RL-TR-93-17
Final Technical Report
March 1993

AD-A267 051

ADVANCED CONCURRENT INTERFACES FOR HIGH-PERFORMANCE MULTI-MEDIA DISTRIBUTED C3 SYSTEMS

MIT Media Lab

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. 6474

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

Rome Laboratory
Air Force Materiel Command
Griffiss Air Force Base, New York

This report has been reviewed by the Rome Laboratory Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RL-TR-93-17 has been reviewed and is approved for publication.

APPROVED: Richard T. Slavinski

RICHARD T. SLAVINSKI
Project Engineer

FOR THE COMMANDER: John A. Graniero

JOHN A. GRANIERO
Chief Scientist
Command, Control and Communications Directorate

If your address has changed or if you wish to be removed from the Rome Laboratory mailing list, or if the addressee is no longer employed by your organization, please notify RL(C3AB) Griffiss AFB, NY 13441-4505. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

ADVANCED CONCURRENT INTERFACES FOR HIGH-PERFORMANCE MULTI-MEDIA DISTRIBUTED C3 SYSTEMS

Nicholas P. Negroponte
Dr. Richard A. Bolt

Contractor: MIT Media Laboratory
Contract Number: F30602-89-C-0022
Effective Date of Contract: 3 January 1989
Contract Expiration Date: 30 June 1992
Short Title of Work: Advanced Concurrent Interfaces for HP Multi-Media Distributed C3 Systems
Period of Work Covered: Jan 89 - Jun 92

Principal Investigator: Nicholas P. Negroponte, Dr. Richard A. Bolt
Phone: (617)253-5960 (617)253-5897

RL Project Engineer: Richard T. Slavinski
Phone: (315)330-7764

Approved for public release; distribution unlimited.

This research was supported by the Defense Advanced Research Projects Agency of the Department of Defense and was monitored by Richard T. Slavinski, RL(C3AB), 525 Brooks Road, Griffiss AFB NY 13441-4505 under Contract F30602-89-C-0022.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 1993	Final Jan 89 - Jun 92

4. TITLE AND SUBTITLE
ADVANCED CONCURRENT INTERFACES FOR HIGH-PERFORMANCE MULTI-MEDIA DISTRIBUTED C3 SYSTEMS

5. FUNDING NUMBERS
C - F30602-89-C-0022
PE - 62301E
62708E
61101E
PR - F474
TA - B1
WU - 01

6. AUTHOR(S)
Nicholas P. Negroponte
Dr. Richard A. Bolt

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
MIT Media Laboratory
20 Ames Street
Cambridge MA 02139

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Projects Agency (CSTO)
3701 N. Fairfax Dr.
Arlington VA 22203

Rome Laboratory (C3AB)
525 Brooks Road
Griffiss AFB NY 13441-4505

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
RL-TR-93-17

11. SUPPLEMENTARY NOTES
Rome Laboratory Project Engineer: Richard T. Slavinski/C3AB/(315)330-7764

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; Distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Under this contract, the MIT Media Laboratory explored state-of-the-art display techniques, together with interactive means, on the theme of advanced mapping displays for the commander. Achievements - including some "world firsts" - occurred in three major research areas:

Enhanced Display Qualities, in particular: the readability of tactical maps through enhanced resolution, unique graphical techniques such as transparency, spatio-temporal trade-offs, and removal of spatio-temporal clutter; map usability, through readily-annotated large-format displays, and visualization of terrain through low-cost, real-time 3-D computer graphics techniques.

Broadened User Actions, through multimodal natural dialogue via concurrent speech input, gesture and gaze.

Advanced Input/Output, in particular: synthetic terrain holography from mapping data, and the innovation and development of real-time holographic video; the extraction of range data from camera depth-of-field information, both to make likelihood estimates of scene structure, and to evaluate user expressive facial output.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Large Format Displays, Terrain Visualization, Dynamic Maps, Multi-Modal Interaction, Holographic Mapping	88
	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

Table of Contents

EXECUTIVE SUMMARY

Under this contract, The MIT Media Laboratory explored state-of-the art display techniques, together with interactive means, on the theme of advanced mapping displays for the commander. Achievements—including some "world firsts"—are summarized below:

PRINT QUALITY MAPS

- investigated algorithms for the display and integration of video image sequences into a very high resolution (2Kx2K) graphics adapter. A light-pen was used for making annotations, and to control movies from drone over-flights placed as "inserts" over a high-resolution background map image

LARGE FORMAT DISPLAYS

- built unique high-resolution display with total pixel count of 6k by 2k, measuring over 60 by 20 inches. Display employs an original virtual framebuffer concept wherein three physical framebuffers supporting the display may, for programming purposes, be treated as a single buffer.

  - the prototype features an underlying software architecture and window tool kit that allows rapid and flexible prototyping of experimental graphic environments

- prototype imagery includes a news and weather browsing/query application, using as backdrop a 4k by 8k pixel cloudless image of the earth derived from over 1000 4km landsat images; e. g., news stories from around world are arranged according to their geographic locale

TERRAIN VISUALIZATION

- demonstrated utility of emerging *virtual environment* (VE) technologies for visualizing and interacting with computational models of terrain, vehicles and aircraft engaged in simulated operations. Explored VE software, interface and modeling technologies, including:

  - displaying synthetic human figures using kinematic simulation of gait

  - study of whole-hand input, including a taxonomy of hand gestures for input to VE systems

  - a *constraint package* for describing causal dependencies among objects, agents and processes in the virtual world;

  - a modular and flexible system for dynamic simulation of mechanisms and vehicles

  - a task level VE system for mission planning ("mission visualizer")

  - an *intelligent camera controller* for maintaining visual context and ensuring smooth visual transitions

Z-AXIS OF INFORMATION

- developed closed-form approximate solution to the problem of *extracting shape information from image shading,* with improved method for estimating the illuminant direction

- developed unique object-oriented "window" toolkit which provides the highest graphic quality anti-aliased fonts, and allows every object to have variable translucency and focus. Layout relationships between these graphic objects can be controlled automatically by networks of constraints invoked from a case library. This system:

  • uses the parameters of color, transparency, and focus in display with as many as 30 salient layers of information to precisely control which objects are brought to the users attention, and by what amount

  • uses Intelligent Layout system so that presentations can be generated automatically which maintain their legibility, structure and style independent of the allocated screen real estate

DYNAMIC MAPS

- in image processing area, developed "steerable filters" to analyze local orientation of textured images, find edges, perform image enhancement, and to extract information

from moving imagery. Also, developed "steerable pyramids" to break down image interactively for analysis and enhancement

MULTI-MODAL INTERACTION & SEMANTIC AND DISCOURSE STRUCTURE

- implemented and demonstrated, for first time ever:

  - speech, gaze, and manual input combined at a single interface

  - *two-hand, free-handed gestural input,* with accompanying speech and gaze, to indicate, rotations, shifts, translations, re-scalings, of solid 3-D items

  - *modulation of free-hand gestural input by gaze*, e.g., user looking at hands or not while gesturing alters interpretation of that gesture by the system

HOLOGRAPHIC TERRAIN MAPPING

- Achieved dramatic advances in basic holographic science, and significant steps in adapting holograms to display applications:

  - for first time anywhere in world, achieved full-color moving holographic video images

  - developed holograms with wide viewing angle: beyond +/- 50 degrees for holograms 300 mm square (frontal area), and only slightly less for holograms over a meter square

- demonstrated interaction with holographic images generated in less than two seconds

- developed holographic 3-D relief maps with overlain topological data and feature data (roads, landmarks, etc.), either viewable by horizontal shift in viewing position; also "pop in" display inserts, viewable by horizontal head-shift, which, e.g., bring in specified regions of 3-D map surface forward at higher magnification

- achieved information and bandwidth reduction for computed holograms via the development of a lensless-Fourier-transform optical geometry that produces a particularly straightforward interference fringe pattern for computation and transmission.

- achieved increased image quality, decreased computation time, and production of surface-shaded images via pre-computational approach

- developed a conceptual model of a one-step optical holographic hard-copy printing process, toward goal of attaining high quality holographic images without requiring the current two-step process of first generating a "master," then a white–light viewable "copy"

LOOKING AT THE USER

- developed a near-real-time computer system which can locate and track a subject's head, and then recognize the person by comparing characteristics of the face to those of known individuals

INTRODUCTION

This report attempts to summarize detailed research progress documented in various publications, reports, articles, and/or academic theses which are identified herein.

Purpose of project:

The purpose of this project was to research and develop in prototype display technologies and interface techniques to serve the interests of the commander. The work was in three broad categories:

- To enhance **display qualities**, in particular: the readability of tactical maps through enhanced resolution, graphical techniques of transparency, spatio–temporal trade-offs, and removal of spatio-temporal clutter; map usability, through readily-annotated large-format displays, and visualization of terrain through low-cost, real-time 3-D computer graphics techniques

- To broaden **user actions**, through multimodal natural dialogue via concurrent speech input, gesture, and gaze

- To explore **advanced i/o**, in particular: synthetic terrain holography from mapping data, and the innovation and development of real-time holographic video; the extraction of range data from camera depth–of–field information, both to make likelihood estimates of scene structure, and to evaluate user expressive facial output.

Project personnel:

Personnel at the MIT Media Laboratory directly involved in the conduct of this research were:

Dr. Edward Adelson, Associate Professor of Visual Sciences and co-Director (with Dr. Alex Pentland) of the Lab's Vision Science Group.

Walter Bender, Principal Research Associate at the Media Laboratory.

Dr. Stephen A. Benton, Professor of Media Technology, Head of the Media Arts and Sciences Section, and Director of the Lab's Spatial Imaging Group.

Dr. Richard A. Bolt, Senior Research Scientist and Director of the Media Lab's Advanced Human Interface Group.

Prof. Muriel R. Cooper, Professor of Visual Studies and Director of the Lab's Visual Language Workshop.

Andrew Lippman, Lecturer, and co-director of the MIT Media Laboratory.

Mr. Ronald L. MacNeil, Principal Research Associate at the Media Laboratory and co-founder with Prof. Muriel Cooper of the Visible Language Workshop.

Dr. Alex Pentland, Associate Professor of Computers, Communication, and Design Technology, and co-Director (with Dr. Edward H. Adelson) of the Lab's Vision Science Group.

Dr. David L. Zeltzer, Associate Professor of Computer Graphics and Director of the Lab's Computer Graphics and Animation Group.

PRINT QUALITY MAPS

Relevant Personnel:

Work under this topic was conducted by **Walter Bender** under the general supervision of **Andrew Lippman**, co-director of the MIT Media Lab.

* * * * * * *

Paper Quality Maps with Video Overlay

We investigated algorithms for the display and integration of video image sequences into a very high resolution (2Kx2K) graphics display. A light-pen was used for making annotations, and to control movies placed as "inserts" over the high-resolution background image.

In our prototype graphic, geographic feature data was overlain on elevation maps. Then, digital video data from a drone "fly over" was displayed on the map in a position and orientation which corresponded to the drone in flight. The interface to the application enabled the user to interrupt the motion sequence on any frame by touching the video window with the light pen. The next level of the decomposition was retrieved from disk, with the subsequent still image doubled in size.

In our approach to movie coding we used vector quantization for simple decoding, and asymmetric quadrature mirror filters (QMF) for fast reconstruction of the sub-bands. Off-line, we created a spatial sub-band representation of the luminance channel of the video, and applied a vector quantizer to the chrominance channels. We achieved motion by

retrieving the lowest level of the sub-band representation while decoding the color in real-time. Subsequent levels of spatial detail are synthesized on-line, although, only by sacrificing motion.

In the application, the lowest level of the spatial decomposition was stored in virtual memory. We were able to store 85 frames per megabyte. These images were applied to the screen directly, from 10 to 12 Hertz. The application affords both motion and spatial resolution, albeit simultaneously.

The video images contribute to both the detail and accuracy of the map information. Much of the video includes close-up footage of moving vehicles and details of the terrain not evident in the map. In its present state, the system could be useful as a means of verifying and augmenting digital terrain maps. With real-time processing of the video it could well find use in tactical situations.

However, given the particular source materials used, the disparity between the very detailed video and the coarse terrain data made it difficult to correlate drone position with map features. The dead-reckoning system used by the drone proved not very useful, the coordinates being up to 5000 meters off true position; before such a system could be used in actual reconnaissance, more accurate position tracking of the drone must be established. Perhaps the correspondence between known and observed features can be put to use in future systems.

The techniques employed in our prototype did not take advantage of special purpose hardware. In particular, no provisions were made for either real-time capture or full screen motion. However, by providing the simultaneous display of multiple, scalable, moving

movies, "multi-media" applications which use preprocessed, window-sized video sequences can be designed.

There are great advantages in using digital video over analog video in when creating multi-media systems. Problems occurring when applying analog video include:

- limited means of distribution: analog video is precluded from many digital distribution channels, such as computer local area networks and digital storage media
- the fixed raster format makes analog signals difficult to scale, and puts upper bound on image quality
- integration of video into applications requires special hardware
- Access to the video material is indirect and limited

In contrast, digital video is inherently flexible:

- directly accessible in units of time or space, thus amenable to manipulation
- scalable in both duration and resolution
- production, distribution, and reconstruction are sufficiently decoupled, so that image quality can be made directly proportional to size of bandwidth and cost of processing

A primary goal of our approach was to minimize the necessary computation while maintaining a reasonably low bandwidth. We accomplished this by using a great deal of data reduction together with a small amount of data encoding. Our movies were displayed small, in windows, but "stills" can be displayed at full screen resolution. The video we

provide is not of as high a quality or as low a bandwidth as is available by adding special decompression hardware, but is still clear and manageable. Future considerations include the use of hardware assist in compression/decompression in order to provide real-time input to the system, as well as more rapid response when reconstituting the video.

* * * * * *

References/Publications/Theses

Bender, Walter and Robert Mollitor. Digital movies for "off-the-shelf displays. *Proceedings, SPIE*, Vol. 1258, February 1990.

LARGE FORMAT DISPLAYS

Relevant Personnel:

Work under this topic was conducted by **Ronald L. MacNeil**, a Principal Research Associate at the Media Laboratory and co-founder (**with Muriel R. Cooper**) of the Lab's Visible Language Workshop.

* * * * * * *

6k by 2k Display Prototype

Command, control and communications display systems must support local and remote groups of viewers at varying viewing distances with extraordinary amounts of dynamic visual information. Conventional large scale displays trade away the resolution needed to provide local context. We have built a prototype display which is the largest high-resolution display extant with a total pixel count of 6k by 2k over 60 by 20 inches. It has an underlying software architecture and window tool kit that allows rapid and flexible prototyping of experimental graphic environments

The display prototype is constructed using three 2k by 2k Sony CRT monitors oriented around a first surface 50/50 beamsplitter so that the vertical edges of the center and adjacent monitors overlap slightly to create a perceptually seamless display. A virtual framebuffer software layer makes the display seem like a continuous buffer to the programmer, taking care of gradient overlapping of the seams and reversing the center image. [Masuishi, MacNeil, & Small, 1991]

A news and weather browsing and query application was prototyped using as a backdrop a 4k by 8k pixel cloudless image of the earth derived from over 1000 4km landsat images [VanSant 91]. News stories from around the world are arranged according to their geographic position.

The graphics toolkit includes video and sound. The display has been linked through a local fiber optic network to a remote research lab and remote teleconferencing experiments have been conducted using multiple video windows and stereo sound on the large high-resolution display.

* * * * * *

References/Publications/Theses

T. Masuishi Tetsuya, Ronald MacNeil, and David Small. (1992) "6,000 x 2,000 Display prototype." *Proceedings of SPIE/IS&T Symposium on Electronic Imaging Science & Technology.* San Jose, CA, February.

Van Sant, Tom. (1991) Satellite Composite view of Earth, by Tom Van Sant and the GEOSPHERE PROJECT, Santa Monica CA. With assistance from NOOA, NASA, EYES ON EARTH. Technical direction: Lloyd Van Warren. Source data derived from NOAA/TIROS-N Series Satellites. Completed April 15, 1990. Tom Van Sant, Inc. 146 Entrada Drive, Santa Monica, CA.

TERRAIN VISUALIZATION

Relevant Personnel:

Work in this area is under the direction of **Dr. David L. Zeltzer**, Associate Professor of Computer Graphics and Director of the Lab's Computer Graphics and Animation Group.

* * * * * * *

Virtual environment (VE) systems for mission visualization and rehearsal

The importance of simulation and virtual environments (VEs) has been amply demonstrated in the training domain, in the form of flight simulation systems [9], and networked land vehicle simulators for group training exercises [5]. The central focus of the work reported here is the design and implementation of VEs for command and control. In particular, we wished to demonstrate the utility of emerging VE technologies for visualizing and interacting with computational models of terrain, vehicles and aircraft engaged in simulated operations. Along the way we have explored VE software, interface and modeling technologies. These include:

- the bolio VE system, which is capable of displaying synthetic human figures using kinematic simulation of gait;
- a study of whole-hand input, including a taxonomy of hand gestures for input to VE systems;
- a constraint package for describing causal dependencies among objects, agents and processes in the virtual world;

- the *virtual erector set*, a modular and flexible system for dynamic simulation of mechanisms and vehicles;
- a task level VE system for mission planning, which incorporates
- an *intelligent camera controller* for maintaining visual context and ensuring smooth visual transitions.

We will briefly describe each of these in the following sections.

The *bolio* virtual environment system

An *integrated graphical simulation platform* (IGSP) is the software platform for implementing a VE. Our first generation IGSP—which we call *bolio*—has served as the testbed for developing the terrain and mission visualizers. The *bolio* system has been reported at technical conferences and symposia [2, 6, 14, 21] and will not be described in detail in this document.

Work on the *bolio* VE system was begun in 1986 by researchers in the Computer Graphics and Animation Group of the MIT Media Lab. Initial design goals were twofold: to provide an intuitive editor for graphical objects at multiple levels of detail; and to serve as a transparent, modular and device-independent developers' interface for a variety of applications [2].

Bolio has since evolved into a VE system that supports multi-modal interaction with distributed, concurrent simulations [13, 21] while it has continued to serve as a devel-

opment testbed with a modular and consistent programmers' interface. This developers' interface includes a global object data base; a mechanism for passing messages among distributed processes; and a function library for access to the primitive operations for kinematics, dynamics, and the constraint definition and satisfaction tools that *bolio* provides [19]. *bolio* has been used by a number of researchers to develop standalone applications packages [3, 6, 8, 12, 15, 20]. Among these is the application dubbed *sa* [18], which was designed for the kinematic simulation of human gait; here, *sa* was interfaced to the *bolio* system, and was used to depict human figures walking over uneven terrain.

The *bolio* system architecture is built around a process scheduler and an associated constraint network that link concurrent simulation processes, displays and input devices [14]. Displays have included conventional monitors as well as VPL LX and HRX Eyephones, a specially-constructed 3-axis force output joystick [10, 17], and stereo audio output using an Apple Mac II to store and retrieve pre-digitized sound samples. Input devices have included Polhemus and Ascension trackers, the six degree of freedom Spaceball, the Articulate Systems *Voice Navigator* for spoken input, VPL DataGloves, the EXOS Dexterous Hand Master, the Mattel Powerglove, and the above-mentioned force-output joystick.

Bolio is implemented in C and runs on several different workstations, including Stardent Titan and HP 800 and 700 series workstations. Control of most of the I/O devices is offloaded onto smaller desktop processors or remote workstations with either Ethernet or serial links to the main workstation. Experience has shown that the overhead for interprocess communication with remote applications and I/O devices is minimal [14].

Whole-hand input

Several new devices have become commercially available for measuring hand motion, including the DataGlove from VPL Inc., and the Dexterous Hand Master from EXOS, Inc. Using these hand-motion measurement devices, we have investigated whole-hand interaction with virtual objects and agents of varying levels of autonomy. In this work we have developed a taxonomy of hand motions useful for direct manipulation in virtual environments. In particular, this taxonomy allows us to use the human hand to implement the conventional logical input devices, *button, valuator*, and *locator*. Rather than building graphical interfaces with virtual devices (i.e., graphically displayed buttons, dials and sliders that are manipulated using a mouse or stylus), we can implement these eminently useful functions as postures and motions of the user's hand. We have found this to facilitate efficient and natural interaction with virtual environments. This work is reported in detail in [13] and [15].

The *bolio* constraint network

By associating events and processes in sets of causal relationships, a *constraint network* defines the top-level behavior of a VE system, as well as behaviors of other components and sub-processes [14]. For example, a program which controls the physically based behavior of a mechanical assembly would send commands to a mechanical simulation module to update the state of the assembly when one of the parts is "moved" by a human wearing a whole-hand input device (e.g., DataGlove). While these associations can be

hard-coded into a VE system, a language for dynamically specifying dependencies and attachments is a very general tool for describing and modifying virtual worlds [21].

A constraint package was a key element of Sutherland's classic work, *Sketchpad* [16], and of Borning's *Thinglab* [1]. Our work with constraints is similar in spirit to both of these systems, although both of them were restricted to 2D graphics. All three systems incorporate rather general mechanisms for defining constraints and constraint satisfaction methods. However, the two earlier constraint systems incorporate an analysis stage, and Borning's work included two additional satisfaction techniques beyond one-pass solutions and relaxation.

Manus was developed initially to handle position and orientation constraints on the motion of rigid objects, and non-rigid motion of polygonal meshes. Thus, unlike the earlier *Sketchpad* and *Thinglab* systems, which were intended to satisfy multiple, interacting constraints encountered in geometric and mechanical design problems, *bolio* does not perform preliminary analysis of the constraint network. Since relaxation is time-consuming and may not converge, the purpose of this constraint planning step is to identify constraints that can be satisfied by simpler, direct means, so that relaxation is invoked only when necessary. However, *bolio* supports an interactive, time-varying virtual environment, perhaps with active agents whose behavior may not be known *a priori*. Thus, constraint satisfaction has to proceed in parallel with forward simulation, and a constraint pre-planning stage is not feasible.

The *manus* constraint network is composed of *bolio* data objects—bOBJECTs—in the *bolio* world connected by instances of constraints. Each instance of a constraint contains information specific to the objects it is connected to, and pointers to the code necessary to process the constraint. Thus, constraint instances share procedures but maintain private copies of relevant data structures. Each time a constraint instance connects to a bOBJECT which should trigger it, it adds a pointer to itself into the bOBJECTs *who-cares* list (part of the `constraints` structure). Later, when a constraint instance modifies the bOBJECTs, the bOBJECT notifies all constraint instances in its *who-cares* list.

Those constraint instances then execute, modifying other bOBJECTs which trigger constraint instances in their *who-cares* list, etc. This process proceeds in an manner managed by the *manus_renormalize* function. When a bOBJECT triggers constraint instances in its *who-cares* list it actually just puts a pointer to each instance on the end of a global *pending constraint instance* list (*pending queue*). The *manus_renormalize* function goes sequentially through the queue (in effect a breadth-first search of the constraint network) invoking constraint instances as they are pulled from the list. As constraint instances execute, objects they affect place new items at the end of the queue. This procedure continues until the queue is empty. Details of the bolio constraint system can be found in [14, 21].

The Virtual Erector Set (VES)

Dynamic simulation is fundamental to modeling the realistic behavior of agents and objects in simulated worlds. Point-mass dynamics, however, is clearly inadequate for modeling objects such as aircraft and vehicles. Therefore, we require methods for simulating the

dynamics of articulated rigid-body assemblages. Unfortunately, we encounter increasing computational expense as the objects to be simulated using forward dynamics increase in complexity.

Work in robotics by Featherstone has resulted in a forward dynamics algorithm that is of linear time complexity [4]. We have implemented and extended this algorithm in two ways. First, Featherstone expresses his algorithm in terms of *spatial notation*, which is based on the *screw calculus*. We have extended this notation such that the mathematical expressions are simpler and more efficient to execute. Secondly, we have incorporated a technique suggested by Lathrop which allows kinematic constraints—such as fixed endpoints—to be freely mixed with specifications of forces and torques acting on arbitrary linkages.

This system has been integrated into our *bolio* IGSP, and provides us with a forward dynamics engine with a consistent and uniform interface. That is, it is now straightforward to incorporate additional simulation packages into *bolio* which can rely on the Virtual Erector Set to perform dynamics computation.

The Virtual Erector Set has provided us with a robust and general tool for computing forward dynamics with endpoint constraints. Development was speeded up due to our use of C++, which facilitated easy coding of mathematical routines using operator overloading.

Actual timings of the system indicate that it does indeed yield linear complexity in the number of degrees of freedom to be simulated. This work is reported in detail in [11, 12].

The mission visualizer

Emerging *virtual environment* (VE) technologies offer powerful and productive human interface solutions for command and control problems, and we have set out to demonstrate the utility of VE technology in this domain.

Our work follows a *task-level* analysis of the goals and requirements of a mission planning system, in which workload and stress levels are likely to be very high. It is particularly important in such situations that the computer interface be as transparent as possible, requiring a minimum of computer expertise and programming skills. Operators and aircrews should deal directly with the objects and processes associated with the task, using a vocabulary and sensorimotor skills that are already familiar to them [19]. We assume that some other module may perform optimal route calculations; our system could input these proposed solutions, display them dynamically and in 3D, and make it easy for personnel to interactively modify flight paths, for example, to take advantage of local terrain masking. For this reason, we think of the system as a mission *visualizer*.

Unlike a flight simulator, in which interaction is restricted to cockpit tasks, we have designed the system to emphasize *task level* interaction with the environment: the operator can easily change not only his viewpoint, but all the objects and their positions within the environment. We provide models of aircraft, terrain, threats and targets, and users interact directly with these models—voice recognition for speech input, a VPL DataGlove for

positional input and gesture and posture recognition, and a more conventional mouse and keyboard interface are all supported. Finally, we provide a range of sensory displays, including wide field of view (FOV) visual displays, and a force output joystick—a device that can generate force cues for the operator.

Using 3D input devices, the operator can specify waypoints through which an aircraft should travel, or prespecified waypoints can be read in from an external source. Once defined in either fashion, the aircraft module will generate a flight path which will pass through each of the waypoints, if possible, based on a simplified aerodynamic model (actually, an A-4 Skyhawk). The system can represent and display an unlimited number of aircraft, each following its own flight path. Motion of the aircraft along flightpaths can be interactively controlled using "VCR" controls that allow the operator to stop action, back up, or fast forward motion as necessary.

The operator can also locate points of interest on the terrain, specify their identity and set the view to originate or terminate at that point, so that lines of sight (LOS) to or from aircraft can be checked. The view can be made to track the motion of any specified aircraft, track a location on the ground from any of the aircraft, or track the motion of a moving object on the ground. This makes it extremely easy to examine LOS between any objects in the system.

Camera control in the system is designed to be both flexible and powerful. Out-the-window views can be easily directed to any object in the environment, including other aircraft, targets or threats. Views may simply display straight-ahead out-the-window imagery, or can be made to follow objects as they move through space. At any time, the point-of-regard

and the viewpoint can be swapped. Voice control is incorporated, so all the operations described in this section are available via hands-free operation.

Besides simple viewpoint and point of regard control, camera control based on conventional camera movements are possible. For instance, the camera may pan, truck (camera view point moves toward the point of regard), dolly (camera viewpoint moves perpendicular to the vector connecting the viewpoint and point of regard), crane, and zoom (field of view is increased or decreased). Certain points of view can be saved and retrieved for instantaneous changes to standard reference views.

A more sophisticated level of view control maintains the positions of selected objects at certain points in the frame. For example, rather than generating an out-the-window view of a target, it may be more informative to construct a view external to both the aircraft and the target.

Finally, a system for automatically tracking the closest target or threat is available. This system can be used to merely indicate the closest target via a line between an aircraft and the target, or the point of regard can be automatically changed and tracked so that the object of regard is *always* the closest target.

Terrain may be derived from aerial imagery using a depth-from-shading algorithm (Pentland, Alex. P. Shape information from shading: a theory of human perception. *Proceedings of the Second International Conference on Computer Vision,* Tampa, FLA, Dec. 5-8, 1988, 448-455.), or terrain objects may be defined and modified interactively using solid modeling tools. In principle, terrain information can be derived from other sources, e.g., DMA DTED, but we have not implemented the necessary interfaces. Terrain

masked radar coverage is generated and displayed automatically, and the locations of aircraft with respect to a radar site are continuously tracked. When an aircraft is "visible" to a radar, the color of the aircraft is changed.

Mission scenarios can easily be created, stored and retrieved using the system. Waypoints are created using the DataGlove, and paths through these waypoints can be displayed and edited using the DataGlove. These paths can then be saved in separate files and loaded at a later time. Positions of targets and threats can also be saved and retrieved. "Menu buttons" can be created that load aircraft, threats and targets with a single selection. Complicated scenarios can thus be accomplished by simple scripting without any need for recompilation. Scripting is performed using Tcl ("tool command language"—a public domain front-end scripting language [7]), along with an in-house interface to OSF/Motif.

We have developed a mission visualizer based on a VEsystem that we think enhances situational awareness and provides for easy interaction with proposed mission routes using 3D perspective graphics, and voice and gesture interaction. While we have not conducted formal evaluations of the system, many visitors, including DoD and Air Force personnel, have seen the system in operation and have commented favorably. Further details can be found in [20].

References/Publications/Theses

1. Borning, A. Thinglab -- A Constraint-Oriented Simulation Laboratory, Tech. Report No. SSL-79-3, July 1979, Xerox PARC: Palo Alto CA.

2. Brett, C., S. Pieper, and D. Zeltzer. *Putting It All Together: An Integrated Package for Viewing and Editing 3D Microworlds* in Proc. 4th Usenix Computer Graphics Workshop, October 1987, pp. 2-12.

3. Drucker, S., T. Galyean, and D. Zeltzer. *CINEMA: A System for Procedural Camera Movements* in Proc. 1992 Symposium on Interactive 3D Graphics, March 29-April 1, 1992, Cambridge MA: ACM Press, pp. 67-70.

4. Featherstone, R., Robot Dynamics Algorithms. 1987, Kluwer Academic Publishers.

5. Johnston, R.S. *The SIMNET Visual System* in Proc. Ninth ITEC Conf., Nov. 30 - Dec. 2 1987, Washington DC, pp. 264-273.

6. McKenna, M., S. Pieper, and D. Zeltzer. *Control of a Virtual Actor: The Roach* in Proc. 1990 Symposium on Interactive 3D Graphics, March 25-28, 1990, pp. 165-174.

7. Ousterhout, J.K. *Tcl: An Embeddable Command Language* in Proc. 1990 Winter USENIX Conference, 1990.

8. Pieper, S., J. Rosen, and D. Zeltzer. *Interactive Graphics for Plastic Surgery: A Task-Level Analysis and Implementation* in Proc. 1992 Symposium on Interactive 3D Graphics, March 29-April 1, 1992, Cambridge MA: ACM Press, pp. 127-134.

9. Rolfe, J.M. and K.J. Staples, ed. Flight Simulation. 1986, Cambridge University Press: Cambridge, England.

10. Russo, M.A. The Design and Implementation of a Three Degree of Freedom Force Output Joystick, M.S. Thesis, May 1990, Massachusetts Institute of Technology.

11. Schröder, P. The Virtual Erector Set, M.S. Thesis, January 1990, Massachusetts Institute of Technology.

12. Schröder, P. and D. Zeltzer. *The Virtual Erector Set: Dynamic Simulation with Linear Recursive Constraint Propagation* in Proc. 1990 Symposium on Interactive 3D Graphics, March 25-28, 1990, pp. 23-31.

13. Sturman, D., D. Zeltzer, and S. Pieper. *Hands-On Interaction with Virtual Environments* in Proc. ACM SIGGRAPH Symposium on User Interface Software and Technolgy, Nov. 13-15, 1989, pp. 19-24.

14. Sturman, D., D. Zeltzer, and S. Pieper. *The Use of Constraints in the bolio System* in ACM SIGGRAPH 89 Course Notes, Implementing and Interacting with Realtime Microworlds, July 31, 1989, Boston MA, pp. 4-1 - 4-10.

15. Sturman, D.J. Whole-hand Input, Ph.D. Thesis, February 1992, Massachusetts Institute of Technology.

16. Sutherland, I.E., *Sketchpad: A Man-Machine Graphical Communication System*. Proc. AFIPS Spring Joint Computer Conf., Spring 1963, 23329-346.

17. Tadros, A.H. Control System Design for a Three Degree of Freedom Virtual Envirnomnet Simulator Using Motor/Brake Pair Actuators, M.S. Thesis, February 1990, Massachusetts institute of Technology.

18. Zeltzer, D., *Motor Control Techniques for Figure Animation*. IEEE Computer Graphics and Applications, November 1982, 2(9), pp. 53-59.

19. Zeltzer, D., *Task Level Graphical Simulation: Abstraction, Representation and Control,* in Making Them Move: Mechanics, Control and Animation of Articulated Figures, N. Badler, B. Barsky, andD. Zeltzer, 1991, Morgan Kaufmann: San Mateo CA, pp. 3-33.

20. Zeltzer, D. and S. Drucker. *A Virtual Environment System for Mission Planning* in Proc. 1992 IMAGE VI Conference, July 14-17, 1992, Phoenix AZ, pp. 125-134.

21. Zeltzer, D., S. Pieper, and D. Sturman. *An Integrated Graphical Simulation Platform* in Proc. Graphics Interface '89, June 19-23, 1989, pp. 266-274.

THE Z-AXIS OF INFORMATION

Shading Analysis of Terrain

Relevant Personnel:

Work under this topic was conducted by **Dr. Alex Pentland**, Associate Professor of Computers, Communication, and Design Technology, and co-Director (with **Dr. Edward H. Adelson**) of the Lab's Vision Science Group.

* * * * * * *

Shape From Shading

Summary

We have developed a closed-form approximate solution to the problem of extracting shape information from image shading, given standard assumptions. Neither integration nor iterative propagation of information is required. When boundary conditions (e.g., edges, singular points) are not available, good estimates of shape may still be extracted by using the assumption of general viewing position. An improved method for estimating the illuminant direction was also developed.

Project Description

The extraction of shape from shading has a relatively long history within the field of computer vision. There have been two general classes of algorithm developed: *local algorithms*, which attempt to estimate shape from local variations in image intensity, and

global algorithms, which attempt to propagate information across a shaded surface starting from points with known surface orientation.

Local algorithms, originally suggested by Dr. Alex Pentland, use strong assumptions about surface shape in order to extract estimates of surface orientation from the shading information within a small image neighborhood. A subsequent integration step is required to convert estimated surface orientation into an estimate of surface shape. These local methods of estimating surface orientation have been shown capable of producing good estimates of shape, however they do not produce exact estimates except under certain limited situations.

Global algorithms, primarily due to Dr. Berthold K. P. Horn of MIT and his students, also make use of an assumption about surface shape in order to extract estimates of surface orientation. In these algorithms, a smoothness assumption is used to relate adjoining points, enabling spatially-isolated information about absolute surface orientation (which must be derived using some other technique) to be iteratively propagated across the surface. The use of a smoothness assumption, however, implies that the algorithms will not produce exact solutions except under certain limited situations. As with the local estimation algorithms, integration is normally required to obtain the surface shape.

In this project, we found that by approximating the image reflectance function using the linear terms of a Taylor series, we were able to derived a simple closed-form expression that relates surface shape to image intensity. This new result may be used in several ways. First, this approximation provides a new method of estimating the illuminant direction that makes weaker assumptions about the viewed scene. Second, this approximation is invertible in closed form, so that shape information may be recovered directly from image

shading without the necessity of integration or iterative propagation of shape information. An finally, because the technique can be implemented by use of linear filters similar to those thought to exist in biological visual systems it may serve as a model for human perception.

Special aspects of this approach are that it makes no assumption about surface smoothness or shape, and that it does not require (but can make use of) boundary conditions to obtain an estimate of shape. To avoid assumptions about surface smoothness or shape, we have simplified the shape-from-shading problem by using a linear approximation of the reflectance function. To avoid requiring known boundary conditions, we have used the assumption of general viewing position to fill in missing boundary conditions with default values. The use of these default boundary conditions seems to produce the most accurate shape estimates for complex, highly-textured surfaces. When boundary conditions are available, of course, they can be directly incorporated into the shape estimate as described above.

We believe that this approach to shape-from-shading seems best suited to the recovery of the high-frequency shape details that are difficult to recover by use of stereo or motion cues. Because there is no smoothness assumption the technique can be directly applied to complex (but still continuous) natural surfaces such as hair, cloth or mountains. Further, experimental results indicate that the recovery process is stable and—especially for the high-frequency shape details—can be quite accurate. One natural method of integrating coarse stereo (or motion) information with this shape-from-shading technique would be to combine their shape estimates in the frequency domain, weighting the stereo information

most heavily in the low frequencies and the shading information most heavily in the higher frequencies.

* * * * * * *

References/Publications/Theses

Horn, B. K. P. and Weldon, E. J. (1988) Direct methods for recovering motion. *International Journal of Computer Vision*, Vol. 2, No. 1, pp. 51-76.

Pentland, Alex. Photometric motion. (1991) IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13, No. 9, September.

Alex Pentland. (1990) *Fast Surface Estimation Using Wavelet Bases*. M.I.T. Media Lab Vision and Modeling Group Technical Report No. 142, June 1990.

Pentland, Alex. P. (1988) Shape information from shading: a theory of human perception. *Proceedings of the Second International Conference on Computer Vision,* Tampa, FLA, Dec. 5-8, 448-455.

Alex Pentland and Stan Sclaroff. (1991) Closed-Form Solutions for Physically-Based Shape Modeling and Recognition. In *IEEE Transactions on Pattern Analysis & Machine Vision,* Vol. 13, No. 7, 715-730. 1991.

Managing Visual Complexity

- Adaptive Typography for Dynamic Mapping Environments
- Development of User Interfaces for Multi-Plane Graphics
- Multi-Plane Image Compositing System

Relevant Personnel:

Work under these three sub-topics of managing visually complexity was done under the supervision of **Prof. Muriel R. Cooper**, Professor of Visual Studies and Director of the Lab's Visual Language Workshop, and **Mr. Ronald L. MacNeil**, a Principal Research Associate at the Media Laboratory and co-founder with Prof. Cooper of the Visible Language Workshop.

* * * * * * *

SUMMARY

The central goal of the work over the contract period has been to build both a hardware and software environment, including a set of example applications, for exploring advanced graphics and AI techniques for controlling visual complexity in large scale high resolution displays for Command, Control and Communications.

To this end we have constructed a prototype display which we believe to be the world's largest high resolution display with a resolution of 6000 by 2000 pixels over a 60 by 20

inch CRT surface. The software framework includes a virtual framebuffer architecture which abstracts away the contorted geometry of the display. Our unique object oriented window toolkit provides the highest graphic quality anti-aliased fonts, and allows every object to have variable translucency and focus. Layout relationships between these graphic objects can be controlled automatically by networks of constraints invoked from a case library.

Applications have been written to explore various C3 scenarios such as:

- querying an on-line world news and weather events database on the large hi-resolution display using a cloudless 4km landsat image of earth as the backdrop which enables many related stories to be seen in their geographic context

- Using the parameters of color, transparency, and focus in a display with as many as 30 salient layers of information we can precisely control which objects are brought to the users attention and by what amount

- Using the Intelligent Layout system presentations can be generated automatically which maintain their legibility, structure and style independent of the screen real estate allocated for them

Adaptive Typography for Dynamic Mapping Environments

When typography moves across a geographic situation display map it will pass over different colors, densities and textures which can obscure or make ambiguous its meaning. We have developed image processing techniques for maintaining discernibility and semantic consistency of typographic elements.

Our approach to maintaining discernibility of the font character is as follows. The critical discernibility zone of fonts lies between the X height line and the base line. Our adapted text is a composite of three elements 1) a "fuzzy box" made by low pass filtering the background under the critical discernibility zone 2) a translucency buffer made by sampling a symmetric Hermite cubic spline, 3) the original anti-aliased character bits. The resultant type can be placed over map backgrounds which are visually very noisy and confusing without impairing the discernibility of the type. [Bardon 91]

Development of User Interfaces for Multiplane Graphics

Color Management in Dynamic Mapping Environments:

Improper color management in a situation display can result in unintentional ambiguities or erroneous cognitive groupings. We have developed an environment which allows graphic objects to be grouped according to their current relative salience into foreground, context and background families, and for these groupings to be modified as the situation demands.

Our approach starts by mapping each object's color to a 3-component vector in Munsell color space, a perceptually derived and therefore perceptually consistent color space. The

interface shows the user where each object lies within this 3-D color space, and provides guides and containers for grouping objects. A heuristic color constraint management system allows the user to modify the local color relationships while maintaining global ordering. [Bardon 92]

Layout Intelligence:

Communication designers practice an effective form of complexity management when they lay out graphic elements against a grid system. The size, relative position and style of the elements and their relationships tells the user about the information hierarchy, making quick, unambiguous reading possible. Dynamic information environments, where the size and disposition of elements and their relationships must change rapidly must be able to function with the same graphic quality as static print graphics.

Our answer to this challenge is to employ the Artificial Intelligence technologies of *constraints* and *case-based reasoning* to create an environment in which good graphic layouts are represented as networks of constraints between graphic elements, with rules for legibility. New layouts are generated by searching the case library for the closest applicable layout and adapting it to the exigencies of the current layout task. [Colby 92]

Multi-Plane Image Compositing System

Information displays for command and control can become dangerous when the amount and complexity of the information becomes too high, or the meaning of the organization of graphic elements is ambiguous. Conventional techniques for dealing with complexity

issues typically lose context when lower priority information is removed or overload the user by adding warning signals. We have developed methods which can maintain context while bringing the level of complexity of such displays under control, either manually or automatically. [Colby, Scholl 91; Scholl 91]

Our approach is to apply image processing techniques that enable the system to control both the gradients of focus and transparency within an image. A real time demonstration application was developed on the massively parallel CM2 using a million virtual processors, one per pixel from 16K real processors. Each layer of a 30-feature layer map was blurred using pyramid coding techniques, interpolated and composited on the fly, using Talbot's Transparency algorithm. An multi-level interface was built which allows the user to control each parameter or to aggregate and prioritized for quick, high-level decision making.

* * * * * *

References/Publications/Theses

Baron, Divider. (1991) Adaptive typography for dynamic mapping environments. To appear in *Proceedings of Image Handling and Reproduction Systems Integration Conference*, San Jose, CA, February 24-March 1.

Bardon, Didier. (1992) *Adaptive color in dynamic mapping: a method for predictable color modifications*. Unpublished M^S VS thesis, Media Arts and Sciences Section, MIT.

Colby, Grace. (1992) *Intelligent layout for information display: an approach using constraints and case-based reasoning*. Unpublished MSVS thesis, Media Arts and Sciences Section, MIT.

Grace Colby and Laura Scholl. (1991) Transparency and Blur as Selective Cues for Complex Visual Information. To appear in *Proceedings of Image Handling and Reproduction Systems Integration Conference*, San Jose, CA, February 24-March 1, 1991.

MacNeil, Ronald. Generating Multimedia Presentations Automatically using TYRO, the Constraint–Based Designer's Apprentice. *Proceedings of the IEEE Workshop on Visual Languages*, Koobe, Japan, October 1991.

MacNeil, Ronald. Adaptive perspectives: Case-based reasoning with TYRO, the Graphic Designer's Apprentice. Published in *Proceedings of the IEEE 1990 Workshop on Visual Languages*.

MacNeil, Ronald. TYRO: a Constraint Based Graphic Designer's Apprentice. *Proceedings of IEEE Workshop on Visual Languages*. 1989.

Scholl, Laura. (1991) *The transitional image.* Unpublished MSVS thesis, Media Arts and Sciences Section, MIT.

DYNAMIC MAPS

Relevant Personnel:

Work in this area was under the direction of **Dr. Edward Adelson**, Associate Professor of Visual Sciences and co-Director (with **Dr. Alex Pentland**) of the Lab's Vision Science Group.

* * * * * * *

Optimal usage of image data requires that the imagery be of highest visual quality and that it be stored and retrieved with the greatest possible efficiency. For this reason we have investigated a number of techniques for image processing and image coding for image data compression.

Image compression

For image data compression, it is desirable to have a method that reduces the data as much as possible while maintaining visual fidelity that is as good as possible. We have developed and implemented a set of coding methods based on quadrature mirror filter (QMF) pyramids and subband coding. The QMF kernels have been specifically designed so as to be compact in space and have good frequency selectivity; orientation selectivity has been a criterion for some of these kernels as well. In addition, the kernels are designed to minimize the visibility of coding artifacts.

QMF pyramids, it should be noted, are very closely related to the "wavelet" methods that have been receiving so much attention recently. QMF's were developed by workers in the signal processing community; the wavelets were developed independently by mathematicians. While there have been claims that the wavelets offer revolutionary advantages, there seems to be little difference in actual performance, and the two classes of techniques may be regarded as equivalent in most real world applications to image coding.

The image data compression with QMF pyramids has been quite successful. We find that the 9-tap separable QMF pyramids offer an excellent combination of properties, including efficient computation and good image coding at low bit rates. We have applied them to many kinds of imagery, including color maps, range imagery, and human faces.

We have also developed a subband data compression scheme that allows for very fast reconstruction while giving very good data compression. In order to allow easy use of the scheme, and in order to allow evaluation and comparison with other systems, we have put together a C program called "EPIC," for Efficient Pyramid Image Code. EPIC was developed on Suns but should run on most UNIX systems with minor modifications. The code is optimized for rapid reconstruction. A mixed Huffman/run-length coding scheme is used for entropy coding.

Steerable filters

For image analysis and image enhancement, we have developed a new family of 2-D filters that we call "steerable filters," which are particularly suitable for extracting information about oriented structures such as lines or edges, as well as for analyzing texture. In

addition, the steerable filters can be generalized to 3-D for use in analyzing motion sequences, where time is considered to be the third dimension (X-Y-T). The steerable filters offer a family of image processing tools that allow for efficient computation.

For image analysis and enhancement we are quite pleased with the results of the steerable filters. They enable use to find and enhance oriented features in natural scenes.

We have also designed a "steerable pyramid," in which the steerable filter concept is generalized to a multiscale format. The steerable pyramid allows for image decomposition and reconstruction. We have used it successfully in image noise removal.

Motion analysis

For motion analysis we have adopted a least-squares approach. A patch-wise least-squares method using a uniform translation model is preferred in cases where the motion field is spatially complex, while an affine least-squares method over a larger region is preferred when the affine transformation model is valid.

We have implemented a multiscale version of a least-squares gradient method for analyzing image motion. We build a multiscale image pyramid, and begin the analysis at the coarse scale. This allows the initial computations to proceed quickly and deals with the problem of multiple matches. We use the extracted flow field to produce an initial warp map of the image. This procedure is iterated from one level to the next, ending with a high resolution warp map. We have run it on synthetic image sequences and have retrieved a good approximation to the true warp field.

This motion analysis has been applied to perform motion-compensated noise reduction on image sequences. For example, we begin with a noisy image sequence of a van driving away. The license plate is not legible in any individual frame, but after the noise reduction process one can clearly read it.

* * * * * * *

References/Publications/Theses

William T. Freeman and Edward H., Adelson. "The Design and Use of Steerable Filters." *IEEE Transactions of Pattern Analysis and Machine Intelligence*, Vol. 13, No. 9, 891-906. September 1991.

Eero P. Simoncelli and Edward H. Adelson. Subband Transforms. In: John W. Woods (ed.), *Subband Image Coding*. Norwell, MA: Kluwer Academic Publishers, 1990.

Eero P. Simoncelli and Edward H. Adelson. "Non-separable extensions of quadrature mirror filters to multiple dimensions." *Proceedings of the IEEE*, Vol. 78, No. 4, 652-664. April 1990.

Michael A. Sokolov. *Visual motion: algorithms for analysis and application*. Unpublished M.S. thesis, Media Arts and Sciences Section, MIT, February, 1990.

MULTI-MODAL INTERACTION & SEMANTIC and DISCOURSE STRUCTURE

Relevant Personnel:

Work in this area is under the direction of **Dr. Richard A. Bolt**, Senior Research Scientist and Director of the Media Lab's Advanced Human Interface Group.

* * * * * * *

Overview

The goal of this research is to enable commanders to communicate directly with displayed information via concurrent speech, gestures, and gaze. This work has two aspects. One is making technologies to capture speech, gesture, and eye outputs work in concert. The other is developing the software intelligence to interpret such inputs in context and to map to an appropriate response in graphics and speech or non-speech sound. Such multi-modal interaction is not necessarily to the exclusion of other modes such as keyboards and mouse or any other i/o modes, but complementary thereto.

This project stresses *multi-modal* natural dialog with displayed information. To now, work in natural language dialog has concentrated on processing strings of words in a natural language (e.g., English, French, German, etc.) either typed or spoken. Some few studies have combined spoken input with manual pointing; none to our knowledge have combined *all three* modes of speech, gestures, and gaze concurrently and in a single system.

Language in everyday life, particularly where people are discussing things right in front of them—as would be the case with commanders before map displays—is ordinarily accompanied by gestures and glances. Such gestures pointing out this or that item, describe this or that movement or action. Glances serve the multiple roles of referencing things (by glancing at them), alerting a dialog partner that this or that gesture is important (looking at their hands), and in general serving to orchestrate the "social" side of the dialog. Thus, traditional natural language work, by not involving all three input modes, has left out major parts of what "naturally" occurs in natural language.

We have made a strong beginning toward making it possible for commanders to communicate with displayed data directly via such multi-modal natural language. Three achievements of note in this work, specifically, are that we have implemented and demonstrated for the first time ever:

- speech, gaze, and manual input combined at a single interface. Prior interface work, both our own and that of others, has use speech-plus-pointing, gaze alone, speech alone, but not before with all three modes working concurrently.

- *two-hand, free-handed gestural input,* with accompanying speech and gaze, to indicate, rotations, shifts, translations, re-scalings, of solid 3-D items in free-space.

- *modulation of free-hand gestural input by gaze*, specifically the case where the user is looking at the hands thus makes the hands referential, not merely for emphasis.

We have demonstrated concurrent speech, gaze, and pointing gesture input in an interface that simulated a forest-fire situation. "Icons" representing bulldozers, water-carrying planes and helicopters, firefighter units, and current conflagrations were set against a background

map of forest terrain. The user could address an icon and inquire the status of the unit or fire that it represented. A typical inquiry took the form: "What is the status of that (looking and/or pointing)?" Icons representing units could be moved freely about, e.g., "Move that unit (looking and/or pointing) south of that unit (looking or pointing)." Units could be created ("Make a helicopter unit...there (looking and/or pointing), and named ("Call that (looking and/or pointing) unit 'Bravo.'"

While projects in the past have combined speech plus manual pointing (including our own lab's pioneering "Put-That-There" project), and several investigations (again, including some at our lab) have featured eye input (Cf. Bolt, 1992), this demonstration is the first in our knowledge to freely admix all three modes of input.

Free-hand manual input has been used to date mostly as an adjunct in "virtual reality" demonstrations where the user is visually immersed in a 3-D graphically rendered space displayed via helmet-mounted monitors. The user, wearing a special glove, indicates with the index finger of one hand the direction in which they would like to "fly," and raises or lowers the thumb as a kind of accelerator/brake. Additionally, in a recent CAD/CAM application, researchers used the single hand with limited speech to stretch and mold graphically rendered solid shapes. However, to our knowledge, our lab been the first to use *two* hands in *free-hand* manual input, that input accompanied by speech, and modulated by gaze.

In our demonstration system (Bolt and Herranz, 1992), the user could use two-handed gestures to scale, rotate, and translate 3-D solid objects (airplanes) on the graphics display. The amount of scaling or movement was gauged by the distance the hands were held apart in front of the user: that is, the distance between the hands was taken to reflect the relative

size of the item. Where there was more than one item on display; the item to be re-scaled or manipulated was the item the user was *looking at.*

Two general types of "co-verbal speech" were implemented in this two-handed gestural input situation. The first was *kinemimic* gestures, used with speech to indicate graphical manipulations or to describe dynamic situations. An example for graphical manipulation might be: "Turn that (looking at some specific item)...this way (the two hands describing a twisting motion, one hand "pinned" in place, the other sweeping an arc)." Another type of kinemimic gesture implemented was that of describing dynamic situations (much like the air ace describing an aerial "dogfight"): for example, "The tank went this way (describing a path with one hand), and the truck came along this way (describing a path with the other hand relative to the path described by the first hand)." The second type of gesture was *spatiographic*, used to specify static layouts or placement of items. An example might be: "The command post was here (indicating with one hand)...and the tower was here (indicating a spot with the other hand relative to the first)."

We have demonstrated what to our knowledge is the first instance of a speech-and-gesture input system which interprets such input based upon whether or not the user is *looking at their hands.* Consider someone describing how good the fishing is in a certain pond. They hold their hands up before them and "chop the air" while saying something like "The best fishing around these parts is at Moosehead Lake." Now, should the speaker be looking us in the eye, we spontaneously interpret their hand gesturing as merely lending emphasis to their words; the gestures carry no additional information. However, should they instead be looking down at their hands as they speak, we take that self-same gesture to indicate the *size of the fish.* What happens in this second instance—where we see another looking at

their hands as they speak—is that we interpret their hands as *referential*. We see their hands not simply performing "baton movements" along with their speech, but as pointing to some meaning beyond themselves.

Hardware and technologies

A Hewlett-Packard 835 workstation with Turbo SRX graphics served as the central machine of our mini-network. It receives inputs from several microprocessors in turn support eyetracking, speech recognition, and gesture sensing. This response is some action on the HP 835's graphics display, possibly accompanied by synthesized speech output on a DecTalk™ speech synthesis system.

It is vital that the user's inputs in speech, gesture, and gaze be synchronized in real-time. For instance, when the user utters "...that...," the system must be able to determine what item the user is looking or pointing to at that moment. Accordingly, all data arriving at the HP 835 is passed through an HP28667A Real-Time Interface Board to be "time-stamped" as to its time of arrival; much of our networking effort was devoted to resolve this synchronization issue.

Eyetracking is via head-mounted ISCAN™ eyetracking optics processed through ISCAN™ RK-426 image-processing logic board housed in an IBM AT PC. Hand-tracking is done by a left and right pair of DataGloves™, the raw output from either glove's logic box going into a DEC 5000 workstation; the DEC consolidates and pre-processes the raw glove data and sends it on to the central HP 835.

Over most of our effort we used discrete-word speech recognition, initially a small 50–word vocabulary system, (TI–Speech by Texas Instruments). We then went to a larger, more sophisticated 25K-word vocabulary system (DragonDictate by Dragon Systems), which also proved to be more robust in the presence of ambient noise. However, both discrete–word systems presented the problem of having to pause about a third of a second between spoken works, making stilted what should be naturally flowing. Later, we acquired a connected speech system (the Ruby system by BBN, Inc.) which permits spoken input without a pause required between words, not only to improve the "flow" of words, that of gesture and gaze as well. (Work with this system will be carried forward in our follow-on project ("Virtual Environments in Command and Control for Ultra-High Definition Displays," sponsored by DARPA/RL, Contract No. F30602-92-C-0141, effective 7/1/92).

The software "agent

The central HP 835 machine hosts the software "agent," implemented in LISP, that interprets the inputs in looking, speech, and gesture, and plans an appropriate response. The strategy in interpreting the multiple streams of speech, gesture, and looking is, first, to parse the incoming speech stream. That is, nothing occurs by virtue of looking or gestures, alone; actions in gestures and/or eyes by themselves are insufficient to set off any system action. In future versions of our system, we may introduce the case of system's responses where a gesture or glance is sufficient to set off some system action. This may be some aspect of personalization of the system, where some deep acquaintance between system and user has been built up with sufficient mutual understanding to allow such to occasionally

happen. Meanwhile, this convention avoids the situation where a random look or gesture may set off some unintended consequence.

First, the agent would parse the spoken input to see if the string of spoken words met an acceptable syntax. If the string of words did not meet an acceptable syntax—usually that of an interrogative sentence ("What is the status of that unit?") or a command ("Color that blue."), then the user was so informed via our system's speech synthesizer.

If the spoken sentence passed the input syntax check, then the input would be examined to see whether or not it contained a complete command at the level of speech. An example is: "Place the blue tank to the east of the red tower," where there is on the display only one blue tank and a single red tower. Under these circumstances, there is no ambiguity. Should there be no red tower on the display, however, there is a semantic error in the input: the situation given in words does not correspond to the state of affairs as given by the display. The error message in this case might be the computer saying to the user "Which tower?" (if more than one tower on display), or perhaps "Do you mean the green tower?" (if there were only one tower on view and its color were green).

If the input is syntactically complete and semantically consistent with the material on the screen, the system simply executes the command as give in speech. It will not insist that the user either be looking or pointing by hand at the items specified in speech. This is in contrast to the situation where the machine might be explaining to the user some aspect of an item on display; should the user not be looking at the item, or portion of the item the machine is discussing or explaining, it would make sense for the computer to insist that the user be looking at the appropriate spot on the display while it is trying to explain something or to describe some process.

Gestalts

When we use speech, gesture and gaze, we automatically distribute what we intend to say over those three modes. For instance, we say "Look at that!" while glancing in some direction, perhaps pointing at some particular thing. Correspondingly, our listener has a *complementary* skill in reconstituting our meaning by integrating what we say with what they observe in our gestures and gaze. It is this latter skill—the skill of the listener to put together the separate threads of meaning in speech, gesture, and gaze—which we are trying to impart to the machine. One aspect of this skill lies in people's ability, both as the sender and receiver of bits of multi-modal dialog—to see the world around them as organized in "gestalts": to see certain items as "belonging together" on the basis of common color, proximity, size, shape, and so on. This perceptual ability makes it possible for people to reference things as *groups*, thus facilitating communication.

We have taken steps to have the machine deal with items on its display in terms of such "gestalts." The user says "Delete those," while sweeping their hand over a set of items. Which items to delete? Only those in the direct path of the sweep? Or, should the system insist that the user point (or gaze) at each and every item they wish to delete? The premise of the "gestalt" approach is that the user tends perceptually to cluster certain items together and to perform gestures on the basis of such groupings. For the message to be effective, however, the machine must also "see" those very same items as similarly grouped.

Accordingly, we have programmed the computer to examine the display and attempt to discern groupings of items which seem to belong together on the basis of proximity, and

similar color, size, and shape. The program takes sets of items arranged at will by the experimenter, proceeds to "look" at them, and then display candidate groupings of items it decides "belong together" figurally. This exercise in endowing the machine with rudimentary principles of gestalt perception is an significant step in endowing the machine with sets of frameworks held in common with the human user, the better to aid reference.

Gesture interpretation

Inputs from speech and eye seem to have fundamental units: the *word* and the *fixation*, respectively. The elements of gestural input are less clear. We found that we needed a way to segment gestural input into basic units, or "proto-gestures," without prematurely imputing this or that "meaning" to any movement or position.

Our approach to gesture recognition is neither to pattern-match data from our gesture-sensing gloves with "templates" of canonical gestures, nor the development of "taxonomies." Rather, our software "agent" uses a combination top-down, bottom-up approach to interpret the incoming information from the user in speech, gesture, and gaze.

The top-down part of the approach consists of looking first at the incoming speech stream, checking it for syntax, and then checking whether it contains all of the information to form a coherent request or command. An example might be "What is the status of the blue unit?" where there is in fact only one blue unit-marker on display. There is thus no ambiguity in

the inquiry—as opposed to where there were several blue markers present on the display.† In contrast, if the inquiry were "What is the status of that unit?" and several unit-markers were shown on the display, then the computer would begin to "search the situation" in order to disambiguate the referent—the "situation" including where the user was looking or pointing when they uttered "...that unit...".

The bottom-up part of the approach is the ongoing aggregation, for both hands, of the raw data from the gesture-sensing gloves into "raw features" and further into gestural segments dubbed "gestlets." Raw features includes such discriminations as "palm facing right," "thumb relaxed," "index finger straight," and so forth. Gestlets include such characterizations as "palm rotation clockwise," and dynamic components such as "attack," "sweep," and "lift." These raw features and gestlets thus are characterizations of the attitude and dynamics of the hands and its parts that are more condensed than the raw data, but stop short of interpretation of what may or not be "meant.

Thus, when the user says "Delete that!" the computer looks amidst the gestlet data occurring at that time to see whether there is evidence to support the hypothesis that the user might be pointing at or looking in some distinct spot at that time: Is the index finger extended, the other fingers flexed? Is the hand oriented such that the index finger aims

---

† This example inquiry is complete and well-formed at the level of speech, and no additional information from gaze or gesture is required. However, in the case of a command with serious or irrevocable consequences, the program might well require that the user also look, point, or both, at the item(s) referenced in speech

toward the screen. Failing to fine contextually meaningful data in the domain of gestural input, the agent program then considers the inputs for eye: specifically, does the eye data suggest that the user was "dwelling" on some particular item at the time of saying "...that..."?

* * * * * *

References/Publications/Theses

Bolt, Richard A. Eye movements in human/computer dialog. Advanced Human Interface Group Report 92-1, 1992.

Bolt, Richard A. and Edward Herranz. Two-handed gesture in multi-modal natural dialog. Proceedings of UIST '92, Fifth Annual Symposium on User Interface Software and Technology, Monterey, CA, November 15-18, 1992. In press.

Kristinn Thorisson, David B. Koons, and Richard A. Bolt. Multi-modal natural language. Video exhibited at the formal video program, CHI '92 ACM Annual Conference on Human Factors in Computing Systems, Monterey, CA, May 3-7, 1992.

Koons, David B. and Kristinn R. Thorisson. Unconstrained eye tracking in multi-modal natural dialogue. Advanced Human Interface Group Report 92-4, 1992.

Starker, I. and R. A. Bolt. A gaze-responsive self-disclosing display. In *Proceedings of CHI '90 Human Factors in Computing Systems* (Seattle, Washington, April 1-5, 1990), ACM press, New York, 1990, 3-9.

HOLOGRAPHIC TERRAIN MAPPING

Relevant Personnel:

Work in this area is under the direction of **Dr. Stephen A. Benton**, Professor of Media Technology, Head of the Media Arts and Sciences Section, and Director of the Lab's Spatial Imaging Group.

* * * * * * *

HOLOGRAPHIC TERRAIN MAPPING: soft copy (holographic video)

Interaction with holographic images generated in less than two seconds

A representative digital data object may consist of 10,000 polygonal facets, of which half are usually visible from the viewing zone. Each of these is populated with points at a density that produces a surface that just appears solid (but not a density so high as to cause "laser speckle" to appear). A typical density (depending on the resolution of the optical system) is 5 points/mm, which may result in an average of 10 points per elemental surface polygon. Thus about 5000 points may contribute to the holographic fringe pattern at any one time. Table look-up methods implemented on the Connection Machine (Model 2) have reduced the computation time to less than 100 microseconds per point (in June 1991; this number has steadily decreased over time as experience has been gained with the CM2), so that total recomputation times of about 0.5 seconds can be reached (another 0.5 second or

more can be added by system and network overhead, and I/O rate limitations). This brings the image response time well within what we consider to be "interactive times" for holographic video images. Only simple interaction interfaces have been explored so far, and much work remains to be done in this area.

Full-Color Moving Holographic Video Images

The holographic video display system is based on an acoustical-optical modulator and carefully matched horizontal scanning mirror. The modulator has three parallel channels that can independently and simultaneously be illuminated by three different wavelengths of light, such as red, green, and blue light. Each of these can be modulated by a separately computed holographic pattern, scaled to bring the images in each wavelength to the same location and size, so that they are registered in three-dimensional space. A specially-fabricated diffraction compensating plate is also needed to bring all three beams back into approximate register in order to match the optical axis of the overall system. This was first demonstrated for still images in October, 1991, and has since been extended to complex moving images.

HOLOGRAPHIC TERRAIN MAPPING - hard copy

Holograms with Wide Viewing Angle

Hard-copy holograms made by conventional optical techniques present only a modest range of side-to-side viewing angles, typically +/- 15 degrees (30 degrees total). Thus, only a limited amount of "look around" (an important source of distance information) can be offered by the image. Attempts to increase the viewing angle ordinarily require a scale-up of several large optical components, especially collimating lenses, greatly increasing the size and cost of a holographic hard-copy printer. Exposing a computer-generated hologram with small-size and closely-spaced optics introduces severe spatial distortions that are ordinarily unacceptable in a precision image. However, because those distortions arise from well-understood physical processes that may be modeled with considerable accuracy, the input images (the data fed to the holographic printer) can be pre-distorted in a mathematically complementary way so as to carefully "balance out" the optical distortions, providing an undistorted image with a wide angle of view. In this way, we have been able to increase the angle of view beyond +/- 50 degrees for holograms 300 mm square (frontal area), and only slightly less for holograms over a meter square.

Holograms of Paper-Map Size (Meter-Square)

The mass and complexity of a holographic laser printer usually grows as the third power of the linear size of the resulting hologram. Thus the scale-up from 300 mm square to meter-square images would be expected to produce a printer 30 times as massive and complex, likely to be prohibitively large for field use. Instead, using the pre-distortion concepts

mentioned above and a novel one-step optical method for hologram printing, we have demonstrated a remarkably compact printer for holographic film over a meter in width. These holograms can be made in almost any length required, but the samples to date have been about a meter square in size. These holograms are intended for illumination with a single-color point source (typically a laser), and subsequent work has been aimed at producing white-light viewable holograms of the same size.

Utilization of 3-D Data Spaces

Most of our holographic map images provide the optical equivalent of a plastic relief model, although of arbitrarily exaggerated depth scale. However, the variation of the views with horizontal location can also be used to introduce more profound variations of the data that is presented. For example, the map may change format from a topological display to a features display (roads, landmarks, etc.) that is mapped over the same three-dimensional relief surface, so that the spatial correspondence between several different sets of data can be made obvious to even a hasty observer. Another option for the use of the new holographic degree of freedom is the use of "pop in" display inserts that bring specified regions of the 3-D map surface forward to the viewer at much higher magnification, with a pointer dropping back toward the relevant location on the map. Although each "pop in" obscures the view of a portion of the map "ground" from any one viewing position, moving from side to side reveals features under the "pop in" so that the entire underlying map information may be appreciated.

Information and Bandwidth Reduction for Computed Holograms

Since the late `60's, researchers have speculated about the generation of hologram fringe patterns by computer, and their electronic transmission to remote displays. Unfortunately, a cursory analysis reveals that many gigapixels are needed to reconstruct a conventional hologram, and most research was abandoned in despair. Such holograms, however, reconstruct images with detail far beyond what the human eye can appreciate. At the Media Lab, we have concentrated on reducing the information content of holograms to meet only the minimum informational needs of the human observer. These reductions imply, in information theoretical terms, significant reductions in the computation and bandwidth needed to generate and transmit holograms. This part of our research program has been directed toward exploring these promised reductions, and finding practical implementations that take advantage of them.

One aspect of this work has been the development of the lensless-Fourier-transform optical geometry that produces a particularly straightforward interference fringe pattern for computation and transmission. The elemental hologram corresponding to each image point is, in this case, a cosinusoidal intensity pattern with a spatial frequency that varies between specific low and high frequencies that vary only in the rate of change (or "chirp") that reflects their distance from the hologram plane. That is to say that the holograms are "spatially invariant" with respect to the x- and y-coordinates, and have no terms higher than second order. These signals resemble those of chirped radar systems, and allow similar analytical tools to be brought to bear. In practical terms, they have allowed the minimization of the bandwidth of the transmitted signals by the elimination of low-frequency noise

terms, and the implementation of a simplified table look-up approach to hologram generation to permit image recalculation at high speed.

Increasing image quality (image realism); decreasing computation time; production of surface-shaded images, with new approaches to solid-scene rendering

The "objects" that are simulated in simple computed holograms consist of collections of thousands of individual point sources, each "radiating" independently of the others to produce a total wavefield for interference with a coherent reference wave. The simplest images are wire frames, with distant image features being visible through nearer ones. Increasingly laborious computations can provide more realistic image features, and more effective aids to the comprehension of the spatial complexity of the data. Our hologram computation software has introduced, in succession, occlusion or opacity clues to depth, surface shading by filling facets with continuous-looking collections of points, physically-realistic light and shadowing of surfaces, full color images, and glossy (specular) highlights. Each elaboration increases the complexity of the calculations, and the corresponding computation time, so that a tradeoff of interactivity and realism must be accepted at every step.

A key issue is, then, the speeding-up of the underlying calculations of the hologram fringes corresponding to a single object point. A propitious choice of optical geometry produces a particular type of fringe pattern (described above) that is amenable to a table look-up approach. This requires the pre-computation of all possible fringe patterns, and their loading into memory, which can take several minutes. However, subsequent image calculations take less than a tenth as much time as calculating the patterns "from scratch."

Holographic Hard-Copy Printer Study Model

In order to build an automatic holographic hard-copy printer, a "holographic laser printer" that responds simply to the push of a button, a one-optical-step holographic printing technique has to be invented. The high-visual-quality holographic images seen today are the result of two-step processes that first generate a "master," and then a white-light viewable "copy" hologram, requiring human operator intervention. Our research has been aimed at producing, in a single-stage automatable step-and-repeat optical printing process, a white-light viewable reflection hologram. In addition, we require this technology to be scalable from 300-mm/side to 1-meter/side, in order to produce holograms comparable to paper map in size, with impressive depth included. Our laser printer prototypes have combined high-resolution liquid-crystal spatial light modulators for real-time input, heat-processed high-resolution photopolymer recording materials for real-time output, and a novel optical system in between to produce near-image-plane images with a wide angle of view.

The key to the novel optical system is the fabrication of a holographic optical element that scatters light in a single direction, such as up-to-down, without scattering from side-to-side, and without the creation of low-resolution intensity variations in the vertical direction. That is to say that the laser "speckle" patterns that conventional diffusers produce must have their low-spatial-frequency components removed for this diffuser to work properly. The design and fabrication of this diffuser element have been the central research topics of this part of the research program. So far, visually acceptable results have been produced at the 300-mm/side scale, as demonstrated in monochrome images in the Autumn of 1991. However, the scale-up of this particular diffuser design has also scaled up the

corresponding speckle patterns, producing objectionable mottling in the large-size holograms. Thus research on the statistical optical design of th se key components has continued, delaying the building of a second full-size holographic laser printer prototype.

* * * * * * *

References/Publications/Theses

Stephen A. Benton. Experiments in holographic video imaging. To appear in: T. H. Jeong (ed.), *Proceedings of the Fourth International Symposium of Display*, 1991, (SPIE, Bellingham, WA, 1992.

Stephen A. Benton. Experiments in holographic video imaging. To appear in: P. Greguss (ed.), *Proceedings of the SPIE Institute on Holography* (IS#08, SPIE, Bellingham, WA, 1991), paper #5.

Michael W. Halle. *The generalized holographic stereogram*. Unpublished Masters Thesis, Media Arts and Sciences Section, MIT, February 1991.

Pierre St. Hillaire. *Real Time Holographic Display: improvements using higher bandwidth electronics and a novel optical configuration*. Unpublished Masters Thesis, Media Arts and Sciences Section, MIT, May 1990.

Pierre St. Hillaire, Stephen A. Benton, Paul Hubel, Mark Lucente. "Color images with the MIT holographic video display." *SPIE, Practical Holography VI*, Vol. 1667, No. 33, 1992.

Pierre St. Hillaire, Stephen A. Benton, Mark Lucente, John Underkoffler, Hiroshi Yoshikawa. "Real Time Holographic Display: improvements using a multichannel acousto-optic modulators and holographic optical elements." In *SPIE, Practical Holography V*, Vol. 1461, No. 37 February 1991.

Pierre St. Hillaire, Stephen A. Benton, Mary Jepson, Joel Kollin, Mark Lucente, John Underkoffler, and Hiroshi Yoshikawa. "Electronic display system for computational holography." In *SPIE, Practical Holography IV*, Vol. 1212, No. 20. January 1990.

LOOKING AT THE USER

Relevant Personnel:

Work under this topic was conducted by **Dr. Alex Pentland**, Associate Professor of Computers, Communication, and Design Technology, and co-Director (with **Dr. Edward H. Adelson**) of the Lab's Vision Science Group.

Looking at the User: Face Recognition

Summary

We have developed a near-real-time computer system which can locate and track a subject's head, and then recognize the person by comparing characteristics of the face to those of known individuals. The research paper describing this system won a Best Paper Prize from the IEEE for the year 1991 (Turk, M., and Pentland, A., (1991) "Face Recognition Using Eigenfaces," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586-591, Maui, HI., June 1991).

The computational approach taken in this system is motivated by both physiology and information theory, as well as by the practical requirements of near-real-time performance and accuracy. Our approach treats the face recognition problem as an intrinsically two-dimensional recognition problem rather than requiring recovery of three-dimensional geometry, taking advantage of the fact that faces are normally upright and thus may be described by a small set of 2-D characteristic views.

The system functions by projecting face images onto a feature space that spans the significant variations among known face images. The significant features are known as "eigenfaces," because they are the eigenvectors (principal components) of the set of faces; they do not necessarily correspond to features such as eyes, ears, and noses. The projection operation characterizes an individual face by a weighted sum of the eigenface features, and so to recognize a particular face it is only necessary to compare these weights to those of known individuals.

Some particular advantages of our approach are that it provides for the ability to learn and later recognize new faces in an unsupervised manner, and that it is easy to implement using a neural network architecture.

Project Description

Computational models of face recognition are interesting because they can contribute not only to theoretical insights but also to practical applications. Computers which recognize faces could be applied to a wide variety of problems, including criminal identification, security systems, image and film processing, and human-computer interaction. For example, the ability to model a particular face and distinguish it from a large number of stored face models would make it possible to vastly improve criminal identification.

Unfortunately, developing a computational model of face recognition is quite difficult, because faces are complex, multidimensional, and meaningful visual stimuli. They are a natural class of objects, and stand in stark contrast to sine wave gratings, the "blocks world," and other artificial stimuli used in human and computer vision research. Thus,

unlike most early visual functions, for which we may construct detailed models of retinal or striate activity, face recognition is a very high level task for which computational approaches can currently only suggest broad constraints on the corresponding neural activity.

We therefore focused our research towards developing a sort of early, preattentive pattern recognition capability that does not depend upon having three-dimensional information or detailed geometry. Our goal, which we believe we have reached, was to develop a computational model of face recognition which is fast, reasonably simple, and accurate in constrained environments such as an office or a household. In addition the approach is biologically implementable and is in concert with preliminary findings in the physiology and psychology of face recognition.

The scheme is based on an information theory approach that decomposes face images into a small set of characteristic feature images, called "eigenfaces," which may be thought of as the principal components of the initial training set of face images. Recognition is performed by projecting a new image into the subspace spanned by the eigenfaces ("face space") and then classifying the face by comparing its position in face space with the positions of known individuals.

Automatically learning and later recognizing new faces is practical within this framework. Recognition under widely varying conditions is achieved by training on a limited number of characteristic views (e.g., a "straight on" view, a 45 degree view, and a profile view). The approach has advantages over other face recognition schemes in its speed and simplicity, learning capacity, and insensitivity to small or gradual changes in the face image.

The early attempts making computers recognize faces were limited by the use of impoverished face models and feature descriptions (e.g. locating features from an edge image and matching simple distances and ratios), assuming that a face is no more than the sum of its parts, the individual features. Recent attempts using parameterized feature models and multiscale matching look more promising, but still confront severe problems before they are generally applicable. Current connectionist approaches tend to hide much of the pertinent information in the weights which makes it difficult to modify and evaluate parts of the approach.

The eigenface approach to face recognition was motivated by information theory, leading to the idea of basing face recognition on a small set of image features that best approximate the *set of known face images, without requiring* that they correspond to our intuitive notions of facial parts and features. Although it is not an elegant solution to the general recognition problem, the eigenface approach does provide a practical solution that is well fitted to the problem of face recognition. It is fast, relatively simple, and has been shown to work well in a constrained environment. It can also be implemented using modules of connectionist or neural networks.

It is important to note that many applications of face recognition do not require perfect identification, although most require a low false positive rate. In searching a large database of faces, for example, it may be preferable to find a small set of likely matches to present to the user. For applications such as security systems or human-computer interaction, the system will normally be able to "view" the subject for a few seconds or minutes, and thus will have a number of chances to recognize the person. Our experiments show that the eigenface technique can be made to perform at very high accuracy, although with a

substantial "unknown" rejection rate, and thus is potentially well suited to these applications.

We are currently investigating in more detail the issues of robustness to changes in lighting, head size, and head orientation, automatically learning new faces, incorporating a limited number of characteristic views for each individual, and the tradeoffs between the number of people the system needs to recognize and the number of eigenfaces necessary for unambiguous classification. In addition to recognizing faces, we are also beginning efforts to use eigenface analysis to determine the gender of the subject and to interpret facial expressions, two important face processing problems that complement the task of face recognition.

* * * * * * *

References/Publications/Theses

Pentland, A. "Photometric motion." In: Proceedings, *IEEE Third International Conference on Computer Vision*, December 4-7, 1990, Osaka, Japan.

Turk, M. *Interactive–time vision: face recognition as a visual behavior*. Unpublished doctoral dissertation, Media Arts and Sciences Section, MIT, 1991.

Turk, M., and Pentland, A., (1991) Eigenfaces for recognition, *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, pp. 71-86.

Turk, M., and Pentland, A., (1991) Face recognition using eigenfaces, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586-591, Maui, HI., June 1991.

INVENTIONS/PATENTS

The following patent applications arose out of this sponsored work:

A Method and Apparatus for Solving Finite Element Method Equations

- MIT Case No. 5349 - Prof. Alex P. Pentland

A new method for solving finite element method (FEM) and/or interpolation (lofting) problems with greatly reduced computational cost. This method uses wavelet functions to compute solutions to FEM, interpolation at a cost proportional to the number of nodes defining the problem, whereas previous solutions have a cost proportional to the square of the number of nodes. For large problems, therefore, this solution can be many orders of magnitude cheaper (or faster) than standard methods.

Method to Enhance Images

- MIT Case No. 5684 - Prof. Edward H. Adelson

(the work underlying this invention was also supported by Goldstar, Ltd.)

We describe a new class of techniques to remove noise from images and to enhance their visual appearance. The general approach we describe is to 1) convert an image into a set of coefficients in a multi-scale image representation; 2) modify each coefficient based on its amplitude, and the amplitude of coefficients of related orientation, position, or scale; 3) convert the modified coefficients back to a pixel representation to make the enhanced image.

CONCLUSIONS & RECOMMENDATIONS

Our 2K x 6K pixel display work demonstrated in dramatic fashion the potential of very high resolution computer-based displays. Text, graphics, and dynamic images can be shown with great flexibility and—at 100 pixels per inch—a visual quality that rivals print media. Further directions for research to support such displaying include experimentation in novel approaches to frame-buffer organization and the exploration of networking techniques, including fiber optic transmission, to support such display density.

Virtual environment systems for terrain visualization and mission planning appear to hold much promise. Our work in this area concentrated upon modeling technologies for terrain, vehicles, and aircraft, and developing approaches to dynamic simulation. We also explored whole-hand input as a means of communicating gesturally with the simulation system. Approaches to task-level interaction need refinement; needed as well are expanded means for user interaction with the virtual environment.

Displayed mapping information can involve many concurrent informational "layers": terrain features, structures, changes over time, weather, movement of equipment and personnel, and so forth. Our work in multiplane graphics demonstrates that through the application of special techniques in color, transparency, and focus (blur), it is possible to control in a very precise way what subsets of such layered information are brought to the user's attention. We have also demonstrated how fonts can move across different backgrounds yet retain high legibility. Also, we have shown how intelligent layout systems, can automate the "look" of graphical presentations independent of allotted display space, and yet maintain legibility and figural integrity.

Our work in image coding and compression techniques demonstrates the high value of such techniques as command and control systems develop ever larger volumes of data traffic. Approaches such as we have researched can also help to eliminate display bottlenecks.

Our work in computer face-recognition with its eigenvector ("eigenface") approach is necessarily preliminary. However, with development, it could have considerable utility in areas of law enforcement such as drug traffic interdiction where many faces, as in crowded airport terminals, might have to be screened.

In our multi-modal natural dialogue work the value of *connected* over *discrete* speech recognition was made clear: the need deliberately to pause between uttered words interferes strongly with the pace and rhythm of any concurrent gesture and gaze. Connected speech input is thus strongly to be preferred, even at the cost of a smaller active vocabulary. We also note the potential value of *directed audio output*, as when, toward the end of our project, we went from a 20-inch diagonal to a 6-foot diagonal display screen. Where the visual arena is greater, the value of apparent directional source of sound gains importance.

The invention and first demonstration of color video holography was a prime technical achievement under this contract. Follow–up in this area includes enlarging the size of the image, and increasing the quality of color effects achieved.

DISTRIBUTION LIST

addresses	number of copies
RL/C3AB ATTN: Richard T. Slavinski Griffiss AFB NY 13441-5700	25
MIT Media Lab ATTN: Dr. R. Bolt 20 Ames Street Cambridge MA 02159	5
RL/SUL Technical Library 26 Electronic Pky Griffiss AFB NY 13441-4514	1
Administrator Defense Technical Info Center DTIC-FDAC Cameron Station Building 5 Alexandria VA 22304-6145	2
Defense Advanced Research Projects Agency 1400 Wilson Blvd Arlington VA 22209	1
SAF/AQSC Pentagon Rm 4D 269 Wash DC 20330	1
Naval Warfare Assessment Center GIDEP Operations Center/Code QA-50 ATTN: E Richards Corona CA 91718-5000	1
HQ AFMC/XTH Andrews AFB MD 20334-5000	1

HQ ACC/SCPT
LANGLEY AFB VA 23665-5001 2

HQ ACC/DRIY
ATTN: Maj. Divine
Langley AFB VA 23665-5575 1

HQ ACC/XP-JSG
Langley AFB VA 23665-5000 1

Wright Laboratory/AAAI-4
Wright-Patterson AFB OH 45433-6543 1

Wright Laboratory/AAAI-2
ATTN: Mr Franklin Hutson
Wright-Patterson AFB OH 45433-6543 1

AFIT/LDEE
Building 642, Area B
Wright-Patterson AFB OH 45433-6583 1

Wright Laboratory/MTEL
Wright-Patterson AFB OH 45433 1

AAMRL/HE
Wright-Patterson AFB OH 45433-6573 1

AUL/LSE
Bldg 1405
Maxwell AFB AL 36112-5564 1

HQ ATC/TTOI
ATTN: Lt Col Killian
Randolph AFB TX 78150-5001 — 1

US Army Strategic Def
CSSD-IM-PA
PO Box 1500
Huntsville AL 35807-3801 — 1

Commanding Officer
Naval Avionics Center
Library D/765
Indianapolis IN 46219-2189 — 1

Commanding Officer
Naval Ocean Systems Center
Technical Library
Code 96428
San Diego CA 92152-5000 — 1

Cmdr
Naval Weapons Center
Technical Library/C3431
China Lake CA 93555-6001 — 1

Superintendent
Code 524
Naval Postgraduate School
Monterey CA 93943-5000 — 1

Space & Naval Warfare Systems Comm
Washington DC 20363-5100 — 1

CDR, U.S. Army Missile Command
Redstone Scientific Info Center
AMSMI-RD-CS-R/ILL Documents
Redstone Arsenal AL 35898-5241 — 2

Advisory Group on Electron Devices
Attn: Documents
2011 Crystal Drive, Suite 307
Arlington VA 22202 — 2

Los Alamos National Laboratory 1
Report Library
MS 5000
Los Alamos NM 87544

AEDC Library 1
Tech Files/MS-100
Arnold AFB TN 37389

Commander/USAISC 1
ATTN: ASOP-DO-TL
Bldg 61801
Ft Huachuca AZ 85613-5000

AFEWC/ESRL 1
San Antonio TX 78243-5000

Software Engineering Inst (SEI) 1
Technical Library
5000 Forbes Ave
Pittsburgh PA 15213

Director NSA/CSS 1
W157
9800 Savage Road
Fort Meade MD 21055-6000

NSA 1
T5122/TDL
Fort Meade MD 20755-6000

NSA 1
ATTN: D. Alley
Div X911
9800 Savage Road
Ft Meade MD 20755-6000

DoD 1
R31
9800 Savage Road
Ft. Meade MD 20755-6000

DIRNSA
R509
9800 Savage Road
Ft Meade MD 20775 1

Director
NSA/CSS
R08/R & E BLDG
Fort George G. Meade MD 20755-6000 1

ESC/IC
50 GRIFFISS ST
HANSCOM AFB MA 01731-1619 1

ESC/AV
20 SCHILLING CIRCLE
HANSCOM AFB MA 01731-2816 1

FL 2807/RESEARCH LIBRARY
OL AA/SULL
HANSCOM AFB MA 01731-5000 1

Technical Reports Center
Mail Drop D130
Burlington Road
Bedford MA 01731 1

Defense Technology Sec Admin (DTSA)
ATTN: STTD/Patrick Sullivan
400 Army Navy Drive
Suite 300
Arlington VA 22202 1

AL/LRG
ATTN: Mr. M. Young
Wright-Patterson AFB OH 45433-6503 1

AL/CFHV
ATTN: Dr. B. Tsou
Wright-Patterson AFB OH 45433-6503 1

AL/HED
ATTN: Maj M. R. McFarren
Wright-Patterson AFB OH 45433-6503 1

WL/XPK
ATTN: Dr. D. Hopper
Wright-Patterson AFB OH 45433-6503 1

AFIT/ENG
ATTN: Maj P. Amburn
Wright-Patterson AFB OH 45433-6503 1

USA ETL
ATTN: Mr. R. Joy
CEETL-CA-D
Ft Belvoir VA 22060 1

SPAWAR
ATTN: Mr. J. Pucci
Code 3241D
2511 Jefferson Davis Highway
Wash DC 20363-5100 1

ATZL-CDC-B
ATTN: Capt Charles Allen III
Ft Leavenworth KS 66027-5300 1

NUSC
ATTN: Mr. L. Cabral
Newport RI 02841-5047 1

NOSC (Code 414)
ATTN: Mr. P. Soltan
271 Catalina Blvd
San Diego CA 92151-5000 1

NTSC (Code 251)
ATTN: Mr. D. Breglia
12350 Research Parkway
Orlando FL 32826-3224 1

MIT Media Lab
ATTN: Muriel Cooper
Visible Language Workshop
20 Ames Street
Cambridge MA 02139 — 1

Dr. M. Nilan
4-214 Center for Science & Technology
Syracuse NY 13244-4100 — 1

The MITRE Corp
ATTN: Dr. H. Veron
MS ED73
Burlingotn Rd
Bedford MA 01750 — 1

NASA
ATTN: Dr. J. Robertson
Langley Research Center
Mail Code 152E
Hampton VA 23665-5225 — 1

Decision Sciences & Engineering Systems
ATTN: Dr. A. Wallace
5025 CII/RPI
Troy NY 12180-3590 — 1

Naval Ocean Systems Center
ATTN: Glen Osga, Phd
Code 444
San Diego CA 92152 — 1

ARI Fort Leavenworth
ATTN: Maj Rob Reyenga
P. O. Box 3407
Ft Leavenworth KS 66027-0347 — 1

Army Research Institute
ATTN: Sharon Riedel, Phd
P. O. Box 3407
Ft Leavenworth KS 66027-0347 — 1

ARI Field Unit
ATTN: Jon J. Fallesen
P. O. Box 3407
Fort Leavenworth KS 66027-0347 — 1

RL/ESOP 1
ATTN: Mr. Joseph Horner
Hanscom AFB MA 01731

Joint Nat'l Intel Develop Staff 1
ATTN: Capt John Hilbing
4600 Silver Hill Road
Wash DC 20389

DREXEL College of Info Studies
ATTN: Dr. S. Andriole
32nd and Chestnut Street
Philadelphia PA 19104

George Mason University 1
ATTN: Ms Lee Ehrhart
Center of Excellence in C3I
School of Info Tech & Engrg
Fairfax VA 22030

*U.S. GOVERNMENT PRINTING OFFICE

MISSION

OF

ROME LABORATORY

Rome Laboratory plans and executes an interdisciplinary program in research, development, test, and technology transition in support of Air Force Command, Control, Communications and Intelligence (C^3I) activities for all Air Force platforms. It also executes selected acquisition programs in several areas of expertise. Technical and engineering support within areas of competence is provided to ESD Program Offices (POs) and other ESD elements to perform effective acquisition of C^3I systems. In addition, Rome Laboratory's technology supports other AFSC Product Divisions, the Air Force user community, and other DOD and non-DOD agencies. Rome Laboratory maintains technical competence and research programs in areas including, but not limited to, communications, command and control, battle management, intelligence information processing, computational sciences and software producibility, wide area surveillance/sensors, signal processing, solid state sciences, photonics, electromagnetic technology, superconductivity, and electronic reliability/maintainability and testability.