Natural Communication with Information Systems

IVAN MARSIC, MEMBER, IEEE, ATTILA MEDL, AND JAMES FLANAGAN, FELLOW, IEEE

Invited Paper

Pervasive networking and sophisticated computing open opportunities for collaborative information processing independent of time and space. In this instance the information system becomes an enhancer of human intellect, as well as a mediator for communication among participants. The human user favors the sensory dimensions of sight, sound, and touch as primary channels of communication. Machines that can accommodate these modes promise flexibilities and functionalities that transcend the traditional mouse and keyboard.

This paper describes research to establish human–computer interfaces that capture attributes of natural face-to-face communication. An experimental multimodal system is developed to study several aspects of natural style human–computer communication. While as yet primitive, the technologies of image and gaze processing, hands-free conversation, and force feedback tactile transduction are combined and used simultaneously for manipulating objects in a shared workspace. Software agents fuse the sensory signals to estimate and interpret user intent. Current areas of experimental application include disaster relief/crisis management, telemedicine/rehabilitation, and mobile office/wearable computers.

Keywords*—Collaborative work, multimedia communication, multimodal interfaces, natural language interfaces, sensory fusion, user interface human factors, user interfaces.*

I. INTRODUCTION

The "face" seen by the user of an information system is the computer port to the system. The underlying complexities may be great, and the desire is to unburden the human to the extent possible. This means making human/computer interaction approach the naturalness of human/human communication. Sensory modalities that are allocated the greatest information load are (typically) sight, sound, and touch. Conversational interaction in particular plays a central role, and current research focuses on emulating these sensory channels. The modalities are normally employed simultaneously and in combination. Humans perform a fusion of these information channels, and machines must have the intelligence to perform similarly. In consequence, the research challenge includes not only transducers that adequately serve sight, sound, and touch signals but also software agents that fuse and interpret these simultaneous data.

Integration of multiple modalities in human/computer interfaces has long been viewed as a means for increasing ease of use. An early multimodal system was Bolt's *Put-That-There* [1]. The system fused speech and hand gesture and was applied to simple management of a small set of virtual objects. Naturalness of interaction was hampered by limitations of the interface technologies at that time. Individual technologies for more natural human/computer communication have only recently matured enough to be employed more effectively in freeing computer users from the constraints of the keyboard and the mouse [2]–[4].

Recent examples of multimodal human/computer interfaces include integration of modalities such as speech and gesture [5]–[8], or language understanding, gaze tracking, lip reading, and gesture recognition [9]–[11]. Carnegie-Mellon University does research on gesture and speech integration, face- and eye-tracking, and lip reading [12], [13]. A multimodal interface was developed for an appointment scheduling task on a computerized calendar. The user can use a combination of spoken input, gesturing with a pen on a touch-sensitive screen, or handwritten words to interact with the system. The role of natural language is currently investigated by researchers at the Oregon Graduate Institute of Science and Technology [5], [6], where speech and hand-gesture integration is employed in a military planning and simulation scenario. The SRI Multimedia Interfaces Group is also building a series of prototype map-based applications that accept handwritten, verbal, and gestural requests [14]. These applications are distinguished by comparatively rich natural language capabilities, access to existing data sources

Manuscript received November 11, 1999; revised April 18, 2000. This work was supported by NSF Stimulate under Contract IRI-96-18854, by NSF KDI under Contract IIS-98-72995, by DARPA under Contract N66001-96-C-8510, and by the Rutgers Center for Advanced Information Processing and its corporate affiliates.

The authors are with Rutgers University, Piscataway, NJ 08854-8088 USA.

Publisher Item Identifier S 0018-9219(00)08103-2.

including the World Wide Web, and a mobile handheld interface. Leading research efforts on human/computer communication exist in Europe, especially at LIMSI in Paris [15], as well as in Japan, especially at ATR in Kyoto.[1] Other efforts address conversational interaction [16].

The research reported here establishes technology for synergistic fusion of multimodal data from gaze, speech, and tactile interaction. We believe that this evolving direction heralds natural-like communication between the human user and the complex information system.

II. Architecture of the Multimodal Interface

The system architecture reported here is determined by practical considerations of task-oriented interaction. We assume that the purpose of dialog is to accomplish a serious task, such as application control or cooperative problem solving where the system helps the user plan activity. Our goal is to interface software applications, rather than building an interface to interact with a computer "brain." Contrary to systems like Eliza [17] and Julia [18], the application plays a central role in the design. The application defines a set of application programming interfaces (APIs) that can be invoked to cause different actions. The APIs determine the user's command vocabulary and grammar for both speech and gesture.

We focus on one particular type of human/computer interaction: dialog-based applications that include a significant multimodal component. This implies that the interaction can benefit from forms of communication other than spoken language, such as gesture and the manipulation and transformation of geometric objects. In this context, multiple sensory modalities can be used to simplify interaction and to disambiguate user-generated information. For instance, pointing can obviate a verbal description of the target. On the other hand, applications that are text-based, such as dictation, access and search of databases, or question-answering protocols may not significantly benefit from multimodality. We deal with the following modalities and related technologies.

1) *Sight*: face finding; eye tracking; visual gesture; image segmentation and recognition.
2) *Sound*: hands-free sound capture; automatic speech recognition; text-to-speech synthesis.
3) *Touch*: force feedback glove; virtual grasp; manual gesture.

These capabilities used in combination permit flexibility and functionality that transcend traditional mouse and keyboard. For example, the act of rotating a graphical object 36.5° clockwise is clumsily done with mouse and keyboard. A brief spoken command easily accomplishes the task.

The multimodal techniques, used in concert, support more natural communication between machine and human user. But simultaneous input means that individual modalities may sometimes reflect information that is redundant, ambiguous, or even contradictory. A software agent that can accept the inputs, fuse the data, and, at least in a primitive way, estimate user intent and meaning (within the existing context) and initiate appropriate action is a key component, as shown in Fig. 1. Even though the input may be nondeterministic, the agent's output to the application is deterministic: it is a call to one or several of the application APIs. The fusion agent invokes the operations on the application only upon reaching accurate interpretations. If the interpretation cannot be achieved or the command does not apply in the current context, the user is notified about it. These and some other types of interactions, e.g., system help, are not directed to the application. Desired functionality of the agent includes fusion of information from multiple sensors; feedback generation; maintaining discourse knowledge; switching between concurrent applications; opportune employment of modalities; and various operating-system-level controls for generation of conversational behaviors. The following sections describe how our design approaches these requirements.

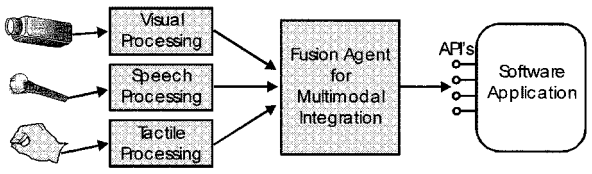


Fig. 1. Architecture of the multimodal system. Most of the user's interactions are directed to the application while some are short-circuited within the agent when the command cannot be interpreted or does not apply in the current context. The agent also generates conversational feedback to the user.

Interaction between the intelligent interface agent and the application is structured so that the agent communicates the user's requests to the application by calling the corresponding APIs. The agent also acts as a listener monitoring the changes in the application's state. These software interactions are best implemented via the command and observer design patterns [19], respectively.

III. Technologies for Natural Communication

Technologies that emulate sensory communication cannot as yet aspire to human virtuosity. But, while primitive in implementation, some dimensions of natural communication can be usefully incorporated into client stations. We focus on the sensory dimensions of sight, sound, and touch.

A. Sight Modality

One step toward natural communication is to enable the machine to know at what point in its visual display the human user is looking at any time. A new face-tracking, eye-tracking technique provides this information without the encumbrance of body-worn equipment. The ingredients are a gimbal-mounted video camera, and a colocated infrared source and ultrasonic ranger, all positioned on the client's desk (Fig. 2). The eye tracker is a commercial instrument produced by ISCAN,[2] which we have modified and integrated into the workstation. A real-time face-tracking algorithm points and focuses the camera and infrared source on the user [20]. Image segmentation locates the eye and the tracker illuminates the eye and computes the angle

[1]See http://www.atr.co.jp/.

[2]ISCAN Inc., 1998. http://www.iscaninc.com.

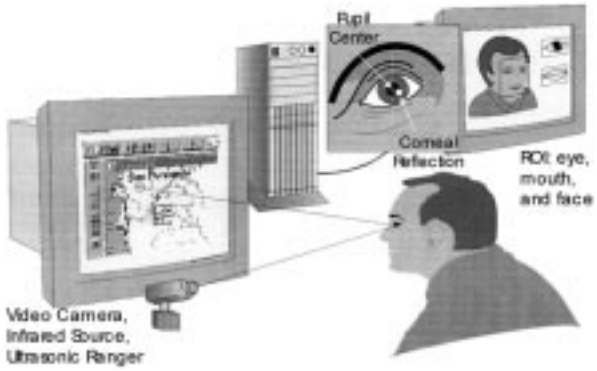


Fig. 2. Integrated gaze and face tracking system. A gimbal-mounted camera and infrared (IR) light source tracks gaze by computing the angle between the corneal reflection of the IR light and the centroid of the pupil. The ultrasonic ranger determines the user's eye distance from the screen [20].

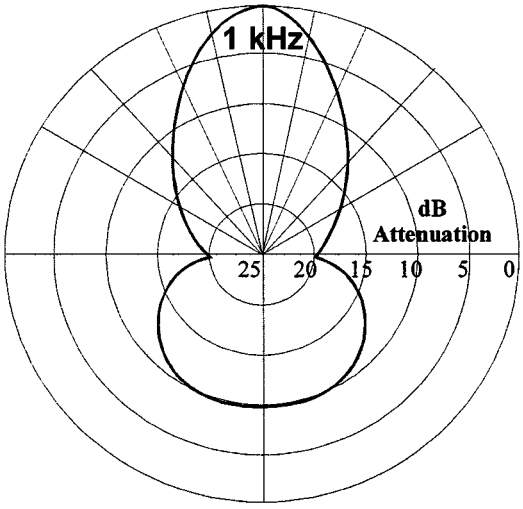
between the corneal reflection and the centroid of the pupil. This direction determines the point on the screen, which is being attended, and a cursor can be displayed at this point. The face-tracking algorithm utilizes skin tone and physiognomy, and eye segmentation is based upon form and luminance. Initial calibration requires about 10 s of observation. Ancillary signal-processing suites for displayed images can provide region-of-interest segmentation based upon color, form, luminance, and position. This capability is useful in identifying objects in complex scenes, such as satellite pictures, blood-cell microscopy, and MRI analysis. The sight modality may include other components, such as visual recognition of hand and body gesture; a review of this research can be found in [21]. Our experiments with capturing "eye movement gestures" are not yet fully refined, so gaze is currently used as a passively tracked modality. In the present implementation, the sight module generates a continuous stream of screen coordinates that signify the focus of the user's attention.

The speed and accuracy of the gaze tracker was evaluated in comparison with conventional mouse inputs. Accuracy of the presently used gaze tracker is 0.45° of arc, and visual feedback can make vernier adjustments of position. On a monitor screen of width 42 cm, for a travel distance of 10-cm gaze, movement is 1.7 times faster than a mouse, and for a travel distance of 20 cm it is 2.1 times faster than a mouse.

B. Sound Modality

Conversational interaction is a preferred means for human information exchange. This is typically accomplished in a "hands-free" manner, without body-worn or hand-held equipment. Microphone array technology for beam-steered high-quality capture of sound at a distance combines with automatic speech recognition and text-to-speech synthesis to support this interface [22], [23]. For a single user station, the microphone array is fix-focused on the stationary client position (Fig. 3). For group conferencing, the microphone array is implemented to track a talker moving about the meeting space and beam-steer the acoustic selectivity to the

(a)



(b)

Fig. 3. (a) Microphone line array used to capture speech from the user while mitigating interfering acoustic noise and reverberation. The workstation array is fixed focused to the user position. (b) Spatial directivity for the array at the frequency 1 KHz.

relevant position. The same steering command points and focuses the video camera.

A variety of speech-recognition systems are now commercially available. Most utilize acoustic features (cepstral coefficients), which are statistically classified by a hidden Markov model (HMM) algorithm. We have used Microsoft Whisper for both speech recognition and synthesis, and currently we are employing IBM ViaVoice. In our present system, the recognizer delivers the text of a recognized phrase along with time stamps for the utterance. The latter are important for temporal synchronization across modalities. The utterances are interpreted by a parser and forwarded to the multimodal integration module described in Section IV.

C. Touch Modality

The Rutgers force-feedback tactile glove is a portable haptic interface designed for interaction with virtual environments [24]. It is shown in Fig. 4. Previous uses of the tactile or gesture modality have usually not included a force-feedback capability. Nevertheless, this capability is essential to grasp, move, and place virtual objects [25]. The glove can read

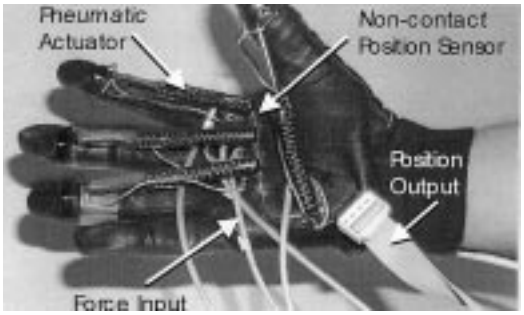


Fig. 4. The Rutgers force-feedback tactile glove [24].

hand gestures (fingertip positions relative to the palm) and apply forces to the fingertips corresponding to interaction. A six-dimensional Polhemus[3] tracker mounted on the back of the hand provides wrist position and orientation.

The hand gesture module is implemented to enable three-dimensional (3-D)-environment navigation (changing the viewpoint of the virtual environment) and virtual object manipulation. The gestures designed for object manipulation include: [*grasp*] used for grasping and moving the selected object; [*thumb up*] associated with resizing the selected object; [*open hand*] corresponding to the "unselect" or "drop object" command and also used as a reset position; [*curling the thumb*] corresponding to mouse clicks; and [*pointing*] for identification and object selection. For pointing in a two-dimensional (2-D) environment, coordinates are calculated for the point where a virtual ray along the user's index finger intersects the screen plane.

Force feedback from the tactile glove provides the user additional information about grasped objects. Some semantic information can be associated with compliance. For instance, a hard virtual helicopter might mean that it is fully loaded, as opposed to an empty softer one.

IV. Multimodal Workstation Design

A user station that incorporates these interface technologies is illustrated in Fig. 5. Utilizing multiple modalities requires software agents able to fuse the error-susceptible sensory information into reliable interpretations that are responsive to (and anticipatory of) human user intentions. To answer this need, we have developed a multimodal input manager (MIM). The manager interprets multimodal information and handles related problems, such as automatic context- and grammar-switching between applications, and hierarchical/selective use of available modalities. The structure of the multimodal input manager is shown in Fig. 6.

Modal interfaces preprocess information from the speech recognizer, the gaze tracker, and the tactile glove modalities. Each modal interface communicates with the modality handler using text. The gaze tracker sends (x, y) positions to the modality handler, while the speech recognizer sends the transcript of the utterance along with time stamps. The tactile glove interface transmits text representation of the actual gesture along with spatial information, such as *grasp* $[x, y, z]$.

[3]Polhemus Inc., http://www.polhemus.com/.

The modality handler detects currently available individual modalities and links them with the fusion agent. It is also able to turn the modalities on/off at the user's spoken command (e.g., "Computer, connect gaze tracker").

The customizer stores different grammars and gesture sets for different applications and loads them to the fusion agent depending on what application(s) the user is actually working on. The fusion agent interprets the multimodal input information and directs it to the applications. The agent's structure is described next.

A. Multimodal Fusion

Fusion of sensory information can be accomplished at three levels: data, features, or decisions (commands) [26]–[28]. Our fusion agent is implemented as a variation of the frame-based method, familiar in artificial intelligence practice [29], [30], and used in several other multimodal systems [10], [1], [8]. It fuses information at the level of decisions, based on the text outputs received from individual modalities. Assume that the user wants to create a helicopter icon at a specific location on a 2-D terrain map. The task can be done using speech and gaze, or by speech and manual gesture. The user can say "create helicopter here" and simultaneously look at a location, or simultaneously look and hand-point to a location. As default, gaze information is considered as the position information. However, if a manual pointing gesture is made, it overrides gaze information; similar to human–human communication. The frame-based fusion agent utilizes a slot-filling procedure illustrated in Fig. 7.

1) Slot Filling: An important component of the architecture is the *slot buffer*. It stores the incoming values for all possible facts defined by the command vocabulary. This is essential for anaphora resolution because present information is often used or referred to later. Consider this example: the user selects an object by saying, "Select the helicopter." The slot buffer maintains *discourse knowledge*, so, the user can say, "Move it." If the object from the previous utterance is stored in a slot buffer, then the "Move it" command has sufficient information about what to move. We use the most recent compatible reference. Other techniques are available and are reviewed in [31].

The agent fills the slots in the slot buffer. It reads the tactile glove and gaze positions if required by the command frame. For example, the utterance "Create a helicopter here," while only looking at a position on the map, causes the following slots to be filled in the slot-buffer: the position of the gaze cursor at the end of the utterance (x_1, y_1), the object's type (helicopter icon), and the operation or command type (create). On the other hand, if the manual gesture recognition module is providing information, i.e., the user manually points to a location while uttering the command, the agent overrides the gaze information and relies on the dominant hand gesture. Gesture information fills other slots too, not only the positional information. For example, the [*grasp*] gesture determines the command type, i.e., "grasp" or "move." Our experimental evaluation of the temporal relationship between gaze pointing and speech

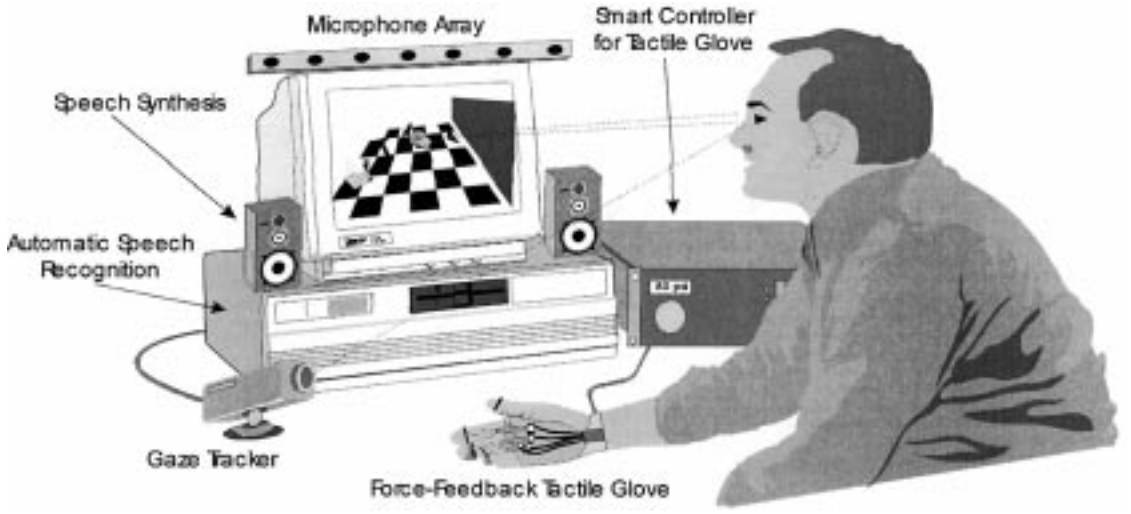


Fig. 5. Experimental client workstation incorporating sight, sound, and touch modalities for human/machine communication. The eye tracker provides a gaze-controlled cursor for indicating objects in the display. The tactile force-feedback glove allows displayed objects to be grasped, "felt," and moved. Hands-free speech recognition and synthesis provides natural conversational interaction.

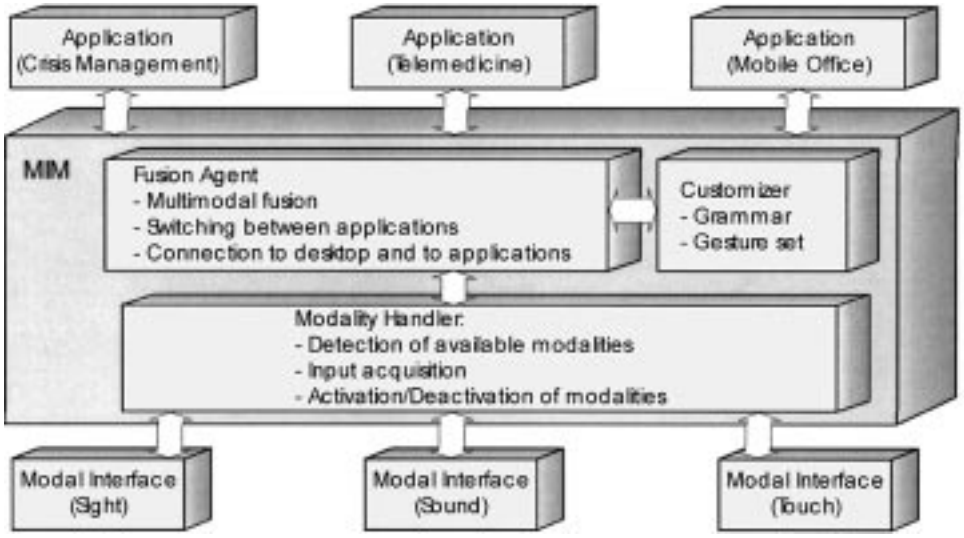


Fig. 6. Structure of the multimodal interface manager (MIM).

shows that gaze tends to precede spoken commands [32], and this is accounted for in the system design.

2) Command Frame Instantiation: Most of the user's utterances represent commands to the application (Fig. 1). Each command corresponds to an application's API, which has a specific signature. The signature may be overloaded or contain default values and may be object-specific. Each API has a corresponding *command frame* with slots corresponding to the function arguments. The frame slots may contain additional information [29], such as the argument type, constraints on values (e.g., range), and default values. The instantiation of a particular command frame is performed by analysis of information in the slot buffer. The agent monitors the slot buffer to determine whether the command slot is filled. If it is not filled, the system waits for more input information provided by any modality. If the command slot is filled, the fusion agent instantiates the corresponding command frame and examines whether there is sufficient information in the slot-buffer to fill each predefined slot of that particular frame. The agent must of course wait until all slots are filled. Then the command is executed through the API invocations.

B. Voice Feedback

For conversational interaction, the system must generate acknowledgment to verify that commands are understood and an intelligent response made to advise the user of actions needed or taken. Our system provides several types of feedback such as screen messages and tactile force, but we mainly focus on synthetic voice answerback. Text-to-speech synthesis is the appropriate technology to supply answers to queries related to the dynamic state of the workspace, requests for confirmation when necessary, general error messages and warnings about the commands that are not applicable in the current context, and notifications about semantic

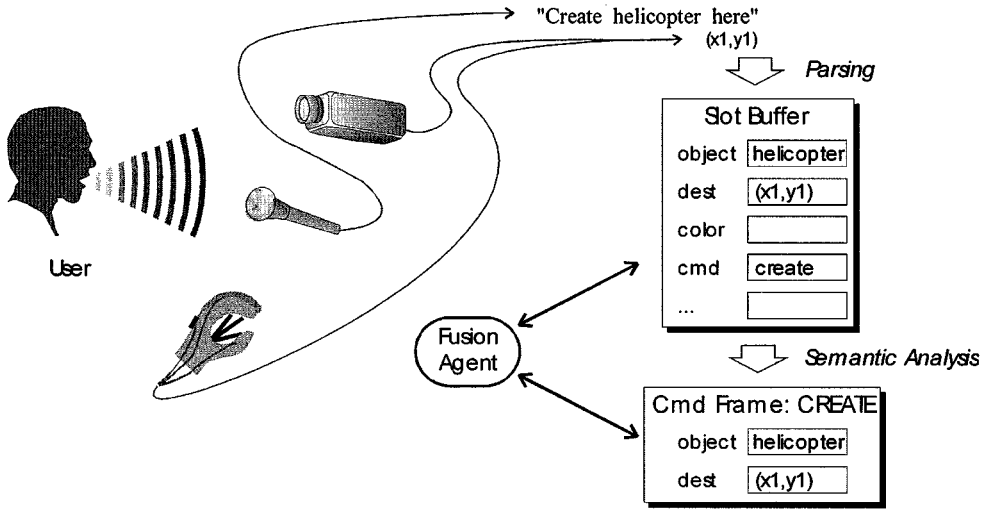


Fig. 7. "Slot-filling" technique used to coordinate, fuse, and interpret simultaneous speech, visual, and tactile inputs. The agent fills the generic Slot Buffer with the pieces of information coming from the modalities. This information is used to construct the specific command to the application. Only when the command frame slots are filled, the agent invokes the corresponding application's API.

or user-related errors. Richer feedback functions and behaviors (such as conversation initiation and giving and taking turns) are permissible [16].

Using vendor-supplied software for speech synthesis, we currently provide voice feedback to the user based on predefined rules. For example, identification of objects on the screen can be done by saying "identify this" while looking at the object, or pointing at it. If, however, neither the gaze nor the glove points at any object, the "Object ID" slot in the slot buffer and the "Identify" frame remain empty, and the command cannot be executed. This will result in the voice feedback, "What should I identify?"

C. Operating System Control

Convenient control of system status and behavior is desirable. Background discussion or noise can cause the speech recognizer to accidentally recognize utterances that were not intended. Two methods alleviate this problem.

1) Use of spoken system commands, like "disable speech" and "computer listen." After the utterance "computer, disable speech" the manager does not execute any task, unless "computer, listen" is first heard.
2) Use of *gaze confirmation* as in human–human communication [33]. Gaze can act as reinforcement in identifying which application an utterance is addressed to. We assume that the user pays attention to an application as he talks. If the user talks but his gaze is not focused on the application for a certain period of time (usually 10 s), the application simply does not listen to the user's speech. This approach is intended to emulate the element of face-to-face communication. During group conversation, people tend to look at the person to whom they are talking (selection by gaze). If a person looks at one person but talks to another, confusion usually results.

Another capability of higher level control is switching between individual applications. Users can employ any combination of speech and gaze. This can be speech-only, as in "select whiteboard," or speech and gaze, saying "select this" and looking at the window, or just focusing on the window for a few seconds. The last option is essentially an extension of gaze confirmation i.e., if the user is staring at a certain application, the system assumes intended commands to follow.

D. Implemented System

Based upon the concepts of Fig. 5, the laboratory implementation of the multimodal client station is pictured in Fig. 8. The gimballed video camera, infrared illumination source, and ultrasonic ranger are colocated just below the monitor screen at the user position. The line microphone array is mounted above the monitor screen and is fix-focused at the user position. This audio pickup supplies the speech recognizer. Voice response from the text-to-speech synthesis system is provided over loudspeaker or earphones. The tactile force feedback glove provides data for manual gesture and grasp. Sensory inputs are fused and interpreted by the multimodal interface manager, as previously described.

V. Experimental Applications

A class of applications that exploits multimodal interaction is telecollaboration. Computer-supported telecollaboration, unlike telephone conversation and e-mail, typically includes significant spatial and graphic content. A multimodal interface that includes gesture, voice, and manipulation is naturally suited for this type of application. It enables natural content manipulation for shared applications.

Previous studies evaluating the effects of different media on communication indicate that collaboration over a shared electronic workspace is effective for certain tasks and has little impact in others [33]–[38]. None of the studies have

Fig. 8. Implemented user station incorporating speech recognition and synthesis for conversational interaction. Speech is combined and synchronized with manual gesture sensing from a force-feedback glove ① and visual gesture sensing from a desk-mounted gaze tracker ②. The fixed-focus microphone array atop the workstation ③ captures speech from the user location while mitigating interfering acoustic noise and reverberation.

Fig. 9. An officer of the U.S. Army National Guard collaborates with the remote mobile task-force commander to position assets using the multimodal interface.

yet investigated the effect of different collaborating environments or input modalities on the collaboration task.

The multimodal interface is viewed as a means for more natural communication between human user and the system mediating collaboration. Toward exploring utility in this venue, we interfaced the multimodal system to an experimental network termed DISCIPLE (for Distributed System for Collaborative Information Processing and Learning) [39]. Current research is exploring application of the networked multimodal collaborative information system to two specific areas:

1) disaster relief/crisis management;
2) telemedicine/telerehabilitation.

A. Disaster Relief/Crisis Management

The mission of the U.S. National Guard is to provide civil security, societal stability, and succor in coping with catastrophic events—such as storms, floods, and national disasters, as well as threats that transcend the capabilities of local protective forces. A central element in maintaining readiness and rapid response is the ability to adapt to a wide variety of demands and to implement resource deployment rapidly. This planning and execution typically falls under the title of "mission planning." In present practice, this planning is typically done by voice communication among staff concerned with operations, logistics, intelligence, and personnel, usually allocating resources and positioning them on a terrain map for the affected region. The layout is graphed by marking icons with a grease pencil on an acetate overlay of the map.

Multimodal interface capabilities suggest that this collaboration and problem solving might be accomplished faster and with higher quality solutions through the use of collaborative networking [40]. An experiment (illustrated in Fig. 9) with the cooperation of officers of the NJ Army National Guard at Fort Dix, NJ, lends support to this hypothesis.

The experimental scenario embraced a domestic crisis situation in which a given area is to be secured and assets deployed to render assistance. Army protocol prescribes the logistic, personnel, and equipment procedures and the means for scoring the solution to the exercise. Two user terminals were incorporated: 1) the task force commander in a mobile command vehicle with only a wireless laptop computer and radio and 2) the command center with the full multimodal interface and database access. The command center officer was given two hours of familiarization with the multimodal interface and networked system (that he had never seen before). Even with the multimodal system in its primitive stage of development, the experimental deployment was accomplished correctly and expeditiously according to army protocol. Participating officers commented that the greater functionality and versatility of the more natural communication was a notable advantage. A general view among the officers was that the system is convenient to use after the brief learning period. Selection of objects was also considerably faster using speech than by using keyboard and mouse, as was the case for gaze tracking as mentioned in Section III-A.

The crisis management system has subsequently been extended to display 2-D and 3-D representations [41], as shown in Fig. 10. The force-feedback data glove definitely improves 3-D navigation and manipulation.

B. Telemedicine/Telerehabilitation

Health-care professionals frequently need to collaborate in diagnosing medical conditions reflected in images and in clinical data. Pathology is especially relevant. Frequently this analysis is aided by computer enhancement of an image. Typical data include blood samples (which may be imaged by remote computer-controlled microscopes in rural areas), X-rays, and MRIs. The collaborative multimodal system described here gives evidence of a new capability in diagnosis [42]. It provides a natural platform for employing newly emerging suites of signal processing. An illustration is Fig. 11, which shows results obtained in collaboration with the University of Medicine and Dentistry of New Jersey. The upper left panel shows blood cells, in this case afflicted by leukemia. Automatic techniques for image segmentation are called up by pen-based gesturing and by speaking

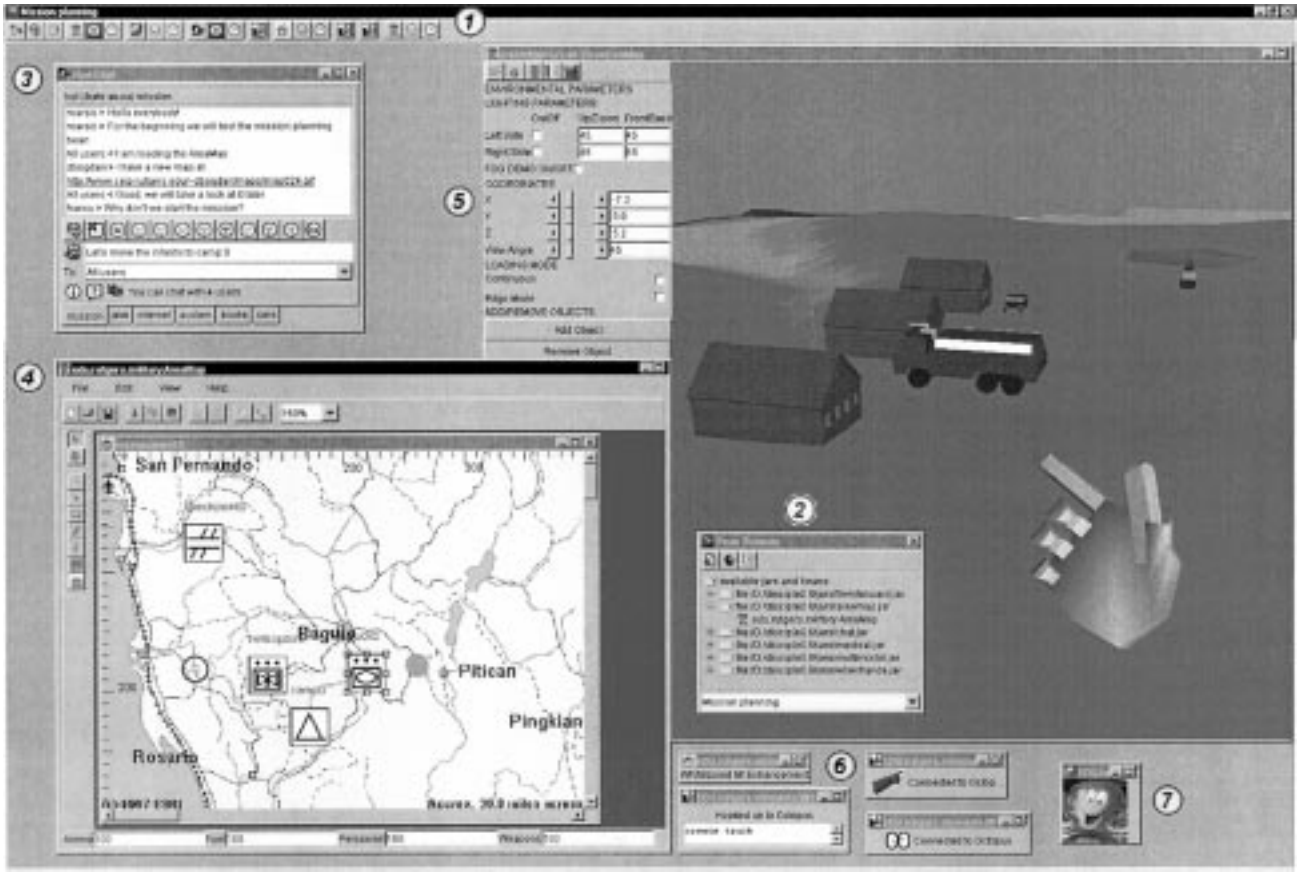


Fig. 10. Snapshot of a user view during collaboration in a meeting place ①. By using the BeanBrowser tool ②, the user can load beans from the Internet into the BeanBrowser, organize, and import them into the meeting place. The following Java beans are present: The Chat ③ can show text, pictures, and active URL links and filter messages by chat topic or user name. The AreaMap ④ and ThreeDeeMap ⑤ are interconnected and display the same geographical area during a military test application. The multimodal enhancement components ⑥ (multimodal connector, speech, tactile glove, and gaze tracker) and the multimodal manager ⑦ provide for interface customization.

voice commands. Image analysis methods—developed in the Robust Image Understanding Laboratory at the CAIP Center—can extract common components on the basis of color and texture (the top middle small panel) and by edge shape (the lower middle small panel) [42]. Additionally, the system may be commanded "Go to the central database and find other samples having similar properties." In this case, the system finds eight such in the database. They are displayed in the lower two rows. Because the database contains confirmed diagnoses, the system can provide tentative diagnoses, which medical specialists can assess. The system evaluation experiments indicate that the system performance is comparable to that of the human experts [43].

C. Evaluation of Modality Combinations

A central issue is the development of reproducible tests and quantitative metrics that reveal the synergies obtained from multimodal human/machine communication [44]. Initial performance experiments, conducted by the cognitive scientists on the research team, compared the combinations of two modalities at a time for the same task and captured the times it took for users to perform the tasks [45]. The command used in the evaluation was: "Create ⟨object⟩ ⟨at location⟩" on a situation map, such as in Fig. 10. The location reflects the variation in the combination of modalities and can be given by speech, e.g., "create camp at Baguio," or a location pointed to by the glove or gaze. For example, "speech and speech" means that both the object and the location are specified by speech. The following measures were obtained during the experiment.

1) Reaction Time (RT): The time it takes for a user to begin executing the task after they have heard the "start" command (spoken by the experimenter). At this point, the user has been given the task to carry out and knows what modalities to use. When subjects are told to begin the task, there is a delay that varies because of the second modality that is used. The reaction time is thus the time between the "start" command and the first speech utterance of the word "Create."
2) Command Specification Time (CST): The time it takes the user to specify the command. This time includes some system overhead. However, we feel that this is a natural part of the total time, as any multimodal interface design is faced with this overhead and we are not doing a psychophysical study measuring human

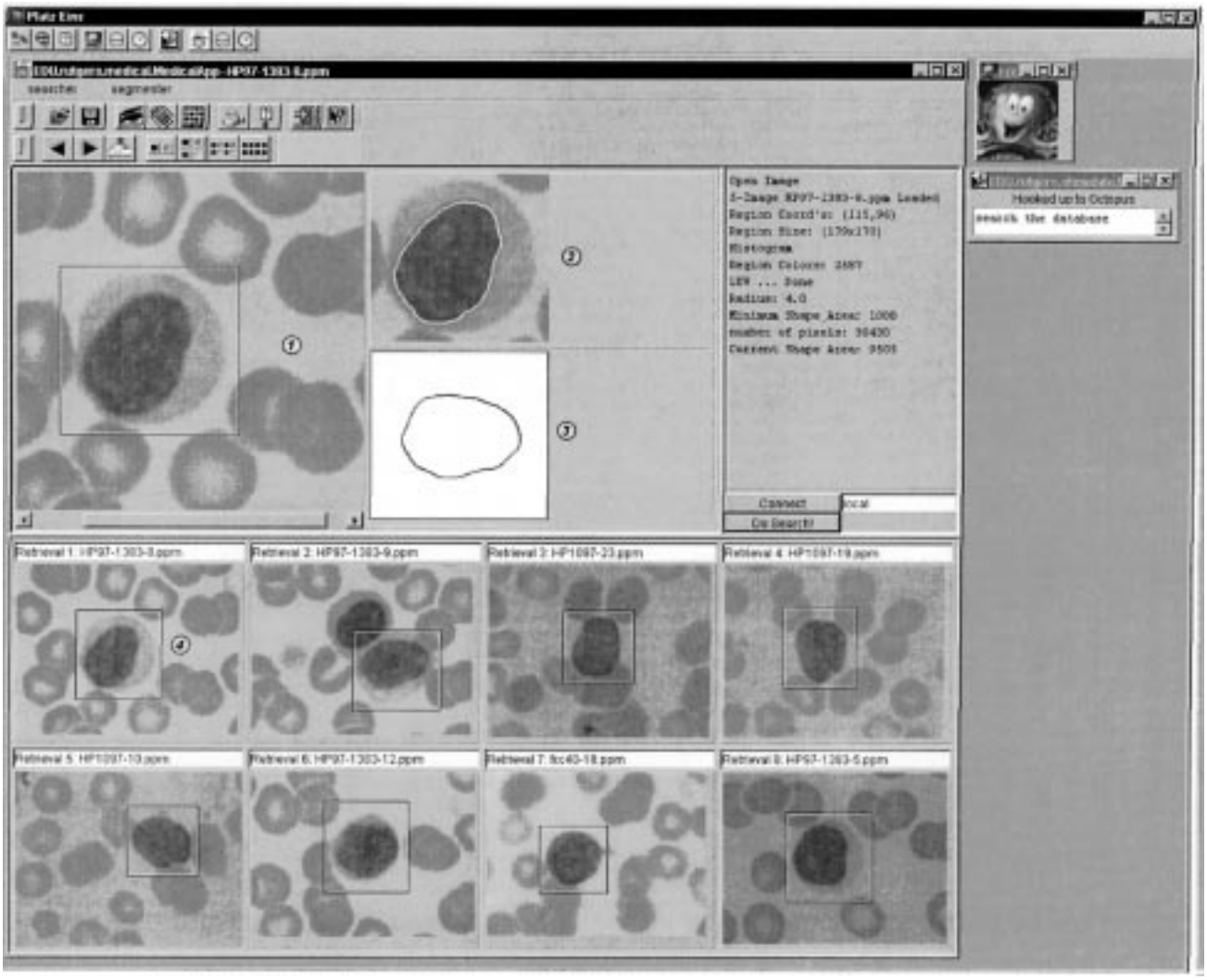


Fig. 11. The medical diagnosis support application as used in the DISCIPLE desktop. The specialist receives an image from a remote microscope and selects a region of interest by a pen-based gesture, ① segments the image, ② selects the cell kernel to extract, ③ searches the database for similar cells, ④ displays similar candidates from a combined database. In this interface speech recognition and text-to-speech synthesis are combined with pen-based gesturing [42].

performance time but rather a system evaluation measuring performance given the constraints of the experimental system.

3) Execution Time (ET): The total command execution time. ET is the time it takes the system to parse and execute the entire command. As revealed in the case of the mouse modality comparisons, there is considerable overhead introduced by the system software that was used in our multimodal system.
4) Precision: Whether an icon is created at the precise location where the user intended the object to be placed or some distance from the desired placement. This value was a subjective decision made by the user and was judged by a value of 1–5, with 5 being exactly on target.
5) Correctness: Whether the command was executed correctly or not, i.e., whether icon and position are both what the user desired. This value was calculated based on the number of errors the user made over the total number of settings the user attempted.
6) Ease of Use: The user's comfort level with the various modalities This value is also a subjective judgment of the user and ranged from 1 to 5, with 5 indicating that the user felt very comfortable using the system.

Preliminary performance data over ten trials for two subjects (one male, one female) are shown in Table 1. Time values are shown in seconds and are given to the closest millisecond. The combination of speech and gaze has a definite speed advantage over the other modality combinations, and this is likely to remain as we remove the various overheads introduced by our experimental system.

Reaction times are of interest because they reflect different behavior patterns of users based on the anticipated secondary modality usage. These data are plotted in Fig. 12(a). Speech and mouse appear to have the shortest reaction time, with the next shortest being speech and glove. In these cases, subjects begin their speech first and then start adjusting their motor activities for the hand–eye coordination task, resulting in greater CST and ET in both cases. In the case of speech

Table 1 Preliminary Performance Comparisons for Command Specification Time (CST) and for Total Execution Time of a Command (ET), see [45]. Command: "create ⟨object⟩ ⟨at location⟩." Speech and Speech Means that Both the Object and the Location are Specified by Speech. (*These Times Were not Recorded Given the Existing Software Configuration)

	Speech and Speech	Speech and Gaze	Speech and Glove	Speech and Mouse	Mouse Only
Average RT	3.51	3.766	3.034	2.037	*
Range CST (s)	0.078–0.157	0.062–0.079	1.844–4.219	*	*
Average CST (s)	0.144	0.071	2.976	16.088	*
Range ET (s)	3.188–4.578	2.391–3.812	4.984–8.062	*	*
Average ET (s)	3.867	2.934	6.245	18.749	11.249
Precision	5	3	4	4	4
Ease	4	4	3	3	3
Correctness	67%	87%	83%	82%	85%

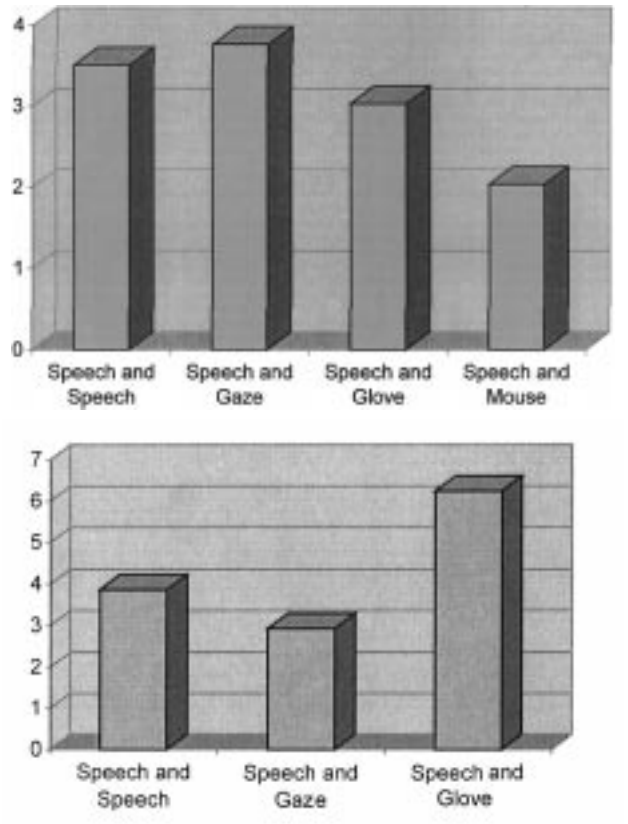


Fig. 12. A comparison of reaction times (RT) and total execution times (ET) for modality combinations [45].

and gaze, a larger RT is seen because the subjects did not initiate the spoken command until gaze position is at the desired location. But the corresponding CST for speech-and-gaze is very small since the command specification is completed very soon after command initiation.

The location precision for icon creation appears to be greatest for speech-and-speech as the (x, y) coordinates are exactly specified. Among pointer-based modalities, precision appears to be better for mouse than for gaze or glove. The speech-and-gaze combination is the most lacking in precision.

We also have the CST and ET data for the combination of speech and mouse modalities, as well as using mouse-only (Table 1). The mouse data in our application reflect not just the use of mouse as a modality but the multimodal integration method used. Also, the mouse-only case requires multiple menu selections and mouse clicks to complete the task; hence, the higher execution time.

Both users felt the mouse-only condition was more difficult to use due to the multiple menu selections and mouse clicks needed to execute the task. The eye tracker, once calibrated, appeared to be the easiest to use for position specification, though both users reported precise cursor control to be a problem. The glove was easy to use in terms of control but was reported to be tiring. The male subject rated the speech-and-speech case high for ease of use. The speech recognizer did not perform well for the female subject.

VI. Discussion

We have described an experimental system for multimodal human–computer interaction and collaboration. Our approach is to approximate—in a primitive way—the natural style of face-to-face communication. In human communication, the modalities of sight, sound, and touch are favored. Therefore, we built a system that tracks the user's gaze, understands voice commands, and is capable of tactile communication. The component technologies, though imperfect, are capable of freeing the user from the constraints of keyboard and mouse when used simultaneously and in combination. We designed and implemented a multimodal interface manager that fuses and interprets input information and provides feedback to the user. In order to demonstrate the functionality of these technologies, two applications have been implemented: 1) a system for mission planning and crisis management and 2) a system for collaborative medical decision-making. Although human factor tests and usability studies are not yet completed, preliminary data suggest that the system is competitive with, or outperforms, keyboard and mouse in graphics-centered tasks.

Multimodal technologies are expected to continue in development and become a key feature of human communication with information systems. Mass deployment of the benefits of computing may depend upon them. Disabled individuals will especially benefit from these technologies. And these methods will be directly applicable to distributed learning, collaborative problem solving, and mobile communication. As yet, these developments are rudimentary.

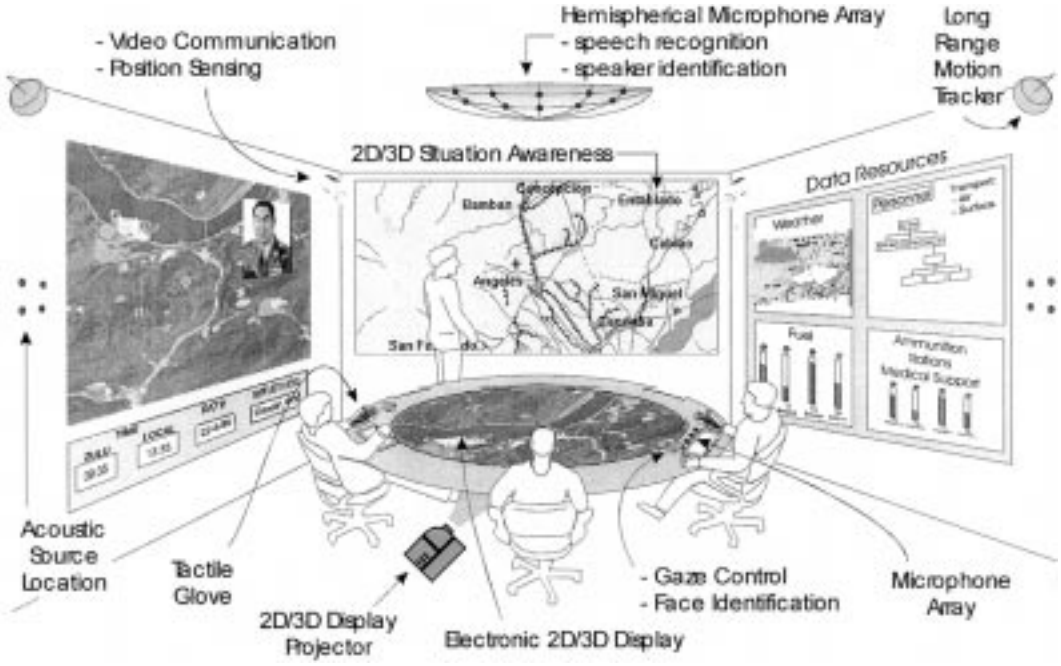


Fig. 13. Envisioned command post of the future. The command group is seated around an electronic table onto which 3-D images (such as terrain maps) and symbols can be rear-projected (from beneath). Each position at the table provides multimodal interaction with the collaborative workspace. An autodirective system for audio and video capture is included to automatically track talkers who are moving in the room [46].

But, used with prudent recognition of their limitations, they can be made to usefully serve the human user.

A. Future Applications

1) Semantics-Preserving Transformation of Information: The task of automatic information transformation across heterogeneous client stations, both within modality and cross-modality, is a long-range goal. Appliquéd onto this task is the desire for multilingual capability in speech and multicultural understanding in context framing. The ability for transformation requires a hierarchical representation of information. In the sight dimension, this means segmentation of image information according to semantically relevant features and coding in space and time that permits progressive transmission according to client resources. In the sound dimension, segmentation is also important. Signal representation preferably employs embedded coding techniques so as to retrieve a range of qualities ranging from high-fidelity PCM, through low-bit-rate LPC, to the text equivalent that can be synthesized by text-to-speech methods. The touch dimension is less bandwidth demanding, and the need for hierarchical representation of tactile information presently seems less critical. Rather, the synchronization of manual gesture, grasp, and signing is a primary issue. Nevertheless, we have resorted to text representation as the lowest common denominator in our experimental system.

Data fusion across multiple modes, as indicated in this paper, remains a major area of software research. Capturing the intelligence to interpret and even anticipate user intent, based upon multiple sensory inputs, context estimation, and semantic analysis, is an area ripe for software progress.

2) Smart Environments: Intelligent multimodal technologies will contribute to smart environments. Examples of future application include: a) mission control centers (as shown in Fig. 13) where controllers will be able to retrieve complex data directly without navigating complicated menu structures, b) airplanes and ships, where voice communication allows commanders to concentrate on tasks in a hands-busy/eyes-busy environment [46], and c) operating rooms where doctors manipulate diagnostic devices and retrieve and display patient data by voice and/or gaze.

3) Wearable Computers: As computing and communication technologies advance, mobile personal systems will become common. Wearable computers are in particular need of nonconventional human/computer interfaces. Some early work [47], [48] has investigated the dependence of the interface upon the user's task. Our focus is in wireless methods for networked collaboration, and this research is in a preliminary stage. We have, however, been using the Xybernaut wearable computer[4] and the multimodal devices described here, and demonstrated effective mobile collaboration over the DISCIPLE network. This research is continuing.

ACKNOWLEDGMENT

Research contributors to this project include Prof. P. Meer, Prof. G. Burdea, Prof. J. Wilder, Prof. M. Mantei Tremaine, and Prof. C. Kulikowski. Graduate researchers involved in the project, B. Dorohonceanu, M. Kaur, B. Sletterink, and H. Trefftz, were also critical to the evaluation of the concepts discussed here.

[4]Xybernaut Inc., http://www.xybernaut.com.

REFERENCES

[1] R. A. Bolt, "Put-that-there: Voice and gesture at the graphic interface," *Comput. Graphics*, vol. 14, no. 3, pp. 262–270, Aug. 1980.

[2] J. L. Flanagan, "Technologies for multimedia communications," *Proc. IEEE*, vol. 84, pp. 590–603, Apr. 1994.

[3] ——, "Research in speech communication," *Proc. Nat. Acad. Sci.*, vol. 92, pp. 9938–9945, Oct. 1995.

[4] ——, "Multimodality," in *Survey of the State of the Art of Human Language Technology*, R. Cole, *et al.*, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1997, pp. 287–298.

[5] P. R. Cohen, M. Johnston, D. McGee, S. Oviatt, and J. Pittman, "QuickSet: Multimodal interaction for simualtion set-up and control," presented at the *Proc. 5th Applied Natural Language Processing Mtg.*, Washington, DC, 1997.

[6] S. Oviatt, A. DeAngeli, and K. Kuhn, "Integration and synchronization of input modes during multimodal human–computer interaction," in *Proc. Conf. Human Factors in Computing Systems (CHI'97)*, Atlanta, GA, 1997, pp. 415–422.

[7] J. Coutaz, L. Nigay, and D. Salber, "The MSM framework: A design space for multi-sensory-motor systems," in *East–West Human Computer Interaction (EWCHI'93)*, L. Bass, J. Gornostaev, and C. Under, Eds. Berlin, Germany: Springer-Verlag, Aug. 1993, vol. 753, Lecture Notes in Computer Science, pp. 231–241.

[8] R. Sharma, T. S. Huang, and V. I. Pavlovic, "Multimodal framework for interacting with virtual environments," in *Human Interaction With Complex Systems: Conceptual Principles and Design Practice*, C. E. Ntuen and E. H. Park, Eds. Dordrecht, The Netherlands: Kluwer, 1996, pp. 53–71.

[9] M. T. Vo, R. Houghton, J. Yang, U. Bub, U. Meier, A. Waibel, and P. Duchnowski, "Multimodal learning interfaces," presented at the *Proc. ARPA Spoken Language Technology Workshop*, Austin, TX, Jan. 1995.

[10] M. T. Vo and C. Wood, "Building an application framework for speech and pen input integration in multimodal learning interfaces," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'96)*, Atlanta, GA, May 1996, pp. 3545–3548.

[11] M. J. Tomlinson, M. J. Russel, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'96)*, Atlanta, GA, May 1996, pp. 821–824.

[12] A. Waibel, "Multimodal human–computer interaction," in *Proc. 3rd Eur. Conf. Speech Communication and Technology (Eurospeech'93)*. Berlin, Germany: European Speech Communication Association, Sept. 1993, p. 39.

[13] A. Waibel, M. T. Vo, P. Duchnowski, and S. Manke, "Multimodal interfaces," *Artif. Intell. Rev.*, vol. 10, pp. 299–319, Aug. 1995.

[14] A. Cheyer and L. Julia, "Multimodal maps: An agent-based approach," in *Multimodal Human–Computer Communication*, Bunt, Beun, and Borghuis, Eds. Berlin, Germany: Springer-Verlag, 1998, Lecture Notes in Artificial Intelligence #1374, pp. 111–121.

[15] Y. Bellik, "Interfaces multimodales: Concepts, modèles et architectures," Thèse Doct. Sci., Univ. Paris-Sud, Orsay, May 30, 1995. See also http://www.limsi.fr/Recherche/IMM/PageIMM.html.

[16] J. Cassell, "Embodied conversational interface agents," *Commun. ACM*, vol. 43, pp. 70–78, Apr. 2000.

[17] J. Weizenbaum, "Eliza—A computer program for the study of natural language communication between man and machine," *Commun. ACM*, vol. 9, pp. 26–45, 1966.

[18] M. L. Mauldin. Chatterbots, tinymuds, and the turing test: Entering the Loebner prize competition. presented at Proc. AAAI-94. [Online] Available: http://www.fuzine.com/mlm/julia-home.html

[19] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*. Reading, MA: Addison-Wesley, 1995.

[20] Y. Liang and J. Wilder, "Real-time face tracking," in *Proc. SPIE Conf. Machine Vision Systems for Inspection and Metrology VII*, vol. 3521, Nov. 1998, pp. 149–156.

[21] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human–computer interaction: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 677–695, July 1997.

[22] Q. Lin, C.-W. Che, D.-S. Yuk, L. Jin, and J. L. Flanagan, "Robust distant talking speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'96)*, Atlanta, GA, May 1996, pp. 21–24.

[23] Q. Lin, E.-E. Jan, B. de Vries, C.-W. Che, and J. L. Flanagan, "System of microphone arrays and neural networks for robust speech recognition in multimedia environments," in *Proc. ICSLP 94*, Yokohama, Japan, Sept. 1994, pp. 1247–1250.

[24] G. Burdea, *Force and Touch Feedback for Virtual Reality*. New York: Wiley, 1996.

[25] F. L. Engel, P. Goossens, and R. Haakma, "Improved effieciency through I- and E-feedback: A trackball with contextual force feedback," *Int. J. Hum.–Comput. Stud.*, vol. 41, pp. 949–974, Dec. 1994.

[26] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proc. IEEE*, vol. 85, pp. 6–23, Jan. 1997.

[27] B. V. Dasarathy, "Sensor fusion potential exploitation—Innovative architectures and illustrative approaches," *Proc. IEEE*, vol. 85, pp. 24–38, Jan. 1997.

[28] R. Sharma, V. I. Pavlovic, and T. S. Huang, "Toward multimodal human–computer interface," *Proc. IEEE*, vol. 86, pp. 853–869, May 1998.

[29] P. H. Winston, *Artificial Intelligence*, 3rd ed. Reading, MA: Addison-Wesley, 1992.

[30] J. F. Allen, *Natural Language Understanding*. Redwood City, CA: Benjamin/Cummings, 1995.

[31] A. Kehler, J.-C. Martin, A. Cheyer, L. Julia, J. Hobbs, and J. Bear, "On representing salience and reference in multimodal human–computer interaction," in *Proc. AAAI'98 (Representations for Multi-Modal Human–Computer Interaction)*, Madison, WI, 1998, pp. 33–39.

[32] M. Kaur, M. Mantei Tremaine, and J. Wilder, "Interactions between speech and gaze in a multimodal system,", submitted for publication.

[33] C. Goodwin, *Conversational Organization: Interaction Between Hearers and Speakers*. New York: Academic, 1981.

[34] M. Stefik, G. Foster, D. Bobrow, K. Kahn, S. Lanning, and L. Suchman, "Beyond the chalkboard: Computer support for collaboration and problem solving in meetings," *Commun. ACM*, vol. 30, no. 1, pp. 32–47, Jan. 1987.

[35] M. Mantei, R. M. Baecker, A. J. Sellen, W. A. S. Buxton, and T. Milligan, "Experiences in the use of media space," in *Proc. Conf. Human Factors in Computing Systems (CHI'91)*, 1991, pp. 203–208.

[36] G. M. Olson, J. S. Olson, M. R. Carter, and M. Storrosten, "Small group design meetings: An analysis of collaboration," *Hum.–Comput. Interact.*, vol. 7, no. 4, pp. 347–374, 1992.

[37] J. S. Olson, G. M. Olson, M. Storrosten, and M. R. Carter, "Groupwork close up: A comparison of the group design process with or without a simple group editor," *ACM Trans. Inform. Syst.*, vol. 11, no. 4, pp. 321–348, Oct. 1993.

[38] S. Whittaker, E. Geelhoed, and E. Robinson, "Shared workspaces: How do they work and when are they useful?," *Int. J. Man–Mach. Stud.*, vol. 39, pp. 813–842, Nov. 1993.

[39] I. Marsic, "DISCIPLE: A framework for multimodal collaboration in heterogeneous environments," *ACM Comp. Surveys*, vol. 31, June 1999.

[40] A. Medl, I. Marsic, M. Andre, C. A. Kulikowski, and J. L. Flanagan, "Multimodal man–machine interface for mission planning," in *Proc. AAAI Spring Symp. Intelligent Environments*, Stanford, CA, Mar. 1998, pp. 41–47.

[41] B. Dorohonceanu, B. Sletterink, and I. Marsic, "A novel user interface for group collaboration," presented at the *Proc. 33rd Hawaiian Int. Conf. System Sciences (HICSS-33)*, Maui, HI, Jan. 2000.

[42] D. Comaniciu, P. Meer, D. Foran, and A. Medl, "Bimodal system for interactive indexing and retrieval of pathology images," in *Proc. 4th IEEE Workshop Applications of Computer Vision (WACV'98)*, Princeton, NJ, Oct. 1998, pp. 76–81.

[43] D. Comaniciu, P. Meer, and D. Foran, "Image guided decision support system for pathology," *Mach. Vis. Applicat.*, 1999.

[44] J. L. Flanagan and I. Marsic, "Issues in measuring the benefits of multimodal interfaces," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'97)*, Munich, Germany, Apr. 1997, pp. 163–166.

[45] M. Kaur, M. Mantei Tremaine, and J. Wilder, "Comparative evaluation of modality combinations in a multimodal system,", submitted for publication.

[46] D. V. Rabinkin, R. J. Renomeron, A. Dahl, J. C. French, J. L. Flanagan, and M. H. Bianchi, "A DSP implementation of source location using microphone arrays," *J. Acoust. Soc. Amer.*, pt. 2, vol. 99, p. 2503, Apr. 1996.

[47] A. Smailagic and D. Siewiorek, "Matching interface design with user tasks," *IEEE Pers. Commun.*, vol. 3, pp. 14–25, Feb. 1996.

[48] J. Yang, W. Yang, M. Denecke, and A. Waibel, "Smart sight: A tourist assistant system," in *Proc. 3rd Int. Symp. Wearabale Computers*, San Francisco, CA, Oct. 1999, pp. 73– 78.

Ivan Marsic (Member, IEEE) received the B.S. and M.S. degrees in computer engineering from University of Zagreb, Croatia, in 1982 and 1987, respectively, and the Ph.D. degree in biomedical engineering from Rutgers University, New Brunswick, NJ, in 1994.

He is an Assistant Professor of Electrical and Computer Engineering at Rutgers University, Piscataway, NJ. He is the Chief Architect for the DISCIPLE system, an advanced groupware system that enables teams, consisting of individuals with specific roles, to collaboratively access, manipulate, analyze, and evaluate multimedia data using geographically distributed networked information systems. He has authored more than 50 journal and conference papers, and three book chapters. His current research interests include groupware, mobile computing, computer networks, and human–computer interfaces.

Attila Medl received the M.S. and the Ph.D. degrees in computer and electrical engineering from the Technical University of Budapest, Hungary, in 1992 and 1996, respectively.

He was involved in the development of a Brain-Computer Interface at the Ludwig-Boltzmann-Institute for Medical Informatics in Graz, Austria. The system predicted hand-movements using EEG signals measured above the subject's motor cortex. As a Hanns-Seidel-Fellow at the Technical University of Munich, Munich, Germany, he developed an adaptive observer for the stable estimation of the anti-tumor immune response. He is currently an Assistant Research Professor at the CAIP Center at Rutgers University, Piscataway, NJ. His research area is multimodal human/machine communication and multimedia systems. He currently participates in two recently funded NSF projects investigating human–computer interaction and collaboration across wired and wireless networks.

James Flanagan (Fellow, IEEE) received the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge.

He is Vice President for Research at Rutgers University, Piscataway, NJ. He is also Board of Governors Professor in Electrical and Computer Engineering, and Director of the Rutgers Center for Advanced Information Processing (CAIP). Flanagan joined Rutgers after extended service in research and research management at Bell Laboratories. He was previously Director of Information Principles Research, with responsibilities in digital communications and information systems. He has specialized in voice communications, computer techniques and electroacoustic systems, and has authored approximately 200 papers, two books, and 50 patents in those fields.

Dr. Flanagan has received scientific awards, including the National Medal of Science, presented by the President of the United States, the L.M. Ericcson International Prize in Telecommunications, presented by the King of Sweden, the Edison Medal of the Institute of Electrical and Electronics Engineers, the Medal of the European Speech Communication Association, the Gold Medal of the Acoustical Society of America, and the Marconi International Fellowship, presented by the Crown Prince of Spain. He is a Fellow of the Acoustical Society of America and the American Academy of Arts and Sciences. He has been awarded Doctor Honoris Causa from the University of Paris-Sud, and from the Polytechnic University of Madrid. He is a member of the National Academy of Engineering, and of the National Academy of Sciences.