. The radiated EM energy is concentrated in the IR band. This "black body" energy can be detected with an IR (Scanning) Array or with an IR (Starring)

Focal Plane Array (FPA).

2.1. IR Experimental Setup

This technique involves placing a lossy non-perturbing IR detection screen in the plane over which the EM fields are to be measured.

2.1.1. Electromagnetic Parameters

The detector screen can be made from a thin sheet of linear, homogenous and isotropic but lossy material. From the complex form of Poynting's Theorem for a linear, homogeneous and isotropic material, the absorbed power P_{abs} within a given volume V of the lossy material is a function of the electric (E) and magnetic (H) field intensities inside the screen and is given by

$$P_{abs} = \int_V (\sigma E^2 + \omega \epsilon'' E^2 + \omega \mu'' H^2)\, dV \quad [W/m^2] \qquad (1)$$

where σ is the conductivity of the detector screen, ϵ'' is the imaginary component of the permittivity of the detector screen, μ'' is the imaginary component of the permeability of the detector screen, and ω is the radian frequency of the incident field. The volume integral is over the illuminated portion of the detector screen. The spectral characteristics of the complex constitute parameters (μ, ϵ,σ) of the detector material must be known (or measured) over the entire frequency bandwidth to be measured.

The incident EM energy is absorbed by the lossy material and is converted into thermal heat energy which causes the temperature of the detector material to rise above the ambient temperature of the surrounding background environment by an amount which is proportional to the local electric and/or magnetic field intensity (energy) at each point (pixel) in the screen material. In regions where the fields are strong, the absorbed energy is large and the resulting pixel temperatures are high; in regions where the fields are weak, the absorbed energy is small and the resulting pixel temperatures are low. The resulting two-dimensional temperature distribution over the surface of the screen can be detected, digitized and stored in the memory of an IR camera. The temperature distribution on the surface of the screen without any EM energy incident on the screen also can be stored in the memory of the IR camera as a ambient background reference temperature distribution. The difference in the temperature distributions at each pixel (between the illuminated and the non-illuminated screen) is due to the effects of the electric or magnetic field incident on the screen at that pixel location.

The EM effects can be visualized by presenting the differenced two-dimensional temperature profile as a false color image, where cool colors (for example shades of blue) represent weak areas of

EM energy and hot colors (for example shades of red) represent strong areas of EM energy. The resulting two-dimensional false color image is called an IR thermogram, ie. an iso-temperature contour map, and is a representation of the electric and/or magnetic field distribution passing through the screen material.

2.1.2. Thermal Parameters

For a planar sheet of detector screen supported by a block of non-conducting material, eg. a styrofoam block, the thermal problem reduces to considering only the radiative and convective heat losses from the surface of the detector material.

The convective heat loss h_c is approximated by

$$h_c = h_o (T - T_{amb})^{1.25} \quad [W/m^2] \tag{2}$$

where h_o varies between 1.4 and 1.6. The radiative heat loss h_r is approximated by

$$h_r = \epsilon_{ir} \sigma_{ir} (T^4 - T_{amb}^4) \quad [W/m^2] \tag{3}$$

where ϵ_{ir} is the detector surface emissivity, σ_{ir} is the Stefan-Boltzman constant in $W/m^2 - K^4$, and the temperatures are in degrees Kelvin. The conductive heat loss is negligibly small.

2.1.3. Thermal Equilibrium

The heat transfer problem in the detector material involves solving a non-linear, second order differential equation in both space and time, while considering radiative and convective heat losses from the surface of the material, conductive heat transfer within the material, and the EM power absorbed in the material as a function of distance into the material. For the case of the thin screens considered here, the temperature is initially considered to be constant in the direction normal to the surface of the material, so that the conductive term normal to the surface of the screen can be ignored and the power absorbed can be considered independent of the direction normal to the surface of the screen. Also, the time dependence of the absorbed heat energy is ignored for the steady-state solution.

Relating the convective and radiative heat losses in equations (2) and (3) to the absorbed power in equation (1), results in the following equation at thermal equilibrium:

$$P_{abs} = h_c + h_r \tag{4}$$

For a properly optimized detector screen, thermal equilibrium is achieved in just a few seconds.

This non-linear thermal/electrical equation can be solved for the electric or magnetic field, as a function of the material temperature T, using approximate techniques.

2.2. Approximate Solution

Equation (4) is highly non-linear for large temperature variations above ambient, due to the thermal processes of convection and radiation. However, for small temperature variations of only a few degrees above ambient, equation (4) can be linearized for small incremental temperature changes $\Delta T = T - T_{amb}$ above the ambient temperature T_{amb}.

This condition of small temperature variations above ambient is a desirable operational constraint, since this is also the requirement for small absorption of the EM energy passing through the screen, which equates to a small perturbation of the incident field when performing the measurement. For this minimally perturbing measurement case, an almost direct linear correlation exists between the incremental surface temperature ΔT and the absorbed electric or magnetic field intensity. The incident electric or magnetic field can then be determined from a solution of Maxwell's Equations (Fresnel's and Snell's laws) for an EM wave incident on a planar film of lossy material. Therefore, it is possible to correlate local surface temperature variations ΔT to E or H field intensities.

For the scanning camera available this summer at Rome Laboratory to make the IR thermograms, temperature differences ΔT as small as 0.09 °K could be detected.

Care is exercised, therefore, in the selection of the screen material not to significantly perturb the electric or magnetic field by the presence of the lossy material. The screen can be designed to absorb from 1% to 5% of the incident power and to produce a temperature change of less than a few degrees. The constitutive parameters of the IR detector screen can be optimized to produce a large temperature rise in the detector material for a small amount of absorbed energy.

Electric and magnetic fields produced by Continuous Wave (CW) sources operated in the sinusoidal steady-state mode are easy to measure because of the large amount of energy contained in the wave. Transients produced by high power microwave (HPM) pulsed sources, especially repetitively pulsed sources, also can be measured, if the average energy content in the pulse is high enough to raise the temperature of the detector screen material above the minimum temperature sensitivity of the IR camera. The thermal mass of the detector material should hold the absorbed heat energy long enough to capture the IR thermogram of the pulse.

2.3. IR Detector Screen

Referring to equation (1), the screen material can be tailored to respond to only one component

of the field, e.g. by optimizing the values of the electrical conductivity σ and the imaginary part of the permittivity ϵ" of the material relative to the imaginary part of the permeability μ" of the material, the detector screen can be made sensitive either to the tangential component of the electric field or to the tangential component of the magnetic field in the plane of the screen.

For example, an *electric* field detector screen can be constructed either

i) from a lossy material with a high conductivity σ and a low imaginary permittivity ϵ" and a low imaginary permeability μ"

or ii) from an electrically polarizable material with a high imaginary permittivity ϵ" and a low conductivity σ and a low imaginary permeability μ".

Alternatively, a *magnetic* field detector screen can be constructed from a magnetically polarizable (magnetizable) material with a high imaginary permeability μ" and a low conductivity σ and a low imaginary permittivity ϵ".

The optimization of the thermal and electrical parameters of the detection screen material should be guided by a thermal/electrical computer code based on a plane wave normally incident on a planar interface between air and the lossy detector material. Other absorptive and re-emittive transducing materials have been studied for use as passive thermal screens for IR thermograms.

2.3.1 Electric Field Detector Screen

For the detection of electric fields, the IR detector screen can be made from a planar sheet of lossy thin-film material. Several different detector screens were used to make electric field thermograms. One was a carbon loaded thin film (eg. Teledeltos Paper) 80 μm thick with a conductivity of 8 mhos per meter. The other screens were made from carbon-loaded Kapton films. The films were loaded with different resistivities per square. These materials are non-polarizable and non-conducting; therefore, the imaginary components of the permittivity ϵ" and the permeability μ" are negligibly small. For these non-conducting, non-polarizable, non-magnetic screen materials, maximum heating occurs due to the electric field and negligible heating occurs due to the magnetic field.

For plane waves normally incident on this carbon loaded electric field detection screen and for an IR FPA with a temperature sensitivity of 0.01 °K, electric fields with a magnitude on the order of 61.4 V/m (1 mW/cm^2 of incident power) can be detected. This result was obtained empirically from experimental data in which the incident power level was incrementally decreased until no electric field contour lines on the IR thermogram were discernable from the ambient background level.

2.3.2 Magnetic Field Detector Screen

For the detection of magnetic fields, an IR detection screen can be fabricated using a ferrite loaded epoxy. Several different ferrite powers have been tested, viz. Nickel Zinc Copper Ferrite, Iron Silicide Ferrite, Ferrite 50, etc. The best detector screen used to make magnetic field thermograms had an 80% by weight mixture of Ferrite 50 with a thickness of 0.5 mm and an imaginary relative permeability μ_r" of approximately 2. The imaginary component of the permittivity ϵ" and the conductivity σ were negligibly small. For this magnetic, non-conducting, non-electric screen material, maximum heating occurs due to the magnetic field and negligible heating occurs due to the electric field.

In the fabrication of a magnetic screen, it is difficult to keep the electric polarizability of the epoxy base material from contributing to the imaginary component of the permittivity ϵ" of the composite and ,thereby, introducing additional electric polarization losses. If the electric properties of the ferrite are not carefully controlled and minimized, the absorbed power will be due to both the electric and magnetic components of the EM field, as given by equation (1). In this case, it is difficult, if not impossible, to separate the two coupling mechanisms from each other and to calibrate the technique.

To detect surface currents on a metallic structure, the magnetic detection screen can be placed in direct contact with the surface of the metal. The thickness of the detector screen must be kept to a minimum, so that the magnetic field that is measured on the outer surface of the screen is a direct reflection of the magnetic field on the inner surface of the screen in contact with the surface of the metal. The material can be held flat against the surface of the metal using a non-conducting glue.

The surface current is related to the tangential component of the magnetic field intensity on the surface of the metal by the magnetic boundary condition on a perfect conductor:

$$\hat{n} \times \vec{H} \mid_s = \vec{J}_s$$

where n is the normal to the surface S. The magnetic field is perpendicular to the direction of the surface current.

2.4 IR Camera

The temperature difference between the screen material and the background is detected, digitized, and stored in the memory of an IR camera on a pixel by pixel basis. The IR imaging system used to take the IR thermograms has approximately 200 by 100 pixels per frame of data. The detector

is scanned over the image. The detector is a Mercury-Cadmium-Teluride (HgCdTe) IR photo detector operated in a photo voltaic mode. The detector operates at liquid Nitrogen temperatures. The camera can detect temperature differences of approximately 0.09 °K, and has a relative accuracy of plus or minus 10% when detecting EM fields.

2.5 IR Images

The stored IR thermogram data represents the temperature distribution over the extent of the detector screen and is a map of the intensity of the electric or magnetic field distribution absorbed in the screen. For small temperature rises less than a few degrees above ambient, the electric and magnetic field intensities are nearly linearly proportional to the temperature change.

2.5.1 Spatial Resolution

The spatial resolution of the IR thermogram is a function of the number of pixel elements in the IR camera, and is fixed by the angular resolution of the telephoto, regular or wide angle lens used on the IR camera when making the IR image. A telephoto lens can be used to look at small details in the field structure on the detector screen; a wide angle lens can be used to look at large scale trends in the field structure in the detector screen.

The telephoto lens also has the added advantage for regular field mapping applications of allowing the IR camera to be located far away from the object under test and, thus, removing any perturbing effects that the metallic structure of the camera might have on the field distribution being measured.

2.5.2 Thermal Resolution

The thermal resolution of the IR thermogram is a function of the digitizer in the IR camera. The IR camera used to take the IR thermograms has a 12 bit digitizer. For a 12 bit digitizer, the temperature range seen by the IR camera is divided into 256 increments. Each digitized increment is assign a unique color, resulting in a temperature resolution of 256 color levels.

2.5.3 Thermogram Errors

The resulting IR image of an EM field depends on the combined EM and thermal properties of the detector material and is subject to several significant, but controllable, errors.

2.5.3.1 Lateral Conduction Effects

Conductive heating in the transverse direction within the screen material causes thermal "bleeding" from the hot spots on the screen to nearby cold spots. This thermal bleeding tends to fill in the nulls (minimums) somewhat, whereas, the areas of maximum heating (peaks) are not effected very much by this effect. This effect can be minimized by operating at small temperature variations above ambient.

2.5.3.2 Lateral Convective Effects

Convective heating of the top of the screen from heat rising from the bottom of the screen causes the top of the screen to appear slightly hotter than the bottom of the screen. This "blurring" of the image can be kept to a minimum by operating at small temperature variations above ambient.

This blurring effect can be eliminated completely by placing the IR detector screen in a horizontal position and observing the image with the IR camera looking down on the screen from above or from the side using an IR mirror.

2.5.4 IR Measurement Accuracy

The accuracy of the IR measurement technique was demonstrated by performing two simple experiments with known theoretical solutions. In one experiment, the diffraction pattern from a "Lloyd's Mirror" was measured. In the other experiment, the diffraction pattern from an "knife edge" conducting half-plane was measured.

In the Lloyd's mirror experiment, the resulting diffraction pattern is due to the antenna interfering with its image in a large ground plane. The near field (Fresnel Zone) antenna pattern of the horn antenna was used to obtain the theoretical results.

In both experiments, the screen material was optimized to measure only the tangential component of the electric field intensity in the plane of the screen.

These experiments were performed in an anechoic chamber. Good correlations between theory and experiment were obtained. The worst errors occurred in the minimums (deep nulls) of the diffraction patterns were thermal bleeding from the surrounding hot areas tended to obscure the real depth of the minimums. Some thermal bleeding out of the maximums into the surrounding areas also occurred, obscuring the real height of the maximums.

Even with conductive bleeding and convective blurring of the image, the measurement error is less than approximately 10% under controlled test conditions.

2.6 IR Advantages and Disadvantages

The IR measurement technique provides a quick and accurate method to observe EM fields in a two-dimensional plane. However, only the magnitude of the electric or magnetic field is measured; no phase information is detected. Also, since this technique is based on the thermal mass of the detector material, high energy is required to produce good thermal images of EM fields.

3. IR Thermograms

Examples of IR measurements taken this summer at Rome Laboratory are now presented.

3.1 Scattering from a Cylinder

Experimental tests were performed on a conducting cylinder irradiated by an incident plane wave. E- and H- field patterns of the scattered/diffracted energy from the cylinder where measured. Tests were conducted at various microwave frequencies relative to the resonant frequencies associated with the cylinder and at numerous angles of incident (eg. end-on, broad-side, oblique-incidence) and for several different polarizations of the incident field relative to the axis of the cylinder (eg. horizontal, vertical, and skewed). Each tests was performed in the near and far fields of the antenna. IR thermograms of the scattered fields were taken. The diffraction patterns of the electric field scattered from the cylinder are clearly indicated in these thermograms.

The equi-temperature contour levels in the IR thermograms are being compared to numerical predictions of the scattered electric field intensity. The numerical predictions are being performed by Dr. Zuffada at Caltech's Jet Propulsion Laboratory (JPL). These results will be used to verify JPL's numerical code which combines the Method of Moments (MoM) with a Finite Element Method (FEM) to determine electric field intensity levels.

3.2 Scattering from an F16 Scale Model Aircraft

The IR imaging technique was also used to map the scattered electric field around a simple scale model of an aircraft. A plastic F16 scale model was constructed and sprayed with several coats of silver paint to made it conductive. IR thermograms of the magnitude of the scattered electric field intensity were taken in the horizontal and vertical longitudinal planes through the fuselage of the aircraft. The diffraction patterns of the electric field scattered from the aircraft and the scattering centers are clearly indicated in these thermograms. Standing waves are setup between the incident

wave and the reflected wave off the aircraft. Surface waves are established along the side of the fuselage and the front edge of the wings between the nose of the aircraft and the missile rails. Diffraction occurs off the wing tips. A shadow zone appears behind the aircraft.

The equi-temperature contour levels in the IR thermograms are being compared to numerical predictions of the scattered electric field intensity. The numerical predictions are being performed with the GEMACS code at Rome Laboratory. These results will be used to validate the accuracy of the GEMACS numerical code.

4. Calibration

The IR measurement technique was calibrated this summer at Rome Laboratory.

4.1 Relative measurements

In the IR measurements described above, only the relative magnitude of the EM fields have been determined. The relative accuracies of the field intensities measured by the IR thermograms taken with the IR scanner were determined using numerical computer codes to predict the normalized field intensities for the same geometries. The relative accuracies were usually no greater than 10% in error. These errors occurred at the minimum (null) field intensity positions in the thermograms, where the effect of thermal conduction (bleeding) from adjacent hot spots tends to fill in the nulls somewhat. Overall, at positions away from nulls, the technique was closer to only 1% in error. Therefore, the technique has the potential to produce extremely accurate field intensity measurements, much less than $\pm$ 1 dB in error.

4.2 Absolute Measurements

To determine the absolute electric or magnetic field strengths from IR thermograms, the IR scanning detectors and detector screens must be calibrated.

Appropriate theoretical models and experimental practices are being developed to permit the absolute determination of electric and magnetic field intensities from thermographic images (IR thermograms) of microwave fields.

4.2.1 Theoretical calibration

A thermal/electromagnetic model is being developed for the interaction of microwaves with a

lossy/complex absorber material. The theory is based on the complex form of Poynting's Theorem for an absorbing material with complex constitutive parameters. The broad-band frequency dependence of the complex constitutive parameters of the IR detection screen material must be determined and used in the model of the detector screen along with its thermal properties.

4.2.2 Experimental calibration

The intensity levels (equi-color levels) of the IR thermograms can be empirically calibrated using standard gain horn antennas at several frequencies, angles of incidence, and polarizations in the near and far fields of the antenna. The predicted field level is simply associated with the resulting color level.

5. Results

The electric/magnetic field distributions can be predicted for a well-known, theoretically tractable problem geometry, e.g. scattering from a linear wire antenna, scattering from a thin rectangular slot aperture, antenna radiation in the near or far field of a rectangular horn antenna, induced cavity modes in a rectangular or cylindrical waveguide, etc. The scattering predictions can be made for a normally or obliquely incident plane wave as a function of frequency, angle of incidence, and horizonal/vertical polarization. These theoretical predictions can be verified experimentally to validate the IR/EM interaction model for each problem geometry.

The results of the initial calibration test for electric field measurements are presented for a lossy Kapton detector screen developed to measure the absolute magnitude of the electric field in the plane of the detector screen.

5.1 Experimental setup

As the first step in calibrating the detector material, an experiment was performed on a planar sheet of the detector material. In the experiment, the screen was positioned in front of a standard gain horn antenna and oriented in several different planes making different angles relative to the antenna. For one series of IR thermograms the screen was positioned in the axial plane in the near and far fields of the horn. For this case, the normal to the screen was oriented perpendicular to the direction of the incident radiation (perpendicular incidence). The camera was placed in front of the screen looking at

the thermal pattern on the screen. In other tests, the normal to the screen was rotated to different angles relative to the bore-sight of the horn. The frequency, angle of incidence, and polarization of the incident wave were varied.

5.2 Experimental data

The induced temperature distribution in the plane of the detector screen was mapped with the IR camera. The measured temperature on the bore-sight of the horn antenna was compared to the value of the incident field calculated using the Friis Transmitter/Receiver Formula with the known and/or measured characteristics of the antenna, corrected for all losses and mismatches. Forward and reverse power were measured with calibrated power meters. The distance to the screen from the phase center of the horn and the angle of the screen relative to the bore-sight of the horn antenna were measured for each configuration.

Experimentally obtained temperatures from the IR thermograms were plotted against the incident power densities for each case. A series of thermograms were taken in the 1 to 2 GHz range of frequencies at distances from 1 to 2 meters from the antenna. The background ambient temperatures for each case were also recorded. Many other configurations were tested at different distances and at various other frequencies, polarizations, and angles of incidence.

On the average, the corrected temperature differences between the various cases was less than $\pm$ 0.5 °F. This temperature error translates into a $\pm$ 1 dB field measurement error.

6. Conclusions

Initial empirical results from anechoic chamber tests indicate that the induced temperature distribution across the IR thermogram in a properly designed, lossy, planar IR detection screen placed in the area over which an EM field is to be mapped can be calibrated to measure the incident EM wave using the Joule heating of the screen material.

The IR measurement technique is, therefore, a viable method to aid in the determination of EM fields in various test situations. The IR method allows for rapid observation of EM field activity and interference, resulting in an in-depth understanding of EM scattering phenomena. An experimental technique is being used to measure and calibrate the absolute magnitude of the field intensity. A theoretical approach to this calibration problem is also being undertaken.

7. Future Work

The development of a suitable magnetic detection screen is in progress. The use of spin-on magnetic thin films is being investigated.

It is also possible to estimate the phase of the incident wave by making a numerical model of the thermal/electromagnetic interaction in the detector screen material, and to used energy minimization techniques to estimate the phase of the incident wave. Microwave holographic techniques can also be used to determine the phase of the wave incident on the IR detector screen if the magnitude of the incident electric field is measured at several different positions in the near field of the radiation source. This phaseless measurement work is being done in conjunction with Syracuse University (Prof. Tapan Sakar) and the National Institute of Standards and Technology (NIST) (Dr. Carl Stubenrauch).

8. Publications

The papers published on this project during the AFOSR Summer Research Program are listed in the Appendix.

Acknowledgement

The author would like to acknowledge the support of the AFOSR Summer Research Program on this IR project and the help of Michael Seifert at Rome Laboratory.

APPENDIX

A. Papers Presented:

1. J.D. Norgard, M.F. Seifert, and T. Pesta
   "Infrared Images of Electromagnetic Fields ... Relative and Absolute Calibration"
   Proceedings of the Dual-Use Technology Conference
   SUNY/Utica, NY
   May 1995.
2. J.D. Norgard and M.F. Seifert
   "Infrared Images of Electromagnetic Interference (EMI) Coupled through a Thin Slot Aperture in a Cylindrical Waveguide Cavity / Part I: Theory"
   Proceedings of the IEEE/EMC Conference
   Atlanta, GA
   August 1995.
3. J.D. Norgard and M.F. Seifert
   "Infrared Images of Electromagnetic Interference (EMI) Coupled through a Thin Slot Aperture in a Cylindrical Waveguide Cavity / Part II: Experiment"
   Proceedings of the IEEE/EMC Conference
   Atlanta, GA
   August 1995.
4. J.D. Norgard, R.M. Sega, and M.F. Seifert
   "Electromagnetic Field Measurements using Infrared Imaging Techniques"
   Proceedings of the ICEAA Conference
   Politecnico Di Torino
   Turin, Italy
   September 1995.

B. Seminars Presented:

1. "Calibration of Infrared Images of Radiated Electromagnetic Fields"
   US Air Force Academy
   Department of Electrical Engineering
   Colorado Springs, CO
   June 1995

2. "Infrared Images of Electromagnetic Fields"
   Norwegian Defense Research Establishment (NDRE)
   FFI/Geophysics
   Kjeller, Norway
   June 1995

C. *Journal Articles Published:*

1. "Measurement of Absolute Electromagnetic Field Magnitudes using Infrared Thermograms"
   EuroTherm Quantitative Infrared Measurement Journal

D. *Papers Submitted:*

1. "Measurement of the Relative Phase of Electromagnetic Fields using Infrared Thermograms"
   URSI Winter Meeting
   CU/Boulder, CO
   January 1996
2. "Code Validation of Cylindrical Scattering Parameters using IR Thermograms"
   ACES Symposium
   US NPS/Monterey, CA
   March 1996
3. "Microwave Holography using Infrared Images of Electromagnetic Fields"
   SPIE Conference
   Thermosense XVIII
   Orlando, FL
   May 1996

Flexible Adjustment of Data

Michael A. Pittarelli
Associate Professor
Department of Computer and Information Science

State University of New York
Utica, NY 13504-3050

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base
Washington, D.C.

and

Rome Laboratory

August 1995

FLEXIBLE ADJUSTMENT OF DATA

Michael A. Pittarelli
Associate Professor
Department of Computer and Information Science
SUNY Institute of Technology at Utica/Rome

Abstract

Databases may exhibit many forms of incompleteness. This report explores methods for overcoming incompleteness in the form of missing tuples. Specifically, algorithms are investigated for replacing a relation that is known to be incomplete with a superset. Of particular interest are algorithms that make available at any time an approximation and with the property that the approximate solution improves monotonically with computing time. How to measure the quality of an approximation depends on the use to which a relation is to be put. Although this report concentrates on rule matching in rule-based systems, application of these and similar methods to problems of knowledge discovery in incomplete databases is also discussed.

FLEXIBLE ADJUSTMENT OF DATA

Michael A. Pittarelli

The squirming facts exceed the squamous mind,
If one may say so. And yet relation appears,
A small relation expanding like the shade
Of a cloud on sand, a shape on the side of a hill.
[8, p. 215]

Database instances may exhibit many forms of incompleteness. The set of attributes may be too small to permit adequate characterization of the entities under consideration. Similarly, the set of values for one or more attributes may not be adequate. It may be necessary to replace {OK, Not_OK} with {Too_Low, Acceptable, Too_High}, for example. Some attribute values may be unknown for or inapplicable to some entities. Various extensions of the relational model have been proposed to deal with this last type of incompleteness [3,4]. We consider databases some tuples of which are believed to be missing in their entirety. The tuples may be absent due to a problem in transmitting a copy of a relation. They may be missing because data gathering has been terminated prematurely.

In particular, we explore methods for replacing with some superset a relation instance that is believed to be missing one or more whole tuples. How exactly to proceed depends on the use to which the data are to be put.

In all cases, a relation instance r is replaced with an r' such that $r \subseteq r' \subseteq dom(s(r))$, where: $s(r)$ is the *scheme* of r, the set of attributes on which r is defined; for any attribute $A \in s(r)$, $dom(A)$ is the set of possible values for A; for a set W of attributes (and a fixed but arbitrary linear ordering on the elements of W),

$$dom(W) = \underset{A \in W}{\times} dom(A).$$

A *database scheme* (or *structure*) is a set of relation schemes, i.e., a set of sets of attributes. The scheme for a relational database $D = \{r_1, \ldots, r_n\}$ is $s(D) = \{s(r_1), \ldots, s(r_n)\}$.

The *refinement* relation on database schemes will be important in what follows. A scheme X is a refinement of scheme Y (and Y is a *coarsening* of X), denoted $X \leq Y$, iff for every $V_x \in X$ there exists a $V_y \in Y$ such that $V_x \subseteq V_y$. X is an *immediate refinement* of Y iff $X \leq Y$, $X \neq Y$, and there is no $Z \neq X$ such that $X \leq Z$ and $Z \leq Y$.

Example: $\{\{A,B\},\{C\}\}$ is an immediate refinement of $\{\{A,B\},\{C,D\}\}$; $\{\{A\},\{B\}\}$ is a refinement, but not an immediate refinement, of $\{\{A,B\},\{C,D\}\}$.

Suppose tuple $t \in \mathrm{dom}(W)$ and that $V \subseteq W$. The *restriction* of t to V is the tuple $t[V] \in \mathrm{dom}(V)$ such that t agrees with $t[V]$ on all attributes in V. The *projection* of a relation r onto $V \subseteq s(r)$ is then

$$\pi_V(r) = \{t[V] \mid t \in r\}.$$

Example: The projection of

r	(A	B	C)
	a_1	b_1	c_2
	a_1	b_1	c_3
	a_1	b_2	c_2
	a_2	b_3	c_1
	a_3	b_3	c_3

onto {A,C} is the relation

$\pi_{\{A,C\}}(r)$	(A	C)
	a_1	c_2
	a_1	c_3
	a_2	c_1
	a_3	c_3

The projection of a relation onto a database scheme is a database, the set of projections of the relation onto the elements of the scheme:

$$\pi_X(r) = \{\pi_V(r) \mid V \in X\}.$$

An operation complementary to projection is *extension*. The extension of a relation r to $V \supseteq s(r)$ is

$$E^V(r) = \{r' \in \mathrm{dom}(V) \mid \pi_{s(r)}(r') = r\}.$$

Thus, the extension of a relation is a set of relations; i.e., a set of sets of tuples. The elements of $E^V(r)$ are partially ordered by the subset relation. The unique maximal element is the *cylindrical extension* of r, denoted $C^V(r)$:

$$\text{For all } r' \in E^V(r),\ r' \subseteq C^V(r) \in E^V(r).$$

Alternatively, with the usual relational algebra collapse of

$$\langle\langle a_1, \ldots, a_n \rangle, \langle b_1, \ldots, b_k \rangle\rangle$$

to

$$<a_1, \ldots, b_1, \ldots, b_k>$$

and reordering,

$$C^V(r) = r \times dom(V - s(r)).$$

The *natural join* (or *join*) of a database $D = \{r_1, \ldots, r_n\}$, denoted $J(D)$, is the intersection of the cylindrical extensions of its elements to the union of their relation schemes. Let

$$V = \cup_{r \in D} s(r).$$

Then

$$J(D) = \cap_{r \in D} C^V(r).$$

Theorem 1: $X \leq Y \leq \{s(r)\}$ implies $J(\pi_Y(r)) \subseteq J(\pi_X(r))$.

Proof: Suppose that $X \leq Y \leq \{s(r)\}$. For each $V_x \in X$, there exists a $V_y \in Y$ such that for every $t \in \pi_{V_x}(r)$ there exists a $t' \in \pi_{V_y}(r)$ such that $t = t'[V_x]$. Thus, for each $V_x \in X$, there exists a $V_y \in Y$ such that

$$C^{s(r)}(\pi_{V_y}(r)) \subseteq C^{s(r)}(\pi_{V_x}(r)).$$

Therefore,

$$\cap_{V_y \in Y} C^{s(r)}(\pi_{V_y}(r)) \subseteq \cap_{V_x \in X} C^{s(r)}(\pi_{V_x}(r));$$

i.e., $J(\pi_Y(r)) \subseteq J(\pi_X(r))$. □

Corollary 2: Let $X \leq \{s(r)\}$. Then $r \subseteq J(\pi_X(r))$. □

Consider production rules of the form

IF <database tuple> THEN <action>

Suppose that we have available to us (as a knowledge base) the relational database instance r, above, and that the tuples $a_2 b_3 c_3$ and $a_3 b_3 c_1$ should also be present, but are missing.

Suppose our rulebase consists of:

R1: IF $a_2 b_3 c_3$ THEN A1
R2: IF $a_2 b_2 c_2$ THEN A2
R3: IF $a_3 b_1 c_1$ THEN A3

None of the condition tuples appears in the relation r. Suppose that additional tuples are not forthcoming and that we must execute one of the three actions. How, in the absence of additional information, can we decide which one?

We could select a rule to fire at random. A more principled approach to rule selection utilizes Theorem 1 and Corollary 2: Find a refinement of $\{s(r)\}$ onto which to project r and

join. The resulting relation will be a superset of r and may contain one of the condition tuples. There remains the problem of identifying a suitable database scheme.

The schemes to be considered will be limited to those that are *irredundant covers* of $s(r)$. $X = \{V_1, \ldots, V_n\}$ is an irredundant cover of a set of attributes V iff

(1) $\cup_{V_i \in X} = V$

(2) $i \neq j \rightarrow V_i \not\subseteq V_j$.

Under the refinement relation on database schemes, the set of irredundant covers of any $s(r)$ is a lattice with maximum element $\{s(r)\}$ and minimum element $\{\{v\} | v \in s(r)\}$. A search for a suitable scheme may be conducted as a traversal of the lattice: start with $\{s(r)\}$, replace a scheme X with an immediate refinement Y, and terminate when projecting onto the current candidate scheme and joining yields one or more of the condition tuples (or when Y $= \{\{v\} | v \in s(r)\}$). Progress toward termination is made at each step, by Theorem 1 and the fact that $\{\{v\} | v \in s(r)\}$ is the minimum element of the lattice.

Unfortunately, the lattice of schemes is not a chain. Only $|s(r)|+1$ of the elements have a single immediate refinement. (When $|s(r)|$ is 4, for example, there are 114 irredundant covers.) As an alternative to breadth-first search, consider the algorithm below, where T denotes the set of condition tuples and $d(r, r')$ is the Hamming distance between the characteristic functions of r and r'. (Note that $d(r, r')$ is just the cardinality of the symmetric difference of r and r'; therefore, the domains of the attributes needn't be known to calculate d, and tuples that appear in neither r nor r' can be ignored.)

Algorithm 1.

Input: r and T. Output: Z.

$Z := \{s(r)\}$;
while $J(\pi_Z(r)) \cap T = \emptyset$ and $Z \neq \{\{v\} | v \in s(r)\}$ do
 begin
  $Ref := \{Y \mid Y$ is an immediate refinement of $Z\}$;
  $Ref' := Ref$;
  match:= false;
  while $Ref \neq \emptyset$ do
   begin
    $X :=$ select(Ref);

```
    Ref:= Ref−{X};
    if T∩J(π_X(r)) ≠ ∅ then
      begin
        if match = false
          then
            begin
              Z_least := X;
              match := true
            end
          else
            if |T∩J(π_X(r))| < |T∩J(π_{Z_least}(r))|
              then Z_least := X
      end
  end
 if match
   then Z:= Z_least
   else Z:= arg min_{X∈Ref'} d(r, J(π_X(r)))
end
```

The algorithm will output a structure Z such that exactly one of the following conditions is met:

(a) $Z = \{\{v\} | v \in s(r)\}$ and $T \cap J(\pi_Z(r)) = \emptyset$

(b)

$$T \cap J(\pi_Z(r)) \neq \emptyset;$$

for all proper aggregates X of Z visited,

$$T \cap J(\pi_X(r)) = \emptyset;$$

and, for all siblings Y of Z,

$$|T \cap J(\pi_Z(r))| < |T \cap J(\pi_Y(r))|.$$

Let $c = |T \cap J(\pi_Z(r))|$ at the termination of the algorithm. If $c = 1$, exactly one rule can be fired. If $c > 1$, then some way must be found to resolve the conflict among the rules that now can fire. If $c = 0$, more data must be obtained. In each of the latter two cases, there is nothing more that relational algebra can contribute.

The rationale for the use of a distance measure is twofold. First, it cuts down the size

of the search space. But the same reduction could be achieved by selecting an element of *Ref'* at random. Instead, the frontier element the projection onto which yields a relation closest on a reasonable metric to the original r is picked to refine further. We want to produce tuples that are not elements of the original relation r, but we don't want them produced at random. The assumption is that some tuples are missing and that they can be recovered by processing r appropriately. It is assumed that r is an approximation to a relation r^* and that r^* contains one or more condition tuples. For any scheme X, $J(\pi_X(r))$ is a candidate for the unknown r^*. By Theorem 1, for any immediate refinement Y of X, $J(\pi_Y(r))$ is more likely to contain condition tuples than is $J(\pi_X(r))$. However, the immediate refinements of X differ with respect to the quality of the approximation to r they offer. As a search heuristic, it is assumed that the better the approximation to r, the better the approximation to the unknown r^* (which r is assumed to resemble).

Definition: r is *reconstructable from* X iff

$$r = J(\pi_X(r)).$$

r is *approximately reconstructable from* X to degree ϵ iff

$$d(r,\ J(\pi_X(r))) \leq \epsilon.$$

A relation scheme satisfies the *join dependency* $*[X]$ iff each of its instances is reconstructable from X [9].

Join dependencies are implied by *functional dependencies* (but not conversely). Functional dependencies are more easily grasped intuitively than are join dependencies. An instance r satisfies a functional dependency from A to B (sets of attributes), denoted $A \rightarrow B$, iff, for any tuples t and t' in r,

$$t[A] = t'[A] \text{ implies } t[B] = t'[B].$$

Theorem 3: If r satisfies $A \rightarrow B$, then

$$r = J(\pi_{\{A \cup B, A \cup (s(r) - B)\}}(r)).$$

Proof:

(i) By Corollary 2,

$$r \subseteq J(\pi_{\{A \cup B, A \cup (s(r) - B)\}}(r)).$$

(ii) Suppose $t \in J(\pi_{\{A \cup B, A \cup (s(r) - B)\}}(r))$. Then

$$t \in C^{s(r)}(\pi_{A \cup B}(r)) \text{ and } t \in C^{s(r)}(\pi_{A \cup (s(r) - B)}(r)).$$

Then there exist t' and t'' in r such that

$$t'[A \cup B] = t[A \cup B]$$

and

$$t''[A \cup (s(r) - B)] = t[A \cup (s(r) - B)].$$

Then, since $A \rightarrow B$,

$$t'[B] = t''[B] = t[B].$$

Therefore, $t = t'' \in r$. So,

$$J(\pi_{\{A \cup B, A \cup (s(r) - B)\}}(r)) \subseteq r. \square$$

Any set of database schemes (e.g., the set of immediate refinements of a scheme X) partitions a set of relation instances into equivalence classes. The equivalence class associated with scheme Z is the set of relations for which Z is the best approximation (among all of the schemes under consideration, with ties broken arbitrarily). Scheme Z is a better approximation of r than is Y iff

$$d(r, J(\pi_Z(r))) \leq d(r, J(\pi_Y(r))).$$

A distance measure d is monotonic with respect to refinement iff $Y \leq X$ implies that X is a better approximation (with respect to d), for any r, than is Y. The Hamming distance between (characteristic functions of) relations is monotonic.

Theorem 4: $Y \leq X$ implies $d(r, J(\pi_X(r))) \leq d(r, J(\pi_Y(r)))$.

Proof: If d is Hamming distance between characteristic functions, then, for any relations r and r',

$$d(r, r') = d(r', r) = |\, r \oplus r' \,|,$$

where $\oplus$ denotes symmetric difference:

$$r \oplus r' = (r - r') \cup (r' - r).$$

Suppose $Y \leq X$. Then

$$\begin{aligned} d(r, J(\pi_X(r))) &= |\, r \oplus J(\pi_X(r)) \,| \\ &= |\, (r - J(\pi_X(r))) \cup (J(\pi_X(r)) - r) \,| \\ &= |\, J(\pi_X(r)) - r \,| \\ &\leq |\, J(\pi_Y(r)) - r \,| \\ &= d(r, J(\pi_Y(r))), \end{aligned}$$

since, by Corollary 2,

$$r \subseteq J(\pi_Z(r)),$$

and, by Theorem 1,

$$J(\pi_X(r)) \subseteq J(\pi_Y(r)). \ \square$$

So, with respect to Hamming distance, X is as good an approximation of r as are any of its immediate refinements. Thus, if r is reconstructable from Y and Y is an immediate refinement of X, then r is reconstructable from X also. And, if r is approximately reconstructable from Y to degree ϵ, it is also approximately reconstructable from X to degree ϵ, if $Y \leq X$.

Again, let T denote the set of condition tuples. We want to follow a path in the lattice of schemes from the root, $\{s(r)\}$, to a structure X such that:

(1) $r \neq J(\pi_X(r))$

(2) $d(r, J(\pi_X(r))) \leq \epsilon$

(3) $T \cap J(\pi_X(r)) \neq \emptyset$.

(4) For all $Y \neq X$ on the path from the root to X, $T \cap J(\pi_Y(r)) = \emptyset$.

There is no guarantee that any of these conditions except (2) can be satisfied (by making ϵ large enough). *A fortiori*, there is no guarantee that all four can be met simultaneously. Typically, however, they will be satisfiable.

The given relation r does not equal the unknown relation r^*. However, the assumption is that r resembles r^* closely enough so that, relative to a fixed set of structures (e.g., all those at the same path length from the root of the lattice of structures), r and r^* are in the same equivalence class: the best structure for r^* is also the best structure for r. Let X denote this structure. Ideally,

$$\text{(a) } r \neq J(\pi_X(r)),$$
$$\text{(b) } r^* = J(\pi_X(r^*)),$$

and

$$\text{(c) for all } V \in X,\ \pi_V(r) = \pi_V(r^*).$$

If these conditions hold, then

$$r^* = J(\pi_X(r)).$$

Conditions (a), (b) and (c) imply that the projections of r are better approximations of the projections of r^* than r is of r^*. If in fact $r \subseteq r^*$, then this is guaranteed to be the case:

Theorem 5: If $r \subseteq r^*$, then, for any $V \subseteq s(r^*) = s(r)$, $d(r, r^*) \geq d(\pi_V(r), \pi_V(r^*))$.

Proof: Suppose $r \subseteq r^*$. Then $\pi_V(r) \subseteq \pi_V(r^*)$. Then

$$d(r, r^*) = |r^* - r|$$

and

$$d(\pi_V(r),\ \pi_V(r^*)) = |\pi_V(r^*) - \pi_V(r)|.$$

Suppose $t_1 \in \pi_V(r^*) - \pi_V(r)$. Then there exists a $t_2 \in r^* - r$ such that $t_1 = t_2[V]$. Suppose $t_3 \in \pi_V(r^*) - \pi_V(r)$ and $t_3 \neq t_1$. Then there exists a $t_4 \in r^* - r$ such that $t_4[V] = t_3$ and

$t_4 \neq t_2$. So,

$$|r^* - r| \geq |\pi_V(r^*) - \pi_V(r)|,$$

which implies

$$d(r,\ r^*) \geq d(\pi_V(r),\ \pi_V(r^*)).\ \square$$

If condition (b) holds, and $X \neq \{s(r^*)\}$, then r^* embodies the sort of redundancy that causes the anomalies that normalization by decomposition is intended to eliminate. Of course, in the database design context, functional (and other) dependencies implying decomposability (i.e., join dependencies) are inferred from the "meanings" of the attributes, and hold for *any* instance over the given scheme. Here, we are testing directly for join dependencies (or approximate join dependencies) in a single relation instance.

Redundancy in r^*, coupled with the greater fidelity of $\pi_V(r)$ to $\pi_V(r^*)$ would explain the success of this technique. Further, there is good reason to expect decomposability (redundancy) in the systems that we actually encounter [7].

Let us demonstrate the search algorithm using the relation r and the three rules above. Recall that none of the three rule antecedents is contained in r. $s(r) = \{\{A, B, C\}\}$ and has one immediate refinement, namely $\{\{A, B\}, \{A, C\}, \{B, C\}\}$. The projection of r onto this structure is the database:

$\pi_{\{A,B\}}(r)$	$(A$	$B)$	$\pi_{\{A,C\}}(r)$	$(A$	$C)$	$\pi_{\{B,C\}}(r)$	$(B$	$C)$
	a_1	b_1		a_1	c_2		b_1	c_2
	a_1	b_2		a_1	c_3		b_1	c_3
	a_2	b_3		a_2	c_1		b_2	c_2
	a_3	b_3		a_3	c_3		b_2	c_2
							b_3	c_3

r equals the join of these projections. The structure $\{\{A, B\}, \{A, C\}, \{B, C\}\}$ has three immediate refinements: $\{\{A, B\}, \{A, C\}\}$, $\{\{A, C\}, \{B, C\}\}$, and $\{\{B, C\}, \{A, B\}\}$. The join of the projections onto $\{\{A, B\}, \{A, C\}\}$ adds a tuple, but it is not an antecedent. The structure $\{\{B, C\}, \{A, B\}\}$ adds two tuples, neither of which is an antecedent. However, one of the two extra tuples contained in the join of the projection of r onto $\{\{A, C\}, \{B, C\}\}$ is an antecedent:

$J(\pi_{\{\{A,C\},\{B,C\}\}}(r))$	(A	B	C)
	a_1	b_1	c_2
	a_1	b_1	c_3
	a_1	b_2	c_2
	a_2	b_3	c_1
	a_3	b_3	c_3
	a_2	b_3	c_3
	a_3	b_3	c_1

The search terminates, and rule R1 fires.

Each of the approximations of r generated using the method just discussed is a superset of r. However, there are more supersets $r' \supseteq r$, $r' \in dom(s(r))$, than there are refinements of $\{s(r)\}$. In the example above, suppose that

$$dom(A) = \{a_1, a_2, a_3\},$$
$$dom(B) = \{b_1, b_2, b_3\},$$
$$dom(C) = \{c_1, c_2, c_3\}.$$

Since $|r| = 5$ and $|dom(s(r))| = 27$, r has $2^{22}-1$ (proper) supersets. $\{s(r)\}$, on the other hand, has only 8 (proper) refinements. (The ratio of the number of refinements to the number of supersets increases with $|s(r)|$. Suppose $|dom(s(r))| = 4$, $|dom(A_i)| = 2$, and $|r| = 8$. Then r has $2^8-1 = 127$ proper supersets and $\{s(r)\}$ has 113 proper refinements.)

These superset approximations can be combined in various ways. The intersection of any collection of supersets of r, e.g.

$$\{J(\pi_X(r)) | X \text{ is at path length } k \text{ from } \{s(r)\}\}$$

is also a superset of r. Likewise, the union of the $J(\pi_X(r))$ is a superset of r. Furthermore, there may not exist a $Y \leq \{s(r)\}$ such that $J(\pi_Y(r))$ equals this union or the intersection.

Example: Suppose that the data are instead

$r^{\#}$	(A	B	C)
	a_1	b_1	c_1
	a_2	b_2	c_2
	a_3	b_3	c_3

Suppose $|dom(A_i)| = 3$. Then

$$J(\pi_{\{\{A,B\}\},\{C\}\}}(r^{\#})) = \pi_{\{A,B\}}(r^{\#}) \times dom(C),$$
$$J(\pi_{\{\{A,C\}\},\{B\}\}}(r^{\#})) = \pi_{\{A,C\}}(r^{\#}) \times dom(B),$$
$$J(\pi_{\{\{B,C\}\},\{A\}\}}(r^{\#})) = \pi_{\{B,C\}}(r^{\#}) \times dom(A).$$

The cardinality of each of these relations is 9. However, $r^{\#}$ is contained in each of them. Hence, the cardinality of their union is at most 24. Each of these structures has the same

immediate refinement, namely

$$\{\{A\}, \{B\}, \{C\}\},$$

the universal lower bound. The join of the projection onto this structure is

$$J(\pi_{\{\{A\},\{B\},\{C\}\}}(r^{\#})) = dom(A) \times dom(B) \times dom(C),$$

the cardinality of which is 27. So there is no structure in the lattice the join of the projection onto which equals this union.

Note, however, that if none of

$$J(\pi_{X_1}(r)), \ldots, J(\pi_{X_k}(r))$$

contains an element of T, then neither will their intersection or union. In other contexts (expanding a relation for purposes of detecting patterns), forming these unions and intersections may be useful.

The adjustment method discussed is flexible in that the number of refinement steps taken depends on the contents of T, the set of condition tuples. Let $Str(r, T)$ denote the structure identified by Algorithm 1 for relation r and set T. Let $Seq(r, T)$ denote the sequence of structures

$$<\{s(r)\}, \ldots, Str(r, T)>$$

generated by Algorithm 1 with r and T as input. Then, for any T and T', either $Seq(r, T)$ is a subsequence of $Seq(r, T')$ or conversely.

Algorithm 1 is not, however, flexible in the sense introduced by Horvitz [2]. The term 'flexible algorithm' is currently understood to mean an algorithm that provides approximate solutions to difficult problems in such a way that, minimally:

(1) an approximate solution is available at any point in the execution of the algorithm;

(2) the quality of the approximate solution increases with an increase in execution time.

Additional properties may include the following:

(3) preemptibility;

(4) continuity of the function from time to quality;

(5) diminishing marginal improvement of quality with an increase in computing time.

(See also Boddy and Dean [1], in which the term 'anytime' is used instead of 'flexible' to describe such algorithms.)

A tuple search procedure that is flexible in this sense begins with the projection of r, not onto an immediate refinement of $\{s(r)\}$, but instead onto $\{\{v\} | v \in s(r)\}$. Then, if

$$|T \cap J(\pi_X(r))| > 1,$$

and there is time to continue the search, r is projected onto immediate *aggregates* of the sibling(s) Z of X for which $d(r, J(\pi_Z(r)))$ is minimized.

Algorithm 2.

Input: r and T. Output: Z.

```
if T∩r = ∅
 then begin
  Z:={{v}|v∈s(r)};
  stick := false;
  while | T∩J(π_Z(r)) | > 1 and not stick do
   begin
    Agg:= {Y | Y is an immediate aggregate of Z};
    X_least := Z;
    done := false;
    while Agg ≠ ∅ and not done do
     begin
      X := select(Agg);
      Agg := Agg − {X};
      if |T∩J(π_X(r))| = 1
       then done := true
       else if 0 < |T∩J(π_X(r))| < |T∩J(π_X_least(r))|
          then X_least := X
     end;
    if done
     then Z := X
     else if Z = X_least then stick := true
        else Z := X_least
  end
 end
```

At any point in the execution of Algorithm 2, the current value of Z is the structure identified by the outer loop as generating the fewest condition tuples. By Theorem 1, successive values of Z generate progressively fewer condition tuples. The inner while loop of

Algorithm 2 identifies the structure among the current value of $\bar{Z}$ and its immediate refinements that generates the fewest condition tuples. It is possible for the current Z to generate more than one condition tuple, while each of its immediate refinements generates zero. In this case, the variable 'stick' is set to true. Ideally, one of the immediate refinements of the current Z generates exactly one tuple (in which case 'done' is set to true). In the worst case (if not interrupted and the outer while loop is entered), the algorithm terminates with Z equal to the single immediate refinement of $\{s(r)\}$, which consists of each of the $|s(r)|$ ($|s(r)|-1$)-element subsets of $s(r)$.

Similar methods may be used in other situations. We may wish to analyze the data in a relational database for patterns. For example, whenever Humidity $>$ 88 and Barom_Pressure $<$ 28.5 and Temp $>$ 75 (F), Rain = True. *Knowledge discovery in databases* is the term currently used for the subfield of machine learning concerned with the identification of such rules and patterns in databases [5]. Can a relation that is known to be incomplete be expanded by projection and join to yield a superset that exhibits more of the patterns in the real world entity that the database is intended to represent?

Suppose we decide to record which of the possible combinations of SEX, HEIGHT, and WEIGHT appear among the members of some population (e.g., residents of Oneida County, NY). Suppose we *discretize* HEIGHT and WEIGHT: we recognize only 3 values for HEIGHT (viz., tall, medium, short) and 3 for WEIGHT (heavy, medium, light). there are 262,144 different possible relation instances. Suppose the actual combinations are:

SEX	HEIGHT	WEIGHT
male	tall	heavy
male	medium	medium
female	tall	heavy
female	medium	medium

We give someone a summer grant to measure and record all of the combinations of values that appear in the population. By the end of the summer, the only combinations that have been observed are:

SEX	HEIGHT	WEIGHT
male	tall	heavy
female	medium	medium

The actual relation is the join of the projections of the observed relation onto the structure

{{SEX},{HEIGHT, WEIGHT}}.

SEX	HEIGHT	WEIGHT
male	tall	heavy
female	medium	medium

A search for an appropriate structure can proceed either by refinement, starting with the universal upper bound in the lattice of structures, or by aggregation, starting with the universal lower bound. Unlike the situation with rule-matching (or analogous situations in flexible adjustment of probabilities, in the context of a particular decision problem [6]), we need a problem-independent stopping criterion. For example, termination may occur when a relatively large increase in distance is detected in going from one level of the lattice to the next. (If refining, back up to the best structure at the previously visited level.)

Under what conditions do patterns appear in adjusted data that are not apparent in "raw" relational data? Intuitively, it seems that the adjusted relation needn't coincide exactly with the "actual" relation in order for interesting patterns in the adjusted relation to coincide exactly with the interesting patterns in the "actual" relation. Two objects (e.g., relations) may be equivalent with respect to important qualities without being identical in every respect. (For example, a crash-test dummy is equivalent to a human being for crash-testing purposes.) These and other questions (e.g., the accuracy, using simulated relational data, of the stopping criterion suggested above) remain for future research.

REFERENCES

[1] Boddy, M. and T. Dean, "Solving time-dependent planning problems," *IJCAI 89*, 1989.

[2] Horvitz, E., "Reasoning about beliefs and actions under computational resource constraints," *Proc. of the Third Workshop on Uncertainty in Artificial Intelligence*, 1987.

[3] Imielinski, T. and W. Lipski, "Incomplete information in relational databases," *JACM, v. 31*, pp. 761-791, 1984.

[4] Keller, A. and Winslett, M., "On the use of an extended relational model to handle changing incomplete information," *IEEE Trans. on Software Engineering, v. 11*, pp. 620-633, 1985.

[5] Pietetsky-Shapiro, G. and W. Frawley, *Knowledge Discovery in Databases*, MIT Press, 1991.

[6] Pittarelli, M., "An algebra for probabilistic databases," *IEEE Trans. on Knowledge and Data Engineering, v. 6*, pp. 293-303, 1994.

[7] Simon, H., "How complex are complex systems?" *PSA 76, v. 2*, Philosophy of Science Association, 1976.

[8] Stevens, W., *The Collected Poems of Wallace Stevens*, Knopf, New York, 1954.

[9] Ullman, J., *Principles of Database and Knowledge-base Systems, v. 2*, Computer Science Press, Rockville, MD, 1988.

Massala Reffell Report not available at time of publication.

A STUDY OF IMPEDANCE MATCHING OF MICROSTRIP PATCH ANTENNA ARRAYS OVER WIDE SCAN ANGLES

Sembiam R. Rengarajan
Professor
Department of Electrical and Computer Engineering

California State University
Northridge, CA 91330-8346

Final Report for:
Summer Faculty Research Program
Rome Laboratory/ERAA

Sponsored by:
Air Force Office of Scientific Research
Hanscom Air Force Base, MA

and

Rome Laboratory

August 1995

A STUDY OF IMPEDANCE MATCHING OF MICROSTRIP PATCH ANTENNA ARRAYS OVER WIDE SCAN ANGLES

Sembiam R. Rengarajan
Professor
Department of Electrical and Computer Engineering
California State University
Northridge, CA 91330-8346

Abstract

The concept of impedance matching of an infinite phased-array antenna over wide scan angles by means of a coupling circuit in the feed network was originally proposed by Hannan et al. To the best of our knowledge this concept was not implemented in practice. We investigated the feasibility of implementing this technique for microstrip patch arrays. Computed values of input impedance as a function of scan angles in E-plane, H-plane, and D-plane (45°) were used as input to an optimization program, which minimized the reflection coefficient magnitude over a range of scan angles. The output of the program are the line lengths and coupling susceptances in the E-plane and in H-plane. We designed coupling networks for matching three microstrip patch array geometries over wide scan angles. Preliminary studies show that it is possible to realize such coupling networks in microstrip-type transmission medium.

A STUDY OF IMPEDANCE MATCHING OF MICROSTRIP PATCH ANTENNA ARRAYS OVER WIDE SCAN ANGLES

Sembiam R. Rengarajan

Introduction

An excellent review of different methods of matching phased-array antennas over wide scan angles was presented by Knittel [1]. Among other methods, the design of a coupling circuit in the feed network [2,3] has the advantage that the array geometry and the element may be designed for optimum radiation pattern performance with the array elements matched only at broadside. Hannan et al. assumed that the input impedance of an infinite planar array of dipoles as a function of scan angles in the E-plane, H-plane, and D-plane (45°) were the same as the corresponding values measured for the central element of a 7x9 array [3]. They obtained theoretical values for the line lengths and coupling susceptance in the E-plane and that in the H-plane, in order to match the infinite array of dipoles. To the best of our knowledge this concept has not been implemented experimentally.

Pozar states that there is a need for matching phased arrays of microstrip patch elements over wide scan angles [4]. Although this has been addressed by Herd [5] by a proper design of the physical parameters of the element and the unit cell size for an infinite array, there are two reasons for investigating the coupling circuit concept for microstrip arrays in this work. In microstrip patch arrays, a coupling circuit may be conveniently implemented in the same layer as that of the patches if there is enough real estate. Otherwise the coupling network may be printed in another layer with a "tile" architecture for the array, which is common these days. A second advantage of having a coupling network for matching an array over wide scan angles is that the element parameters and the unit cell size may be optimized for radiation pattern performance with the elements matched only at broadside, without any consideration on how the scan impedance varies. It may be pointed out that if the array has a scan blindness angle, we can only match to scan angles less than the blindness angle.

Methodology

Three patch array geometries were considered for this study. The first two arrays were those studied by Pozar and their scan reflection coefficients are presented in Figs. 3 and 8 in [4] respectively along with their physical parameters, are reproduced in Fig. 1. Unfortunately the phase values are not available in [4]. Therefore the scan reflection coefficient magnitude and phase were recomputed using the computer program developed by Herd for an infinite array of stacked patch antennas [5]. The third array shown in Fig. 2 has a stacked-patch geometry and it consists of a superconducting layer at the bottom (ε_r=24, h=0.0169 λ_0) with a square conducting patch of side 0.096 λ_0, with an air gap of h=0.03336 λ_0 above, followed by another dielectric layer (ε_r=3.8, h=0.0667 λ_0) with a square patch of side 0.2325 λ_0 located at the lower surface of the top dielectric layer. Note that ε_r is the relative permittivity, h is the thickness of the dielectric layer and λ_0 is the free space wavelength. An inset feed centered at one edge of the lower patch and 0.015 λ_0 from the edge is used. A square unit cell of size 0.4 λ_0 is assumed for the infinite array. The scan reflection coefficients for this were also computed using the same code [5].

The design of the coupling impedance follows along the lines of the procedure given by Hannan et al [3]. In [3] wide angle match was accomplished by considering 6 parameters for an infinite array. First a length of transmission line l_a is introduced behind the element port where the scan impedance is known. Then a coupling susceptance b_e connected between adjacent elements in the E-plane is determined. Subsequently another line length l_b and coupling susceptance b_h connected between adjacent elements in the H-plane is determined. Finally a third line length l_c and a susceptance b are connected at the input of each feed element to achieve a match. We found that it is not necessary to have the last two parameters to achieve a match over wide scan angles. A good match is achievable with only four parameters l_a, b_e, l_b, and b_h whereas if we use three parameters l_a, b_e, and b_h the input reflection coefficient is found to be slightly higher. Fig. 3 shows a schematic representation of the coupling network in E-plane. A similar arrangement may exist in the H-plane. The design

of the coupling network is treated as a four (or three) parameter optimization problem where the objective is to minimize the cost function. The cost function is specified as the sum of the squares of the reflection coefficient magnitudes. A variation of the definition of cost function in terms of fourth powers of reflection coefficient magnitude etc. were experimented with to put a greater penalty on high values of the reflection coefficient magnitude. The minimization is carried out in different directions using numerically evaluated gradients until the solution converged.

Table 1 shows the reflection coefficient magnitude and phase (Γ) for different scan angles in E, H, and D planes for the first array (see Fig.1). Tables 2 and 3 show the four parameters of the coupling network and the resulting Γ for a maximum scan angle of 75° and 70° respectively. In the first case the maximum voltage standing wave ratio S_{max} before and after introducing the matching network are 6.7 and 1.86 respectively. When we consider a maximum scan angle of 70° the corresponding figures are 4.7 and 1.58 respectively. Table 4 shows similar results for a maximum scan angle of 70° with only three parameters for the coupling circuit. Here S_{max} is 1.78 after introducing the matching network, a value slightly greater than the corresponding figure in Table 3.

Table 5 shows the reflection coefficient magnitudes and phases of the second microstrip array. We note that there is a good match in the H-plane. It is sufficient to have only two parameters for the coupling circuit, l_a , and b_e. Table 6 shows these parameters and Γ in the E-plane and D-plane after introducing the matching network. A maximum scan angle of 35° is assumed for matching. The values of S_{max} before and after matching are 9 and 2.1 respectively. The reflection coefficient magnitudes in the H-plane are unaffected by this matching network.

A study of the third array showed that if the maximum scan angle is 70°, S_{max} before and after introducing the 4-parameters matching network are 6.1 and 1.8 respectively. The latter figure becomes 3.2 if we have only three parameters in the coupling circuit. However, if the maximum scan angle is only 60°, S_{max} before and after introducing the 4-parameter matching

network are 4.3 and 1.4 respectively. The latter figure becomes 2.05 if we have only 3 parameters in the coupling network.

Realization of the coupling circuit

Capacitive and inductive susceptances required in the coupling circuit may be realized in a microstrip transmission medium by printed structures such as interdigital capacitors and spiral inductors in rectangular or circular configuration. In order to realize a series inductor there is a need to do a wire-bonding between one of the inductor terminals and a microstrip line. Computer aided designs (CAD) of these elements are possible with packages such as [6]. Initial studies with [6] showed that *equivalent lumped* coupling susceptance values obtained in our study of three microstrip arrays may be realized. Further work is required by including the complete scattering parameters of such elements and the connecting transmission lines. We need to study the antenna input matching by the required coupling circuit as well as the realization of coupling elements together as an optimization problem.

Conclusion

We have revisited Hannan's approach to matching a phased array over wide scan angles by means of a coupling circuit and shown that it is suitable for infinite arrays of microstrip patch elements. In this initial study we assumed that the coupling circuit consists of equivalent lumped capacitive or inductive suscepance. Further work is required to match the array scan impedance by considering the two-port scattering matrix parameters of the coupling circuit with the use of CAD programs for such elements. It would be desirable to implement this for a linear array or for a planar array such as the second array studied in this work. Further work on the design of coupling circuits for elements near the edge of a large array where the environment is quite different from an infinite array will be interesting.

Acknowledgment

The author wishes to express his appreciation to the staff of Rome Laboratory/ERAA for all their help and assistance during his stay at the laboratory. In particular, Drs. M. Davidovitz, J.

Herd, R. Mailloux, and H. Steyskal are gratefully acknowledged for technical discussions and help in carrying out this study.

References

[1] G.H. Knittel, "Wide angle impedance matching of phased-array antennas - a survey of theory and practice," in Digest of Phased Array Antenna Symposium (A. Oliner and G. Knittel, Eds.), 1970, pp. 62-65.

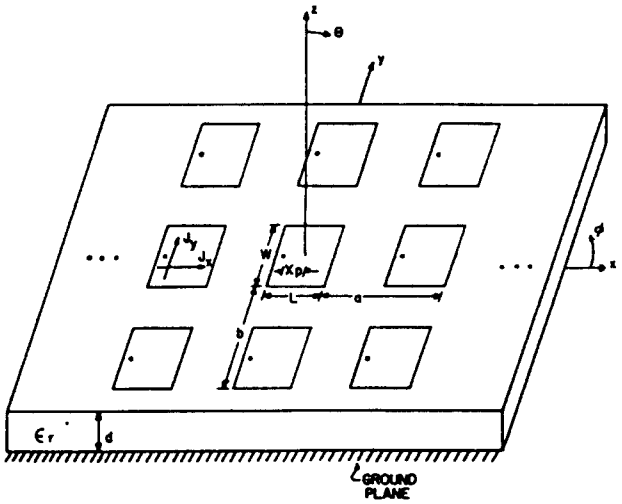
[2] N. Amitay, "Improvement of planar array match by compensation through contiguous element coupling," IEEE Transactions on Antennas and Propagation, vol. AP-14, no. 5, pp. 580-586, Sep. 1966.

[3] P.W. Hannan, D.S. Lerner, and G.H. Knittel, "Impedance matching a phased-array antenna over wide scan angles by connecting circuits," IEEE Transactions on Antennas and Propagation, vol. AP-13, no. 1, pp. 28-34, Jan. 1965.

[4] D. Pozar, "Analysis of an infinite array of rectangular microstrip patches with idealized probe feeds," IEEE Transactions on Antennas and Propagation, vol. AP- 32, no. 10, pp. 1101-1107, Oct. 1984.

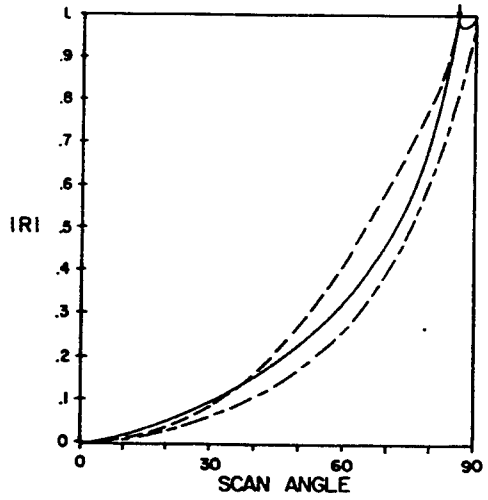
[5] J. Herd, "Full wave analysis of proximity coupled rectangular microstrip antenna arrays," Electromagnetics, vol. 11, pp. 21-46, Jan. 1991.

[6] Hewlett Packard Company, Microwave and RF Design Systems, HP 85150D, Santa Rosa, CA, May 1994.



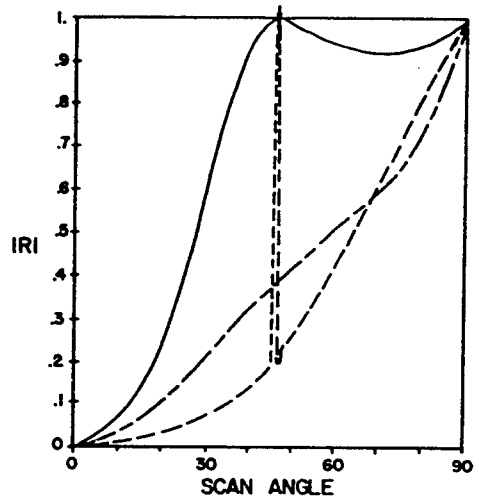
. Geometry of an infinite array of microstrip patches. The ground plane is in the $z = 0$ plane and the top of the substrate is in the $z = d$ plane.

(a) Geometry



Reflection coefficient magnitude of an infinite microstrip array.

(b) Parameters and Γ for array 1 from fig. 3 of [4]



Reflection coefficient magnitude of an infinite microstrip array.

(c)) Parameters and Γ for array 2 from fig. 8 of [

Fig.1 Geometry and parameters of patch arrays 1 and 2

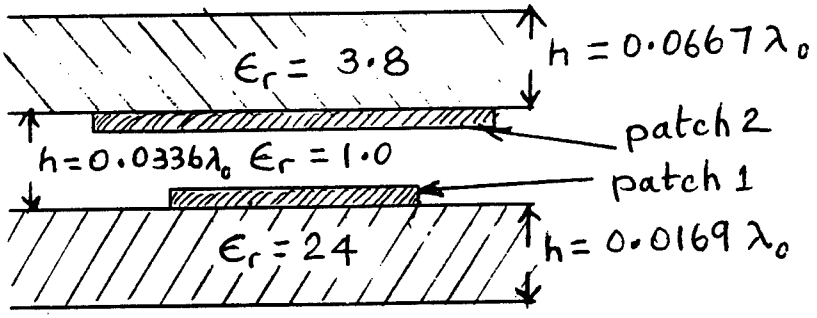


Fig. 2 Geometry of the patch array 3

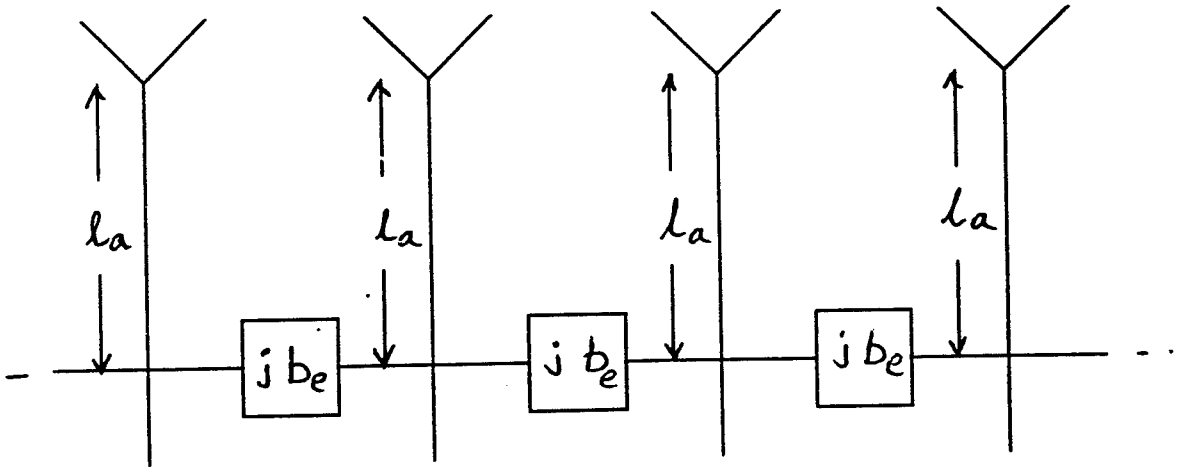


Fig. 3 Coupling network shown in the E-plane

Scan Angle (deg.)	E-Plane Reflection Coefficient		D-Plane Reflection Coefficient		H-Plane Reflection Coefficient	
	magnitude	phase (deg.)	magnitude	phase (deg.)	magnitude	phase (deg.)
0.0	0.0	0.0	0.0	0.0	0.0	0.0
5.0	0.002	-57.57	0.003	-49.33	0.003	-43.44
10.0	0.009	-58.27	0.011	-49.56	0.013	-43.92
15.0	0.020	-59.69	0.024	-50.12	0.029	-44.69
20.0	0.036	-61.76	0.042	-50.93	0.051	-45.77
25.0	0.056	-64.52	0.065	-52.04	0.081	-47.16
30.0	0.079	-68.05	0.091	-53.46	0.117	-48.88
35.0	0.105	-72.42	0.121	-55.27	0.160	-50.92
40.0	0.134	-77.75	0.153	-57.51	0.210	-53.29
45.0	0.166	-84.15	0.188	-60.25	0.268	-56.00
50.0	0.201	-91.76	0.225	-63.56	0.332	-59.05
55.0	0.239	-100.70	0.266	-67.46	0.403	-62.43
60.0	0.282	-111.07	0.311	-71.89	0.480	-66.13
65.0	0.334	-122.89	0.366	-76.70	0.562	-70.13
70.0	0.401	-136.02	0.435	-81.54	0.649	-74.40
75.0	0.500	-150.09	0.526	-85.92	0.739	-78.90

Table 1 Reflection coefficient as a function of scan angle for the microstrip patch array 1 (before matching) .

Scan Angle (deg.)	E-Plane Reflection Coefficient		D-Plane Reflection Coefficient		H-Plane Reflection Coefficient	
	magnitude	phase (deg.)	magnitude	phase (deg.)	magnitude	phase (deg.)
0.0	0.0	0.0	0.0	0.0	0.0	0.0
5.0	0.007	-41.05	0.003	42.80	0.010	90.95
10.0	0.026	-44.71	0.013	46.37	0.037	92.50
15.0	0.055	-47.24	0.030	48.73	0.077	94.81
20.0	0.091	-51.15	0.051	51.13	0.126	97.57
25.0	0.128	-55.47	0.075	53.67	0.174	100.26
30.0	0.163	-60.00	0.100	56.91	0.217	102.60
35.0	0.191	-64.92	0.124	60.06	0.248	104.18
40.0	0.210	-70.37	0.145	63.49	0.264	104.68
45.0	0.216	-76.60	0.161	66.86	0.262	103.63
50.0	0.209	-84.08	0.168	70.54	0.242	100.70
55.0	0.186	-93.76	0.164	74.74	0.200	94.36
60.0	0.146	-108.96	0.147	80.85	0.138	79.81
65.0	0.097	-143.18	0.113	92.41	0.078	30.12
70.0	0.109	144.95	0.071	128.99	0.144	-44.07
75.0	0.249	105.11	0.110	-158.96	0.308	-69.28

Table 2 Reflection coefficient as a function of scan angle for the microstrip patch array 1 after introducing the coupling network for matching upto 75 deg. with 4 parameters: l_a=0.159 λ_0, b_e normalized = 0.162, l_b=0.407 λ_0, b_h normalized = -0.340)

Scan Angle (deg.)	E-Plane Reflection Coefficient		D-Plane Reflection Coefficient		H-Plane Reflection Coefficient	
	magnitude	phase (deg.)	magnitude	phase (deg.)	magnitude	phase (deg.)
0.0	0.0	0.0	0.0	0.0	0.0	0.0
5.0	0.005	-31.38	0.003	45.13	0.008	91.00
10.0	0.020	-35.69	0.012	48.66	0.029	92.28
15.0	0.043	-38.08	0.027	50.86	0.062	94.10
20.0	0.070	-42.08	0.046	53.03	0.100	96.29
25.0	0.098	-46.50	0.067	55.33	0.138	98.32
30.0	0.123	-51.20	0.089	58.35	0.170	100.02
35.0	0.142	-56.54	0.109	61.36	0.191	100.95
40.0	0.152	-62.85	0.126	64.80	0.199	100.77
45.0	0.151	-70.64	0.136	68.44	0.188	98.86
50.0	0.138	-81.05	0.139	72.91	0.159	94.46
55.0	0.112	-97.00	0.132	78.97	0.110	83.75
60.0	0.081	-129.59	0.114	89.40	0.052	40.81
65.0	0.087	167.99	0.088	112.46	0.095	-49.29
70.0	0.172	126.65	0.088	162.13	0.225	-72.75

Table 3 Reflection coefficient as a function of scan angle for the microstrip patch array 1 after introducing the coupling network for matching upto 70 deg. with 4 parameters: l_a=0.180 λ_0, b_e normalized = 0.135, l_b=0.387 λ_0, b_h normalized = -0.290)

Scan Angle (deg.)	E-Plane Reflection Coefficient		D-Plane Reflection Coefficient		H-Plane Reflection Coefficient	
	magnitude	phase (deg.)	magnitude	phase (deg.)	magnitude	phase (deg.)
0.0	0.0	0.0	0.0	0.0	0.0	0.0
5.0	0.003	-117.81	0.006	-114.35	0.010	-107.21
10.0	0.013	-123.57	0.025	-113.72	0.037	-110.84
15.0	0.027	-125.70	0.054	-115.40	0.077	-114.01
20.0	0.043	-129.93	0.090	-118.17	0.125	-117.99
25.0	0.058	-134.49	0.128	-121.60	0.172	-123.18
30.0	0.069	-139.27	0.165	-124.79	0.211	-128.43
35.0	0.073	-145.03	0.197	-128.16	0.239	-134.08
40.0	0.068	-152.75	0.222	-131.08	0.251	-140.30
45.0	0.052	-165.20	0.236	-133.70	0.248	-147.92
50.0	0.028	161.12	0.239	-135.59	0.230	-157.84
55.0	0.035	64.31	0.228	-136.65	0.200	-173.10
60.0	0.092	31.50	0.202	-135.87	0.173	161.34
65.0	0.175	16.65	0.157	-131.58	0.183	124.54
70.0	0.284	4.75	0.094	-112.97	0.260	91.75

Table 4 Reflection coefficient as a function of scan angle for the microstrip patch array 1 after introducing the coupling network for matching upto 70 deg. with 3 parameters: l_a=0.221 λ_0, b_e normalized = 0.114, b_h normalized = 0.280)

Scan Angle (deg.)	E-Plane Reflection Coefficient		D-plane Reflection Coefficient	
	magnitude	phase (deg.)	magnitude	phase (deg.)
0.0	0.0	0.0	0.0	0.0
5.0	0.012	-115.38	0.006	-113.95
10.0	0.050	-117.76	0.026	-115.03
15.0	0.118	-122.03	0.057	-116.77
20.0	0.225	-128.76	0.098	-119.07
25.0	0.381	-138.76	0.148	-121.79
30.0	0.584	-152.83	0.203	-124.75
35.0	0.797	-170.58	0.260	-127.71

Table 5 Reflection coefficient as a function of scan angle for the microstrip patch array 2 (before matching)

Scan Angle (deg.)	E-Plane Reflection Coefficient		D-plane Reflection Coefficient	
	magnitude	phase (deg.)	magnitude	phase (deg.)
0.0	0.0	0.0	0.0	0.0
5.0	0.027	-117.84	0.014	-116.68
10.0	0.101	-124.58	0.054	-121.64
15.0	0.199	-134.66	0.113	-125.82
20.0	0.288	-148.20	0.181	-131.00
25.0	0.327	-167.16	0.250	-136.85
30.0	0.264	153.88	0.309	-142.50
35.0	0.355	24.08	0.356	-147.66

Table 6 Reflection coefficient as a function of scan angle for the microstrip patch array 2 after introducing the coupling network with only 2 parameters: l_s=0.091 λ_0, b_e normalized = 0.659)

VERTICAL-CAVITY SURFACE-EMITTING LASERS FOR "SMART PIXEL" ARRAYS AND OPTOELECTRONIC INTERCONNECTS

Dean Richardson
Assistant Professor and Photonics Program Coordinator
Department of Electrical Engineering Technology

State University of New York Institute of Technology at Utica/Rome
P.O. Box 3050
Utica, NY 13504-3050

Final Report for:
Summer Faculty Research Program
Rome Laboratory Photonics Center

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory Photonics Center

October 1995

VERTICAL-CAVITY SURFACE-EMITTING LASERS FOR "SMART PIXEL" ARRAYS AND OPTOELECTRONIC INTERCONNECTS

Dean Richardson
Assistant Professor and Photonics Program Coordinator
Department of Electrical Engineering Technology
State University of New York Institute of Technology at Utica/Rome

Abstract

Vertical-Cavity Surface Emitting Lasers (VCSELs) for "Smart Pixels" and optoelectronic interconnects have been studied under the RDL Summer Faculty Research Program. The objective was to provide reliable and robust processing recipes through iterative fabrication/characterization cycles. A working 4x4 VCSEL array has been demonstrated.

1 Introduction

Semiconductor diode lasers are key components in a vast range of applications such as photonic switching, optical communication and optical interconnects. A revolutionary form of these devices, the VCSEL, is currently undergoing extensive study. VCSELs have the advantages of high packing density with narrow optical beam[1] and low threshold current.[2] The desirable round geometry of the devices produces circular radiation patterns which are superior for fiber coupling over lateral lasers. The circular beam may also be better suited for vertical optical interconnections. The circular beam is much more compact than the LEDs used in previous vertical optical interconnect studies.[3] Higher packing density can result from the narrower beam diameter. The small size also facilitates 2-D addressable array fabrication for parallel communications. VCSELs are also more easily manufactured than present commercial laser diodes since they are amenable to wafer scale fabrication and testing.

To this point, much research effort has been concentrated on novel designs and processing approaches for VCSELs, since their power conversion efficiency is currently much lower than that of conventional lasers. VCSEL devices integrated with novel optoelectronics have been also been investigated, but the lengthy and sophisticated processing procedures necessary to produce efficient VCSELs has so far limited the success of these integration efforts due to the daunting complexity of the resulting heterostructures. The aim of the research project summarized here was to continue efforts to develop reliable, robust, and relatively simple fabrication recipes for individually addressable VCSEL arrays to allow practical integration with other optoelectronics.

2 Overview of Work Performed

The following tasks were accomplished during my term under the 1995 Summer Faculty Research Program:

- an MBE growth recipe for InGaAs-based 980-nm VCSEL arrays was formulated, based on previous UCSB-generated designs.

- A VCSEL mask set for device isolation, contacting, and related processing was created.

- A VCSEL wafer supplied by PRI, Inc. was processed into devices and device arrays using photolithography, ECR etching, chemical etching, metallization, and other relevant techniques.

- Optical and electronic properties of the as-grown devices were characterized

- Code for locating VCSEL cavity resonances was developed; code for designing DBR mirrors was obtained from other sources; and extensive work on a model treating gain, threshold behavior, and strain effects in VCSELs was completed.

In the sections that follow, we review these accomplishments. Section three discusses our theoretical modeling efforts; section four briefly summarizes device fabrication work performed; section five presents device testing results; and section six contains our conclusions.

3 Theoretical Modeling

Our modeling efforts focused on predicting gain and threshold behavior in 980 nm strained InGaAs VCSELs. Since we had earlier encountered surprising hurdles in obtaining CRAY cpu time needed for a full many-body treatment of strained InGaAs gain dynamics, it was decided to pursue the conceptually simpler Fermi-golden rule approach.[4] This allowed modeling efforts to coalesce around the UCSB-developed treatment of Scott Corzine et al..[5] This approach offered the advantage of providing a more intuitively clear connection to electrical pumping properties like threshold current, slope efficiency, etc. than more physics-oriented many-body techniques. The Corzine approach to modeling key VCSEL behavior is diagrammed schematically in Fig. 1.

This approach has four basic steps. First and foremost, the subband structure for the quantum-well gain region must be determined. This is the most involved step from the standpoint of theoretical complexity and computational effort. Once this has been accurately determined, a loop is entered in which optical gain and current density are calculated as a function of carrier density. During each pass through the loop, three steps are accomplished in a self-consistent fashion: the quasi-Fermi levels are determined for a given N and P; the gain corresponding to these quasi-Fermi levels is calculated, and the current density required to maintain the quasi-Fermi level separation is solved for.

Next to the subband structure determination, finding the spectral gain is the most complicated step in the Corzine approach. The gain is defined as the product of three terms: the band-to-band interaction strength, the squared field strength of one photon, and the number of interacting transition pairs. The second term is a function of the waveguiding and confinement characteristics. The first and third terms, which primarily depend on evaluating fermi functions and quasi fermi levels, rely fundamentally on knowledge of the subband structure in the material. This is also the point at which the effects of strain are most accurately treated.

Because of its crucial importance, we tackled the subband structure first in our modeling efforts. Extensive effort was devoted to developing a Luettinger-Hamiltonian-based model allowing for the calculation of InGaAs valence and conduction subband structure for both strained and unstrained materials. As it turned out, in the limited time frame available, this was as far as we got. In fact, we were only successful in developing code which predicted the subband structure of unstrained materials. This code has yet to be incorporated within the larger framework required to predict overall device gain and threshold current. It is likely that another 3-6 months of effort will be required to produce the working optimization code originally envisioned.

3 Theoretical Modeling (cont.)

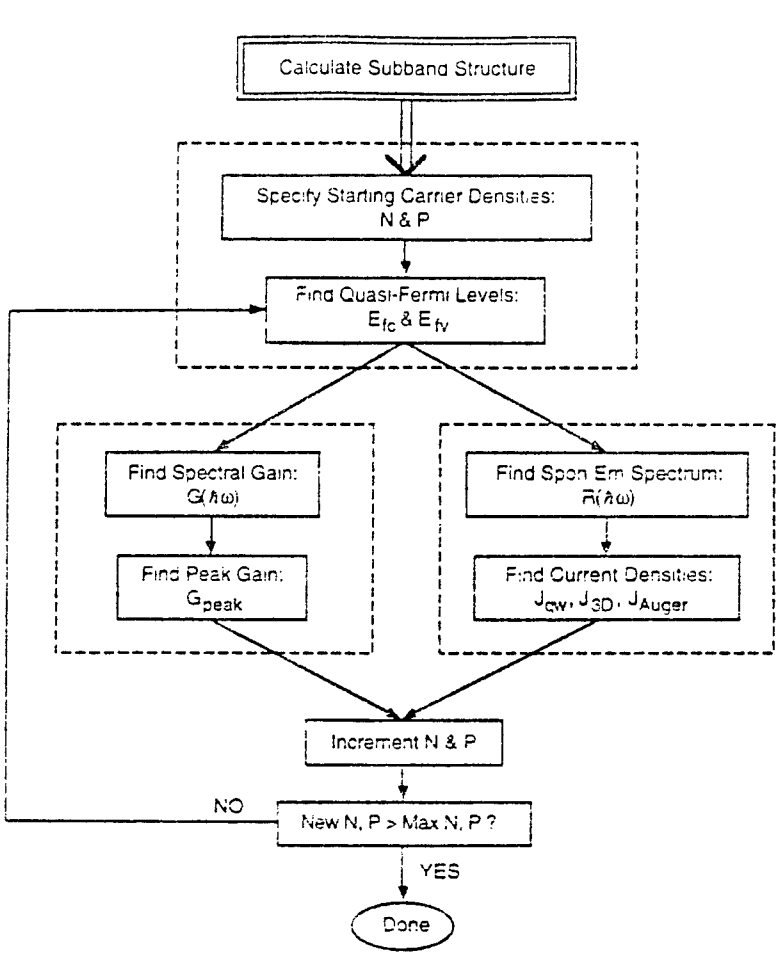


Figure 1 Flow chart for QW gain calculations developed by S. Corzine et al..

The program immediately came up against accuracy problems and round-off error effects in computing the matrix determinants needed to evaluate the various subbands. We were thus forced to import "canned" LAPACK routines to perform these operations. As of yet, stability problems in the combined routines have not been completely overcome. The code calculates subband structure, but the accuracy is questionable far from the center of the 1st Brillouin zone (i.e., far from k=0). Since these regions are crucial for strain-related effects, extensive additional work will be necessary to resolve these issues.

3 Theoretical Modeling (cont.)

A basic model of VCSEL facet reflectivity was developed for use in DBR mirror stack design. Additionally, a copy of the Randy Geels' original MultiLayer code written for the Macintosh was also obtained from Dr. Geels for comparison with the DBR code we developed. Unfortunately, the Geels program lacks the capability to treat material absorption, which is definitely relevant in obtaining accurate mirror designs.

To summarize our theoretical efforts, we made strong progress in developing a flexible and device-oriented model suitable for realistic evaluation and optimization of gain and threshold behavior in 980-nm strained InGaAs VCSEL devices. However, additional effort will be required during the next phase of the project to bring the model to completion. Several models of DBR mirror performance and associated cavity resonances were developed and/or acquired. The resulting suite of theoretical tools should put us in excellent shape to tackle design and characterization of the next generation of devices in future contracts.

4 Fabrication

The VCSEL wafer used in this work was supplied by PRI of Longmont, Colorado. The wafer was grown by molecular beam epitaxy on a GaAs buffer layer. The bottom mirror is Si doped AlAs/AlGaAs with 29 periods. The active region consists of 0.19 μm $Al_{0.25}Ga_{0.75}As$ confinement layers surrounding a gain region with two 80 Å $Al_{0.25}Ga_{0.75}As$ quantum wells (QW) and 80 Å GaAs barriers. The p-type of mirror is capped with a 250 Å p^+ GaAs contact layer.

There are quite a few different fabrication approaches used in making VCSELs. One relatively simple scheme requiring comparatively few processing steps is the "etched pillar" VCSEL design, similar to those used in research groups at UCSB and elsewhere.[6] Fabrication typically involves several different steps of masking, deposition, and etching as illustrated in Figure 2. Step 1 is used to define etch-post and etch masks. Once the desired thickness of photoresist is spun on and developed, the wafer is ready to be etched by electron cyclotron resonance (ECR) in step 2. The etched mesas basically serve to define the current injection path, and the etching depth is just below the active region. Then polyimide is used in step 3 to passivate the side wall surface of the VCSEL etched posts, and polyimide is baked for a few cycles. Polyimide which covers the top of the mesas will be removed by reactive ion etching (RIE) since the laser emission is from the top. In step 4, a second mask is patterned on the etched posts for the p-metal deposition. Metal deposition uses standard lift-off techniques. Finally, n-type ohmic contact material is evaporated. The detailed process we used to fabricate VCSELs is given in Appendix I.

4 Fabrication (cont.)

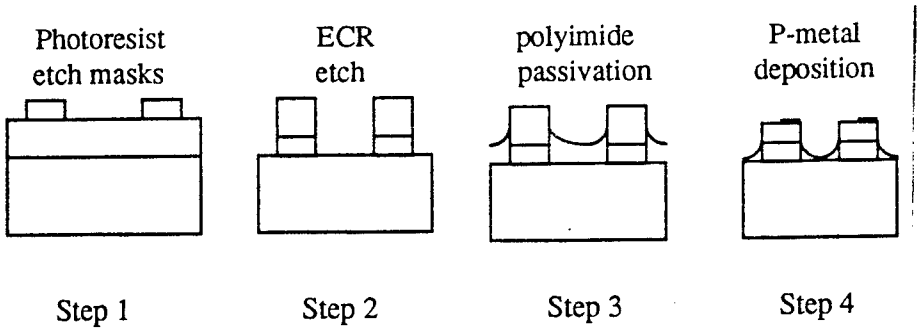


Figure 2 Schematic of etched-pillar VCSEL processing flow

The process described above utilizes dry etching; wet etching has been also carried out for comparison. For wet etching, step 1 is the same as described in Figure 2. In step 2, a mixed solution of 3:1;50 of H_3PO_4:H_2O_2;H_2O has been used for etching. However, the etching rate is difficult to control and the resulting mesa surfaces are very rough.

In our work, VCSELs were fabricated using the processes described above into both 3x4 and 4x4 arrays. Each VCSEL is designed to be of a different diameter ranging between 5 and 30 μm so that the VCSELs' performance can be compared. The pitch (laser to laser spacing) is also varied from 10 to 40 μm for comparison. A scanning electron microscope (SEM) micrograph of the completed arrays is shown in Figure 3.

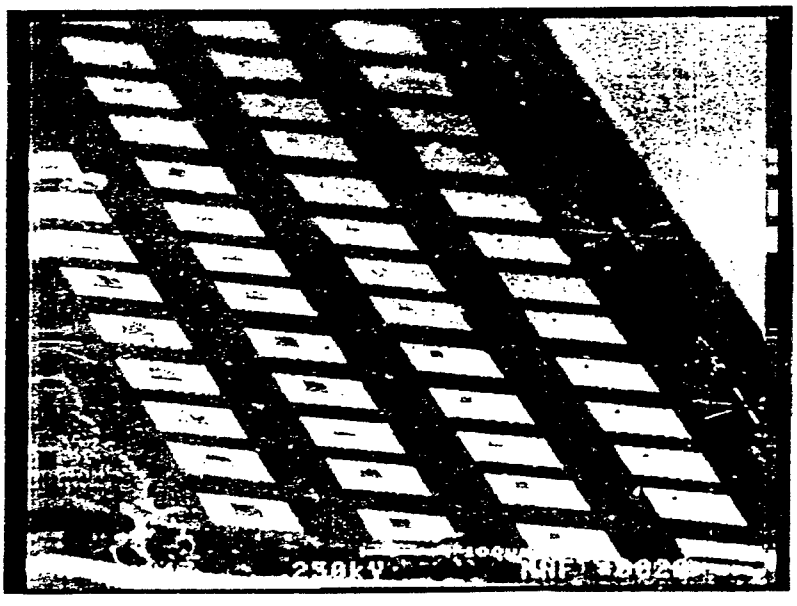


Figure 3 SEM image of completed VCSEL array

5 Device Testing

The testing of the VCSELs arrays was carried out at the Rome Laboratory Photonics Center.[7] A schematic of the probe station set-up used to perform I-V measurements is shown in Figure 4. A Cambridge Instruments MicroZoom II is the central part of the probe station, and a CCD camera and monitor are connected to the microscope. The microscope is mounted on a large X-Y translation stage and a laboratory jack. The devices to be tested are electrically mounted on copper blocks using silver epoxy to maintain good common ground and heat sinking properties. The devices have X-Y translation, two-axis tilt, and rotation for proper positioning. Electrical probing is accomplished using Alessi high-frequency and DC probes mounted on X-Y-Z translation stages.

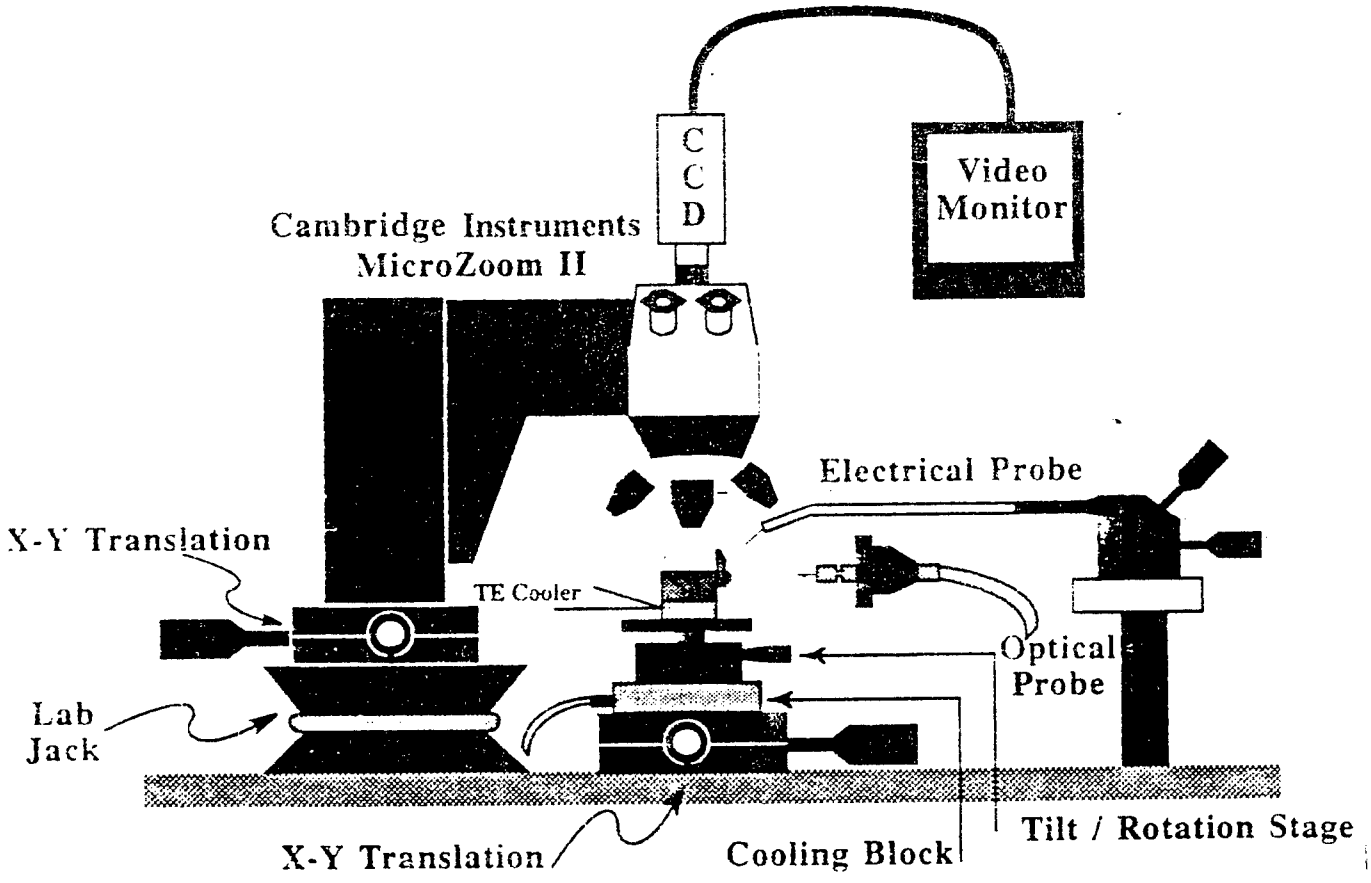


Figure 4 Probe Station set-up

5 Device Testing (cont.)

The purpose of the electrical probing is to supply the necessary bias and drive signals to the devices and to pick off the electrical signals at the desired points. The tip of the probe is of great concern when contacting the devices. The Alessi tungsten probe tips usually make the better electrical contact with the devices, provide a more stable electrical connection and have a lower contact resistance. An electrical probe contact indicator is made to ensure the physical contact between the tip and the device.

Electronically, the device is driven by the HP 8116A Pulse/Function Generator (PFG). The PFG supplies either triangular or sine pulses of the desired frequency and amplitude. The HP 54111D Digital Storage Oscilloscope (DSO) measures the I-V and L-I curves of the devices. A typical example of I-V characteristics of individual VCSELs tested is shown in Figure 5.

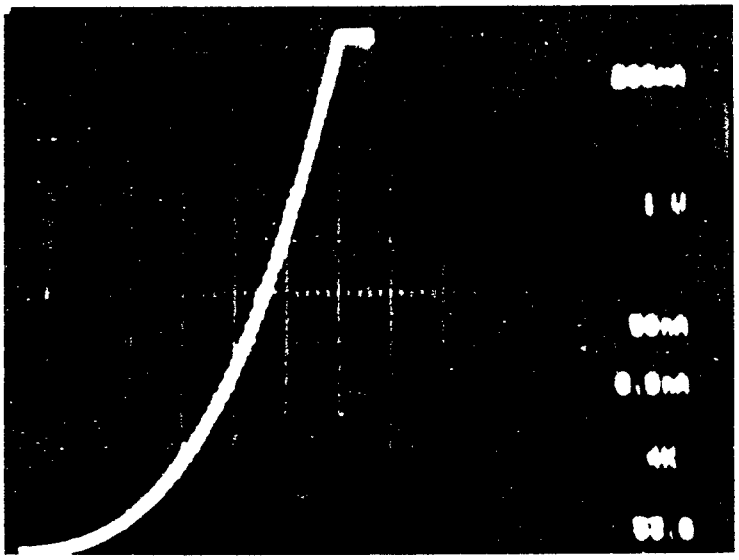


Figure 5 Typical I-V characteristic of individual VCSEL device.

6 Conclusions

We have demontrated the production of individually-addressable AlGaAs/GaAs VCSEL arrays. The fabrication procedure has been simplified and contains only two levels of masks. The quality of the side walls of the VCSELs etched by ECR is good. As a comparison, wet etching produces less desirable device quality. The reduced complexity of the VCSEL array fabrication scheme bodes well for the use of VCSELs as "Smart Pixels" in novel optoelectronic interconnect applications.

7 References

1. J.L Jewell, A. Scherer, S.L. McCall, Y.H. Lee, S. Walker, J.P. Harbison, and L.T. Florez, "Low threshold electrically pumped vertical-cavity surface emitting microlasers," *Electron. Lett.* **25,** 1213 (1989).

2. R.S. Geels, S.W. Corzinne, J.W. Scott, D.B. Young, and L.A. Coldren, "Low threshold planarized vertical cavity surface emitting lasers", *IEEE Photon. Tech. Lett.*, **2,** 234 (1990).

3. H.F. Bare, F. Haas, D.A. Honey, D. Mikolas, H.G. Craighead, G. Pugh and R. Soave, "A simple surface-emitting LED array useful for developing free-space optical interconnects", *IEEE Photon. Tech. Lett.*, **5,** 172 (1993).

4. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*. Wiley, New York (1991).

5. S. W. Corzine, "Design of vertical-cavity surface-emitting lasers with strained and unstrained quantum well active regions," Ph.D. thesis, University of California at Santa Barbara (1993).

6. M.G. Peters, D.B. Young, F.H. Peters, B.J. Thibeault, J.W. Scott, S.W. Corzine, R.W. herrick, and L.A. Coldren, *SPIE Proceedings, Vertical-Cavity Surface Emitting Laser Arrays,* **V-2147,** edited by Jack Jewell, Los Angeles, CA (1994) and references therein.

7. Michael A. Parker, Paul D. Swanson, Stuart I. Libby, " Photonic devices and systems for optical signal processing", RL-TR-93-157, August 1993.

Appendix I Processing Log

1 Wafer clean
aceton, methanol, isopropanol and DI water
ammonia hydroproxide
vapour prime 30 min.

2 Zero level mask
170C hot plate 10-15 minutes prior to photoresist (PR)
Shipley primer 3000rpm 10sec.
CS50 6000rpm 30sec.
pre-bake 90C 2 min.
expose on 10:1 stepper 1.3sec. for positve tone
post-bake 90C 1min.
OCG 945 40-60sec.

3 Mesa definition
glue chips to 3" Si wafer by FSC-M, bake 1 hour @ 90C
ECR setting: 10 Torr He @ 0C chamber temperature, 16 sccm Cl2, 4 sccm BCL3, 2 mT chamber pressure, 16 A upper, 20 A lower magnet current, 400 W microwave, 80 W RF
pre-run of ECR 300-500sec.
1165 30min. to strip PR

4 Polyimide (PI) passivation
Q3289:QZ3290=1:9 for PI adhesion
spin 20sec. @ 5000rpm
287 PI 20sec. @ 5000rpm
post-bake 90C 30min., 120C 30min., 150C 15min., 180C 15min., 240C 30min.
strip of PI at the top of mesas in RIE 8min.

5 p-type metallization
Shipley primer 3000rpm 10sec.
CS50 6000rpm 30sec.
pre-bake 90C 2 min.
expose on 10:1 stepper 0.6sec. for image reversal
Yes oven 85min.
HTG flood 60 sec.
OCG 945 4min.

Appendix I Processing Log (cont.)

6 p-type metal evaporation
low power O2 descum 1min.
ammonia hydroproxide dip 10sec
CHA evaporator
400A Ti, 200A Pt, 3000A Au, 3000A Cr, 1500A Ni
soak in 1165 >45min for lift off

7 n-type metallization
lap 2μm off from substrate
CHA evaporator 100A Ni, 400A Ge, 800A Au, 500A Ag, 700A Au
300 RTA

IMAGE EXPLOITATION: WAVELETS RESEARCH AND IMAGERY TOOLKIT DEVELOPMENT

Frank Y. Shih
Associate Professor
Department of Computer and Information Science

New Jersey Institute of Technology
Newark, NJ 07102

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D. C.

July 1995

IMAGE EXPLOITATION: WAVELETS RESEARCH AND IMAGERY TOOLKIT DEVELOPMENT

Frank Y. Shih
Associate Professor
Department of Computer and Information Science
New Jersey Institute of Technology

Abstract

The exploitation of digital reconnaissance imagery requires efficient storage, transmission, image processing, browsing, retrieval and analysis of large volumes of image data. An image compression scheme for multiresolution coding by wavelets to alleviate the transmission problem of high resolution images over low speed channels is developed. This scheme has the advantage of progressive transmission such that it can transmit the lowest resolution image to the decoder and the finer details are transmitted progressively. A block variance detector on the lowest subband to predict if the high subbands contain edge information is used to achieve efficient compression. The morphological close-minus-open operation is used for target recognition to eliminate background clutter whose size is larger than the targets. A macro wavelet function with a Gaussian low-pass filter is used to clutter with particle sizes similar to or smaller than the targets. The graphical user interface of an imagery toolkit is developed under X window in SUN Solaris system. It includes the often-used techniques in image processing such as image manipulation, enhancement, feature extraction, filtering, degradation, statistics, segmentation, and morphology.

IMAGE EXPLOITATION: WAVELETS RESEARCH AND IMAGERY TOOLKIT DEVELOPMENT

Frank Y. Shih

1. Introduction

The introduction of wavelets in signal and image analysis was a major breakthrough because of their ability to describe a signal or image in time and in frequency simultaneously, thus overcoming classical limitations of Fourier analysis. If tracing back, some of the pioneering work of wavelet transform was introduced by Haar (1910) [6], Gabor (1946) [5], and Morlet (1982) [10]. As it came into the attention of other scientists it was recognized to be useful for other signal analysis applications. Nowadays, there exist more than ten widely-used wavelet transform algorithms (see [4], [6], [9], [14]).

Like sines and cosines in Fourier analysis, wavelets are used as basis functions in representing other functions by decomposition. These wavelet functions are generated by dilations and translations of a single basis function (called *the mother wavelet*). The Haar wavelet has been used in various mathematical fields. It is well known that any continuous function can be approximated uniformly by Haar functions. Dilations and translations of the mother wavelet ψ,

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k), \tag{1}$$

where j and k are integers, define an orthogonal basis in $L^2(R)$ (the space of all square integrable functions). The simplest of all wavelets is the Haar wavelet $\psi(x)$. It is a step function taking values 1 and -1 on [0, 1/2) and [1/2, 1], respectively.

The 2-D wavelet transform of an input image $I(x, y)$ is expressed as

$$W(a_x, a_y, b_x, b_y) = \frac{1}{\sqrt{a_x a_y}} \iint I(x, y)\, \psi\left(\frac{x - b_x}{a_x}, \frac{y - b_y}{a_y}\right) dx\, dy, \tag{2}$$

where (a_x, a_y) are dilation parameters in x and y directions, and (b_x, b_y) are translation parameters that apply the wavelet filter function ψ to different local regions of the scene.

Wavelet transform methods for image data produce a multiresolution representation in which the spatial structure of the image is preserved within every level. The quadrature mirror filters were developed for 1-D and 2-D wavelet decomposition and reconstruction in Mallat's algorithms [8]. The filtering analyzes each dimension of the input data into two subbands: the first consists of the average information of the signal (low frequency components) and the second contains the detailed information (high frequency components). The low frequency information can be considered as the approximation of the image and the high frequency information

as the details lost at this step of the approximation process.

Let $\phi(x)$ is a 1-D low-pass filter and $\psi(x)$ is the mother wavelet function (a high-pass filter) associated with the scaling function. In two dimensions, the 2-D wavelet transform can be equivalently computed with a separable extension of the 1-D decomposition algorithm [8]. For a matrix of 2-D image data $I(x, y)$ of dimension $2^n \times 2^n$, where n is an integer, we first convolve the rows of $I(x, y)$ with the 1-D wavelet filter and two matrices of dimension $2^n \times 2^{n-1}$ are obtained as shown in Fig. 1, where $\phi(x)$ and $\psi(x)$ are a low-pass filter and a high-pass filter, respectively.

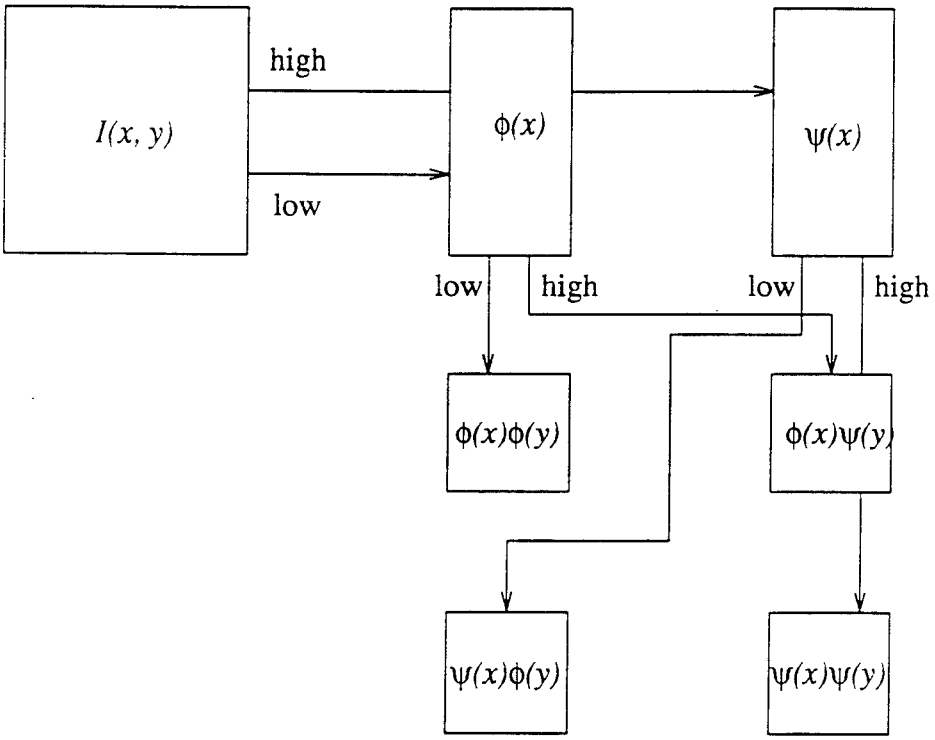


Fig. 1. The 2-D image wavelet transform by a separable 1-D decomposition algorithm.

We then convolve the columns of the resulting two matrices with the 1-D wavelet filter and the four resulting matrices $\phi(x)\phi(y)$, $\phi(x)\psi(y)$, $\psi(x)\phi(y)$, and $\psi(x)\psi(y)$ of dimension $2^{n-1} \times 2^{n-1}$ are obtained. The matrix $\phi(x)\phi(y)$ is a half-resolution down-sampled image of low frequencies (referred to as *the thumbnail image*), the image $\phi(x)\psi(y)$ gives the vertical high frequencies (horizontal edges) and $\psi(x)\phi(y)$ the horizontal high frequencies (vertical edges) and $\psi(x)\psi(y)$ the high frequencies in both directions (the corners as well as diagonal edges).

The filtering is repeated successively on the first, low-frequency subband, producing a pyramid of subbands. Part of the rationale for focusing on the low-frequency subband is that

most of the energy of the image is contained in that portion of the data. The output image of the wavelet representation is arranged as shown in Fig. 2, where the top-left quadrant image is the output for the wavelet filter $\phi(x)\phi(y)$ (the low-frequency components), the top-right for $\phi(x)\psi(y)$ (the horizontal edges), the bottom-left for $\psi(x)\phi(y)$ (the vertical edges), and the bottom-right for $\psi(x)\psi(y)$ (the diagonal edges). Note that the subscript index indicates the level of resolution. The wavelet transform procedure can be recursively repeated on the thumbnail image until a 2×2 image is occurred or a predefined criterion is satisfied. The criterion for continuing the decomposition can be based on the minimization of a cost function, such as total energy of the image representation [3]. Fig. 3 shows the original Lenna image and its three level decomposition of the wavelet representation.

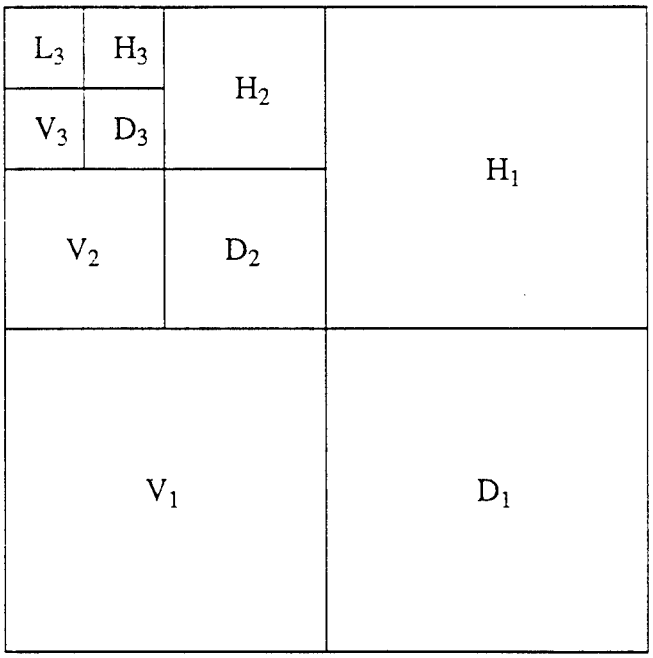
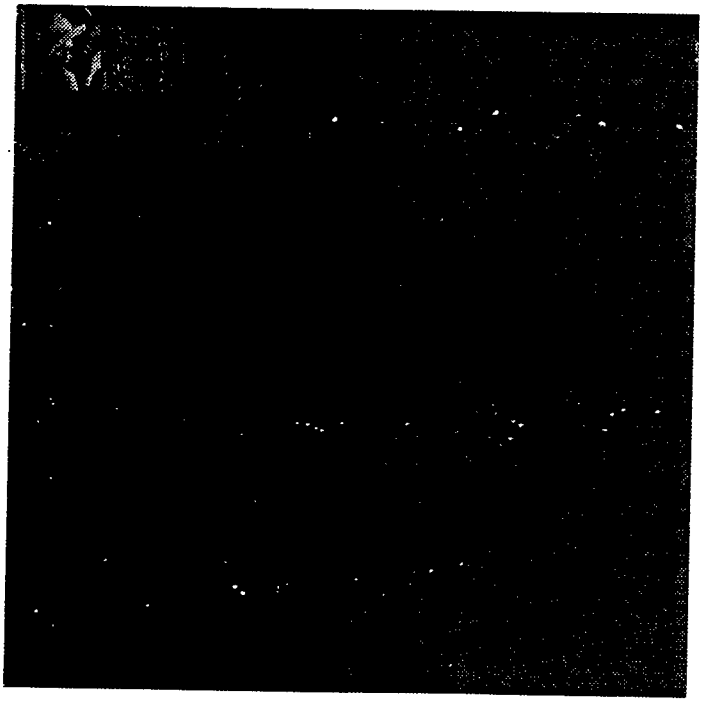


Fig. 2. Three level resolutions in a 2-D wavelet transform.

2. Multiresolution Decomposition and Progressive Transmission

Our approach for multiresolution decomposition is described as follows. At the encoder, the input image is decomposed into several frequency subbands. The lowest subband is encoded by Differential Pulse Code Modulation (DPCM) [16] or some other loseless coding algorithms. The other high subbands are partitioned into blocks to be vector quantized. Not all blocks in high subbands are needed to be vector quantized. As we can expect, the probability density function (PDF) of the high subbands are closely approximated by a generalized Gaussian distribution

(a)



(b)

Fig. 3 (a) Original Lenna image, (b) Wavelet representation on three level resolutions.

with $\rho = 0.7$ [1]. That is, there are few nonzero values which contain the information of edge in high subbands. In Fig. 3b, most of nonzero values almost concentrate on the edge of original image. From [1], the variance of the real PDF of high subbands and PDF after discarding the smooth corresponding blocks in high subbands are almost equal. Here we refer the corresponding blocks in high subbands are blocks in high subband whose positions are corresponding to the block in the lowest subband.

Transmission of high resolution images over low speed channels always presents a problem, owing to the lengthy transmission time. For instance, transmitting a picture of 512×512 pixels at 8 bits/pixel over a 9600-baud line can take 4 minutes. An approach to alleviate the problem is the multiresolution coding scheme which decomposes the signal into the lowest resolution and the difference between the approximation of original signals at each successive resolution. This scheme has the advantage of progressive transmission [12]. That is, it can transmit the lowest resolution signal to the decoder, and the finer details are transmitted progressively. This property is a very useful technique for image database query. We can determine that the image is needed or not at the first glance into the lowest resolution image. It can save the transmission time and increase the bandwidth of a network.

We apply a block variance detector on the lowest subband to predict the corresponding blocks in high subbands whether they are high variation or not. If the block in the lowest subband is determined to be smooth, then we think the corresponding blocks in high subbands contain all zero values. Therefore, the corresponding blocks in high subband are discarded. Otherwise, the corresponding blocks are vector quantized and transmitted to decoder.

When the decoder receives the lowest subband completely, it applies the same block variance detector as the encoder's on the lowest subband. If the block in the lowest subband is determined to be high variation, then the incoming index to look up codebooks is obtained for patching the codevectors on the corresponding blocks in high subbands. Otherwise, the decoder pads zero values on the corresponding blocks in high subbands.

The variance detector of a block to be smooth or high variation is defined as

$$\sigma = \left[\sum_{m=0}^{M} \sum_{n=0}^{N} \frac{(x(m, n) - \bar{x}(m, n))^2}{M \times N} \right]^{1/2}, \tag{3}$$

where the $M \times N$ is the blocksize, the $\bar{x}$ is the mean of this block. If σ is greater than the predefined threshold, then the block is determined to be high variation. Otherwise, it is considered as

smooth. The scheme can effectively reduce bitrate without degrading image quality visually.

3. Scanning Pattern Techniques

A scanning pattern approach refers to any technique that transmits transform coefficients in the order following a certain scanning pattern through the 2-D transform coefficients. Initially, a small number of coefficients, determined by the scanning pattern, from each block are transmitted to the receiver to generate the first approximation of the image. In each pass of subsequent transmission, a group of coefficients, also determined by the scanning pattern, from each block are transmitted to the receiver to refine an existing reconstruction.

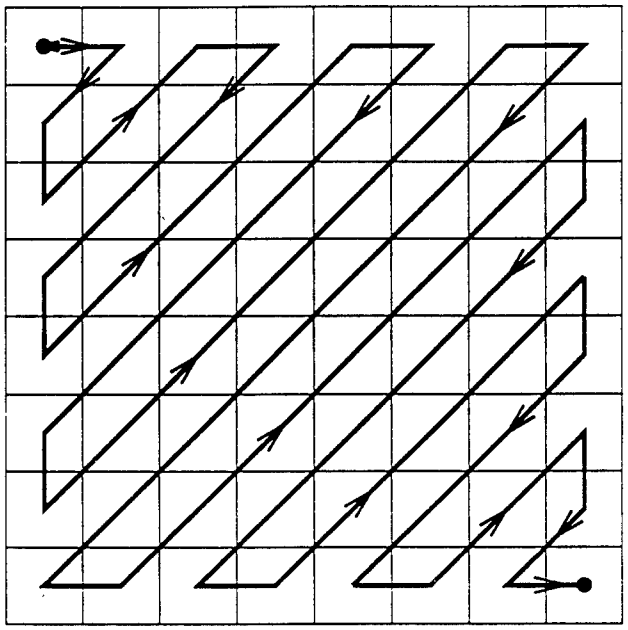
Differences among variations of this approach are generally in the scanning patterns. The optimal scanning pattern should be the one that sends coefficients in the descending order of their variance values since this will minimize the mean-squared distortion due to untransmitted transform coefficients. The zigzag pattern is an often-used scanning pattern in threshold transform coding [2]. This pattern usually generates more consecutive zeros for coefficients under the threshold, and efficient run-length codes can be employed. An example of the zigzag pattern for 8×8 discrete cosine transform (DCT) coefficients is shown in Fig. 4a.

We observe that if the ordering of the transform coefficients are carried out exactly in a zigzag manner, the resulting bit-stream does not have a property which can provide a low complexity multiresolution previewing. We modify the zigzag scanning strategy as shown in Fig. 4b such that the ordering of the transform coefficients follow the priority order of the rectangular zones. That is, the order of the sequence is almost zigzag within each subblock of size 1×1, 2×2, 4×4, etc.

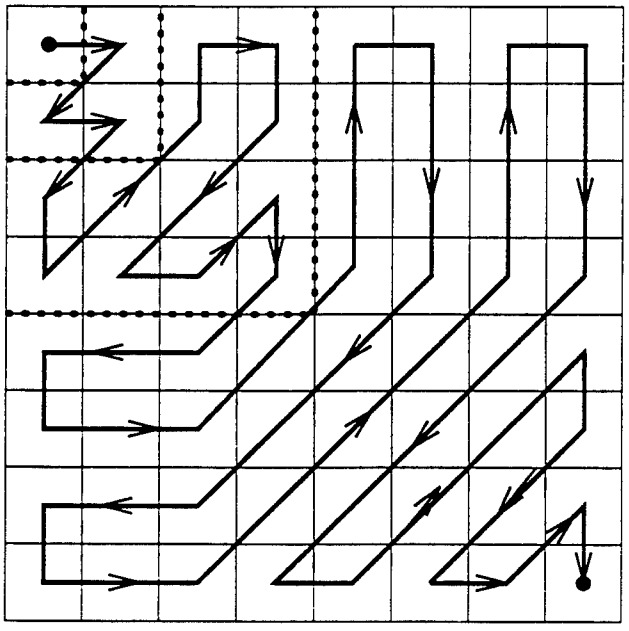
4. Target Recognition

Morphological processing and wavelet transforms are combined together to detect multiple object in a high clutter scene where multiple classes of objects with distortions and contrast variations are present. Mathematical morphology [7], [11] obtains its name from the studies of shape analysis. Use of this theoretical basis for image processing by many algorithmists has shown that it is natural and easy to think in terms of shapes when one is designing computer vision algorithms. Mathematical morphology can extract image shape features such as edges, fillets, holes, corners, wedges, and cracks by operating with various shape-structuring elements.

4.1. Morphological Operations – The morphological operators deal with two images. The image being processed is referred to as the *active image*, and the other image being a kernel is referred to as the *structuring element*. Mathematical morphology represents image objects as sets in a Euclidean space. The basic morphological operations are dilation, erosion, opening, and



(a)



(b)

Fig. 4. (a) Zigzag scanning pattern for the block size of 8×8, (b) A modified scanning pattern.

closing.

Dilation combines two sets using vector addition of set elements. Dilate is the locus of all centers such that the structuring element translated by placing the origin at the centers of B at c) hits the image set. Dilation by disk structuring elements corresponds to the isotropic expansion algorithm. Erosion is the morphological dual to dilation. It combines two sets using vector subtraction of set elements. Erosion is the locus of all centers such that the structuring element translated by placing the origin at the centers is entirely contained within the set.

In practical applications, dilation and erosion pairs are used in sequence, either dilation of an image followed by the erosion of the dilated result (called *closing*), or erosion followed by dilation (called *opening*). In either case, the result of iteratively applied dilations and erosions is an elimination of specific image detail smaller than the structuring element without the global geometric distortion of unsuppressed features.

4.2. Morphological Close Minus Open Operation – The procedures we use in the morphological recognition algorithm [17] are: perform an opening and a closing of the input image with a horizontal structuring element and subtract the opened output from the closed output, repeat this with a vertical structuring element to produce another output, and combine the two outputs using a grayscale intersection (pointwise minimum). The structuring elements used are chosen to be larger than the height and width of the largest object.

Fig. 5a shows an image of cars in a toll highway. Fig. 5b shows the result of a closing with the vertical structuring element of height 13 (i.e. 1×13). All dark cars whose height is smaller than 13 are removed and replaced by the background gray level. Fig. 5c shows the vertical opening of Fig. 5a with the same structuring element; this removes the bright cars. Fig. 5d shows the subtraction of Fig. 5c from Fig. 5b; all cars are bright and those dark cars and shadows of bright cars with distances less than 13 have been merged. Figs. 5e and 5f show the closing and opening respectively with the same horizontal structuring element of width 17 (i.e. 17×1). Fig. 5g shows the subtraction of Fig. 5f from Fig. 5e. The pointwise minimum (Fig. 5h) of Figs. 5d and 5g shows that the merged cars have been separated and all the cars are bright.

4.3. Wavelet Transform Combined with Morphology for Clutter Reduction – The morphological close minus open operation eliminates background clutter whose size is larger than the size of the target objects. However, all clutter with particle sizes similar to or smaller than the largest object size is still present. We use wavelet filters to detect the existence of the high clutter regions in an input scene, and then eliminate these regions in target recognition.

We view a high clutter region as consisting of many clutter particles with different sizes, shapes and with different degree of edge smoothness. Wavelet transforms are well known to be

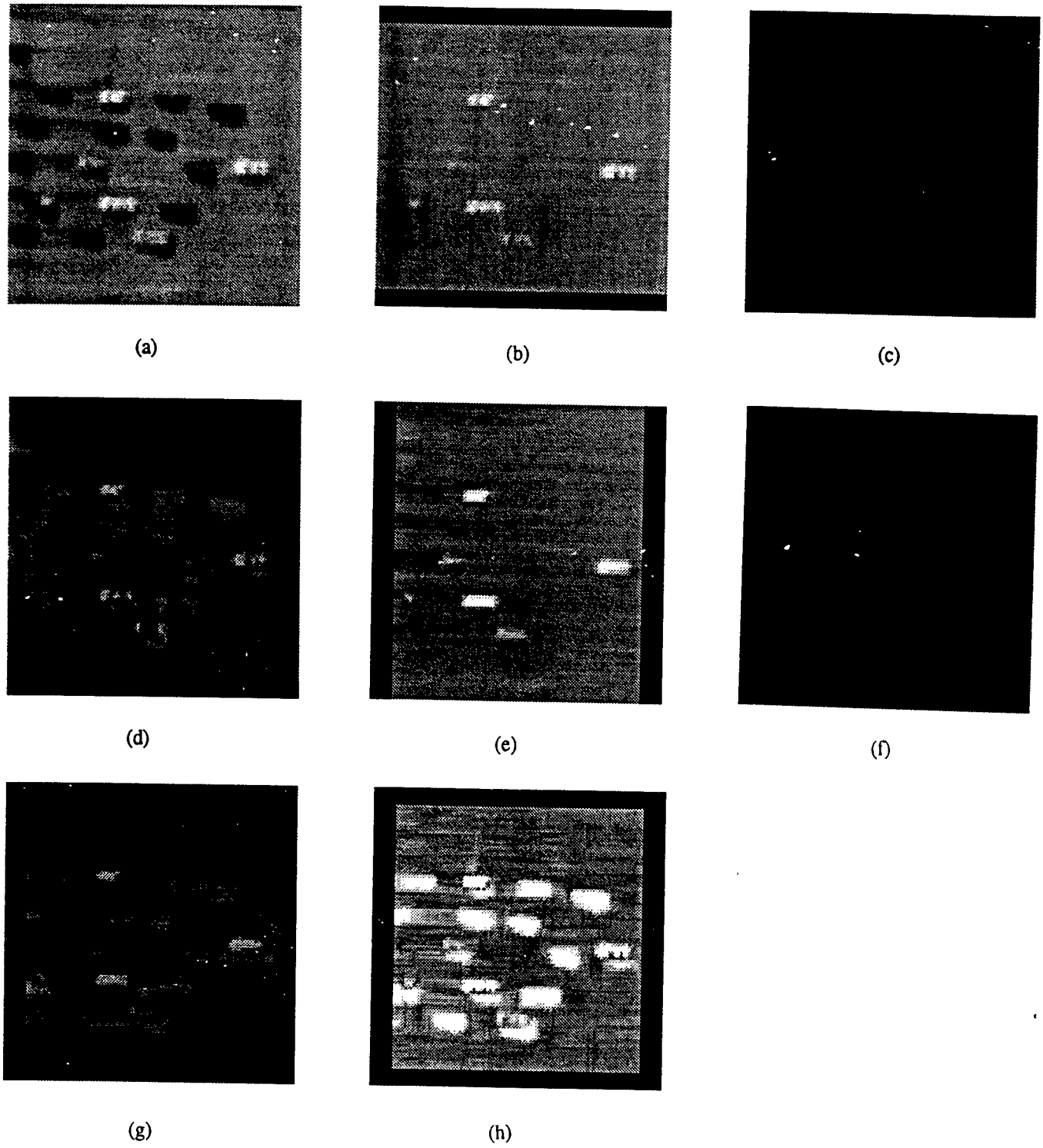


Fig. 5. (a) Input image of cars in a toll highway, (b) Closing of (a) with 1×13, (c) Opening of (a) with 1×13, (d) Difference of (b) and (c), (e) Closing of (a) with 17×1, (f) Opening of (a) with 17×1, (g) Difference of (e) and (f), (h) Minimum of (d) and (g).

able to detect edges of varying smoothness [9]. We use $\phi_a(x)$ to denote $a^{-1/2}\phi(x/a)$ and use similar forms for $\phi_a(y)$, $\psi_a(x)$ and $\psi_a(y)$. For each scale a, there are three output images as shown in Fig. 2. In Fig. 3b, we show the wavelets for three different scales ($a_x = a_y = a = 2, 4, 8$). Note that the topest scale is 2. From observation, the high clutter regions exist dominantly in scales $a = 4$ and $a = 8$. We combine both of the wavelet functions into a single macro wavelet function $w(x, y) = c_1\phi_4(x)\psi_4(y) + c_2\phi_8(x)\psi_8(y)$. Then we smooth the output with a Gaussian low-pass filter.

5. Enhanced Graphical User Interface for Imagery Toolkit and Its Extensions

The Imagery Exploitation 2000 (IE2000) imagery toolkit was initially developed at Rome Laboratory under the 1993 AFOSR summer faculty research program by professors Robert Stevenson [15] and Robert Snapp [13]. Then Audrey Copperwheat added graphical user interface for some functions. In this project, we have extensively implemented the graphical user interface for all the functions of imagery toolkit. Furthermore, some useful image processing rountines such as morphological operations and segmentation are added.

The IE2000 Imagery Toolkit is a general purpose government-off-the-shelf imagery exploitation software package intended for distribution to Department of Defense (DOD) application developers and users. It was developed on SunOs 4 and we upgrade to SunOs 5 or Solaris. It was written in ANSI C to make portable and was developed using Motif. It is intended to be a comprehensive tool, covering all the fields in image processing. An effort to include most of techniques has been attempted, but due to limited time and its large number some remain to be implemented. Later versions will include missing techniques as well as new ones. The toolkit has aimed to make it very easy to upgrade and modify. Users can quickly learn about different techniques by inspecting the codes and can enhance the system by either implementing new techniques not provided or improving the ones provided.

5.1. The Main Window – The principal interface to the program is the main window. It consists of three areas: the menubar, the editing icons and the display area. Fig. 6 shows the main window.

5.1.1. The menubar – There are 11 pulldown menus located in the menubar. The first one "File" is used to call read and write dialogs, and to clear, restore and exit. The following 8 menus carry out often-used image processing routines. The menu "Interface" intends to interact with other application programs. The "Help" menu provides on-line information about the application.

5.1.2. The editing icons – The editing icons include clockwise or counterclockwise rotation, image cutting, zooming, and audio interface.

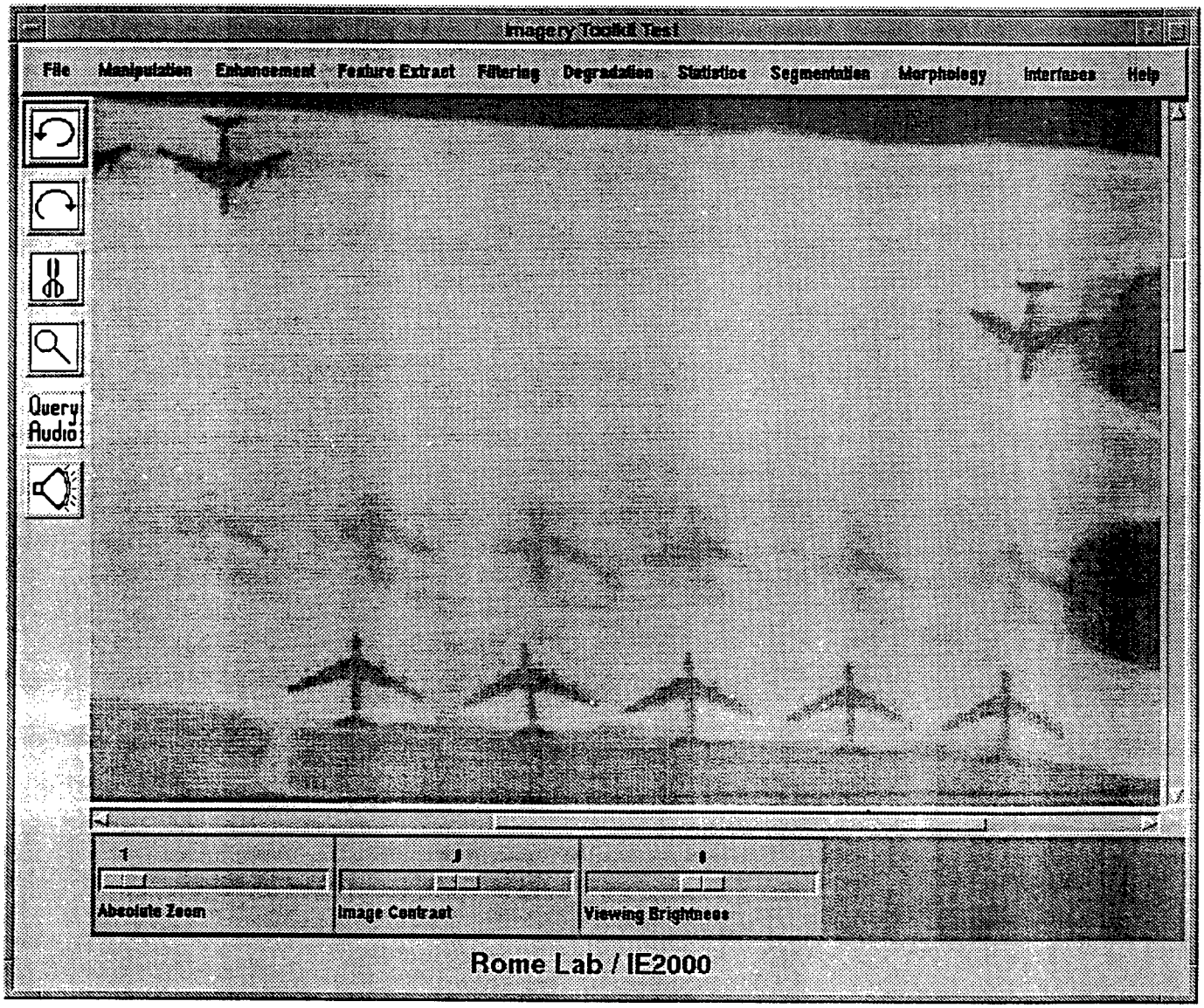


Fig. 6. The main window.

5.1.3. The display area – This area is reserved for image display. If the image size is larger than the display area, horizontal and vertical scroll bars are automatically shown.

5.2. The Read and Write Dialogs – These dialogs are used to select a file to read or write an

image from or to an imagery server or disk. Both use the file selection box widget which encapsulates the task of opening a directory file, reading its entries and traversing the directory tree. The contents are displayed in two list widgets: one holds the directory files within a given directory and the other holds ordinary, link and device files. Selecting a file and traversing the directory tree are done by clicking an item in the appropriate list widget. The format supported currently is TIFF. Fig. 7 shows the read dialog.

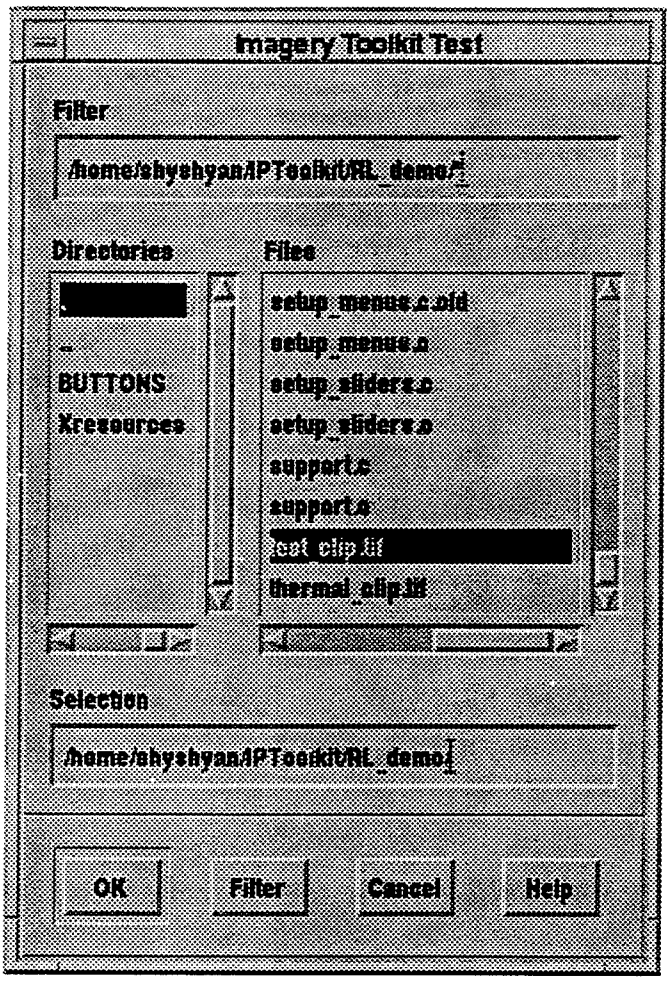


Fig. 7. The read dialog.

5.3. The Image Processing Routines – The image processing routines are grouped into 8 categories: manipulation, enhancement, feature extraction, filtering, degradation, statistics, segmentation, and morphology.

5.3.1. Manipulation – The image manipulation contains zoom, rotation, crop, flip on the horizontal or vertical axis, show an overview window, set parameters, and add or multiply a constant. Fig. 8 shows the munipulation pulldown menus. The "Zoom" menu includes 3 pulldown menus: zoom in, zoom out, and original image restoration. The zoom factor can be set in the

menu of "Set Parameters." The "Rotate" menu includes 3 pulldown menus: rotate right (clock-wise), rotate left (counter-clockwise), and restore to 0 rotation (the original image). The rotation angle (degree) can also be set in the menu of "Set Parameters." The "Show Overview Window" allows the creation of another window showing the image processed. This makes it possible to compare serveral images processed by different techniques. The "Add Constant" menu calls a pulldown menu: constant to add. It is possible to increment each pixel with a given amount to alter the brightness. Similarly, each pixel can be multiplied by a given ratio to alter the contrast by using the "Multiply Constant" menu.

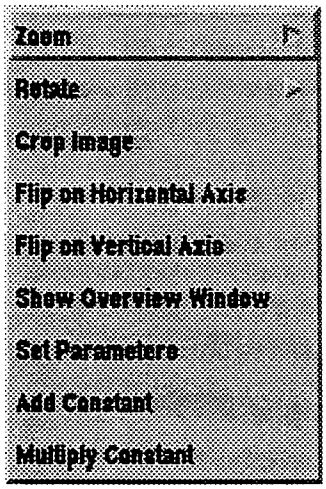


Fig. 8. The manipulation pulldown menus.

5.3.2. Enhancement – The image enhancement contains adjust mean, adjust contrast, adjust Gamma, map intensities, invert intensities, stretch range, mask image, set parameters, and histogram equalization. Fig. 9 shows the enhancement pulldown menus. The contrast, Gamma, and intensity offset can be selected by using the "Set Parameters" menu. The "Mask Image" is to enhance an image using an unsharp mask operator whose window width and height and the alpha value used as the fraction of the highpassed added to the original, can also be selected by using the "Set Parameters" menu.

5.3.3. Feature Extraction – The feature extraction contains threshold count, etc., edge detection, and set parameters. Fig. 10 shows the feature extraction pulldown menus. The "Threshold Count" returns the number of pixels above threshold or below threshold. The "etc." reserves for future extension. The "Edge Detection" includes the Canny, Frei and Chen, Kirsch, Marr Hildreth, Prewitt, Roberts X, Robinson, and Sobel edge operators.

5.3.4. Filtering – The image filtering contains maximum, average, minimum, cross median,

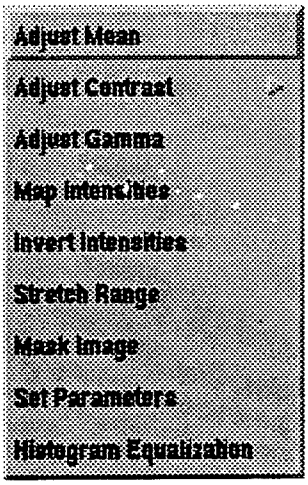


Fig. 9. The enhancement pulldown menus.

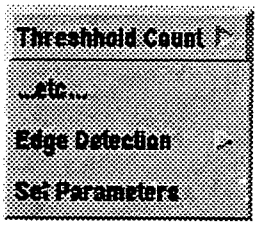


Fig. 10. The feature extraction pulldown menus.

square median, cross mean trimmed, square mean trimmed, convolution, and Gaussian smoothing. Fig. 11 shows the filtering pulldown menus. The "Convolution" menu displays a pulldown menu which allows the interactive input of a 3×3 mask. The "Gaussian Smoothing" menu allows the input of standard deviation. The remaining filters allows the selection of either 3×3 or 5×5 window.

5.3.5. Degradation – The image degradation contains random noise and uniform noise additions. Fig. 12 shows the degradation pulldown menus. The random noise corrupts individual bits in the image with the probability "percentage." The "seed" parameter initializes the random number generator. The uniform noise is distributed uniformly over the range from -range to +range. The "seed" parameter initializes the random number generator.

5.3.6. Statistics – The image statistics contains moment, standard deviation, maximum, mean and minimum. Fig. 13 shows the statistics pulldown menus. These buttons return the statistical values measured for the image.

5.3.7. Segmentation – The image segmentation contains bilevel threshold, half threshold, and multilevel threshold. Fig. 14 shows the segmentation pulldown menus. The "Bilevel Threshold" creates a binary image by mapping all pixel values below "Threshold" to 0 and all above to MAXPIXEL. The "Half Threshold" creates a grayscale image by mapping all pixel values

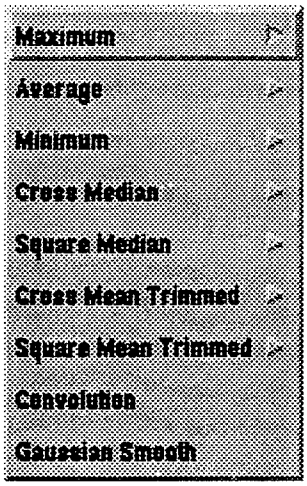


Fig. 11. The filtering pulldown menus.

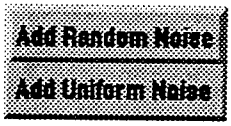


Fig. 12. The degradation pulldown menus.

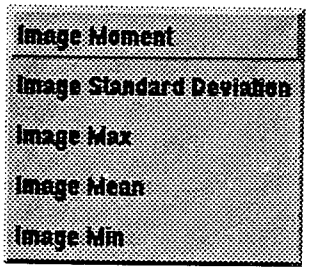


Fig. 13. The statistics pulldown menus.

above "Threshold" to MAXPIXEL and all below to their original values. The "Multilevel Threshold" creates a four-level image by selecting three threshold values and mapping all pixels values in between to 0, 85, 170, and 255, respectively.

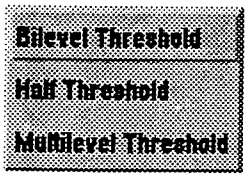


Fig. 14. The segmentation pulldown menus.

5.3.8. Morphology – The mathematical morphology contains binary dilation, erosion, opening, closing, soft dilation, soft erosion, soft opening, soft closing, set parameters, grayscale dilation, and grayscale erosion. Fig. 15 shows the morphology pulldown menus. The binary

morphological opertors allow the selection of a 3×3 or 5×5 flat structuring element. The soft morphological operators besides the same selection have the input "rank" set up in the "Set Parameters" menu. The grayscale morphological operators allow the input values of a 3×3 structuring element.

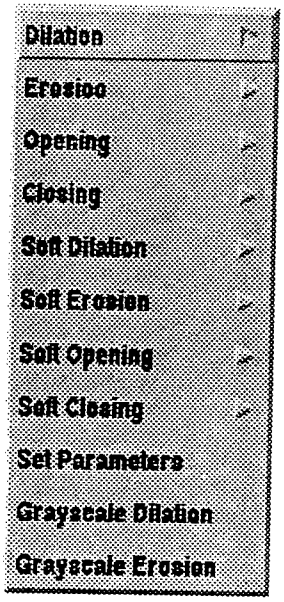


Fig. 15. The morphology pulldown menus.

6. Conclusions

In the 1995 AFOSR summer faculty research program, we have accomplished the graphical user interface programming for IE2000 Imagery Toolkit and the extensive routines of image morphological operators and enhancement. We have demonstated that the toolkit is very easy to upgrade and modify. The future extension will have to be object oriented in order to take advantage of inheritance, polymorphism and encapsulation. The language of choice is C++. The program will not have to be rewritten from scratch, many of the current routines can be modified to be useful. The overall skeleton of the program will be maintained.

We have investigated a new image compression scheme for multiresolution coding algorithm by wavelet transform coding. It is observed that the energy of high subbands are mainly concentrated around the appropriate edges of the original image. A simple direct vector quantization to encode the high subbands does not take full use of the sparsely distributed nature of the high subband, and waste many bits to encode the blocks with very low variance. We apply a block variance detector on the lowest subband to predict whether the high subbands contain edge or not. We have also experimented with the wavelet transform combined with the morphological close minus open operation to extract dark or bright targets compared to the intensity of background.

References

[1] M. Antonini, M. Barland, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Processing*, vol. 1, no. 2, pp. 205-220, Apr. 1992.

[2] W. Chen and W. K. Pratt, "Scene adaptive coder," *IEEE Trans. Communications*, vol. 32, no. 3, pp. 225-232, 1984.

[3] R. Coifman and M. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Information Theory*, vol. 38, no. 2, pp. ?, March 1992.

[4] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Comm. Pure and Appl. Math.*, vol. 41, no. 7, pp. 909-996, 1988.

[5] D. Gabor, "Theory of communications," *Journal of Insti. Elec. Eng.*, vol. 93, pp. 429-457, 1946.

[6] A. Haar, "Zur theorie der orthogonalen Funktionensysteme," *Math. Ann.*, vol. 69, pp. 331-371, 1910.

[7] R. M. Haralick, S. R. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 9, pp. 532-550, July 1987.

[8] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674-693, 1989.

[9] S. Mallat and S. Zhong, "Wavelet transform maxima and multiscale edges," in M. Ruskai *et al.*, ed., (*Wavelets and Their Applications*), Jones and Bartlett, pp. 67-104, 1992.

[10] J. Morlet, G. Arens, E. Fourgeau, and D. Giard, "Wave propagation and sampling theory I, II," *Geophysics*, vol. 47, pp. 203-236, 1982.

[11] F. Y. Shih and O. R. Mitchell, "Threshold decomposition of grayscale morphology into binary morphology," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 31-42, Jan. 1989.

[12] K. R. Sloan and S. L. Tanimoto, "Progressive refinement of raster images," *IEEE Trans. Computer*, vol. 28, no. 11, pp. 871-874, 1979.

[13] R. R. Snapp, *IPToolkit: An Image Processing Environment for the X Window System*, Final Report for Summer Faculty Research Program, Rome Laboratory, Aug. 1993.

[14] J. L. Starck and A. Bijaoui, "Filtering and deconvolution by the wavelet transform," *Signal Processing*, vol. 35, pp. 195-211, 1994.

[15] R. L. Stevenson, *Image Processing Toolkit*, Final Report for Summer Faculty Research Program, Rome Laboratory, Aug. 1993.

[16] J. W. Woods and S. O'Neil, "Subband coding of images," *IEEE Trans. Accoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1278-1288, Oct. 1986.

[17] A. Ye and D. Casasent, "Morphological wavelet transform for distortion-invariant object detection in clutter," *Proc. SPIE Conf. on Wavelet Applications*, vol. 2242, pp. 525-537, 1994.

Connecting CASE to Simulation Development

Jeffrey W Smith
Associate Professor
Computer Science Department

415 Graduate Studies Research Center
University of Georgia,
Athens, GA 30602

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory/IRRE
32 Hangar Road
Griffiss AFB, NY 13441-4114

August 1995

Connecting CASE to Simulation Development

Jeffrey W Smith
Associate Professor
Computer Science Department
University of Georgia, Athens, GA

Abstract:

Object-Oriented Computer-Aided Software Engineering (CASE) tools are needed in the process of simulation development. Simulation development is similar to other forms of software development, though there are differences. We have developed an automated connection of Object-Oriented CASE software development tools to simulation development. Specifically, we have evolved an environment in which the Object-Oriented CASE tool ROSE™ is connected to the Object-Oriented simulation development environment MODSIM™. With this connection, we can use the outputs from the CASE tool to define classes, objects, and relationships and to suggest the modularity of the simulation system. We can provide (stubs of) code from the CASE tool to directly aid coding in the simulation development environment.

Ideally, the connection runs both ways, producing code from CASE files and CASE files from code. This bidirectionality keeps the implementation and documentation of the class and object structures in step, we have made some progress towards this Round Trip connection.

Connecting CASE to Simulation Development

Jeffrey W Smith

I. Introduction:

Software development is an exacting task, and one to which software tools have been late in arriving. Tools for Computer-Aided Software Engineering (CASE), programs to help in the development of programs, have contributed to progress in these areas of software systems development. Software tools can aid in the traceability of design and capture of design assumptions, requirements, and constraints that is required to perform the validation and verification of complex systems. There is increased emphasis on the early stages of system definition due to the expense of missteps in this phase of systems development, the advent of object-oriented technology, and decreased time to market [TP83]. A simulation development is much like any other software development, but there are differences. To develop a simulation, as with any software, data structures must be defined, related, and communicated and procedural steps must be developed, arranged, and ordered. CASE tools which help these operations can be applied to this field. However, there is no CASE tool at present directed at the specifics of simulation development, some extension or adaptation of CASE tools is necessary to apply them to the simulation development environment.

In a recent simulation development [BDA95a, BDA95b, Min92] sponsored by the Rome Laboratory and performed by Nichols Research Corp. (NRC), a process simulation was developed in MODSIM, a simulation language derived from MODULA-2. As a part of their study, NRC prespecified the simulation objects and classes in the CASE tool ROSE from Rational Corp. Once the simulation was specified in ROSE, the specification was passed to the traditional development process, and any further benefits from the definition phase were largely foregone as a functional separation developed between the definition and the progressing implementation. This gap will appear as the system is developed since there are updates and refinements to the function of the system as the implementation proceeds, and these are seldom reflected in updated and refined defining documents.

The gap between the CASE tool and the software development environment can be bridged by an automated tool that connects the definition domain (of the CASE tool) with the implementation domain (of the simulation system). The bridge is formed by reading in a

portable ASCII file format output by the CASE tool, analyzing it syntactically and extracting semantic information from it and using this to output information useful in simulation development. This syntactic analysis has proceeded by the process or creating or adapting a standard-format grammatical notation and system-standard language analysis tools (*lex* and *yacc* [JL78] in this case) to the analysis process. There is some exemplary material in this report on how the bridge was built, and examples from each of the analysis domains is included. This information can be used to generate the definition modules and implementation module headers for the MODSIM simulation system, which serves to define the data structures and their relationships for the development of simulations.

The structure of the translation system which connects the CASE environment (ROSE) to the simulation environment (MODSIM) is shown in Figure 1. The petal files which are output by ROSE are read in by a program named "petal", which translates (portions of) them to MODSIM code. The MODSIM code can be compiled by the MODSIM compiler (mscomp) and will define (portions of) the simulation. The MODSIM code can also be accepted by a program called "simula", which outputs a file that can be reintroduced into the CASE environment. Due to the scale of the translation problem, compiler development tools (*lex* and *yacc* [JL78]) were used to develop them. These tools and their application to the CASE-to-simulation connection are detailed below. Keep in mind in this discussion that the parser is the program that does the translation. When we define a (lexical analyzer and) grammar, we imply a parser which *yacc* will build. We then run this parser to do the translation.

II. Software Definition.

CASE Tools:

A CASE tool is any program that helps in software development, modification, or maintenance. This takes in a lot of territory, but we are most interested in the definition/specification stage that occurs at the very beginning of the process. With the advent of the object-oriented approach to software definition and development, there has been a corresponding development of object-oriented specification and CASE tools. The use of CASE tools for system definition and initiation of development is an idea whose time has come. CACI has a graphics-centered definition environment (SIMOBJECT) and NRC also uses one (GENESys). We will use a more general tool for Object-Oriented definition (ROSE).

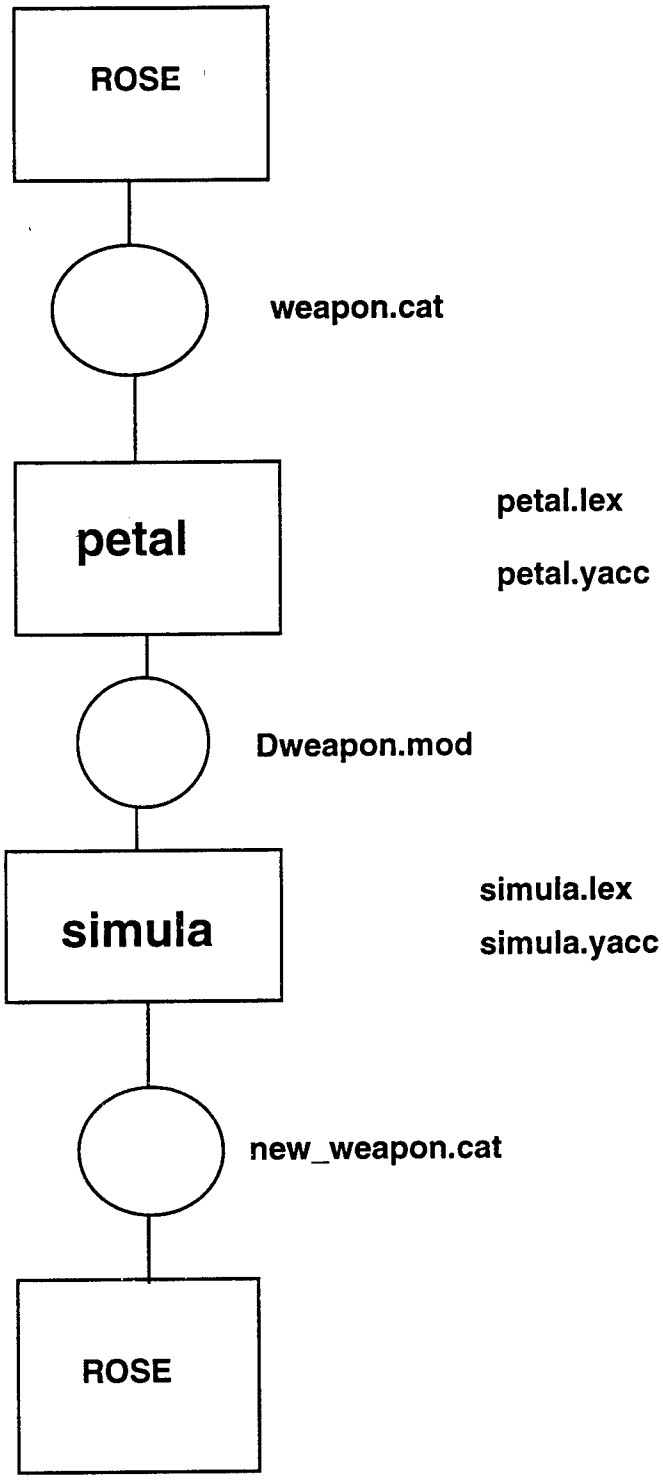

ROSE
weapon.cat
petal
petal.lex
petal.yacc
Dweapon.mod
simula
simula.lex
simula.yacc
new_weapon.cat
ROSE

Figure 1

A tool that can actually be used to reduce the developer's workload or make him more productive is what is needed in the software development process. Once we have found a CASE tool with desirable properties, we want to use it as extensively as possible. The definition phase of software system development is an upstream process in the sense that it has leverage on the subsequent development of the system. It can (and should) also serve as an overview, a document, and an education asset for the system throughout its lifetime. To realize these benefits, the definition information must stay in contact with the implementation.

The ROSE CASE Tool:

Rational Object-oriented Systems Engineering (ROSE) is an approach and a software methodology implemented to support object-oriented analysis and design [Boo91, Whi94a, Whi94b], team development, and code generation and analysis in C++ and Ada. ROSE runs on the PC and compatibles with Windows 3.1+, and does have a code generation capability, in C++ [Rat94]. Such a code generation capability may be labor-saving in development, but it may be more important as a component in the Round Trip concept [Boo91, Whi94a,b], since the generated code produces the defining function headers and data structure boilerplate and documentary annotations that make the Reverse Engineering portion of the cycle possible without loss of information. (In Reverse Engineering, code which may have been updated in implementation is read back into the CASE tool to provide an updated definition and documentation for the system.) The code generated would be directly applicable to simulation if the simulation were done in C++, or indirectly applicable if we choose to translate from the C++ stubs produced by the CASE tool to the simulation language (MODSIM in our case). Instead, we have chosen to make the translation from the ROSE portable ASCII file format (the *petal* format) to MODSIM, and in the process to learn how to approach the translation problem so as to be able to provide a general connection between CASE and simulation, which will bring the benefits of documentation, requirements and constraint statements, and validation capability to the simulation development environment.

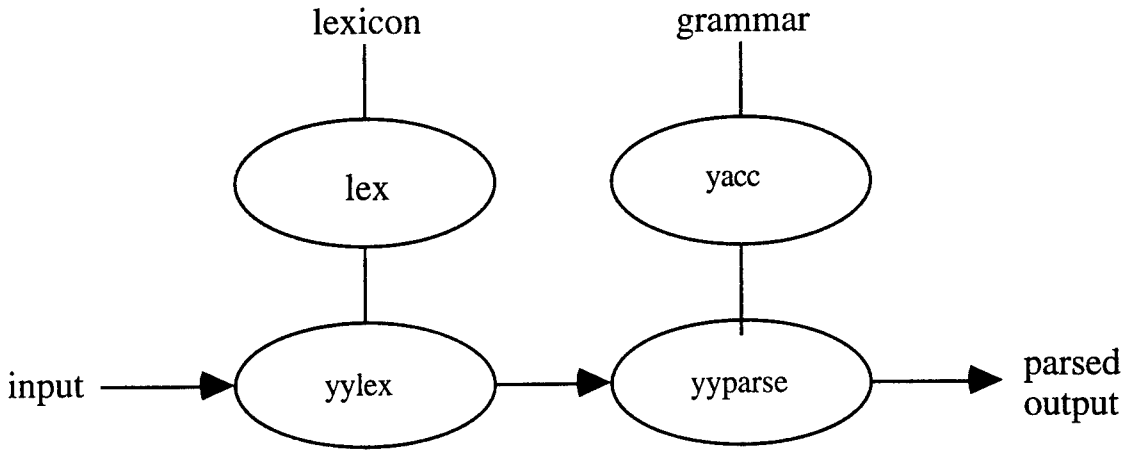


Figure 2

The ROSE approach builds a definition of the system from 4 models which can be specified: logical, physical, static, and dynamic. The ROSE user creates and refines these 4 models of his system within the framework of a general model representing the problem domain and the software system, including classes, class categories, objects, operations, states, subsystems, modules, processors, devices, and relations between them [Boo91]. Icons that represent these aspects of the model appear where appropriate in the ROSE GUI, allowing a pictorial view and manipulation of the model's application, diagram, and specification aspects. This information would be useful in implementation in addition to the planning of the system; it is presented in three formats: the graphical format which the GUI provides to the user for entry and manipulation, a printed template format, and the files (called *petal* files) that the system uses to save the work on the disk between sessions. The petal files are also the portable format of the system information.

III. Software Implementation.

Simulation Development

As the field of computer science advances, it becomes more specialized and fragmented. The simulation community is one such fragment -- simulation developers use their own languages (SIMSCRIPT, SIMULA, GASP, GPSS, SLAM, MODSIM, *etc.*) and their own development environments especially tailored to the simuiation task. However, there would be benefit from access to the more extensive and capable development tools available in the mainstream development environment.

As code development proceeds, implementation concerns are paramount, with an inevitable diminishment in the attention paid to definition. But the definition remains very important, and any gap between the definition and the implementation must be viewed with alarm. We need automated tools that can either/both:

1. Produce useful implementation code from the definition phase
2. Produce an updated definition from a changed implementation.

Since the definition of the system includes the most accessible documentation of the system, useful for understanding the system, training, etc. it is important to keep this up-to-date. The basic idea of Round Trip Gestalt Design (notated as RTE for Round Trip Engineering) [Boo91, Whi94a, Whi94b] is that no development effort should ever be wasted, so all the work that is done must be applicable to the final system design. However, we want a solution (at least a prototype of a solution) as soon as possible. In this model of development, progress is made with each lap around the cycle, but it is important to have the capability to make a lap around the cycle.

In order to do the RTE, the capability to translate reliably and usefully from the definition space into the implementation space. For instance, the ability to produce class definitions and object headers from the definition code make a useful connection; if this is well done, changes to the class structure can be done in the definition domain rather than the implementation domain; which would serve as an immediate bridge over the definition-implementation gap.

The MODSIM Development Environment:

MODSIM II is a general-purpose modular language based on MODULA-2 and with extensions that make it particularly suitable for simulation development [MOD93a, MOD93b, MOD94]. MODSIM is the centerpiece of a group of simulation products that extend and enhance the simulation environment (e.g., SIMGRAPHICS II, SIMMASTER, SIMOBJECT). MODSIM is a very capable simulation development environment which has been used to produce many large and capable simulations. For instance, MODSIM has been used to simulate the traffic flow about the Chunnel, and is nearly a standard in the simulation of airport, air traffic, and military operations [OP95].

IV. Translation Example.

If we want to translate from one environment to another, the level at which the translation should occur is the first issue to resolve. If both the CASE tool and the simulation system have a GUI, it is tempting to attempt a connection at this level, but unless the GUI's are identical we do not have the technology for this level of translation. If both have templates, the corresponding fields can be filled in to provide a translation at this level. In our case, ROSE has both a GUI and templates but MODSIM has neither presently available with the language product. ROSE also has files (the petal files) which provide the ASCII portable statement of the CASE tool output, and MODSIM has a well-defined language which it accepts. For reasons of available tools and definitions, the file-to-language level of translation is the approach we have adopted.

Background on Grammars:

"A programming language can be defined by describing what its programs look like (the *syntax* of the language) and what its programs mean (the *semantics* of the language) [ASU88, p.25]. This statement is also true of files, datasets, network transactions, indeed most computational entities. "For specifying the syntax of a language, we present a widely used notation, called *context-free grammars* or BNF (for Backus-Naur Form).... Besides specifying the syntax of a language, a context-free grammar can be used to help guide the translation of programs... known as *syntax-directed translation.*" [ASU88, p.25]

A grammar is a formal set of rules and procedures for the analysis or synthesis of language constructs. The rules are called *productions*, and have the form
LHS --> RHS
which we read as 'the left-hand side generates the right-hand side', or 'the right-hand side can be generated by the left-hand side'. The grammar has *terminal symbols* that actually appear in the language, and *nonterminal symbols* that represent intermediate steps in grammatical analysis. The terminal symbols are grouped for analytical purposes into *lexical tokens*. A grammar generates member strings of a language by starting at the start symbol and applying productions (replacing the LHS with the RHS until there are only terminals left. This process is said to work *top-down*. A grammar can accept (reject) strings by replacing terminals with tokens, then replacing RHS of productions with LHS until only the start symbol is left. This process is said to work *bottom-up*. Top-down analysis works out readily by hand and appears to be more elegant, but bottom-up methods apply to a larger class of grammars and are more amenable to computer implementation. For practical problems, the describing grammars can become very large and computers

must be used to process them (analyze, check for conflicts, reduce). The UNIX-based tools *lex* and *yacc* [JL78, SF85, ASU88, Pys88] have become a standard in this area, and are the tools we made use of for this translation project. When a defining grammar is input into the computer in *yacc* format, the program will build a parser for the defined language. There are frequently performance or size difficulties with this very general definition, but we have neglected such issues for this prototype development.

Using grammatical analysis on linguistic entities, there are several technical issues which must be addressed -- ambiguity, associativity of operators, precedence, *etc.* Here is a very small example grammar for a subset of arithmetic expressions:

```
expr   -->  expr + term   || expr - term   || term
term   -->  term * factor || term / factor || factor
factor -->  number        || ( expr )
```

In this example grammar, all the symbols mentioned are nonterminals except for ()+-*/. This grammar incorporates the precedence of */ over +- and nested expressions using ().

The process of analyzing an input stream to determine whether or not it is a member of the set acceptable to the grammar is called *parsing*. "A parser can be constructed for any grammar... For any context-free grammar, there is a parser that will parse a string of N tokens in O(N^3)." [ASU88, p.48]. We need to do more than analyze the input, we need to know what it means so that we can translate it into another format. As the symbols are encountered in the parsing process, they can be viewed as having certain *attributes* (this word is the accepted terminology, but an attribute is not crisply defined, it is anything that aids the translation process). The grammar with attributes attached to its symbols is called an *attributed grammar*, which can be used in the process of syntax-directed translation [ASU88, p.34]. Here is a part of the grammar above with attributes attached:

```
expr   --> expr + term      expr.a = expr.a | term.a | '+'
expr   --> expr - term      expr.a = expr.a | term.a | '-'
expr   --> term             expr.a = term.a
term   --> 0 || 1 || ...    term.a = '0' or '1' ...
```

This grammatical stub shows an attributed grammar describing a program that will convert an arithmetic expression in infix notation (3 + 4) to one in postfix notation (3 4 +). Postfix

notation is useful in computers and calculators for arithmetic since expressions can be written unambiguously without parentheses and because they can be evaluated with a single operand stack. The first rule states that if the production (expr + term) is invoked, the output should be (expr term +). Figure 3 shows the *lex* and *yacc* files for a complete implementation of this function (this example was adapted from one in [ASU88]), followed by example input and output illustrating the operation of this program. This method of parser development (given a grammar) is well documented [JL78, ASU88, Pys88].

The size of the parse produced from the lexical and grammatical specification of toy is mentioned here for purposes of comparison with the later parsers we developed:
The iex portion has 12 nodes, 29 partitions; the yacc portion has 9 terminals, 3 nonterminals, 13 grammar rules, 26 states.

ROSE Information for MODSIM Development:

As objects and classes are developed, they do serve as a guide and template for further system development. Properly done, the classes, inheritance and relation structures, and objects that result from object-oriented CASE development will be the framework for all subsequent development, will be the overview of the system, and will provide an education and training tool over the lifetime of the system. At least, that is the idea.

In practice, once the classes are defined and their relationships clearly depicted in whatever sort of diagram the developers may favor, the pressures of performance, function, and time affect the structure and blur the formal lines of definition. The major goal of Round Trip Engineering is to prevent this gap from developing by keeping the initial phases of system development in close contact with subsequent implementation.

```
/* FILE = toy.l  jws:950821 */
%{
#include "y.tab.h"
extern int yylval; /* jws */
%}
/* FILE = toy.l  jws:950821 */

%%
[ \t\n] ;               /* white space */
[0-9]+      {yylval=atoi(yytext); return(INTEGER);}
.     {return(yytext[0]);}    /* other chars */



/* FILE = toy.y  jws:950821 */
/* produce lukasciewicz (postfix) expression from infix */
%token INTEGER
%{
#include "express.h"
extern char yytext[];
%}
%%
expr:  term
      | expr '+' term   {printf(" + ");}
      | expr '-' term   {printf(" - ");} ;
term:  factor
      | term '*' factor {printf(" * ");}
      | term '/' factor {printf(" / ");} ;
factor: '(' expr ')'
      | INTEGER               {printf(" %d ",yylval);}
      | '+' INTEGER           {printf(" %d ",yylval);}
      | '-' INTEGER           {printf(" %d ",-yylval);}
      | '+' '(' expr ')'
      | '-' '(' expr ')'      {printf(" 0- ");} ;
%%
int yywrap(){return(1);}



# sample input/output from the above
lex.yy.c:if(yywrap()) return(0); break;
toy:yywrap
smithjw/YACC: echo 3+4 | toy                    3 4 +
smithjw/YACC: echo 3+(4+2) | toy          Badly placed ()'s.
smithjw/YACC: echo " 3+ (4+2)" | toy            3 4 2 + +
smithjw/YACC: echo "(3+4) + 2 " | toy           3 4 + 2 +
smithjw/YACC: echo " - ( 3 + 4) + 2 " | toy     3 4 + 2 +
smithjw/YACC: echo " - ( 3 + 4) + 2 " | toy     3 4 + - 2
smithjw/YACC: echo " - ( 3 + 4) + 2 " | toy     3 4 + 0- 2 +
smithjw/YACC: echo " + 3 " | toy                3
smithjw/YACC: echo " (3+4) + 2" | toy           3 4 + 2 +
smithjw/YACC: echo " 3 * (4 - 2)/5" | toy 3 4 2 - * 5 /
smithjw/YACC: echo " -3 *(4-2)/5" | toy         -3 4 2 - * 5  /
```

Figure 3

Connecting ROSE to MODSIM:

The last thing a person needs when dealing with a system under development (get it to work, explore implementation alternatives, get it to work faster, get it to work again) is any hitch in the development tools that distract from or even obscure the real problem. In this context, a 'seam' in the development environment is any noticeable break in the uniformity of the process. For (specific) example, suppose that there is a manual procedure that must be performed in the midst of an otherwise automated set of procedures. The fact that the manual procedure is slower and less convenient will serve to discourage its use. If the manual procedure is a part of a development process, then the development process is slowed or undone to the extent that the procedure is distracting, more effort (or perceived as such), more time consuming, and more error-prone. 'Seamless' development lacks discontinuities in the process. The key to seamless development is automation: end-to-end, robust, complete. This is what we want, as does anyone facing similar problems (for instance, the CAD community).

Once the files have been captured, those portions which do not apply to the software development can be dropped from consideration. For instance, there are portions that deal with the position and placement of the class and object representation in the ROSE GUI. While this information is valuable when working with the ROSE environment, there is no need to transfer it to the MODSIM environment.

There are many tools which aid in the structural analysis of files, programs, or other computer artifacts which have structure. These tools, of which *lex* and *yacc* are exemplary [JL78, SF85, Pys88], require a grammar as input to specify the structure. The question then becomes: **how can one find or create a grammar** for a particular structure one is trying to analyze? There are two ways to do this, with compromises and variations possible:

1. Evolve a grammar from an example set of the structures. This will result in an *ad hoc* grammar, since it will describe only the set of examples for which it was developed, but if done in the proper spirit (as generally as possible while getting the problem done), a workable grammar can be produced in this way.
2. If the structure were generated in compliance with some grammar, then that grammar will also serve to analyze the structure.

Each of these methods was used to develop a grammar in the analysis of the ROSE (the petal parser developed with the first method) and MODSIM (the simula parser developed

with the second method) structures that served as the foundation for the CASE-simulation connection.

Inventing a Grammar:

Suppose that you have no structural description at all, but suspect (or know) that there is structure present. **How can you proceed to a grammar** for the structure? The procedure is to develop a system which is as general as possible and obeys certain general principles of operation, but also does the desired example as a special case. The "does the example" part will accomplish the immediate function. The "as general as possible" part will provide a foundation for further development and extensions.

A grammar for the ROSE files was developed following this procedure. The files from the CASE specification of the BDASIM simulation provided the training set used to develop the grammatical rules. Starting with the shortest file (the one that needed the fewest rules to analyze), a grammar for the acceptance of ROSE files was developed by the addition or modification of rules in the grammar as cases arose. This procedure is straightforward, but it is laborious and can become intricate as the size and complexity of the grammar increase. Also, it is not straightforward to be complete and elegant and meaningful and semantic all at once. As a general course, this method should be avoided since the amount of work to an operational grammar is unknown *a priori.* However, it can be made to work when there is no alternative. In this case, the ROSE files were all recognized with the parser "petal". The file set consisted of 22 .cat files (9094 lines) and 2 .mdl files (934 lines) for a total of 10028 lines. The .cat (category) files and .mdl (model) files are both in .ptl (petal) file format, but their contents differ. The grammar went through 14 versions in evolution (though the last two were for demonstration purposes), took approximately three weeks to develop, and ended up with 439 productions. These facts apply basically to syntactic recognition, since only one example file was worked through the semantic portion of the translation. Unfortunately, this process does not guarantee a general grammar for the ROSE files, only a grammar as general as we could make it. Here are summary statistics for the petal grammar parser we developed:

```
2277 nodes, 6802 positions
168 terminals, 209 nonterminals, 439 grammar rules, 622
states
```

Subsequently, Rational sent us a description of their file reader, from which we could infer that they used the yacc approach to a parser. Their grammar appears to be about twice the

size of "petal", though much of the difference may come from the fact that this parser is for a later version of the file format and includes the code generation portions of the file format which were absent from the BDASIM petal set. Here is a set of summary statistics for the petal grammar from Rational Corp:

```
331 terminals, 421 nonterminals, 909 grammar rules, 1806
states
```

Adapting a Grammar:

The MODSIM Reference Manual contains a BNF-format grammar for the MODSIM language. This grammar carries no date or version number and it turns out to be not quite the grammar for the language (statements which are rejected by the grammar are accepted by the compiler. For instance, the production (in BNF format) for a formal parameter list is published as:

FormalParameters --> ([FPSection { ; FPSection }]) [: ident]

This statement includes the information that a formal parameter list must be encased in parentheses -- they are not indicated as optional. In fact (as the compiler implements the language), they *are* optional if the list is NULL. There are several divergences of this sort between the grammar and the compiler. As another instance, the specification of the structure CLASS is not mentioned at all, so this must be added to the grammar (using the *ad hoc* method).) Nonetheless, this grammar is close enough to the current version to make a useful starting point for grammar development. The BNF format is only similar to the yacc input format, so the information must be translated as well as adapted and extended. The working grammar was developed by elaborating the published grammar in much the same way as in the *ad hoc* method used for the petal grammar. The resulting parser, "simula", accepts all 65 Implementation modules (15862 lines), all 65 Definition modules (3364 lines) and all 7 Main modules (1246 lines) included in BDASIM. The simula parser went through 6 versions and took two weeks to develop. We mad no attempt to add semantics to this parser. Again, this does not guarantee that this parser is general, but since the MODSIM grammar that was used as a starting point must reflect the structure of the language, it is probably pretty general.

To complete a lap around the development domain, definition must be extractable from implementation information. We will have made some progress in this area with the analysis of MODSIM code and the generation of readable *petal* files from it. Here is a set of summary statistics for simula.y:

```
666 nodes, 2558 positions
```

```
95 terminals, 166 nonterminals, 300 grammar rules, 489 states
```

V. Conclusions and Further Work:

Conclusions:

We have shown how one specific CASE tool can be applied to one specific simulation development. We have redone exemplary portions of the simulation development to show the improvement in the development process effected by the CASE tool. The Rational Corp. CASE tool ROSE has been applied to the software development for the MODSIM development environment. We used the UNIX development tools to translate the outputs of the ROSE system for Object and Class definitions and refinements to the inputs required by the MODSIM development environment (essentially, MODSIM source code statements; this is mostly a code translation process at the syntactic level, but the semantics have had some influence as well).

This approach can be made to work (as the limited example shows), but the semantics are going to be even harder than the syntax. Also, the *ad hoc* nature of the solution means that the very next file may present a new case that causes the parser to fail. This approach is not robust, so it is not a good idea to use it to build a tool for the critical path of system development.

Further Work:

To proceed to a robust connection between ROSE and MODSIM:
1. Get the attributed petal grammar and associated lexical definition from Rational.
2. Get the most up-to-date MODSIM grammar from CACI.
Proceed to mate these. The methods used in development in this project can serve as a guide in this procedure, they have been shown to be effective.

On a more general note, the idea of a closer connection between the definition and implementation phases of system development must be pursued. A properly defined system can be prototyped, have early simulation results, can be connected to other software to preview system integration, has more credibility throughout the life cycle of the software. The benefits from formal definition are very great, but it is a lot of effort. If that effort can be made to pay off by jump-starting the implementation phase, it is a winner!

REFERENCES and BIBLIOGRAPHY:

[ASU88] Aho, A.V. and R. Sethi and J.D. Ullman, **Compilers: Principles, Techniques, and Tools**, Addison-Wesley, 1988.
QA76.76.C65A37 1988

[BDA95a] BDASIM v1.0 User's Manual (DRAFT), NRC, 1995.

[BDA95b] "Bomb Damage Assessment Study Modeling and Analysis", NRC, 6 June 1995.

[Boo91] Booch, Grady, **Object-Oriented Design with Applications**, Benjamin-Cummings, 1991.
QA76.64.B66 1991

[DEC94] Digital Equipment Corp., DEC OSF/1 Command and Shell User's Guide, Feb 1994.

[Fir93] Firesmith, D.G., **Object-Oriented Requirements Analysis and Logical Design**, Wiley, 1993.
QA76.64.F57 1993

[Fri91] Frisch, A, **Essential System Administration,** O'Reilly & Assoc., 1991.
QA76.76.O63F78 1993

[Ho91] Ho, Y-C (Ed), **Discrete Event Dynamic Systems**, IEEE Press, 1991
T57.6.D564 1992

[JL78] Johnson, S.C. and M.E. Lesk, "Language Development Tools", pp. 2155-2175, **BSTJ 57:6**, July, 1978

[KP84] Kernighan, B.W. and R. Pike, **The UNIX Programming Environment,** Prentice-Hall, 1984.
QA76.6.K495 1984

[KR78] Kernighan, B.W. and D.M. Ritchie, **The C Programming Language**, Prentice-Hall, 1978.
QA76.73.C15K47 1978

[MAJ93] Markey, J. and E.L. Anderson and D.G. Joder, "Coliection Requirements Management Application (CRMA)", RL-TR-93-174, August, 1993.

[Min92] Minster, D.G. "Battle Damage Assessment in the Next War (DRAFT)", RAND Corp., June, 1992.

[MOD93a] MODSIM II Reference Manual, CACI Products Co., 1993.

[MOD93b] MODSIM II Tutorial, CACI Products Co., 1993.

[MOD94] MODSIM II User's Manual, CACI Products Co., 1994.

[Nye92] Nye, A., **Xlib Programming Manual, 3e,** O'Reilly & Assoc., 1992.

QA76.76.W56X216 1993

[ORe91] OReilly, **Guide to OSF/1: A Technical Synopsis**, O'Reilly & Assoc., 1991.
QA76.76.O63G813 1991

[OP95] Oswalt, I. and R Painter, "CACI Simulation Developments Overview", (presentation), 1995.

[Pet81] Peterson, J.L., **Petri Net Theory and the Modeling of Systems**, Prentice-Hall, 1981.
QA76.9.S88P47 1981

[Pys88] Pyster, A.B., **Compiler Design and Construction: Tools and Techniques**, van Nostrand Reinhold, 19898
QA76.6.P9 1987

[Rat94] Getting Started with Rational ROSE rev. 2.5, 1994.

[SF85] Schreiner, A.T. and H.G. Friedman, Jr., **Introduction to Compiler Construction with UNIX**, Prentice-Hall, 1985.
QA76.76.C65S37 1985

[Ste92] Stevens, W.R. **Advanced Programming in the UNIX Environment**, Addison-Wesley, 1992.
QA76.76.O63S754 1992

[TP83] Thurber, K.J. and P.C. Patton, **Computer System Requirements**, Heath, 1983.
QA76.9.S88T45 1982

[Whi94a] White, I., **Rational ROSE Essentials: Using the Booch Method**, Benjamin-Cummings, 1994.
QA76.76.D47W482 1994

[Whi94b] White, I., **Using the Booch Method: A Rational Approach**, Benjamin-Cummings, 1994.
QA76.76.D47W483 1994

[Zei76] Zeigler, B.P., **Theory of Modelling and Simulation**, Wiley, 1976
QA76.9.C65Z44

Application of ATM Networks in Distributed Systems

Scott Spetka
Assistant Professor
Department of Computer Science

State University of New York
Institute of Technology at Utica/Rome
Route 12 North
Utica, New York 13504

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

September 1995

Application of ATM Networks in Distributed Systems

Scott Spetka
Assistant Professor
Department of Computer Science
State University of New York
Institute of Technology at Utica/Rome

Abstract

Every time an order of magnitude improvement occurs in hardware speed, it requires a qualitative reexamination of system design and software architecture. Every major advance in hardware has left software engineers and system designers looking for operating system structures to harness new capabilities and applications that are in need of increased performance. High-Bandwidth ATM Networks are no different than other major advances in this respect. This study examines low level system support for ATM networks. In addition, experimental measurements show that harnessing the potential of ATM in existing applications requires appropriate low-level system support. Two applications are used as examples where the benefits of ATM can be immediately applied.

Application of ATM Networks in Distributed Systems

Scott Spetka

Introduction

We begin by presenting performance measurements that indicate the potential performance advantage of ATM and the importance of appropriate software interfaces to exploit improved speed. In addition, as we look into the future of this technology, we also consider performance measurements which will help us to gauge the potential increased demand for this ATM, to support the types of applications that we propose.

One application that already suffers from an acute shortage of bandwidth is the Internet-based world-wide digital broadcast system. Mbone (the multicast backbone) [http://www.best.com/~prince/techinfo] uses a tree-structure to broadcast real-time audio, video, and other multimedia presentations. To use network bandwidth efficiently, the Mbone multicast routers form a store-and-forward tree, using the underlying Internet for communications between mrouters (see Figure 1). Mbone users have kernel support for receiving broadcast packets directly from an mrouter, through an ethernet interface.

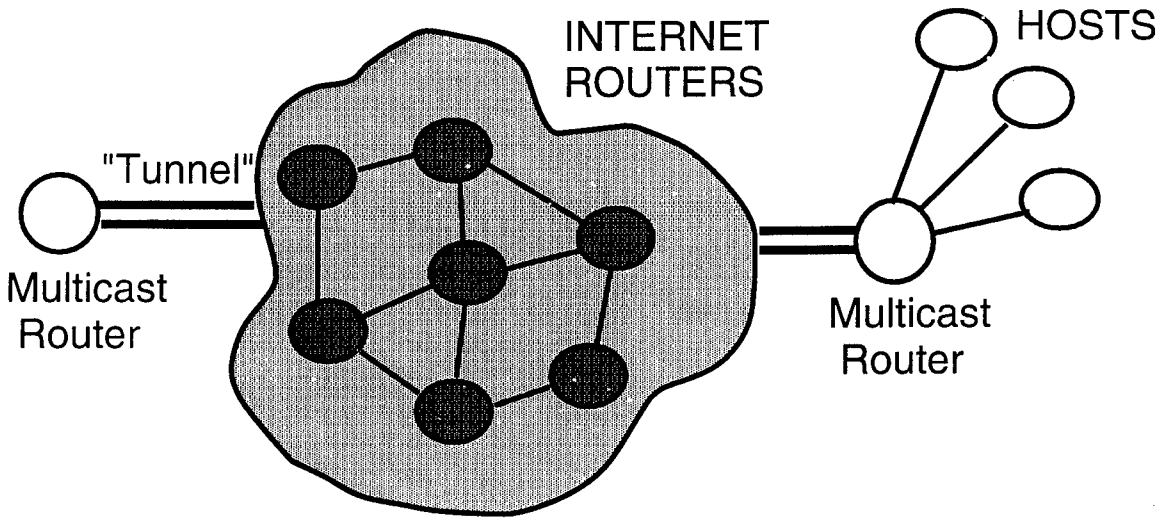


Figure 1 - An Mbone Multicast Tunnel

World-wide distribution of data has grown exponentially over recent years with the World-Wide Web as the primary tool for access. Some automation of the browsing process, using indexes that can be built periodically and active robots [http://www.cs.sunyit.edu/Robots/tkwww.html] gives users the capability to initiate a WWW query that may require access to distributed data resources. The WWW is increasingly being used for access to databases. Interfaces like GSQL [ftp://ftp.ncsa.uiuc.edu/Web/tools/gsql] are used for access to relational databases through WWW browsers. Increased availability of Databases and their usage is clearly related to the drop in price of disk storage from $1.00/MB two years ago to around $.20/MB today. The increase in access of globally distributed information will require increased use of high-bandwidth networks to support the growing demand.

Performance Measurements

A performance study compared two protocols used to access the ATM network interfaces, tcp-ip and the ATM application interface (API). Table 1 shows the difference in performance that was measured for a tcp-ip connection to an ATM interface and a connection made through the ATM API. The measurements show that the ATM API provides a significant improvement over tcp-ip. The measurements were made using the netperf program [http://www.cup.hp.com/netperf/NetperfPage.html]. Performance experiments were conducted using a FORE ATM switch with Taxi interface cards in each of two systems communicating through the switch (a Sparc 20 and an HP 755).

The tcp-ip protocol provides for reliable and robust communication through a relatively unreliable network where congestion and failures can require that each packet, even in a connection oriented protocol, may have to take a different path through the network. Packet reassembly and retransmission requests are supported well by this protocol. The overhead for this processing was small, compared to slow network speeds when networks were limited to 10 Mbits/sec. The functionality of tcp-ip is necessary for unreliable networks.

Protocols for ATM networks assume the reliability of the switches on a communication path. Advance reservation of resources allows a guaranteed availability of resources to satisfy the needs of each application. This closer relationship between the application and the ATM API allows a reduction in overhead and reservation of appropriate network resources to support high-speed communication.

Benchmark Test	Message Size (bytes)	Throughput 10^6bits/sec
TCP STREAM TEST	16384	38.88
FORE UNIDIRECTIONAL SEND TEST	4096	110.88

Table 1 - Comparison of TCP-IP and ATM/API

It is also important to consider the future demand for ATM networks. A 486Dx2/66 processor is not capable of driving an ethernet interface at more than 5Mbits/sec. under the BSD/OS Unix system [http://www.bsdi.com]. However, the table below uses standard benchmarking programs [ftp://ftp.cs.sunyit.edu/pub/BENCH] to demonstrate that a modern PC running BSD/OS can perform comparably to an HP-755. This increase in processor performance for inexpensive PCs is bound to result in increased demand for high-speed applications. Software needs to be developed to allow the global network to meet the expected increase in demand from this market.

Benchmark	HP-755	P-120 Pentium
MFLOPS	13	9
Dhrystones	55,000	125,000

Table 2 - Comparison of Unix Workstation and Modern PC

Conclusion

This report shows the potential for improved application performance through use of ATM networks when application software is adapted to use the ATM application interface and avoid the overhead for TCP/IP support of network reliability. Additional experiments need to be performed in videoconferencing applications and distributed database applications that are described above.

Acknowledgment

The author would like to thank Kurt Wiehenstroer for his help in facilitating access to the Rome Laboratory ATM network through the systems in the DISE Laboratory.

Author Biography

Scott Spetka received his Ph.D. degree in computer science from UCLA in 1989. He is currently an Assistant Professor in the Computer Science Department at the State University of New York Institute of Technology at Utica/Rome. His research interests are in the areas of distributed databases, operating systems and networks. During the last year, Scott has been developing a network of PCs running the Unix operating system. Before becoming involved in WWW research, Scott was developing SUNY Nodes, a Network-Oriented Data Engineering System. The system is used to experiment with query processing techniques.

HEAVY-HOLE SCATTERING BY CONFINED NONPOLAR OPTICAL PHONONS IN A SINGLE $Si_{1-x}Ge_x$/Si QUANTUM WELL

Gang Sun
Assistant Professor
Engineering Program/Physics Department

University of Massachusetts at Boston
100 Morrissey Blvd.
Boston, MA 02125

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

August 1995

HEAVY-HOLE SCATTERING BY CONFINED NONPOLAR OPTICAL PHONONS IN A SINGLE $Si_{1-x}Ge_x$/Si QUANTUM WELL

Gang Sun
Assistant Professor
Engineering Program/Physics Department
University of Massachusetts at Boston

ABSTRACT

Intrasubband and intersubband scattering rates of heavy holes are obtained due to confined nonpolar optical phonons in a $Si_{1-x}Ge_x$ quantum well with Si barriers. Guided and interface Ge-Si and Ge-Ge modes and unconfined Si-Si modes are considered. A continuum model is used for the two components of the ionic displacement of confined vibrations: the uncoupled s-polarized TO mode and the hybrid of the LO and p-polarized TO modes. The guided mode is obtained using the model of a quantum well with infinitely rigid barriers, and the interface mode is derived from the hydrodynamic boundary conditions. While the total intersubband scattering rates are reduced as a result of confinement, the opposite is found for the intrasubband scattering. Depending on the well width and Ge content, the intersubband scattering rates are reduced by a factor of two to four with respect to their values for no confinement. Thus, one would expect comparable enhancement in the intersubband lifetimes crucial to the population inversion in a $Si_{1-x}Ge_x$/Si intersubband laser.

HEAVY-HOLE SCATTERING BY CONFINED NONPOLAR OPTICAL PHONONS IN A SINGLE $Si_{1-x}Ge_x/Si$ QUANTUM WELL

Gang Sun

I. INTRODUCTION

The possibility of lasing due to intersubband transitions in quantum well (QW) structures is crucially dependent on the lifetimes of the involved subbands. In an earlier communication[1] lifetimes were calculated for a $Si_{1-x}Ge_x/Si$ multiple quantum well (MQW) structure due to acoustic and nonpolar optical phonon scattering. It was pointed out that, because of the absence of polar optical scattering for silicon based systems, the lifetimes and their differences are consistently an order of magnitude larger than for the GaAs/AlGaAs system, and do not show the marked decrease when the intersubband energy exceeds the optical phonon energy. For this case, the scattering was calculated assuming bulk propagating phonons corresponding to the average composition. As pointed out in that paper, it is expected that the treatment of bulk-like phonon scattering leads to an overestimate of the intersubband transition rates, and therefore provides a conservative approach in determining the subband lifetimes. In order to more accurately determine the subband lifetimes, it is necessary to take into account of confinement effects on optical phonons in heterostructures. In a $Si_{1-x}Ge_x/Si$ QW structure, one can expect Si-Si, Ge-Si, and Ge-Ge optical modes in the alloy.[2] The Ge-Si and Ge-Ge modes tend to be confined by the Si barriers as has been seen in Raman scattering experiments.[3] The phonon confinement gives rise to guided and interface modes which will scatter the carrier less than will bulk phonons for intersubband transitions, thus resulting in revised subband lifetimes. This is due to the discrete spectrum of the guided modes and the weak intersubband scattering of interface modes. However, the intrasubband scattering process tends to be enhanced due to confinement.

There has been a great deal of effort in dealing with the issue of phonon confinement in heterostructures.[2,4–15] Most of published literature have focused on confined polar optical phonons, because a large body of the heterostructures are constructed from polar semiconductor materials. Recently, the demonstration of an infrared intersubband transition laser in the InGaAs/AlGaAs material system[16] has renewed the interest of investigating the possibility of constructing lasers within the Ge/Si material system, which eventually would allow monolithic integration of optical components with advanced Si microelectronics.[17]

A self-consistent continuum theory of confined nonpolar optical phonons has been developed in an infinite plate with free boundary conditions.[4] In the present paper, we extend that continuum theory to examine the confined nonpolar optical phonons of Ge-Si and Ge-Ge vibration modes with proper mechanic boundary conditions in a $Si_{1-x}Ge_x/Si$ QW. Furthermore, we use these results to estimate the heavy-hole scattering rates by the

nonpolar optical phonons. To the best of our knowledge, there has not been any work in estimating the scattering rates due to the nonpolar optical phonons in $Si_{1-x}Ge_x/Si$ heterostructures taking into account the phonon confinement.

The three different vibration modes, namely, Si-Si, Ge-Si, and Ge-Ge, are considered separately and are given proper weights in heavy-hole scattering calculation. The Si-Si mode propagates freely throughout the structure and is therefore treated as bulk-like, while the Ge-Si and Ge-Ge modes are confined by the barriers, resulting in guided and interface modes. Approximations made in the course of the calculations for the guided and interface modes will be discussed. We will examine both intrasubband and intersuband scattering processes. We will also compare these results with those assuming bulk propagating phonons. We will show that indeed the assumption of bulk-like optical phonons overestimates the scattering rates for intersubband transitions, but tends to underestimate the intrasubband process. The difference is largely attributed to the heavy-hole interaction with interface modes which can be neglected for the intersubband process, but contributes significantly to the intrasubband process. The scattering rates due to various scattering modes will be investigated as a function of the well width and as a function of the alloy composition of the well.

II. PHONON CONFINEMENT

Phonon confinement occurs due to the lack of overlap of the bulk frequency dispersions in the adjacent heterostructure materials. Thus in a Si/Ge/Si short period superlattice, the higher frequency optical modes of Si correspond to an frequency gap in the Ge layer and are therefore confined. The Ge modes in the Ge layers are strictly resonant with the Si acoustic modes and are 'quasi-confined', but the displacement pattern is similar to that of a true confined mode and indeed the Raman intensities are comparable.[2]. The Raman spectrum has revealed that there exist Ge-Ge, Ge-Si, and Si-Si vibration modes in the $Si_{1-x}Ge_x$ alloy layer.[3] The Ge-Ge vibrations have phonon energy $37.4meV$ and the Ge-Si vibrations $50.8meV$. These two modes of the $Si_{1-x}Ge_x$ well are confined by the Si-Si modes of the barrier (photon energy $64.3meV$); however, the Si-Si optical modes of the well are unconfined, as are the Si-Si optical modes of the barrier.

One of the simplest conceptual model is to treat the QW system with infinitely rigid barrier. The boundary condition to be satisfied is the vanishing of the ionic displacement of all confined vibration modes. This is an assumption of strict confinement, yielding guided modes of the confined phonons. This assumption needs to be relaxed in order to admit interface modes. As pointed out in the continuum theory,[4] the ionic displacement of confined vibrations has two components: one is the hybrid of the LO and p-polarized TO (p-TO) modes, and other is the uncoupled s-polarized TO (s-TO) mode. These modes are defined as follows: If we consider a (x, z) plane containing the normal to the layers and the phonon wavevector $\mathbf{Q}$, then

$$\mathbf{Q} = q_x\hat{e}_x + q_z\hat{e}_z \qquad (1)$$

where $\hat{e_x}$ and $\hat{e_z}$ are unit vectors. The p-TO mode has its displacements normal to **Q** and in the plane, while the s-TO displacements are normal to **Q** and perpendicular to the plane ($||\hat{e_y}$). A description of the s-TO mode is

$$u_y = e^{iq_x x}(Ae^{iq_z z} + Be^{-iq_z z}), \tag{2}$$

while the hybrid of the LO and p-TO modes is given by

$$\begin{aligned} u_x &= e^{iq_x x}[q_x(Ce^{iq_L z} + De^{-iq_L z}) + q_T(Ee^{iq_T z} + Fe^{-iq_T z})], \\ u_z &= e^{iq_x x}[q_L(Ce^{iq_L z} - De^{-iq_L z}) - q_x(Ee^{iq_T z} - Fe^{-iq_T z})]. \end{aligned} \tag{3}$$

The z-components of the LO and TO wavevector have been distinguished by q_L and q_T, respectively.

The above choice for the ionic displacement guarantees that

$$\nabla \times \mathbf{u}_L = 0 \quad \text{and} \quad \nabla \cdot \mathbf{u}_T = 0, \tag{4}$$

which hold for isotropic materials (assumed here), allowing **u** to be decomposed into LO and TO components,

$$\mathbf{u} = \mathbf{u}_L + \mathbf{u}_T. \tag{5}$$

Since the LO and TO modes must have the same frequency to be effectively coupled, they have to satisfy, according to the bulk LO and TO dispersions,

$$\omega^2 = \omega_0^2 - \beta_L^2(q_x^2 + q_L^2) = \omega_0^2 - \beta_T^2(q_x^2 + q_T^2), \tag{6}$$

where β_L and β_T are the velocities of LO and TO dispersions, respectively.

Guided Modes

In the case of strict confinement, the boundary condition of course is that the displacements **u** vanish at the boundaries $z = (0, L)$. For the s-TO mode, this leads to

$$u_y = Ae^{iq_x x}\sin(q_z z), \qquad \text{with} \quad q_z = \frac{n\pi}{L} \tag{7}$$

where $n = 1, 2, \cdots$. This mode does not mix with other modes, nor does it give rise to the interface mode discussed later.

For the coupled LO and p-TO modes, applying the boundary conditions of $u_x = 0$ and $u_z = 0$ at $z = (0, L)$ to Eq.(3) leads to a system of four constant-coefficient linear equations for C, D, E, and F. The vanishing of the determinant of the 4×4 constant coefficient matrix gives rise to the following relation,

$$(q_T q_L + q_x^2)^2 \sin(q_T L)\sin(q_L L) = 2q_T q_L q_x^2[\cos(q_T - q_L)L - 1]. \tag{8}$$

Eq.(8) leads to solutions of guided modes consisting of coupled phase-matched LO and TO modes with wavevectors

$$q_L = \frac{n_L\pi}{L} \quad \text{and} \quad q_T = \frac{n_T\pi}{L}, \tag{9}$$

where $n_L = 1, 2, \cdots$, $n_T = 3, 4, \cdots$, and $n_T - n_L = 2, 4, 6, \cdots$. This choice of quantum numbers is due to the constraint Eq.(6) since $\beta_T < \beta_L$.

Two sets of guided mode solutions emerge and they are given in either sine or cosine form referring to the z-component of the ionic displacement. The 'sine' solution is

$$\begin{aligned} u_x &= 2Ce^{iq_x x} q_x [\cos(q_L z) - \cos(q_T z)], \\ u_z &= 2iCe^{ik_x x} [q_L \sin(q_L z) + \frac{q_x^2}{q_T} \sin(q_T z)], \end{aligned} \tag{10}$$

and a 'cosine' solution

$$\begin{aligned} u_x &= 2iCe^{iq_x x} [q_x \sin(q_L z) + \frac{q_L q_T}{q_x} \sin(q_T z)], \\ u_z &= 2Ce^{iq_x x} q_L [\cos(q_L z) - \cos(q_T z)]. \end{aligned} \tag{11}$$

It is worthwhile to point out the similarities between the allowed values of q_z, q_L, q_T and q_x for the guided modes in the case of infinitely rigid barriers assumed here and that of an infinite plate with free boundary conditions.[4] But the displacement patterns differ dramatically between the two situations.

Interface Modes

The assumption of strict confinement, that the vibration amplitudes are zero at the interfaces, rules out the possibility of interface modes. In general such modes exist and are of increasing importance in carrier scattering intrasubband transitions in comparison with guided modes. The assumption of strict confinement must therefore be relaxed to admit such modes. It is obvious that the amplitudes of interface modes at the boundaries should remain small in order to keep the assumption of strict confinement yielding the guided modes. In fact, we need only to relax the vanishing assumption of the x-component of the displacement to admit interface modes while maintaining the zero boundary condition on the z-component. A similar boundary condition has been employed in treating the polar optical phonons for a single GaAs/AlAs QW.[10]

Applying the hydrodynamic boundary conditions[5] at the interfaces to the LO and TO modes independently and considering the fact that the optical phonon frequency in the barrier region is greater than that in the well region, we obtain only a solution with even z-component for the LO mode,

$$\mathbf{u}_L = \begin{cases} Ce^{iq_x x}(q_x \hat{e_x} - q_{L2} \hat{e_z}) e^{-iq_{L2} z} & z < 0, \\ Ce^{iq_x x}(\frac{\rho_1}{\rho_2})^{1/2}(\frac{\eta_2}{\eta_1}\frac{\rho_2}{\rho_1} q_x \hat{e_x} - q_{L2}\hat{e_z}) & 0 < z < L, \\ -Ce^{iq_x x}(q_x \hat{e_x} + q_{L2}\hat{e_z}) e^{iq_{L2}(z-L)} & z > L, \end{cases} \tag{12}$$

where the subscripts 1 and 2 refer to well ($Si_{1-x}Ge_x$) and barrier (Si) regions, respectively, ρ_i is the material density, and

$$\eta_i = \omega_i^2/\omega^2 - 1, \tag{13}$$

where ω_i is the Γ-point optical phonon frequency in layer i $(=1,2)$. For both Ge-Si and Ge-Ge confined modes we have $\omega_2 > \omega_1$. The LO wavevector in the well region $q_{L1} = 0$ and that in the barrier region,

$$q_{L2}^2 = \frac{\omega_2^2 - \omega_1^2}{\beta_2^2} + (\frac{\beta_1^2}{\beta_2^2} - 1)q_x^2 \tag{14}$$

The LO mode solution is of two-dimensional bulk type with constant amplitudes in the well region propagating with wavevectors parallel to the interfaces.

Both odd and even solutions emerge for the TO mode, but the boundary condition that the z-component of $\mathbf{u}_L + \mathbf{u}_T$ vanishes at the interfaces admits only the even solution,

$$\mathbf{u}_T = \begin{cases} De^{iq_x x}(q_{T2}\hat{e_x} + q_x\hat{e_z})e^{-iq_{T2}z} & z < 0, \\ De^{iq_x x}(\frac{\rho_1}{\rho_2})^{1/2}\dfrac{1}{\cos(q_{T1}L/2)} & \\ \quad \times[-iq_{T1}\sin q_{T1}(z - L/2)\hat{e_x} + q_x \cos q_{T1}(z - L/2)\hat{e_z}] & 0 < z < L, \\ -De^{iq_x x}(q_{T2}\hat{e_x} - q_x\hat{e_z})e^{iq_{T2}(z-L)} & z > L, \end{cases} \tag{15}$$

where the wavevectors for the TO mode, according to Eq.(6)

$$\begin{aligned} q_{T1}^2 &= (\frac{\beta_{L1}^2}{\beta_{T1}^2} - 1)q_x^2 && \text{(well)}, \\ q_{T2}^2 &= (\frac{\beta_{L2}^2}{\beta_{T2}^2} - 1)q_x^2 + \frac{\beta_{L2}^2}{\beta_{T2}^2}q_{L2}^2 && \text{(barrier)}. \end{aligned} \tag{16}$$

Combining Eqs.(12) and (15) and applying that the z-component of $\mathbf{u}_L + \mathbf{u}_T$ vanishes at the interfaces, lead to $D = (q_{L2}/q_x)C$, and finally the interface mode is obtained,

$$\mathbf{u} = \begin{cases} Ce^{iq_x x}[(q_x e^{-iq_{L2}z} + \dfrac{q_{L2}q_{T2}}{q_x}e^{-iq_{T2}z})\hat{e_x} & \\ \quad +(-q_{L2}e^{-iq_{L2}z} + q_{L2}e^{-iq_{T2}z})\hat{e_z}] & z < 0, \\ Ce^{iq_x x}(\frac{\rho_1}{\rho_2})^{1/2}\{[\dfrac{\eta_2}{\eta_1}\dfrac{\rho_2}{\rho_1}q_x - \dfrac{i}{\cos(q_{T1}L/2)}\dfrac{q_{L2}q_{T1}}{q_x}\sin q_{T1}(z - L/2)]\hat{e_x} & \\ \quad +q_{L2}[\dfrac{\cos q_{T1}(z - L/2)}{\cos(q_{T1}L/2)} - 1]\hat{e_z}\} & 0 < z < L, \\ Ce^{iq_x x}\{[-q_x e^{iq_{L2}(z-L)} - \dfrac{q_{L2}q_{T2}}{q_x}e^{iq_{T2}(z-L)}]\hat{e_x} & \\ \quad +[-q_{L2}e^{iq_{L2}(z-L)} + q_{L2}e^{iq_{T2}(z-L)}]\hat{e_z}\} & z > L. \end{cases} \tag{17}$$

The dispersion of bulk Si-Si mode has little overlap with that of Ge-Si mode. The scattering process involves only long wavelength in-plane wavevectors, and the z-components of phonon wavevectors in the barriers, q_{L2} and q_{T2}, are approximately the dimension of the Brillouin zone. This is because the dispersion relations overlap only at large phonon wavevectors in the barrier region. Since $q_{L1} = 0$ and $q_x \approx 0$, the interface mode frequency $\omega \approx \omega_1$, therefore, $\eta_2/\eta_1 >> 1$. The major scattering contribution from

the interface mode is from the displacement in the well region where the heavy-holes are confined. The amplitude of the LO-component is much greater than that of the TO-component in Eq.(17) for $0 < z < L$, since

$$\frac{\eta_2}{\eta_1}\frac{\rho_2}{\rho_1} >> \frac{q_{L2}q_{T1}}{q_x^2} \tag{18}$$

considering q_{T1} also small. Therefore, the exact values of q_{L2} and q_{T2} are not crucial in determining the scattering rates because the TO mode contribution is at least two orders of magnitude less than the LO mode. The above displacement pattern Eq.(17) has been used in evaluating the interface mode scattering for the confined Ge-Si and Ge-Ge modes.

III. HEAVY HOLE SCATTERING RATES

The nonpolar optical phonon interaction Hamiltonian involving a heavy hole is[18]

$$H = \frac{(M_1 M_2)^{1/2}}{M_1 + M_2}\mathbf{D}\cdot\mathbf{u} \tag{19}$$

where M_1 and M_2 are the masses of the two atoms in the unit cell, and $\mathbf{D}$ is the optical deformation potential.

The normalization of the displacement amplitudes is carried out through the approach of equating the energy of the vibration mode with that of a simple harmonic oscillator[9] as

$$\chi^2 = \frac{S}{\Omega}\int_0^L \mathbf{u}^*\cdot\mathbf{u}dz, \tag{20}$$

where L is the well width, S is the sample surface area (in (x,y) plane), Ω is the volume of the unit cell, and χ is the normal coordinator of the oscillator.

The heavy-hole band offset is calculated taking into account the compressive strain in the $Si_{1-x}Ge_x$ well region.[19] The heavy-hole bands are decoupled from the light-hole and split-off bands at $\mathbf{k} = 0$ and can be treated independently.[20] The heavy-hole energy levels and envelope wavefunctions are obtained by the finite square well model as shown in Fig.1. The heavy-hole state can be characterized by $|\mathbf{k}, n>$ with the in-plane momentum $\mathbf{k}$ and subband index n. In the approximation of constant effective mass for heavy holes, the matrix element for the transition from state $|\mathbf{k}, n>$ to $|\mathbf{k}', n'>$ due to nonpolar optical phonon scattering is

$$< \mathbf{k}', n'|H|\mathbf{k}, n >= \begin{cases} \sqrt{\dfrac{\hbar[n(\omega_o)+1/2\mp 1/2]}{2\rho_1\omega_o SL\Delta_A(q_z)}}\delta_{\mathbf{k}'\pm\mathbf{q}\mathbf{x},\mathbf{k}} D_y G^y_{nn'}(q_z) & \text{(s-TO)} \\ \sqrt{\dfrac{\hbar[n(\omega_o)+1/2\mp 1/2]}{2\rho_1\omega_o SL\Delta_C(q_L,q_T)}}\delta_{\mathbf{k}'\pm\mathbf{q}\mathbf{x},\mathbf{k}} & \\ \quad\cdot[D_x G^x_{nn'}(q_L,q_T) + D_z G^z_{nn'}(q_L,q_T)] & \text{(hybrid)}, \end{cases} \tag{21}$$

for the s-TO mode and the hybrid of the LO and p-TO mode, respectively. $n(\omega_o)$ is the number of optical phonons at thermal equilibrium, and the upper and lower signs

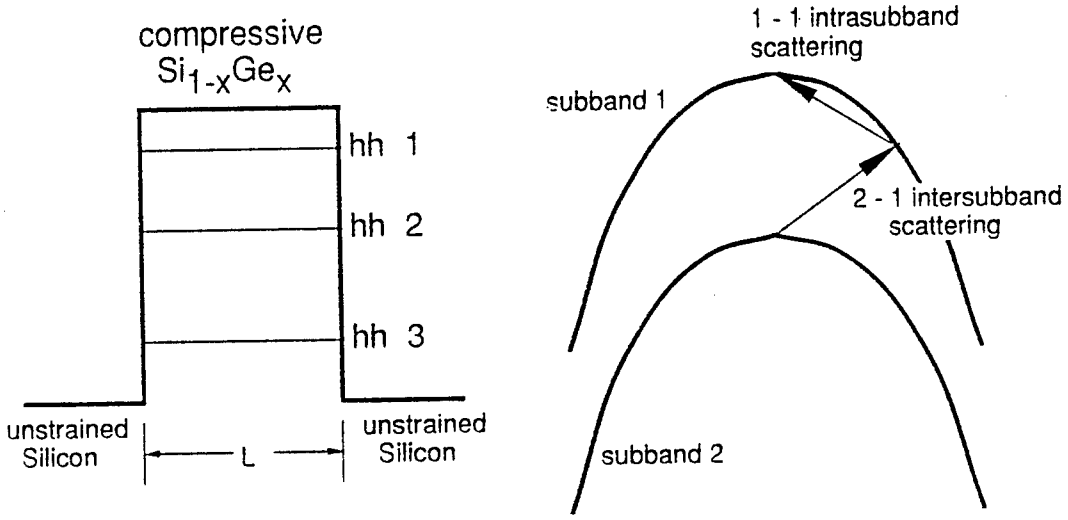


Figure 1: Heavy hole band structure and scattering processes.

refer to phonon absorption and emission, respectively. The three components of the optical deformation potential, D_x, D_y, and D_z are assumed equal to $D_o = D/\sqrt{3}$ in the calculation, in view of the assumption of isotropy. The Kronecker symbol indicates the in-plane (x, y) momentum conservation. The normalization factors are given by

$$\begin{aligned} \Delta_A(q_z) &= \frac{1}{L}\int_0^L u_y^* u_y dz \qquad \text{(s-TO)}, \\ \Delta_C(q_L, q_T) &= \frac{1}{L}\int_0^L (u_x^* u_x + u_z^* u_z) dz \qquad \text{(hybrid)}. \end{aligned} \tag{22}$$

The $G_{nn'}$-functions contain envelope wavefunctions, ψ_n and ψ'_n, from which interference effect can be obtained. Specifically,

$$G^y_{nn'}(q_z) = \int_0^L \psi_n \psi_{n'} u_y dz, \tag{23}$$

for the s-TO mode, and

$$\begin{aligned} G^x_{nn'}(q_L, q_T) &= \int_0^L \psi_n \psi_{n'} u_x dz, \\ G^z_{nn'}(q_L, q_T) &= \int_0^L \psi_n \psi_{n'} u_z dz, \end{aligned} \tag{24}$$

for the hybrid of LO and p-TO mode. It should be noted that given the proper displacement expressions, Eq.(21) for the matrix element is valid for both the guided and interface modes discussed above.

Obviously, depending on the alloy composition in the well, the three modes (Si-Si, Ge-Si, and Ge-Ge) will have different interaction strengths with the carriers in the QW

structure. Specifically, each interaction should be weaker than that for the case where it is the only existing mode and therefore needs to be accounted for properly. A crude model to approximate the relative strength of each individual mode is to assign a proper weight in the calculation of the phonon scattering rate. If we assume bonds formed between Si-Si, Ge-Si, and Ge-Ge are purely random, then we can give the following weights according to the Ge content, x, to each of the vibration modes in the $Si_{1-x}Ge_x$ well: $w_{Si} = (1-x)^2$ for Si-Si mode, $w_{GeSi} = 2x(1-x)$ for Ge-Si mode, and $w_{Ge} = x^2$ for Ge-Ge mode.

Applying the Fermi golden rule, we obtain the scattering rate due to the guided modes,

$$W^j_{nn'} = \begin{cases} \dfrac{w_j m^*_{hh}[n(\omega_o)+1/2\mp 1/2]D_o^2}{2\hbar^2\rho_1\omega_o L}\sum_{q_z}\dfrac{|G^y_{nn'}|^2}{\Delta_A} & \text{(s-TO)} \\ \dfrac{w_j m^*_{hh}[n(\omega_o)+1/2\mp 1/2]D_o^2}{2\hbar^2\rho_1\omega_o L}\sum_{q_L,q_T}\dfrac{|G^x_{nn'}+G^z_{nn'}|^2}{\Delta_C} & \text{(hybrid),} \end{cases} \tag{25}$$

where we have assumed that for the intersubband process ($n \neq n'$) the heavy holes are scattered from the bottom of their original subbands, and for the intrasubband process ($n = n'$) the heavy holes have just enough kinetic energy (equal to the Si-Si optical phonon energy) to emit an optical phonon to reach to the bottom of the same subband. w_j (j =Ge-Si, Ge-Ge) is the weight assigned to a particular vibration mode and m^*_{hh} is the heavy-hole effective mass.

For the s-TO guided mode, the summation in Eq.(25) is over all $q_z = n\pi/L$ limited by

$$q_z^2 < (\frac{\pi}{a_j})^2 - q_x^2, \tag{26}$$

where a_j is the lattice constant and q_x is given by the constraint of the in-plane momentum conservation. For the hybrid guided mode, the summation is restricted to those combinations of $q_L = n_L\pi/L$ and $q_T = n_T\pi/L$ which, according to Eq.(6), yield discrete values of q_x satisfying the in-plane momentum conservation. It should be pointed out that an unique final state heavy-hole wavevector $\mathbf{k}'$ (initial $\mathbf{k} = \mathbf{0}$) will result if the phonon frequency dispersion is neglected, corresponding to a constant optical phonon energy. Then the in-plane momentum conservation would be impossible since the discrete values of q_x will not be able to exactly match the unique value of $\mathbf{k}'$. However, taking into account of the phonon dispersion, a region of heavy-hole wavevector $\mathbf{k}'$ will be obtained and the scattering process is permitted as long as q_x lies in that region.

There is no s-TO interface mode, and the scattering rate due to the hybrid interface mode can be given, similarly,

$$W^j_{nn'} = \frac{w_j m^*_{hh}[n(\omega_o)+1/2\mp 1/2]D_o^2}{2\hbar^2\rho_1\omega_o L\Delta_C}|G^x_{nn'}+G^z_{nn'}|^2, \tag{27}$$

where $G^x_{nn'}$ and $G^z_{nn'}$ are functions of q_{T1}, q_{L2}, and q_{T2} with $q_{L1} = 0$. The determination of the phonon wavevectors in the well and barrier regions have been discussed in section

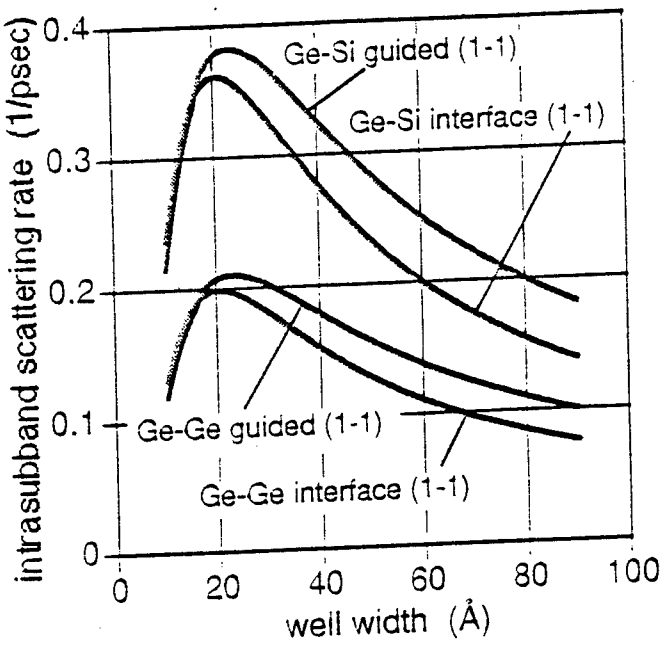


Figure 2: Intrasubband scattering rates in the lowest subband ($n = 1$) due to Ge-Si and Ge-Ge guided and interface modes as a function of well width with $x = 0.5$.

II. Eqs.(25) and (27) are used to evaluate both the intrasubband and intersubband transitions due to the confined phonons of Ge-Si and Ge-Ge modes. However, since the Si-Si vibration should be bulk-like, the calculation of its scattering rate is trivial.[21]

Intrasubband Scattering

The intrasubband scattering rate was calculated assuming that the transition originated from the heavy-hole state with a kinetic energy equal to the Si phonon energy to the bottom of the same subband. Therefore, only the process of phonon emission is considered in the calculation. All results are obtained at room temperature. The structure parameters are varied within the limits for producing metastable strained $Si_{1-x}Ge_x$ alloy on the Si substrate.[22] Fig.2 shows the intrasubband transition rates (1-1) within the ground-state heavy-hole subband ($n = 1$) due to the Ge-Si and Ge-Ge guided and interface modes as a function of the well width for a $Si_{0.5}Ge_{0.5}$/Si QW. It can be seen that for the intrasubband process the strength of interface mode scattering is about same order of magnitude as that of corresponding guided mode. This is in contrast with the intersubband process discussed later. The scattering rates of both guided and interface modes for Ge-Si and Ge-Ge vibrations increases with the increase of the well width when it is narrow ($L < 25\text{Å}$). This is due to the increase of the interference G-function with $n = n' = 1$ given in Eqs.(23) and (24) as the envelope function for subband 1 becomes more confined in the well region for larger well width. However, the G-function increase becomes negligible as the well width increases further since the envelope function has

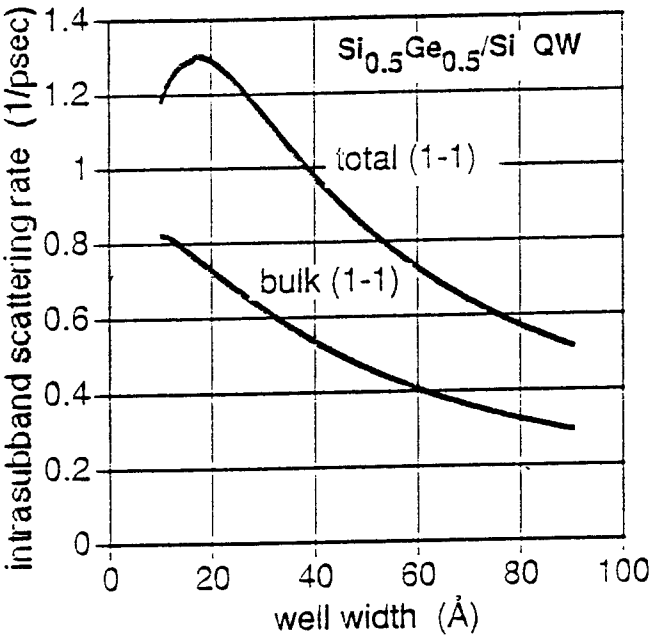
mostly been confined within well region. Examining the interface scattering rate given in Eq.(27), one would expect it to decrease with the further increase of the well width, which is clearly demonstrated in Fig.2 for both Ge-Si and Ge-Ge phonon interactions. A similar decrease of the interface scattering rate in a GaAs/$Al_xGa_{1-x}As$ QW has been reported.[7]

For the guided mode, the scattering rate is given by summing the contributions from the s-TO and hybrid of LO and p-TO modes. However, the hybrid interaction with the heavy holes is much weaker than the s-TO mode, because of the requirements on the hybrid mode to simultaneously satisfy the in-plane momentum conservation and the frequency dispersion Eq.(6). This excludes the interaction of most hybrid modes with the heavy holes, leaving the s-TO modes as the dominant scattering mechanism. In fact, the number of allowed s-TO vibration modes will increase with the increase of the well width. It is easy to see from Eq.(23) that for the intrasubband process ($n = n'$) the interference G-function is nonzero only for even s-TO modes ($n_{q_z} = 1, 3, 5, \cdots$), and the effective contribution is from the lowest order mode $n_{q_z} = 1$ with higher order modes being at least two orders of magnitude less. The guided mode scattering for the intrasubband process is therefore practically a single s-TO mode interaction and behaves similarly to the interface mode. The difference between the Ge-Si and Ge-Ge vibrations is mainly due to the different weights assigned to them according to the Ge content in the Si_xGe_{1-x} alloy. Our calculation also showed that the (2-2) intrasubband scattering process has very similar behavior with the interface mode scattering slightly higher than the guided mode, but overall smaller rates compared to the (1-1) process.

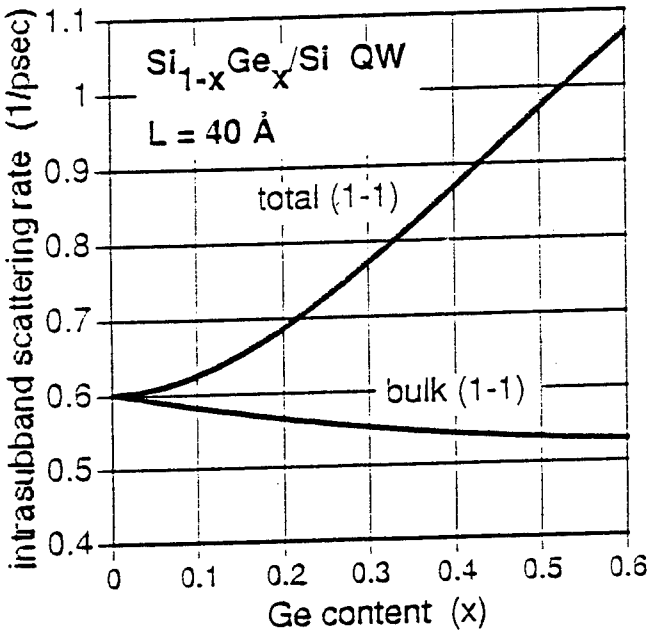
Fig.3(a) shows the total scattering rate as a function of the well width, which is given by summing the contributions from the Ge-Si, Ge-Ge, and Si-Si (treated as bulk-like) vibrations. Also shown in Fig.3(a) is the result calculated neglecting the phonon confinement and treating all vibration modes as bulk-like for comparison. It can be seen that the phonon confinement actually enhances the intrasubband scattering process. However, the opposite is true for the intersubband process to be shown below. Fig.3(b) shows the same result as a function of the Ge content in the alloy for the well width of $40\mathring{A}$. The difference increases with the increase of the Ge content since the contributions from the confined phonons will increase while that from the unconfined Si-Si vibration decreases.

Intersubband Scattering

The intersubband scattering rate was calculated assuming that the transition originated from the bottom of a subband n with zero kinetic energy to another subband ($n' \neq n$). Both phonon absorption and emission processes are considered in the calculation. Fig.4(a) shows the (2-1) scattering rate with the Si-Si, Ge-Si, and Ge-Ge components as a function of well width for a $Si_{0.5}Ge_{0.5}$/Si QW. Subband 2 appears at the well width of $20\mathring{A}$. The total scattering is obtained by summing the contributions derived from the Ge-Si, Ge-Ge, and Si-Si vibrations including both guided and interface modes. But the interface mode scattering is at least two orders of magnitude weaker than



(a)



(b)

Figure 3: Total intrasubband scattering rates in the lowest subband ($n = 1$) due to Ge-Si, Ge-Ge confined modes and Si-Si unconfined modes, compared to the bulk intrasubband scattering rate assuming no confinement. (a) as a function of well width with Ge content $x = 0.5$; (b) as a function of x for a well width of 40Å.

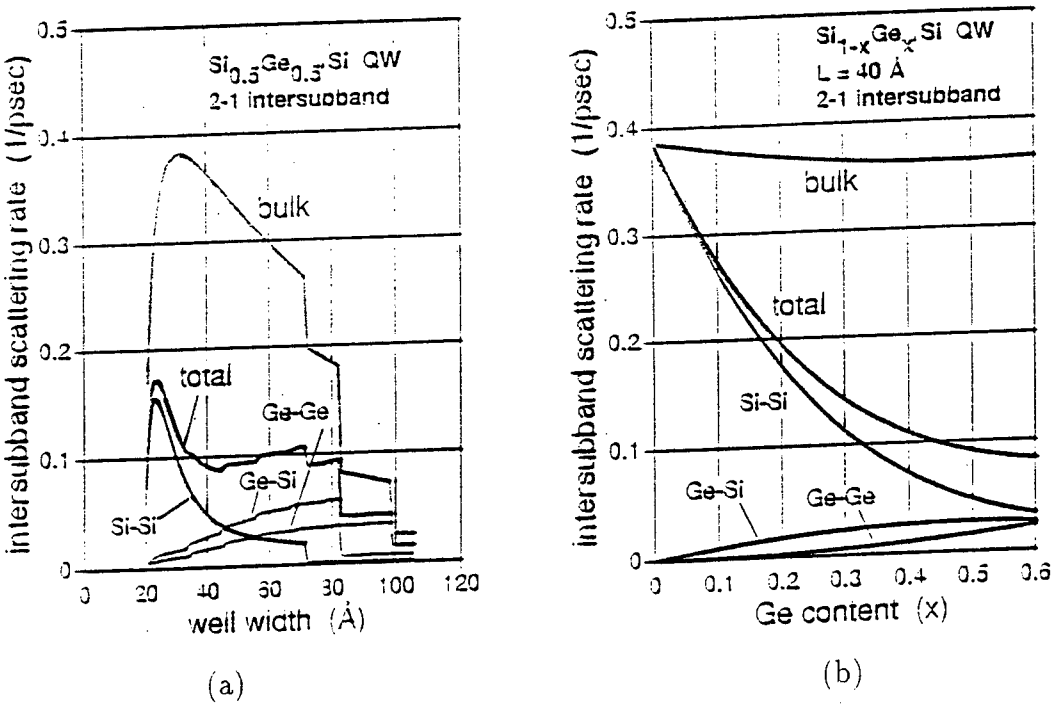


Figure 4: Total intersubband scattering rates from subband 2 to 1 due to Ge-Si, Ge-Ge confined modes and Si-Si unconfined modes, compared to the bulk intersubband scattering rate assuming no confinement. (a) as a function of well width with Ge content $x = 0.5$; (b) as a function of x for a well width of $40\mathring{A}$.

the guided mode for the intersubband process. Similar weakness of the interface mode has been shown in a $GaAs/Al_xGa_{1-x}As$ QW.[7] The result assuming bulk-like phonons is also shown in Fig.4(a) for comparison. The sharp increase of the scattering rates in the small well width region ($L < 25\mathring{A}$) is once again due to the increasing confinement in the well region of subband 2 which leads to an increase of the interference G-function. Both scattering intensities from Si-Si and bulk-like phonons reduces with further increase of the well width. Since initially the narrow well width only allows a small number of guided modes, the guided mode scattering is weak compared to the Si-Si scattering and the total scattering rate follows the dependence of Si-Si vibration. As the number of allowed guided modes increases with increasing well width, the guided mode scattering actually surpasses the Si-Si component, leading to an increase of the total scattering rate. The small discontinuous incremental steps in the Ge-Si, Ge-Ge, and therefore total scattering curves, are due to the discrete nature of the increase in the number of allowed guided modes as the well width increases. The sudden drops in the curves occur when the energy separation between the two subbands reduces to less than one of the three phonon energies corresponding to different vibrations. Specifically, the drops occurred at $L = 72,\ 83,\ 99\mathring{A}$ correspond to Si-Si phonon ($64.3meV$), Ge-Si phonon ($50.8meV$), and Ge-Ge phonon ($37.4meV$), respectively. It can be seen in Fig.4(a) that the phonon confinement reduces the scattering rate by a factor of two to four as compared to the bulk-like phonons. It is therefore important to take into account the phonon confinement

effect in estimating the subband lifetimes.

Fig.4(b) shows the (2-1) scattering rates as a function of Ge content in the $Si_{1-x}Ge_x$ alloy for a well thickness of $40\mathring{A}$. The contributions from the guided modes increases with the Ge content while the unconfined Si-Si component reduces. As a result, the difference between the total scattering rate and the bulk-like rate widens with Ge content.

IV. SUMMARY AND DISCUSSION

In an earlier paper we treated phonon-induced intersubband transitions in $Si_{1-x}Ge_x$/Si MQW structures.[1] The absence of polar optical scattering resulted in subband lifetimes an order of magnitude larger than those in MQWs of III-V semiconductors and without the precipitous decrease at the optical phonon threshold. This earlier work assumed bulk-like, three dimensional phonons. Noting strong experimental evidence for phonon confinement in $Si_{1-x}Ge_x$/Si MQWs, the present paper extends the earlier treatment to the case in which phonons are confined.

A single $Si_{1-x}Ge_x$ QW with Si barrier is considered. The decoupled heavy-hole subbands and wavefunctions are determined by the finite square well model. Assigning statistical weights to the Si-Si, Ge-Si, and Ge-Ge vibrations in the QW assuming a purely random bond model, the Ge-Si and Ge-Ge modes are treated as confined while the Si-Si modes are extended. The confined modes are of two kinds, guided modes and interface modes. The guided modes consist of the s-TO and coupled LO and p-TO modes, with the boundary condition that the displacements vanish at the interfaces. The interface mode is obtained applying hydrodynamic boundary conditions to the LO and TO modes separately and then requiring that the z-component of their sum (not the x-component) vanish at the interfaces. The LO mode is a two-dimensional bulk-type solution with constant amplitude in the well and propagating parallel to the interface; the TO mode is an even function (referring to the z-component) with a sinusoidal spatial dependence in the growth (z) direction so that the boundary conditions at the interfaces can be satisfied.

The Ge-Ge optical phonon spectrum in the well region has no overlap with that of Si-Si optical phonons in the barriers, but instead overlaps with that of the acoustic Si phonons. The model for the interface mode derived from the optical phonon dispersion relations using the hydrodynamic boundary conditions is not without criticism. Further investigation of this subject is underway to provide a more accurate estimate of the intrasubband scattering rates. But we would still expect the intersubband process to be dominated by the guided-mode scattering, which ultimately determines the subband lifetimes.

The intrasubband and intersubband scattering rates are calculated as a function of the QW structure parameters: Ge content, x, in the $Si_{1-x}Ge_x$ well and the well width, L. These parameters are varied within the limits reportedly to produce the

metastable strained layers of $Si_{1-x}Ge_x$ alloy on Si substrates without introducing a large number of misfit defects. We choose the Ge content of $x = 0.5$ as we vary the well width to give a fair weight of the Ge-Si and Ge-Ge confined modes. We use the fixed well width of $40\mathring{A}$ to allow at least two confined heavy-hole subbands as we vary the Ge content. Our results are as follows. For intrasubband scattering, the scattering by the interface mode is nearly the same as that for the guided mode (Fig.2), and the total intrasubband scattering with confinement is larger than that without confinement (Fig. 3). However for intersubband scattering, the interface mode scattering is at least two orders of magnitude smaller than that for the guided mode. Finally, as can be seen from Fig.4, the total intersubband scattering rate with confinement is a factor of two to four smaller than that which would be obtained assuming bulk phonons (i.e. no confinement). Thus, one would expect a factor of two to four enhancement in the intersubband lifetimes which determine population inversion in an intersubband laser. Since phonon confinement is well documented for the GaAs/AlAs system,[23] lifetime enhancements should also be present here, though the scattering due to the polar modes is stronger and where the splitting of the LO and TO modes will require a treatment different from the present nonpolar case.

ACKNOWLEDGEMENTS

The author wish to thank Dr. Richard A. Soref for his support during the period of this summer research program, without which this research accomplishment would not have been possible. The author would also like to acknowledge Dr. Lionel Friedman for many helpful discussions.

References

[1] G. Sun, L. Friedman, and R.A. Soref, Appl. Phys. Lett. **66**, 3425 (1995)

[2] A. Fasolino, E. Molinari, and J.C. Mann, Phys. Rev. B **39**, 3923 (1989)

[3] S. C. Jain and W. Hayes, Semicond. Sci. Technol. **6**, 547 (1991)

[4] B. K. Ridley, Phys. Rev. B **44**, 9002 (1991)

[5] M. Babiker, J. Phys. C: Solid State Phys. **19**, 683 (1986)

[6] M. Babiker, A. Ghosal, and B. K. Ridley, Superlattices and Microstructures **5**, 133 (1989)

[7] B. K. Ridley, Phys. Rev. B **39**, 5282 (1989)

[8] M. P. Chamberlain, M Cardona, and B. K. Ridley, Phys. Rev. B **48**, 14356 (1993)

[9] B. K. Ridley, Phys. Rev. B **47**, 4592 (1993)

[10] N. C. Constantinou and B. K. Ridley, Phys. Rev. B **49**, 17065 (1994)

[11] B. K. Ridley, Appl. Phys. Lett. **66**, 3633 (1995)

[12] F. A. Riddoch and B. K. Ridley, Physica **134B**, 342 (1985)

[13] A. Fasolino, E. Molinari, and A. Qteish, Condensed Systems of Low Dimensionality, Edited by J. L. Beeby et al., Plenum Press, New York, 495 (1991)

[14] E. Molinari and A. Fasolino, Appl. Phys. Lett. **54**, 1220 (1989)

[15] E. Molinari, A. Fasolino, and K. Kunc, Superlattices and Microstructures **2**, 397 (1986)

[16] J. Faist, F. Capasso, D. L. Sivco, A. L. Hutchinson, C. Sirtory, and A. Y. Cho, Science **264**, 553 (1994)

[17] R. A. Soref, Proc. IEEE **81**, 1687 (1993)

[18] B. K. Ridley, Quantum Processes in Semiconductors, Clarendon Press, Oxford, Chapter 3, (1982)

[19] A. Kahan, M. Chi, and L. Friedman, J. Appl. Phys. **75**, 8012 (1994)

[20] G. Sun and L. Friedman, Superlattices and Microstructures **17**, No.3, (1995)

[21] B. K. Ridley, J. Phys. C: Solid State Phys. **15**, 5899 (1982)

[22] J. Bean, Proc. IEEE **80**, 571 (1992)

[23] D. Levi, Shu-Lin Zhang, M. V. Klein, J. Klem, and H. Morkoc, Phys. Rev. B **36**, 8032 (1987)

A HYBRID MM/GTD NUMERICAL TECHNIQUE FOR LOSSY DIELECTRIC ROUGH SURFACE SCATTERING CALCULATIONS

James C. West
Associate Professor
School of Electrical Engineering

Oklahoma State University
202 ES Stillwater, OK 74078

Final Report for
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

August 1995

A HYBRID MM/GTD NUMERICAL TECHNIQUE FOR LOSSY DIELECTRIC ROUGH SURFACE SCATTERING CALCULATIONS

James C. West
Associate Professor

James Michael Sturm
Graduate Research Assistant

School of Electrical and Computer Engineering
Oklahoma State University

Abstract

A hybrid numerical technique combining the moment method and the geometrical theory of diffraction has been extended to allow the calculation of electromagnetic scatter from lossy dielectric surfaces. The hybrid technique eliminates the non-physical edge effects that are introduced in standard moment method implementations, thereby allowing the application at extreme grazing angles. The dielectric surface is represented using impedance boundary conditions. Sample calculations demonstrate the reduction in scattering from a rounded-apex wedge when the surface conductivity is reduced. The technique should allow more realistic calculation of the scattering from land and water surfaces than can be obtained using a perfectly conducting surface.

I. INTRODUCTION

Traditional implementations of the moment method for calculating the electromagnetic scattering from rough surfaces have been limited to application at moderate to large illumination grazing angles [1, 2]. Finite computer resources limit the length of the surface that can be numerically modeled, introducing non-physical edges in the scattering surface that can lead to unrealistic diffractive scattering. These "edge-effects" are often avoided using an illumination weighting function that reduces the incident field to negligible levels at the edges. Thorsos [3] showed that electromagnetically valid weighting functions can yield unrealistic surface illumination if the modeled surface is insufficiently large. Unfortunately, the required surface length increases dramatically with decreasing grazing angle (increasing incidence angle), limiting the smallest grazing angle at which the technique can be applied.

West [4] implemented a hybrid numerical technique combining the moment method (MM) and the geometrical theory of diffraction (GTD) that overcomes many of the limitations of the traditional moment method when calculating the scattering from perfectly conducting surfaces. In this approach the surface is extended to infinity, thereby eliminating the artificial edges. The technique has been used to investigate the effects of surface self-shadowing on the backscattering from perfectly conducting surfaces approximating rough ocean waves [4] and the effects of multi-path on backscattering from breaking ocean waves [5]. This hybrid technique has been enhanced to allow the application to lossy dielectric surfaces, thereby allowing accurate prediction of the scattering from rough land surfaces as well as sea surfaces. A detailed description of the enhanced technique is given below, as well as sample calculations of the scattering from lossy dielectric objects.

II. OVERVIEW OF TECHNIQUE

A. *Perfectly Conducting Surfaces*

Application of the hybrid MM/GTD numerical technique to scattering from perfectly conducting surface is described in [4]. A detailed review is given here. Adapted from the technique described by Burnside *et al.*[6], it is quite similar to the standard moment method in that scattering from the surface is found by first numerically solving an integro-differential equation to yield the surface current. In both methods, the unknown surface current is represented as a summation of known basis functions. The weighting coefficients associated with each basis function that give the "best" approximate solution are obtained using the moment method. The primary difference between the two techniques is that the hybrid approach uses *a priori* knowledge of the current obtained from GTD to define well behaved basis functions and additional source terms that allow the treatment of special infinitely long surfaces, thereby avoiding the artificial edge effects introduced in the standard MM. The surface current is then radiated to yield the scattered field.

For the one-dimensionally rough surfaces considered here, vertically polarized scattering is best described by the magnetic field integral equation (MFIE) [7]:

$$\begin{aligned} H^i(l) &= 0.5J_s(l) + j\frac{\beta}{4}\int J_s(l')\,(\hat{\mathbf{n}}'\cdot\rho')\,\mathrm{H}_1^{(2)}(\beta|\rho-\rho'|)\,dl' \\ &= L_M[J_s(l)], \end{aligned} \tag{1}$$

where l is the arc length along the scattering surface, $H^i(l)$ is the incident magnetic field at the scattering surface, $J_s(l)$ is the unknown surface current to be found, β is the free space wave number, ρ is the position vector of the observation point, ρ' is the position vector of the source point, $\hat{\mathbf{n}}'$ is the normal unit vector at the source point, and $\mathrm{H}_1^{(2)}$ is the first-order Hankel function of the second type. The integration is the principal value integral (avoiding the singularity where $l = l'$) over the entire surface. Horizontally polarized scattering is more

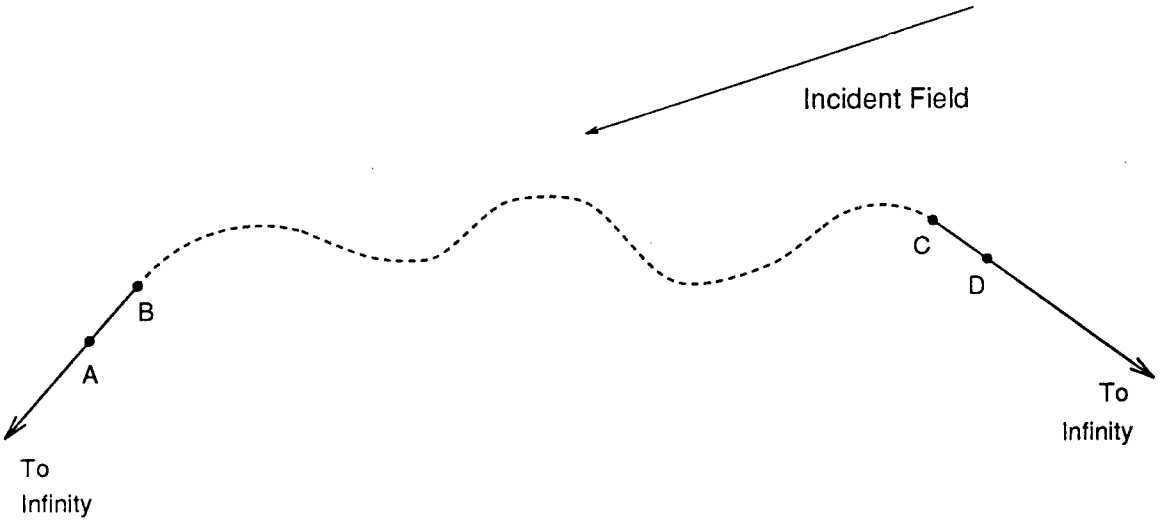


Figure 1: Arbitrary scattering surface.

easily treated by the electric field integral equation (EFIE):

$$\begin{aligned} E^i(l) &= \frac{\beta\eta_0}{4}\int J_s(l')\mathrm{H}_0^{(2)}(\beta|\rho-\rho'|)\,dl' \\ &= L_E[J_s(l)], \end{aligned} \tag{2}$$

where η_0 is the intrinsic wave impedance of free space. The MFIE and EFIE can be written in the single notation

$$F^i(l) = L_X[J_s(l)], \tag{3}$$

where F is either E or H and X is either E or M. In the standard moment method, the infinite integrations in equations (1) and (2) are truncated to be over a finite surface arc length. The current on the modeled length L is then divided into a weighted summation of adjacent pulse basis functions, and the moment method is used to find the associated weighting coefficients. It is the truncation of the integrations that lead to the non-physical edge effects.

The hybrid technique is applied to one-dimensionally rough surfaces of the form shown in Figure 1. The dashed section of the surface represents the actual rough surface while the solid line represents infinitely long, planar extensions. The extensions are chosen such that all

points on the actual surface are shadowed from all points on the extension (except of course at the intersection points B and C). Because the surface is arbitrary, little is known initially about the current between points A and D. Thus, the current in this region is described using standard MM pulse basis functions with impulse testing functions (yielding point matching) centered on the basis functions.

Since the extensions are shadowed from the arbitrary surface points, the fields at the surface of the extensions can be entirely described as the sum of a field diffracted from point B or C plus the geometrical optical (GO) incident and reflected fields:

$$F^t = F^i + F^s = F^{GO} + F^d, \tag{4}$$

where F^t is the total field, F^i is the incident field, F^s is the scattered field, F^{GO} is the gemetrical optics incident and reflected fields, and F^d is the diffracted field. The current on the extension is obtained by applying the surface boundary conditions to equation (4), yielding the physical optics current associated with the GO fields plus an additional current component associated with the diffracted field (the "diffraction-field current"):

$$J_s = J_{PO} + J_d. \tag{5}$$

Since the extension is flat and perfectly conducting, the PO current is known exactly *a priori.* (Note that if the extension is shadowed from the incident field the PO current is simply zero). However, the diffracted field, and therefore the diffraction-field current, is not known initially and must be determined using the moment method. Since it extends to infinity, use of ordinary sub-domain MM basis functions to describe this current would lead to an infinite order system of linear equations that cannot be solved. Instead it is recognized that at distances far enough away from the diffraction point the diffracted field is ray optical. Thus,

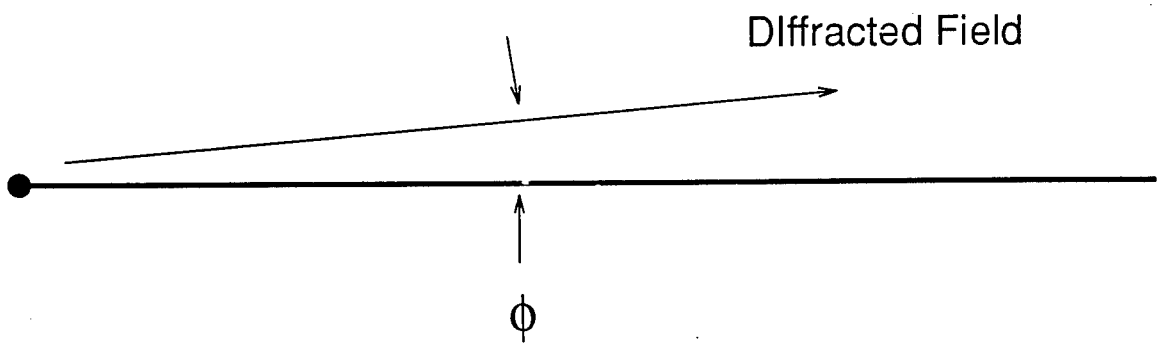


Figure 2: Diffracted field in the vicinity of the extensions.

the form of the diffracted field at the extension beyond points A or D is given by

$$F^d = F_0 \frac{e^{-jkr}}{r} f(\phi), \tag{6}$$

where r is the distance from the diffraction point and $f(\phi)$ is an arbitrary function of the angular cylindrical coordinate with the diffraction point as the origin, as shown in Figure 2. Applying the surface boundary condition $\mathbf{J_s} = \hat{\mathbf{n}} \times \mathbf{H}$ yields the diffraction currents

$$\begin{aligned} J_d &= J_0 \frac{e^{-jkr}}{\sqrt{r}}, \qquad \text{(vertical polarization)} \\ &= J_0 \frac{e^{-jkr}}{r^{1.5}}, \qquad \text{(horizontal polarization).} \end{aligned} \tag{7}$$

We now see that a single basis function of the form of equation (7) can be used to include the diffraction current from the diffraction point to infinity in the hybrid numerical technique. This, combined with the known physical optics currents, entirely describes the current on the infinite extensions. Since there are no discontinuities on the modeled surface, no artificial edge effects are introduced.

The current on the entire surface may now be written as

$$J_s = J_{MM} + J_D + J_{PO}, \tag{8}$$

where J_{MM} is the current between points A and D described by ordinary MM pulse basis functions:

$$J_{MM} = \sum_{m=1}^{N} \alpha_m P(l - l_m), \tag{9}$$

where $P(l - l_m)$ is a pulse function centered at l_m and α_m are unknown weighting coefficients to be found via the moment method. J_D includes both diffraction current terms:

$$J_D = \alpha_{N+1} J_d|_{back} + \alpha_{N+2} J_d|_{front}, \tag{10}$$

and J_{PO} is the physical optics current on the front and back faces given by

$$\mathbf{J_{PO}} = 2\hat{\mathbf{n}} \times \mathbf{H^i}|_{back} + 2\hat{\mathbf{n}} \times \mathbf{H^i}|_{front}. \tag{11}$$

Substituting equation (8) into equation (3) gives

$$F^i = L_X[J_{MM} + J_D + J_{PO}]. \tag{12}$$

Because the J_{PO} is entirely known *a priori* and $L_X[\,]$ is a linear operator, the physical optics term may be moved to the left hand side, giving

$$F^i - L_X[J_{PO}] = L_X[J_{MM}] + L_X[J_D]. \tag{13}$$

Thus, the physical optics current simply appears as a field source term in the hybrid technique. Evaluating equation (13) at the centers of the basis functions (point matching or collation), plus at two additional points on the extensions yields $N + 2$ algebraically linear equations with $N + 2$ unknowns. Solving this system yields the moment weighting coefficients α_m,

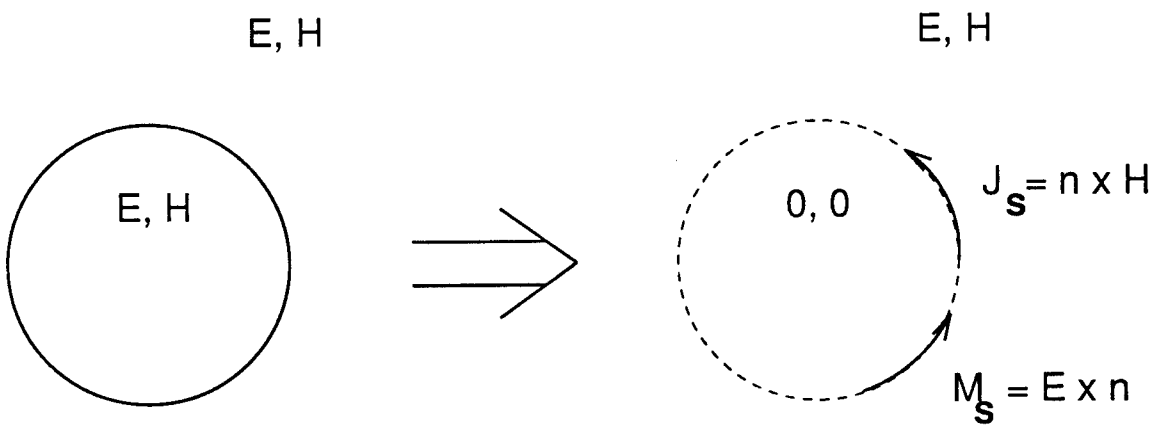


Figure 3: Equivalent problem to be solved with lossy dielectric scatterer.

completing the MM solution of the current. The far field scatter is then determined from

$$F^s = -L_X[J_{MM} + J_D + J_{PO}]\Big|_{r\to\infty}. \tag{14}$$

B. *Lossy Dielectric Surfaces*

When the scattering surface is perfectly conducting a true surface current exists. Thus, the moment method solves the physical scattering problem directly. When the surface is not perfectly conducting a surface current cannot be supported; the field penetrates the surface and a volume current density exists. The moment method is not well suited for direct application to volume current problems. Instead, the equivalence principle [7] is applied as shown in Figure 3, yielding both electric (J) and magnetic (M) surface current densities that radiate the desired scattered field. Although the equivalent problem includes only surface currents, the moment method still cannot be applied directly since the unknown electric and magnetic currents are co-located. Instead, the magnetic current is expressed in terms of the electric current using impedance boundary conditions [8]. Assuming that the conditions

$$|N| \gg 1, \qquad |\mathrm{Im}(N)k\rho_l| \gg 1 \tag{15}$$

where N is the complex refractive index of the scattering medium and ρ_l is the radius of curvature of the surface, are met everywhere on the surface, the field penetrating into the surface propagates as a plane wave in the negative surface normal direction. The two surface current components can then be related by [9]

$$\mathbf{M} = -Z_s \hat{\mathbf{n}} \times \mathbf{J}, \tag{16}$$

where Z_s is the intrinsic wave impedance of the lossy dielectric.

Applying duality to equations (1) and (2) to determine the near-field radiation of the magnetic current density and using equation (16), it is straightforward to show that the appropriate two-dimensional MFIE for determining vertically polarized scattering from a lossy dielectric scatterer is [10]

$$H^i(l) = L_M[J_s(l)] - \frac{Z_s}{\eta_0^2} L_E[J_s(l)], \tag{17}$$

Similarly, with a lossy dielectric surface the EFIE becomes

$$E^i(l) = L_E[J_s(l)] - Z_s L_E[J_s(l)]. \tag{18}$$

Since equations (17) and (18) each include only the unknown surface current J_s (and not M_s) they are well suited to solution using moment method techniques.

The hybrid MM/GTD technique can be extended to apply to equations (17) and (18) to find the scattering from lossy dielectric surfaces of the type shown in Figure 1 with little modification. The surface current between points A and D is again divided into pulse basis functions as described in equation (9), and the diffraction-current basis functions are unchanged from equation (7) since the diffracted field is still ray optical at suitable distances from the diffraction point [11]. The physical optics current does need to be modified slightly

since the surface is no longer perfectly conducting:

$$\mathbf{J_{PO}} = (1-\Gamma)\hat{\mathbf{n}} \times \mathbf{H^i}|_{back} + (1-\Gamma)\hat{\mathbf{n}} \times \mathbf{H^i}|_{front}. \tag{19}$$

where Γ is the appropriate parallel (vertical) polarized or perpendicular (horizontal) polarized reflection coefficient on the front and back extensions. (Note that equation (19) reduces to (11) with a perfectly conducting surface.) Substituting equation (8) (with the modified J_{PO}) into equation (17) and moving the known terms to the left hand (source) side yields

$$H^i(l) - L_M[J_{PO}(l)] + \frac{Z_s}{\eta_0^2} L_E[J_{PO}(l)] \tag{20}$$

$$= L_M[J_{MM}(l) + J_D(l)] - \frac{Z_s}{\eta_0^2} L_E[J_{MM}(l) - J_D(l)]. \tag{21}$$

Similarly, the EFIE becomes

$$E^i(l) - L_E[J_{PO}(l) + Z_s L_M[J_{PO}(l)] \tag{22}$$

$$= L_E[J_{MM}(l) + J_D(l)] - Z_s L_M[J_{MM}(l) - J_D(l)]. \tag{23}$$

Both equations (20) and (22) can be evaluated at the $N+2$ matching points, and the resulting linear system algebraic equations solved to give the unknown coefficients α_n, completing the numerical solutions. The far-field scattering from the surface is then found by evaluating

$$H^s = -L_M[J_{MM} + J_D + J_{PO}]\Big|_{r\to\infty} + Z_s L_M[J_{MM} + J_D + J_{PO}]\Big|_{r\to\infty} \tag{24}$$

or

$$E^s = -L_E[J_{MM} + J_D + J_{PO}]\Big|_{r\to\infty} + \frac{Z_s}{\eta_0^2} L_M[J_{MM} + J_D + J_{PO}]\Big|_{r\to\infty} \tag{25}$$

1) *Implementation considerations:* The majority of the Fellowship was spent deriving efficient implementations of equations (20) through (25). The findings are summarized

here.

Evaluation of $L_X[J_{PO}]$ proved to be the most numerically intensive operation in the hybrid technique. It requires the integration to infinity of an integrand that oscillates rapidly and decays slowly, and therefore converges quite slowly. Moreover, these terms must be evaluated at each incidence angle examined since it appears as a source term. It was found that the convergence can be dramatically increased by evaluating it as an infinite series and applying the epsilon convergence-acceleration algorithm [12].

Evaluation of $L_X[J_D]$ also involves the infinite integration of a rapidly oscillating, slowly converging integrand. This also proved well suited to acceleration via the epsilon algorithm. Evaluation of $L_X[J_{MM}]$ is unchanged from that given by Axline and Fung [1].

The radiation of J_{MM} in equations (24) and (25) was accomplished using the far-field approximations given by Axline and Fung [1]. Also using these approximations the integrations in $L_X[J_D(l)]\Big|_{r\to\infty}$ can be evaluated in terms of Fresnel integrals using the routines of Press et al. [13]. Evaluation of $L_X[J_{PO}(l)]\Big|_{r\to\infty}$ required more consideration. Using the approximation of Axline and Fung yields an integrand that does not decay out to infinity, and therefore technically has no solution. Instead, the exact integral was evaluated at a very large (but finite) observation range using an asymptotic end-point expansion [14].

III. APPLICATION

The numerical routines were first tested by finding the scattering from a lossy dielectric cylinder. While a cylinder is not a surface of the type shown in Figure 1 and does not require use of the hybrid MM/GTD technique, it does give a test case in which the impedance boundary moment-method scattering can be compared with an exact solution. A cylinder with a complex dielectric constant of 70 - j 50 was chosen to approximate the properties of sea water at microwave frequencies. The results are shown in Figure 4. The moment calculations agree with the exact solution to within 1 dB at all scattering angles at both horizontal

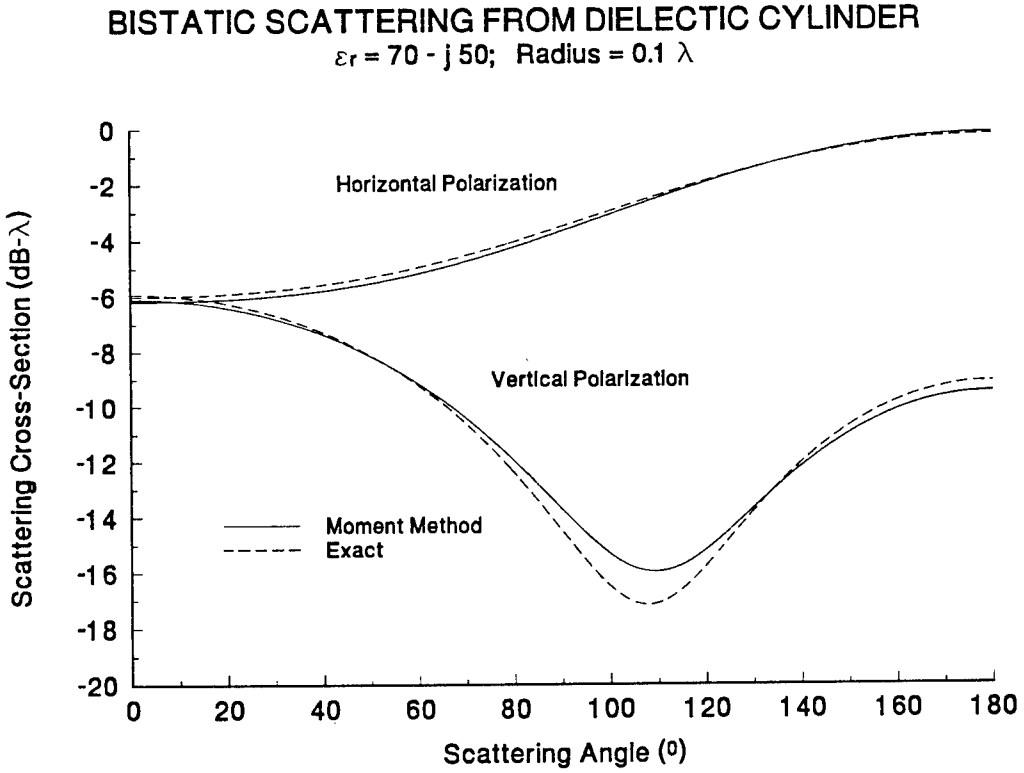


Figure 4: Comparison of moment method and exact calculation of scattering from a lossy dielectric cylinder.

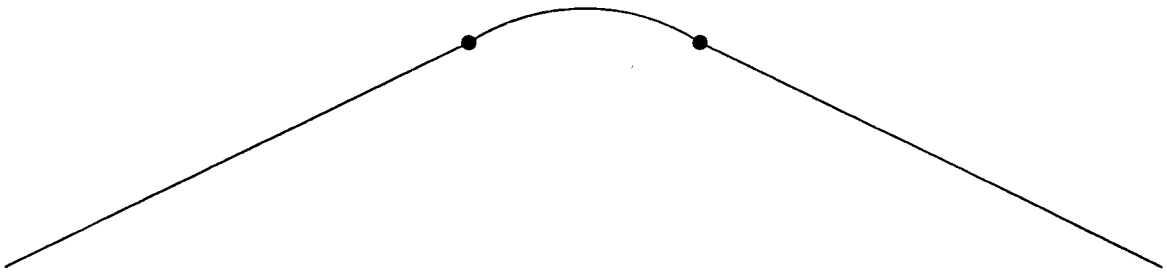


Figure 5: Rounded wedge scattering surface.

polarization (the electric field is parallel to the cylinder axis) and vertical polarization (the electric field is perpendicular to the axis), confirming the validity of the approach at both polarizations.

The scattering from a wedge with a rounded apex, as shown in Figure 5, was calculated using the full implementation of the hybrid MM/GTD technique. The radius of curvature of the apex was set at 0.5λ and the interior angle of the wedge was 120°. The scattering was calculated with the real part of the dielectric constant fixed at 70 and imaginary part ranging from infinity (perfectly conducting case) down to 10. The results are shown in Figures 6 and 7. The incidence angles in the figures are referenced to vertical. At incidence angles ranging from $-30°$ to 30° the backscattering is dominated by specular reflection from the apex. (The singularities occuring at $-30°$ and 30° are due to specular reflection from the infinitely long extensions). Since the conditions of equation (15) are met, at these incidence angles the reduction in the backscattered field with decreasing surface conductivity should be directly proportional to the normal incidence, flat-surface reflection coefficient. When $\epsilon_r = 70 - j10$ the magnitude of the reflection coefficient is 0.788, indicating that the scattering from this surface should be very close to 2.1 dB below that from the perfectly conducting surface in this region. The actual reductions at 0° incidence agree with this value to within 0.15 dB at both polarizations.

At incidence angles beyond 30° ($-30°$) the scattering is no longer dominated by specular reflection, but instead is due entirely to back diffraction, both from the discontinuity in the

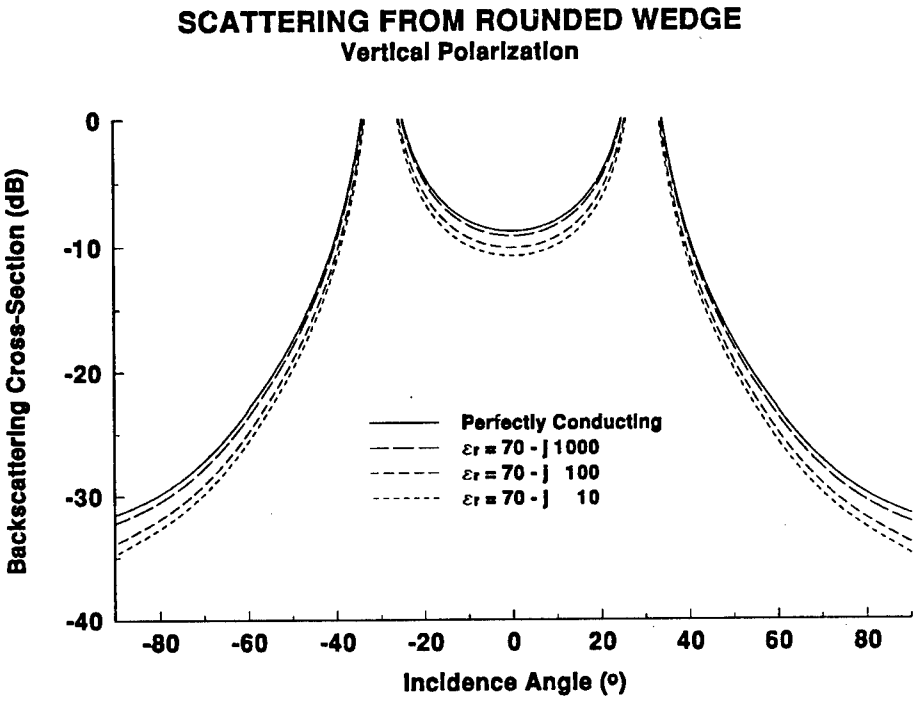


Figure 6: Scattering from rounded-apex wedge: vertical polarization.

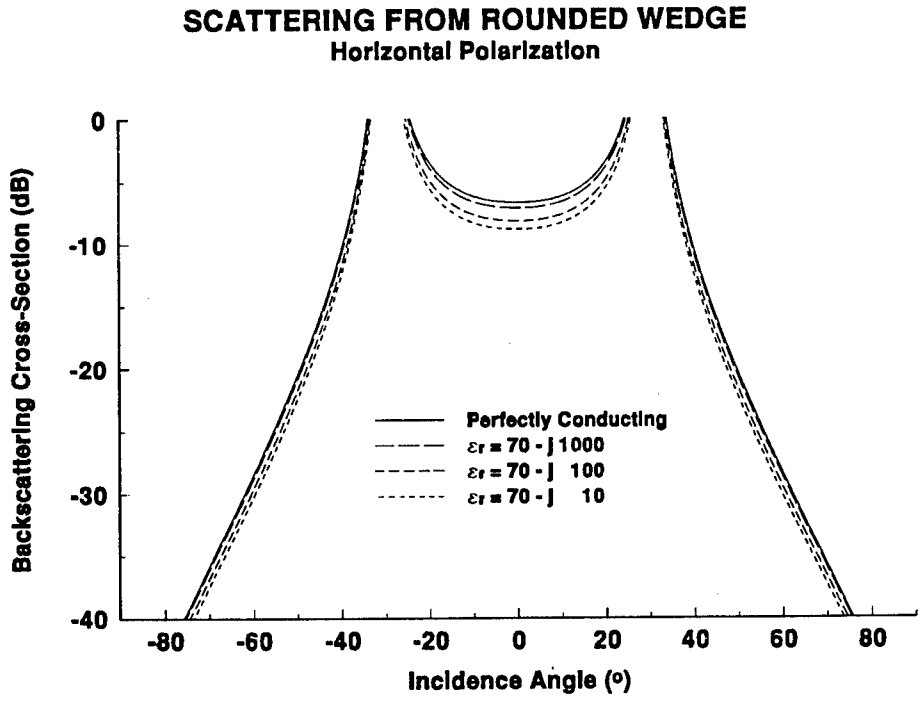


Figure 7: Scattering from rounded-apex wedge: horizontal polarization.

radius of curvature where the rounded apex begins [15] and the point where the incidence vector is exactly tangential to the apex [16]. Thus, the change in the scattering due to a reduction of the surface conductivity cannot be easily predicted. However, the reductions observed in this region are quite realistic. It is interesting to note that the reduction in the region is significantly greater at vertical polarization. This most likely occurs because the transmission into the surface is greater at this polarization when the local incidence angle is not normal.

IV. CONCLUSIONS

A hybrid numerical technique combining the moment method and the geometrical theory of diffraction has been enhanced to allow the calculation of the electromagnetic scattering from a lossy dielectric surface. Sample calculations of scattering from a rounded wedge show that the expected reduction in specular scattering from a lossy surface is accurately predicted. Realistic reductions were also predicted in the diffractive scattering from the surface. This technique should allow more accurate prediction of the scattering from land and sea surfaces than can be obtained using techniques that assume a perfectly conducting surface.

References

[1] R. M. Axline and A. K. Fung, "Numerical computation of scattering from a perfectly conducting slightly rough surface", *IEEE Transactions on Antennas and Propagation*, vol. AP-26, no. 3, pp. 482–488, May 1978.

[2] S. L. Broschat, "The phase perturbation approximation of rough surface scattering from a Pierson-Moskowitz sea surface", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 31, no. 1, pp. 278–283, Jan. 1993.

[3] E. I. Thorsos, "The validity of the Kirchhoff approximation for rough surface scattering using a Gaussian roughness spectrum", *Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 78–82, Jan. 1988.

[4] J. C. West, "Effect of shadowing on electromagnetic scattering from rough ocean-wave-like surface at small grazing angles", *IEEE Transactions on Geoscience and Remote Sensing*, 1995, under review.

[5] J. C. West, M. A. Sletten, and J. M. Sturm, "Comparison of numerically predicted scattering from a breaking wave with experiment", in *Proceeding of the Progress in Electromagnetic Research Symposium*, Seattle, Washington, 1995, p. 759.

[6] W. D. Burnside, C. L. Yu, and R. J. Marhefka, "A technique to combine the geometrical theory of diffraction and the moment method", *IEEE Transactions on Antennas and Propagation*, vol. AP-23, no. 4, pp. 551–558, July 1975.

[7] C. A. Balanis, *Advanced Engineering Electromagnetics*, Wiley, New York, 1989.

[8] T. B. A. Senior and J. L. Volakis, "Generalized impedance boundary conditions in scattering", *Proceedings of the IEEE*, vol. 79, no. 10, pp. 1413–1420, Oct. 1991.

[9] A. W. Glisson, "Electromagnetic scattering by arbitrary shaped surfaces with impedance boundary conditions", *Radio Science*, vol. 27, no. 6, Nov. 1992.

[10] W. V. T. Rusch and R. P. Pogorzelski, "A mixed-field solution for scattering from composite bodies", *IEEE Transactions on Antennas and Propagation*, vol. AP-34, no. 7, 1986.

[11] R. Tiberio, G. Pelosi, G. Manara, and P. H. Pathak, "High-frequnency scattering from a wedge with impedance faces illuminated by a line source, part i: Diffraction", *IEEE Transactions on Antennas and Propagation*, vol. 37, no. 2, Feb. 1989.

[12] H. T. Thacher, "Algorithm 215: SHANKS", *Communications of the ACM*, vol. 6, no. 11, pp. 662, Nov. 1963.

[13] W. H. Press, S. A. Teulkolsky, W. T. Wetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Programming*, Cambridge University Press, Cambridge, 2 edition, 1992.

[14] Jr. C. A. Siller, "Evaluation of the radiation integral in terms of end-point contributions", *IEEE Transactions on Antennas and Propagation*, vol. AP-23, no. 9, Sept. 1975.

[15] G. L. James, G. Tong, and D. A. Ross, "Uniform diffraction solution for a discontinuity in curvature", *Electronics Letters*, vol. 11, no. 23, pp. 557–559, Nov. 1975.

[16] P. H. Pathak, "An asymptotic analysis of the scattering of plane waves by a smooth convex cylinder", *Radio Science*, vol. 14, no. 3, pp. 419–435, May 1979.

DEFECTS IN SPUTTER DEPOSITED MGO FILMS AS CAPPING LAYER FOR DEPOSITING HIGH TEMPERATURE SUPERCONDUCTING FILMS ON GAAS

Peter Ka-Chai Wu
Assistant Professor
Department of Physics

Southern Oregon State College
1250 Siskiyou Blvd.
Ashland, OR 97520

Final Report for:
Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory,
Hanscomb AFB, MA

September, 1995.

DEFECTS IN SPUTTER DEPOSITED MGO FILMS AS CAPPING LAYER FOR DEPOSITING HIGH TEMPERATURE SUPERCONDUCTING FILMS ON GAAS

Peter Ka-Chai Wu
Assistant Professor
Department of Physics
Southern Oregon State College

Abstract

Future developments in superconductor-semiconductor microwave devices depend on depositing high-quality high-temperature superconductor (HTS) thin film on semiconductors. Buffer layer technology using Y-stabilized ZrO_2 (YSZ) as a buffering layer and MgO as a capping layer is one of the most promising processes for growing HTS such as $Yba_2Cu_3O_{7-\delta}$ (YBCO) on GaAs substrate. The superconducting properties of YBCO depend on the quality of the buffer layer, i.e. c-oriented, smooth, and free of defects, which in turn depends on the quality of the capping layer. The defect structure in the MgO capping layer deposited by sputtering is examined in the present study. Defects such as pin holes, precipitates, impurity induced structures, bubbles, and steps are identified. Films with pin holes are found to be ineffective as a chemical barrier. They are also poor templates for growing c-oriented YSZ films. Other defects only have localized effects with no observable structural effects on the YSZ buffer layer. Pin holes in the MgO layer can be eliminated by reducing the target to sample distance.

DEFECTS IN SPUTTER DEPOSITED MGO FILMS AS A GAAS CAPPING LAYER FOR DEPOSITING HIGH TEMPERATURE SUPERCONDUCTING FILMS

Peter Ka-Chai Wu

I. Introduction

The advantages of using high-temperature superconductors (HTS) in devices is demonstrated in microwave devices such as Josephson junction and dc superconducting quantum interference devices fabricated using HTS such as $Yba_2Cu_3O_{7-\delta}$ (YBCO).[1-3] Most of the HTS devices are fabricated on a single crystal oxide surface which serves as a template as well as a chemical buffer layer for the HTS. There is however no reliable and reproducible methodology to grow high-quality HTS thin films on semiconductors which is essential for fabricating integrated superconductor-semiconductor devices.[4] Integration of HTS devices with semiconductor devices can enhance the performance, improve the reliability, and reduce the cost.

The most widely used technique to grow high quality YBCO films on semiconductors is to use a buffer layer.[5,6] Materials such as MgO, CeO_2, Y-stabilized ZrO_2 (YSZ), and $SrTiO_3$ are good buffer layers for YBCO grown on GaAs substrates. GaAs is an attractive substrate because of its low dielectric constant or low loss tangent for microwave applications and the low cost. High quality YBCO thin films with a critical temperature (T_c) of > 90 K and a critical current (J_c) of 5×10^6 A/cm^2 at 77 K are routinely grown on single crystal oxide substrates. T_c and J_c of these YBCO thin films were found to be very sensitive to materials properties such as grain size and orientation. Films with the highest J_c are c-axis oriented because the most favorable transport properties are on the a-b plane. J_c thus increases as c-axis alignment is enhanced.[7] J_c is also very sensitive to structures such as grain size and grain boundary.[8,9]

The difficulties in obtaining high-quality YBCO thin films on GaAs are as follows: (1) GaAs is not stable under the oxidizing environment needed to grow HTS thin films; (2) without a capping layer, GaAs decomposes at temperatures above 600°C, which is required

to grow the HTS film; (3) GaAs reacts with the HTS and destroys its superconducting property; (4) crystalline YBCO film will not form because of the large lattice mismatch between GaAs and YBCO; and (5) there is a significant difference in the thermal expansion coefficients. A buffer layer such as MgO[11] or YSZ[12] can circumvent most of these problems.

The buffer layer must be a good template to grow c-oriented YBCO. It must also remain smooth and free of defects, because the J_c of the YBCO film subsequently deposited is very sensitive to structures such as grain size and grain boundary.[14,15] For small misorientation angles, the ratio of the grain-boundary critical current density to the bulk critical current density is roughly proportional to the inverse of the misorientation angle; for large angles, this ratio saturates to a value of about 1/50.[16] A textured, e.g., c-axis oriented, film with in-plane rotation or mosaic can reduce superconducting properties.

YSZ is superior to MgO as a buffer layer because of a better match in lattice constant and thermal coefficient of expansion with YBCO. T_c as high as 92 K has been reported for YBCO thin films grown on YSZ substrates.[13] It is thus important to develop a method of depositing high quality YSZ, i.e., c-oriented, smooth, free of structural defects, on GaAs substrates. To deposit high quality YSZ films, the minimum substrate temperature (T_s) required is 600°C. This is not acceptable as GaAs will decompose at these temperatures. This problem can be circumvented by first depositing a layer of MgO, ~100 nm thick, to cap the GaAs before depositing the YSZ. MgO is deposited as T_s is raised from room temperature to 650°C, the YSZ deposition temperature. The MgO will prevent the GaAs from decomposing. The quality of the YSZ subsequently deposited is dependent on the quality of the MgO capping layer.

MgO can be deposited using direct sputtering of an MgO target or reactive sputtering of metallic Mg. Both methods require an abient with O_2. The goal of this work is to identify the defect structure of the MgO deposited. The reaction or interdiffusion of chemical species between the GaAs and MgO is examined. The defects formed at the MgO capping layer can propagate through the YSZ layer and ultimately affect the quality of the

YBCO thin film. The effects of the defects on the quality of the YSZ buffer are examined. The optimal deposition parameter necessary to eliminate defects was tested and identified.

II. Experimental

GaAs substrates used are high resistance single crystal wafers. Wafers are dipped in H_2SO_4:H_2O_2:H_2O = 5:1:1 (by volume) solution before introduction into the deposition chamber. No other *in situ* treatment is done prior to deposition.

MgO films are deposited using RF sputtering of either a MgO target or metallic Mg target. Sputtering power of 100 W is used in both cases. T_s at the beginning of deposition is 25°C. It is then raised to a final temperature of 600°C in 1200 sec and held constant throughout the deposition. In the case of the MgO target, a gas mixture of 2.7 Pa of Ar and 0.13 Pa of O_2 is used. In the case of the metal Mg target, a gas mixture of 1.3 Pa of Ar and 0.13 Pa of O_2 is used. The deposition rates for using the MgO target and the Mg target are 0.38 and 0.55 nm/min respectively.

YSZ is deposited using RF sputtering of an YSZ target. The sputtering power is 120 W. The gas mixture is composed of 3.3 Pa of Ar and 3.3 Pa of O_2. T_s is kept at 700°C throughout the deposition.

Film orientation is identified using X-ray Diffraction (XRD). Optical and Scanning Electron Microscopy (SEM), and Atomic Force Microscopy (AFM) are used to identify surface structures, grain shape and size, as well as roughness. Energy-dispersive X-ray Spectrometry (EDX), X-ray Photoelectron Spectroscopy (XPS), and Auger Electron Spectroscopy (AES) are used to determine composition and chemical states.

III. Results and Discussion

Several types of defects are identified on the MgO film. These are pin holes, precipitates, defect induced structures, bubbles, and steps. Pin holes when they are present are distributed throughout the entire surface. The other defects are found to be localized. Pin holes can be eliminated by reducing the target to sample distance.

III.a. Pin holes

One of the most important defects we found on the MgO films is pin holes. The SEM micrograph of a MgO surface with pin holes deposited using an MgO target shows that the pin holes are evenly distributed on the entire surface of the MgO films. An AFM micrograph, Figure 1, shows that the average grain size between pin holes is 500 nm with a mean surface roughness of 28.9 nm. Because the height range, 220 nm, is on the order of the thickness of the film and the AFM cantilever is incapable of reaching into tight areas, It is likely that the pin holes extend all the way to the substrate surface.

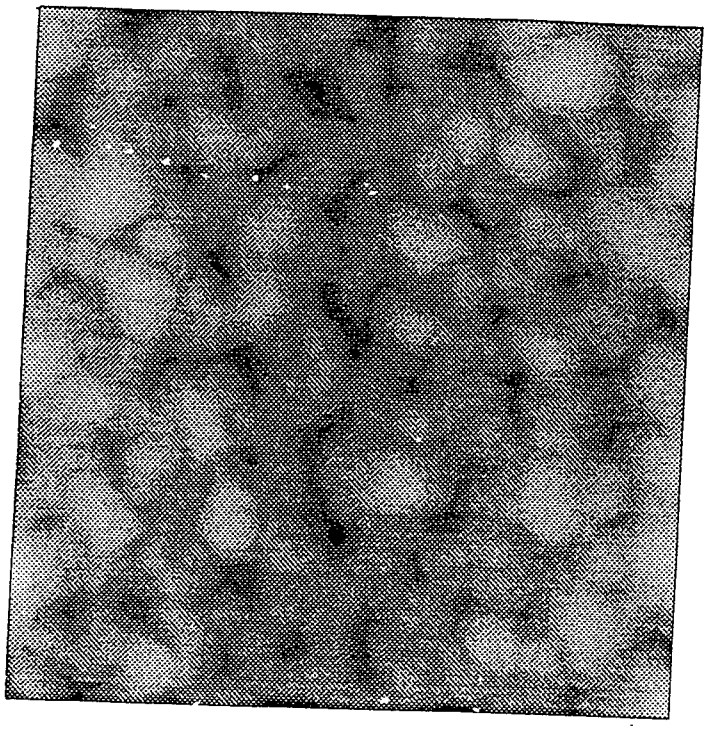


Figure 1. AFM micrograph of a MgO film with pin holes. (height range = 220 nm, size = 5 μm X 5 μm)

Figure 2 shows an expanded view of one of the pin holes. The large grains observed in Figure 2 are composed of several smaller grains. The pin holes are formed where three or more of these grains touched. This type of microstructure indicates a coalescent process where smaller grains combined to form a larger three-dimensional grain. Above a

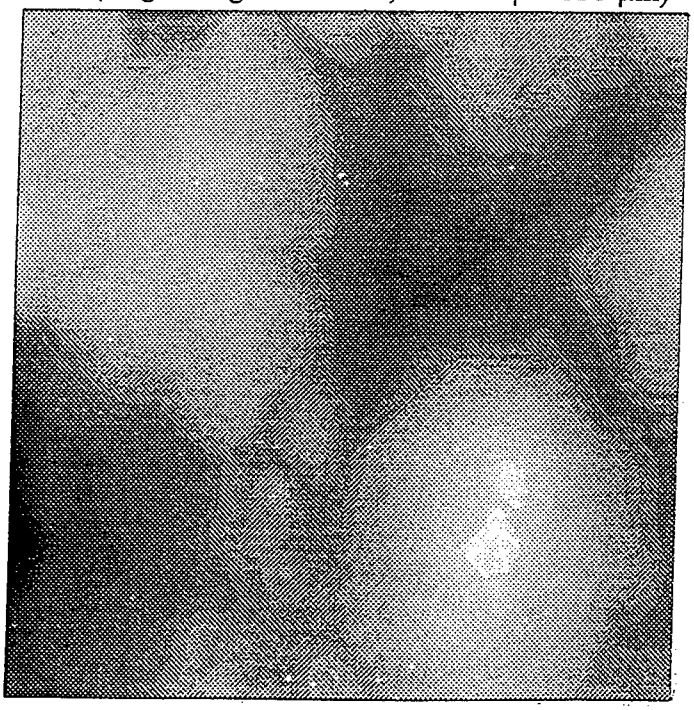


Figure 2. AFM micrograph of a pin holesin the MgO film. (height range = 150 nm, size 1 μm X 1 μm)

certain size, larger grains are less likely to coalesce because of the larger energy barrier for such a process. When neighboring grains grow and touch each other a pin hole is formed.

The micrograph from the surface of a MgO film without pin hole, Figure 3, shows an average grain size of 50 nm and a mean surface roughness of 3.7 nm. The average grain is much smaller than that shown in Figure 1 and is mainly spherical. A large range of grain size is present. This means that no substantial amount of coalescing is present. These results indicate that pin holes are formed when the coalescent rate is higher than the nucleation and growth rate.

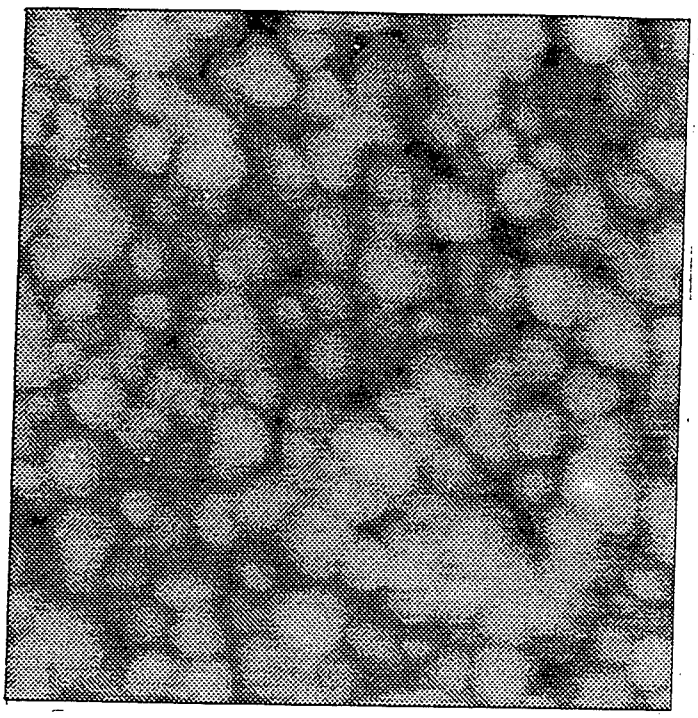


Figure 3. AFM micrograph of a MgO film without pin holes. (height range = 20 nm, size = 1 μm X 1 μm)

The result of an AES depth profile of a 30 nm thick MgO film with pin holes is shown in Figure 4. Figure 4 shows the profile of the entire film plus part of the GaAs substrate. The MgO layer and the GaAs substrate is clearly identified. The MgO/GaAs interface is not a sharp interface. An interfacial region containing Ga is shown extending into the MgO. Thus Ga or oxides of Ga are reacting with or diffusing into the MgO. In contrast no substantial amount of As is found in the MgO matrix.

XRD measurements on the MgO film with pin holes shows a polycrystalline structure. YSZ deposited on these MgO films shows no preferred (001) orientation. This is not surprising as the YSZ is grown on a very rough surface and the orientation of the YSZ grain orientation depends on its nucleation sites on the MgO grain.

To determine the chemical state of the Ga in the interfacial region, an 4.8 nm thick MgO film is deposited and examined with XPS. This thickness is chosen so that the photoelectron from the interfacial region can escape from the surface. Possible new chemical states promoted by the sputtering process are eliminated. Figure 5 shows the Ga(2p 3/2), Mg(2p), and O(1s) spectra from such a film. The O(1s) spectrum is composed mainly of two peaks indicating two major types of oxide in the sample. The main peak in the Mg(2p) spectrum has the correct energy shift for MgO. The Ga(2p 3/2) spectrum shows one peak indicating one major chemical state, which has the correct chemical shift for Ga_2O_3. Both the fitted Mg(2p) and Ga(2p 3/2) peaks are less than 1 eV wide indicating

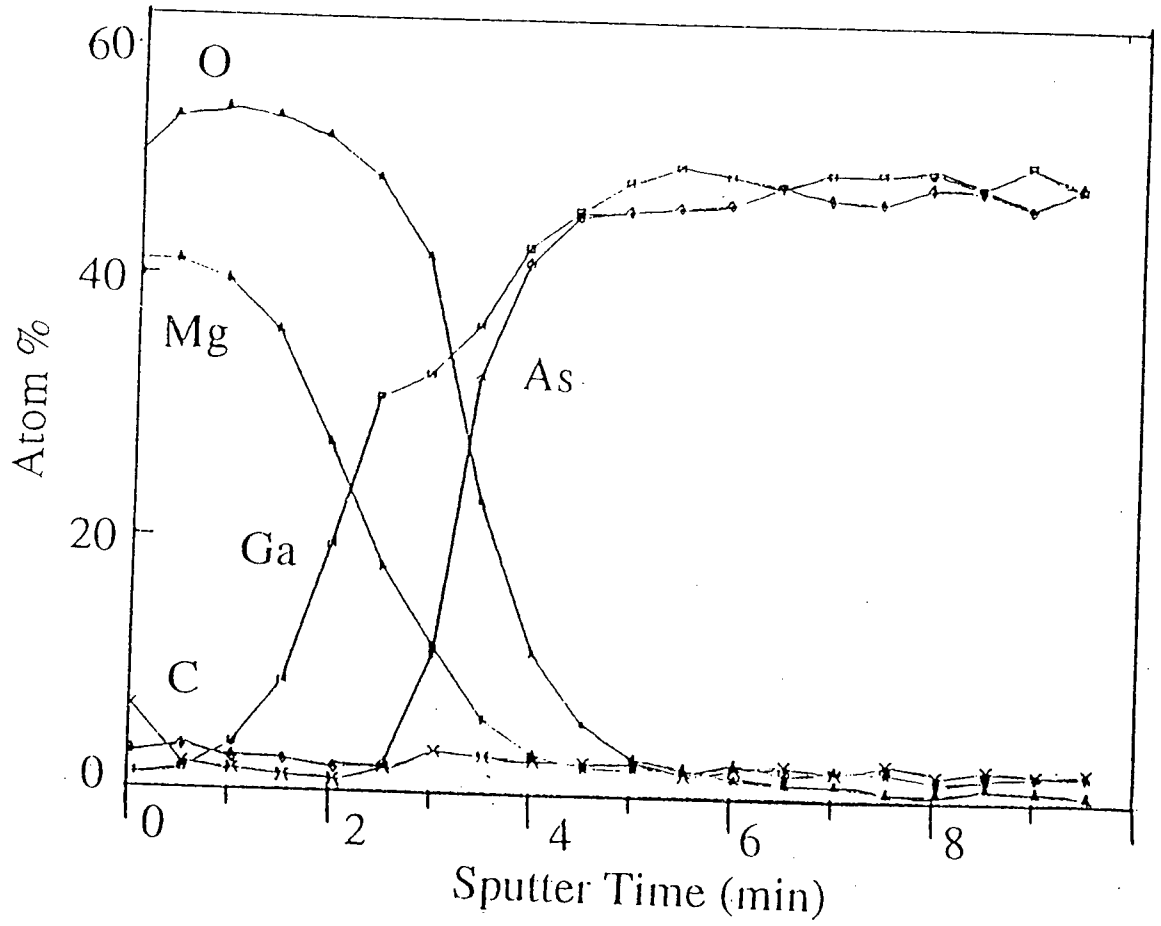


Figure 4. AES depth profile of a MgO film with pin holes showing both the MgO film and the GaAs substrate. ◊ - As, □ - Ga. ☆ - Mg, ▵ - O, and X - C.

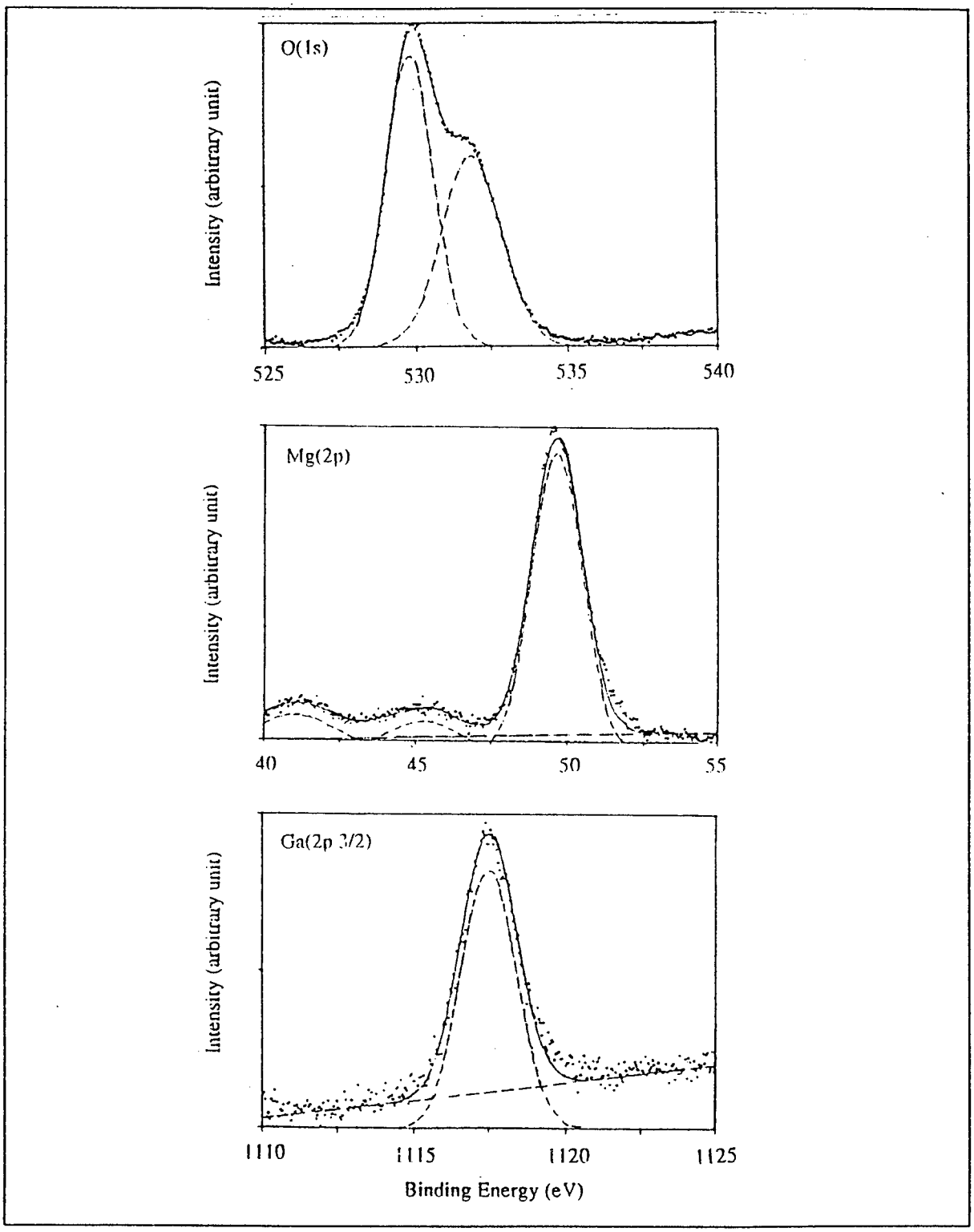


Figure 5. XPS spectra from an MgO film 4.8 nm thick. Dashed lines are fitted results. X-axis is Binding Energy (eV); top spectrum, O(1s); middle spectrum Mg(2p); and bottom spectrum Ga(2p 3/2)

one major chemical state contributes to each peak. The XPS result thus shows the interfacial region composed mainly of Ga_2O_3 and MgO and not a reaction product of the MgO and the substrate.

An AFM micrograph, Figure 6, of an YSZ film grown on MgO an film with pin holes shows the YSZ surface to be smooth as compared to the MgO underneath. The average grain size is 200 nm with a mean surface roughness of 2.3 nm. Thus

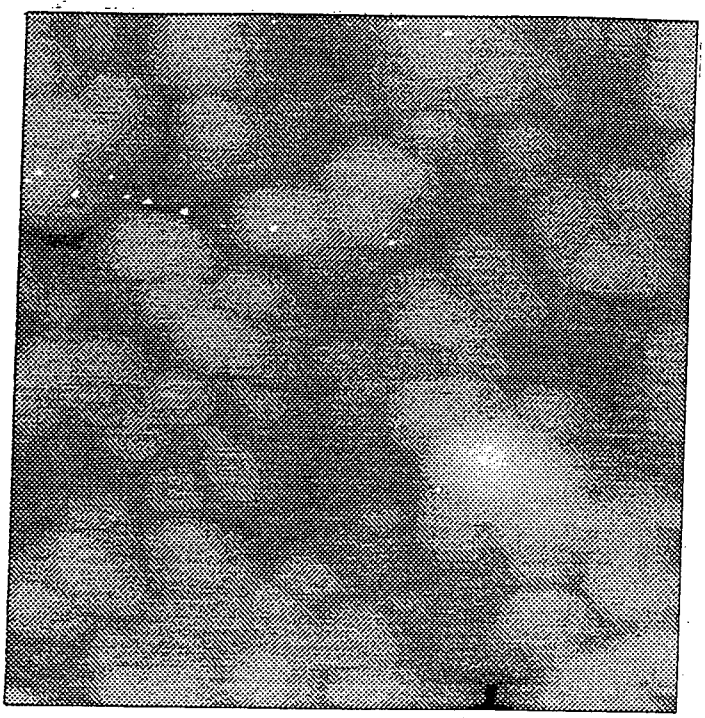


Figure 6. AFM micrograph of a YSZ film grown on an MgO film with pin holes.
(height range = 36.5 nm, size = 1 μm X 1 μm)

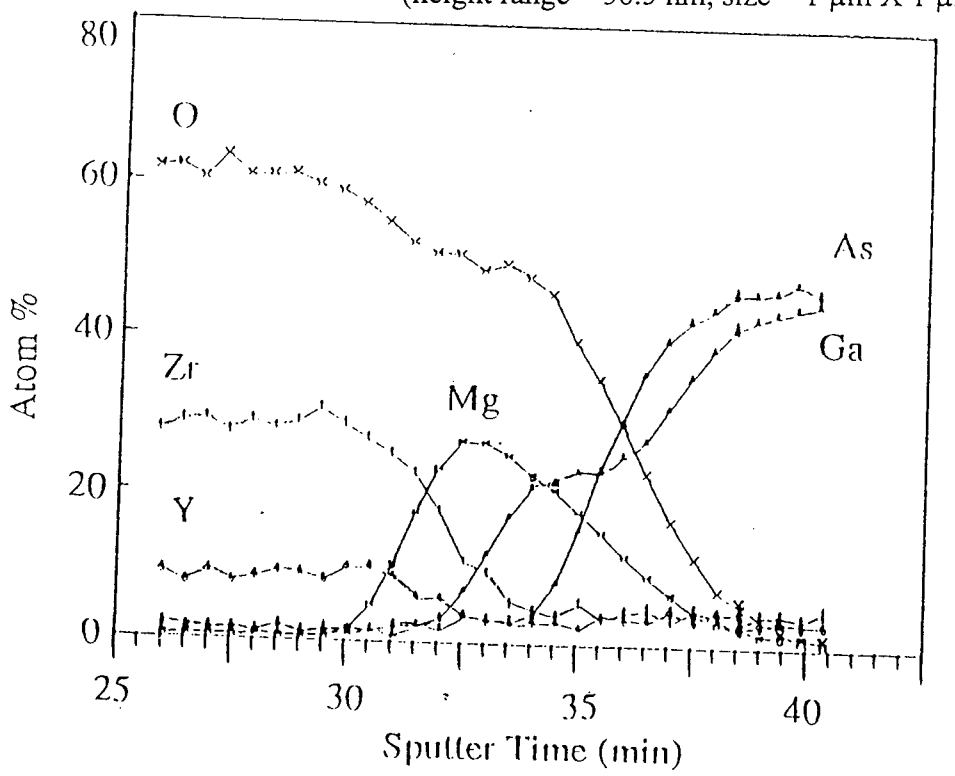


Figure 7. AES depth profile of a YSZ film grown on an MgO film with pin holes.
X - O, □ - Mg, ☆ - As, ◇ - Y, | - Zr, and △ - Ga.

even though the substrate is rough, the YSZ will fill in the pin holes and result in a smoother film.

The AES depth profile from a sample with YSZ deposited on a MgO film with pin holes is shown in Figure 7. Figure 7 represents both films and part of the GaAs substrate. Both the YSZ and the MgO layers and the GaAs substrate were clearly identified. The interfacial region found in Figure 4 is also identified here. The MgO and interfacial region thickness are estimated to be 40 nm. Even though the Ga_2O_3 diffuses almost through the MgO layer, no mixing or reaction is evident in the YSZ layer.

Figure 8 shows an AES depth profile from an MgO film without pin holes. This profile represents all of the MgO films and part of the GaAs substrate. The interfacial region observed in Figure 4 and 7 is not present here indicating no chemical reaction between the MgO and the GaAs substrate. This result agrees with the XPS result in that the Ga present in the interfacial region is a result of Ga_2O_3 diffusing into the MgO instead

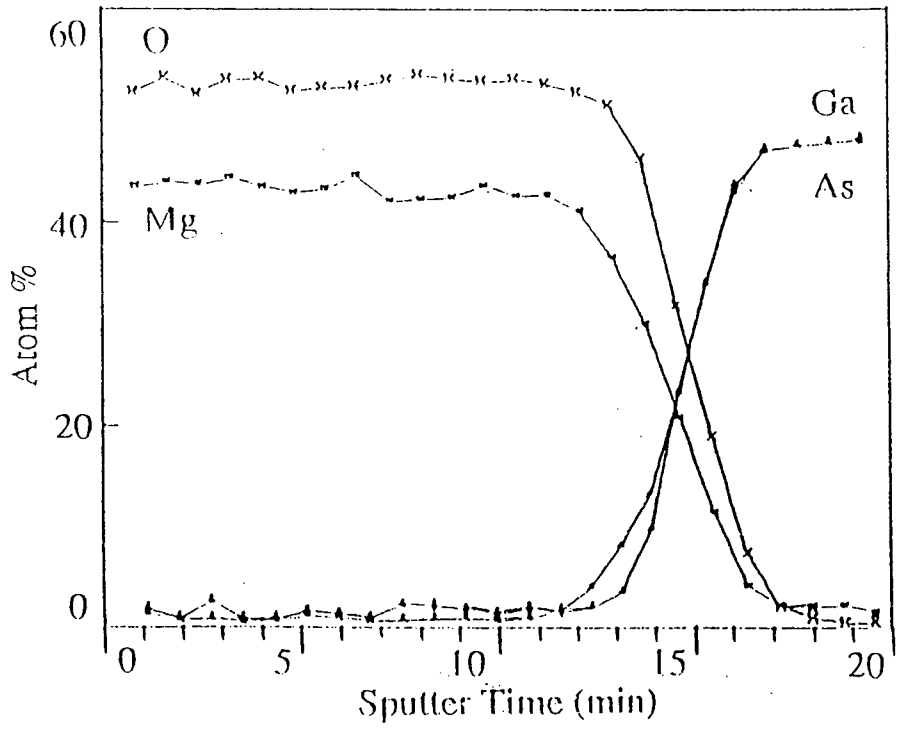


Figure 8. AES depth profile of an MgO film without pin holes.
X - O, ▵ - Ga, □ - Mg, and ☆ - As

of a chemical reaction between the overlayer and the substrate. Furthermore, the dominant diffusion path is through the pin holes. Surface or grain boundary diffusion should not be the main diffusion mechanism. Because of the smaller grain size, the total surface area of MgO grains is much larger in films without pin holes than that with pin holes. If surface diffusion is the dominant transport mechanism, one would expect a thicker interfacial region than that shown in Figure 8. This is evidently not the case. The main diffusion venue is thus through the pin holes.

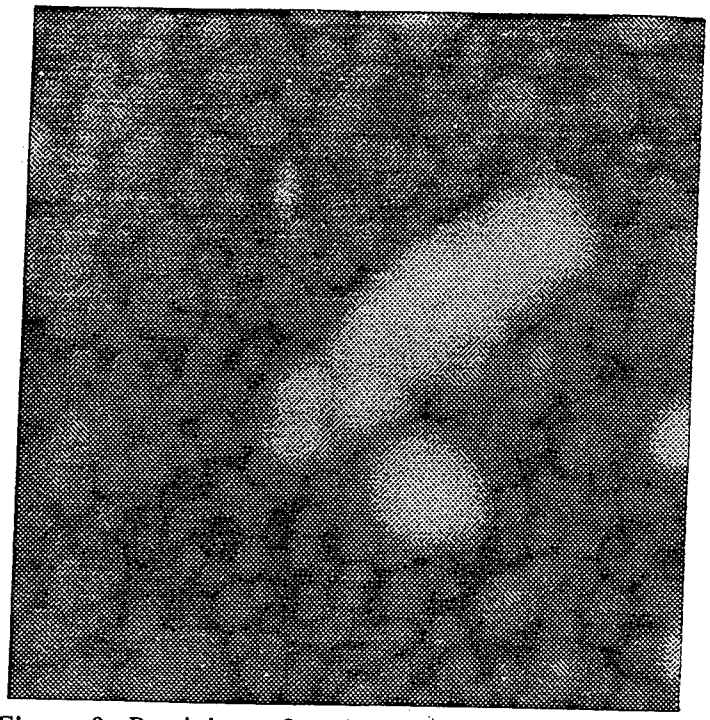


Figure 9. Precipitates found on an MgO film. (height = 150 nm, size = 2 μm X 2 μm)

III.b. Precipitates

Precipitates are found on some films. These are small particles distributed evenly across the surface. The majority of the precipitates are circular in shape with an average diameter of 350 nm and 50 nm in height, see the round particle in Figure 9. Some larger precipitates are also found, see the larger elongated particle in Figure 9. These precipitates are found on reactively sputtered MgO films. It is thus possible that they are a result of segregation from the matrix due to composition differences. However, no chemical identification is available. Work is required to determine deposition conditions which will eliminate these precipitates.

III.c. Impurity induced growth morphology

Figures 10 and 11 show a special structure found on some of the MgO films. Figure 10 shows the morphological structure surrounding a particle in the center. Figure 11 is an enlarged view of the structure near the particle in the center. The extent of this type of structure is 20 to 30 μm in diameter. The morphology of the MgO in this area is very different from that of the normal MgO film, see Figure 1. Several structures like this were found located randomly on the surface. The composition or the origin of the particles in the center of this structure are not known. These structures were only found on a few films and their appearance does not show any trend in any deposition condition. It is thus likely that these particles are contaminants due to careless substrate preparation.

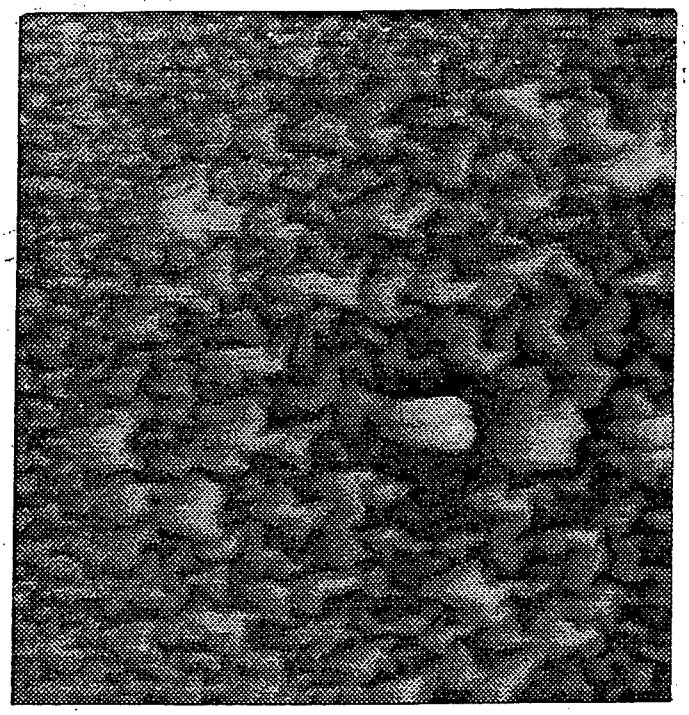


Figure 10 . AFM micrograph of the structure shown in Figure 11 enlarged 5X. (height range = 1 μm, size = 6 μm X 6 μm)

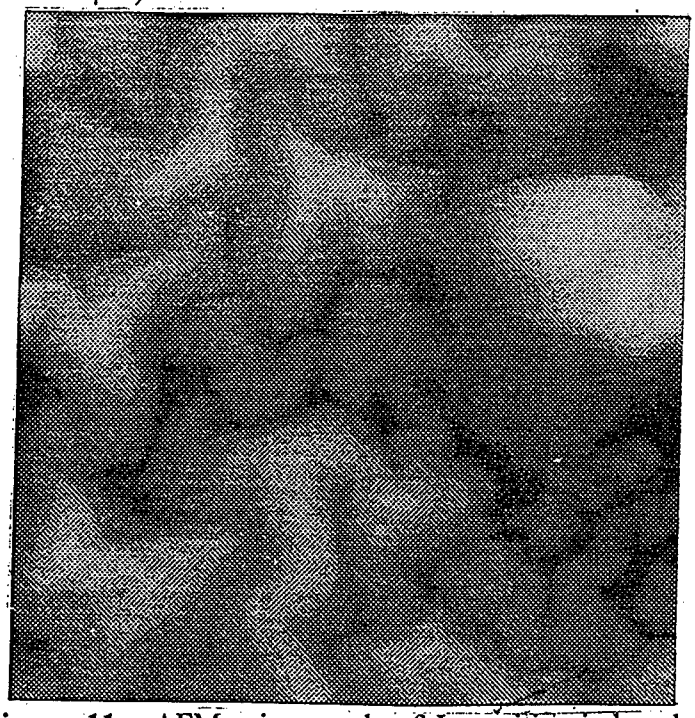


Figure 11. AFM micrograph of Impurities induced structures, see text. (height range = 1 μm, size = 20 μm X 20 μm)

III.d. Bubbles

Bubbles are found on MgO films deposited using reactive

sputtering of metal targets. Figure 12 shows one of these bubbles. This particular bubble has a diameter of 6 μm and a height of 220 nm. Some of the bubbles will burst. Figure 13 shows one of these burst bubbles. This bubble has a diameter of 8 μm. We have used AFM to probe the inside of the burst bubble. It is found that the bottom of the bubble is flat. The bottom of the burst bubble is found to be 280 nm deep. The presence of trapped bubbles in the MgO film is not acceptable. This is because the intended application of these films is in superconducting devices. Bubbles, especially those with trapped gas, can be detrimental to devices during thermal cycling.

Bubbles can be a result of trapped gas during sputter deposition or internal chemical segregation leading to release of gas. Thus varying gas pressures and mixtures as well as changing deposition rates should make a difference. Work is being done to eliminate these types of defects.

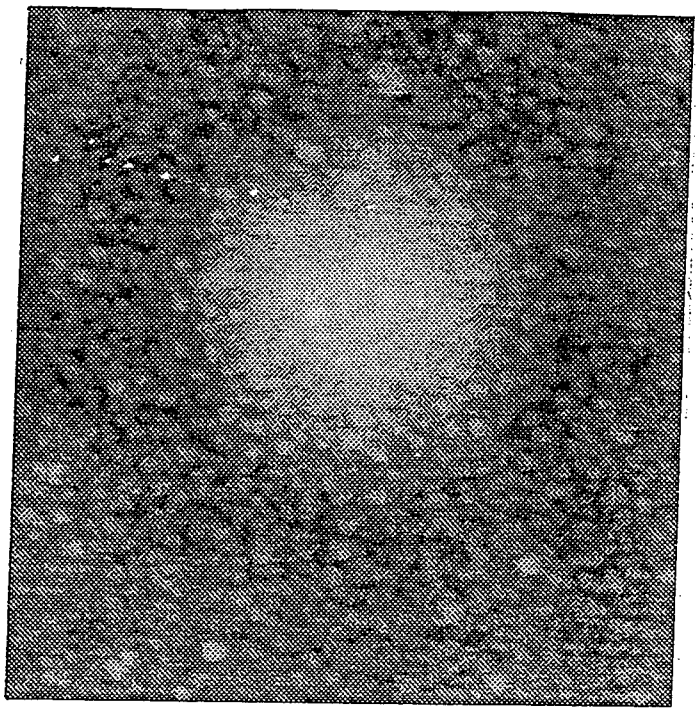


Figure 12. AFM micrograph of a bubbie in a MgO film. (height range = 500 nm, size = 10 μm X 10 μm)

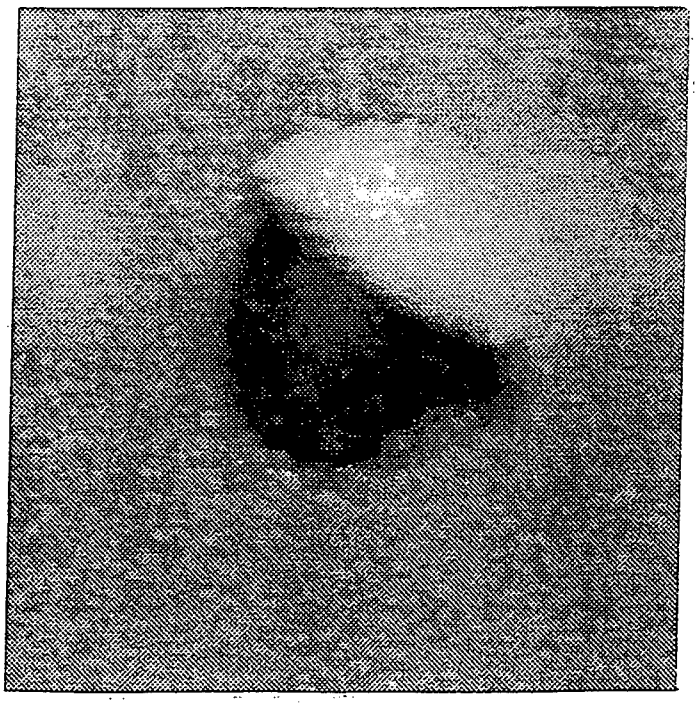


Figure 13. AFM micrograph of a burst bubble on a MgO film. (height range = 900 nm, size = 15 μm X 15 μm)

III.d. Steps

Steps on the MgO surface are found on the edges of the samples. Figure 14 shows some of these steps. Cross section analysis of these structures shows flat terraces between step edges. Terrace height determined by cross-sectional study ranges from 13 to 26 nm . There are holes on some of the terraces. The holes also have flat bottom, the depth of these holes has the same range as the terrace height. Some step edges delaminated from the next terrace down and curled up. This result and the result from the burst bubble, Figure 13, indicates that the MgO is grown in a layer-by-layer fashion. Because steps are only found at the edges of the sample, the cause of the steps can be a result of structural defects or the temperature gradient during deposition at the edge of the sample.

Figure 14. AFM micrograph of steps on a MgO film. (Height range = 300 nm, size = 5 μm X 5 μm)

IV. Conclusion

Several types of defects of sputter deposited MgO films on GaAs are identified. They are pin holes, precipitates, impurity induced structures, bubbles, and steps. MgO films with pin holes are poor chemical buffers because of the diffusion of Ga_2O_3 through the MgO film. These films are also bad templates for growing a c-oriented YSZ buffer layer. YSZ deposited on the MgO surface with pin holes will fill the holes and grow into a comparatively smooth surface. Pin holes can be eliminated by reducing the sample to target distance during sputter deposition. All the other defects were found to be localized and have no detectable long range effects on the YSZ films subsequently deposited. High

quality YSZ films can be deposited on GaAs using the MgO as a capping layer. Good MgO films can be deposited using deposition condition stated earlier and with a sample-to-target distance of less than 5 cm.

Acknowledgments

I wish to thank Dr. M.N. Alexander and Dr. J.T. Schott for their hospitality and support during my stay at Rome Laboratory. Special thanks to Dr. Al. Drehman, H. Jiang P. Yip, B. Demczyk, and Y.-H. D. Li for their technical support, discussion, and hospitality.

Reference:

1. R.H. Ono, MRS Bulletin, **XVII** (8), 34 August (1992).
2. R.H. Hammond and R. Bormann, Physica C, **162-164**, 703 (1989).
3. N. Newman and W.G. Lyons, J. Supercond., **6** (3), 119 (1993).
4. M.J. Burns., P.R. de la Houssaye, S.D. Russell, G.A. Garcia, S.R. Clayton, W.S. Ruby, and L.P. Lee, Appl. Phys. Lett. **63**(9), 1282 (1993).; S.S. Toncich, F.A. Miranda, K.B. Bhasin, and T.J. Kascak, Preprint (1994).
5. MRS bulletin, (August 1992).
6. MRS bulletin, (September 1994).
7. F. Yang, E. Narumi, S. Patel, and D.T. Shaw, Appl. Phys. Lett., **60**, 249 (1992).
8. D. Dimos, P. Chaudhari, J. Mannhart, and F.K. LeGoues, Phy. Rev. Lett., **61** (2), 219 July (1988).
9. S.E. Babcock, MRS Bulletin, XVII (8), 20 August (1992).
10. Y.D. Li, A. Drehman, J. Horrigan, R. Andrews, and D. Bliss, Proceedings of MRS Fall Meeting (1994) and Poster Sesssion of the APS meeting (March 1995).
11. Y.D. Li, A.Drehman, J. Horrigan, R. Andrews, and D. Bliss, Proceedings of the Fall meeting of the Materials Research Society (1994).
12. P. Tiwari, S.Sharan, and J.Narayan, Appl. Phys. Lett., **59** (3), 357 (1991); Q. Jia, S. Lee, W. Anderson, and D. Shaw, Appl. Phys. Lett., **59** (9), 1120 (1991).
13. P. Tiwari, S. Sharan, and J. Narayan, Appl. Phys. Lett., **59** (3), 357 (1991); Q. Jia, S. Lee, W. Anderson, and D. Shaw, Appl. Phys. Lett., **59** (9), 1120 (1991).
14. D. Dimos, P. Chaudhari, J. Mannhart, and F.K. LeGoues, Phy. Rev. Lett., **61** (2), 219 July (1988).
15. S.E. Babcock, MRS Bulletin, XVII (8), 20 August (1992).
16. F. Yang, E. Narumi, S. Patel, and D.T. Shaw, Appl. Phys. Lett., **60**, 249 (1992).
17. "Handbook of X-ray Photoelecton Spectroscopy", ed. G.E. Muilenberg, Perkin-Elmer corp, Physical electronics Division, Eden Prairie, MN (1979).