

Designing a Human-Centered, Multimodal GIS Interface to Support Emergency Management

Ingmar Rauschert, Pyush
Agrawal, Rajeev Sharma

Department of Computer Science
and Engineering, 220 Pond Lab,
The Pennsylvania State University,
University Park, PA 16802, USA

{rauscher, pagrawal,
rsharma}@cse.psu.edu

Sven Fuhrmann, Isaac
Brewer, Alan MacEachren

GeoVISTA Center, Department of
Geography, 302 Walker Building,
The Pennsylvania State University,
University Park, PA 16802, USA

{fuhrmann, isaacbrewer,
maceachren}@psu.edu

Hongmei Wang,
Guoray Cai

Information Sciences and Technol-
ogy, 002K Thomas Building, The
Pennsylvania State University, Uni-
versity Park, PA 16802, USA

{hzw102, gxc26}@psu.edu

ABSTRACT

Geospatial information is critical to effective, collaborative decision-making during emergency management situations; however conventional GIS are not suited for multi-user access and high-level abstract queries. Currently, decision makers do not always have the real time information they need; GIS analysts produce maps at the request of individual decision makers, often leading to overlapping requests with slow delivery times. In order to overcome these limitations, a paradigm shift in interface design for GIS is needed. The research reported upon here attempts to overcome analyst-driven, menu-controlled, keyboard and mouse operated GIS by designing a multimodal, multi-user GIS interface that puts geospatial data directly in the hands of decision makers. A large screen display is used for data visualization, and collaborative, multi-user interactions in emergency management are supported through voice and gesture recognition. Speech and gesture recognition is coupled with a knowledge-based dialogue management system for storing and retrieving geospatial data. This paper describes the first prototype and the insights gained for human-centered multimodal GIS interface design.

Categories and Subject Descriptors

H.5.2 [Information Interfaces and Presentation]: User Interfaces – *graphical user interfaces, input devices and strategies, interaction styles, natural language, user-centered design, voice I/O.*

General Terms

Management, performance, design, reliability, human factors

Keywords

Multimodal human-computer-interface, speech and gesture recognition, human-centered design, collaborative work, GIS, interactive maps.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GIS'02, November 8-9, 2002, McLean, Virginia, USA.

Copyright 2002 ACM 1-58113-591-2/02/0011...\$5.00.

1. INTRODUCTION

Rapid access to geospatial information is crucial to decision-making in emergency situations when decision makers need to work collaboratively, using GIS for hazard mapping and visualization as well as for improving situational awareness. Since current GIS provide mostly unimodal interaction tools and support only a single user, GIS functionalities are not accessible to all participants at the same time and do not allow users to express meaning, extract important information or develop planning scenarios effectively [2].

To support the ways in which humans work and interact in a collaborative emergency task situation, a new paradigm for computing is required that is multimodal, rather than unimodal, collaborative, rather than personal, and dialogue-enabled, rather than unidirectional. In order to meet these goals, we will develop principles for implementation and assessment of natural, multimodal, multi-user dialogue-enabled interfaces to GIS that make use of large screen displays and virtual environment technology. The project stage reported upon here is concerned specifically with the use of computer vision and speech processing as a means of interpreting and integrating information from two modalities, spoken words and free hand gestures.

A multimodal interface using speech and gesture as it is presented in this work has many advantages over systems only using one modality (e.g. speech) or using standard input devices (keyboard, mouse). Researchers [4, 12, 14, 19] suggest that multimodal interfaces are more efficient for interacting with geospatial information than is unimodal interaction. Where speech provides an effective and direct way of expressing actions, pronouns and abstract relations, it fails when spatial relations or locations have to be specified. Speech is not self-sufficient [13] and therefore gestures can provide an effective second modality that is more suitable for expressing spatial relations and is less error prone than if expressed in words alone [15].

Nevertheless, integration of multimodal interfaces in GIS and widespread collaborative use of GIS in emergency situations remains elusive. The goal outlined in this paper, is to develop a Dialogue-Assisted Visual Environment for Geoinformation (DAVE_G) that uses different interaction modalities, domain knowledge and task context for a dialog management that supports collaborative group work with GIS in emergency management situations.

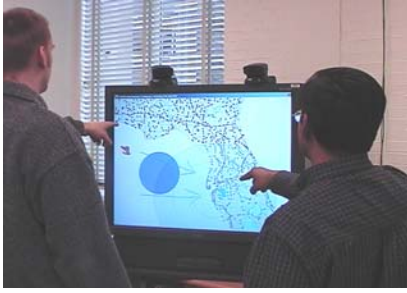


Figure 1. Collaborative interaction of two users with the HCI system for GIS

2. RELATED WORK

User studies and evaluations [8, 17] of recently developed multimodal Human-Computer-Interfaces (HCI) emphasize the use of more natural computer interfaces, such as speech and free-hand gestures. However, in spite of calls for research to make GIS easier to use [11], limited work has focused on multimodal GIS interaction [18].

Lokuge [10] developed a reactive interface display, which uses spoken queries, dialog management and domain knowledge to provide users with access to complex geo-spatial databases. However, the database retrieval system of Lokuge [10] relies only on speech as an interaction modality and thus remains unimodal. Research by Oviatt et al [13, 15] indicates that speech recognition as a single interface modality is error-prone and not efficient, especially if used in context with dynamic maps.

Cohen, et al [3], developed a pen and speech based interface to request dynamically generated maps. With their interface, necessary spatial attributes of queries are indicated directly by applying the electronic pen to the map display. Assessment of this pen-based interaction research showed that users could express spatial relations more efficiently. In combination with non-dominant speech input, users were less confused and made fewer errors.

Sharma et al [7] developed a general framework, for multimodal Human-Computer-Interfaces. This framework uses natural speech and free-hand gestures to enable user interactions with a system using a large screen display. Two implemented applications demonstrated the success of this framework. One prototype is an “intelligent” campus map of Penn State. New visitors to Penn State can post queries and ask for assistance on locations and directions around campus. The second application is a multi-modal crisis management simulation [6] that allows users to direct police cars, fire trucks and other emergency vehicles.

Most published multimodal systems, including the ones by Sharma and colleagues mentioned above, support only a single user at one time or do not use natural gestures for interaction. The research reported here is devoted to multimodal and multi-user interfaces for geo-spatial information systems applying the multimodal interface framework [7] by Sharma et al.

3. PROTOTYPE DESIGN AND SUPPORTED FUNCTIONALITIES

Multimodal user interfaces promise to be more natural and effective than traditional unimodal interfaces. As pointed out by Oviatt [13], speech is utilized much less if it is coupled with gestures to express spatial relations. Similarly, the utilization of a multimodal

Table 1. Supported User Commands

Command Type	Data Query	Viewing	Drawing
Functionality	show/hide layers, buffers	scroll left/right/up/down	circle
	spatial, w/ gestures	zoom in/out/full extend	line
	by attribute, w/o gestures	center at	free-hand
		zoom area	

dal interface is likely to differ from the standard mouse and keyboard interface we are used to. Therefore, special attention has to be applied to the design of such a new interface, to generate effective user interfaces for multi-user applications. Designing a multimodal collaborative GIS application is a “chicken and egg” problem, where neither such a system nor relevant and exhaustive user studies are available. In order to break through this cycle, a prototype multimodal user interface for GIS was created that serves as an initial system for the user studies. The prototype will then be redesigned and extended to provide efficient speech and gesture modalities for more natural interactions.

The initial DAVE_G prototype uses a large screen display, non-intrusive microphone domes (attached to the ceiling) and active cameras that allow users to move freely in front of the system. Supported modalities are speech and natural gestures, whereas spoken commands could be chosen freely within the definition of an annotated grammar. Figure 1 shows this prototype system when used during a collaborative work of two users.

In designing the prototype, we have adopted a cognitive systems engineering approach that involves incorporating domain experts into the earliest stages of system design. See [16, 21] for overviews of cognitive systems engineering and work domain analysis. This approach involves onsite visitations to emergency management operations centers, and a detailed account of the entire approach will be reported elsewhere. As a first step in this work domain analysis procedure, a questionnaire was administered. The questionnaire was sent out to 12 emergency managers in Florida, Washington D.C., and Pennsylvania of which four replied. The objective of the questionnaire was to identify GIS-based response activities and operations on disaster events.

The participants indicated that a GIS-based emergency response would need to support zoom, pan, buffer, display, and locational selection of geospatial data. The emergency tasks for which these operations were used include transportation support, search and rescue, environmental protection, and firefighting. This allowed us to compile the required GIS functionality into three categories: data query, viewing and drawing (see Table 1).

Viewing commands are frequently used and they are generally scenario independent. Thus, in the initial DAVE_G prototype, commands, such as “scroll”, “zoom” and “center” were implemented utilizing combined speech and gesture input. While “scroll” and “zoom in/out” commands are unimodal, “center at” and “zoom here” make use of both gesture and speech recognition.

The initial GIS specific commands can be divided into two sub commands: data query and selection. For data query “show”, “hide” and “locate” are supported. These commands need a pro-

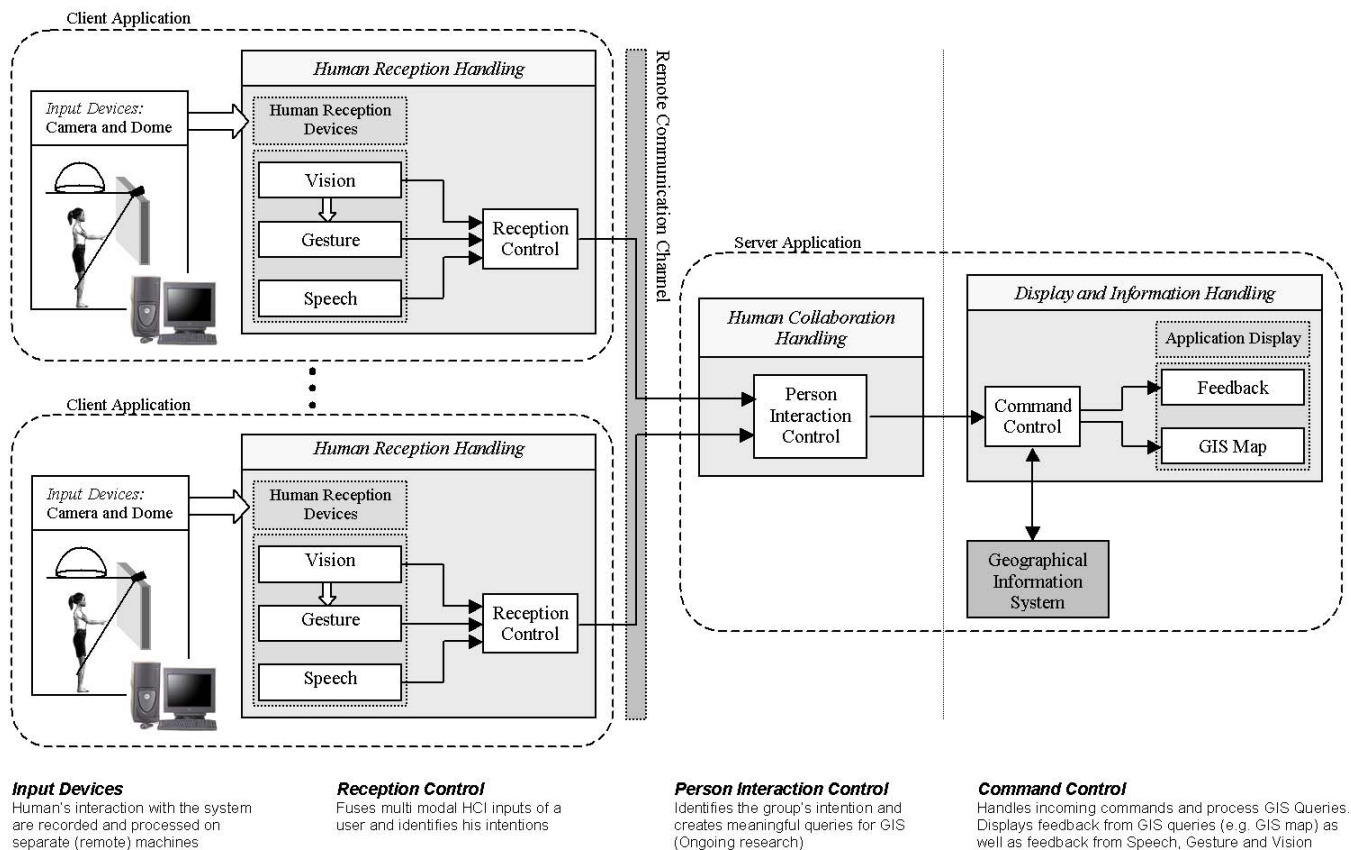


Figure 2. A system overview of the current multi-user human-computer-interface for GIS applications.

noun such as a city or data layer name and thus mainly speech is used to execute the commands. Although speech commands drive query specification and selection (e.g. selection by attribute: “show homes with incomes under 15,000 in Miami-Dade county”, or spatial selection: “show all homes within the hurricane’s surge zone”), data selection events are also controlled by gesture input (e.g. “show all emergency facilities in this area”). Drawing commands, on the other hand, are primarily driven by gestures in the DAVE_G prototype. The current system implementation supports circular, linear and free hand drawing.

4. A PROTOTYPICAL USER INTERFACE FOR GIS

The DAVE_G prototype is based on the multimodal interface framework by Sharma et al [7] and applies ArcObjects [5] for providing necessary GIS functionalities [5]. The prototype is composed of three modules: the Human Reception-, the Human Collaboration-, and the Display and Information Control module (see Figure 2).

The Human Reception Control module uses a single, non-calibrated active camera to find and track a user’s head and hand. The captured hand trajectory is used to recognize gestures and supports the interpretation of a user’s spoken command applying signal fusion. A microphone dome captures the user’s speech and thus allows recognition of spoken commands. The recorded human voice is processed by standard speech recognition software [1]. Figure 2 shows that multiple instances of the Human Reception Control module can be created

The Human Reception Control module is realized as a standalone client application allowing simultaneous usage on distributed client machines. This set-up is flexible, inexpensive, non-stationary and provides sufficient computer processing speed. Communication between components is established through a simple protocol, communicating with the server application over a standard Ethernet infrastructure.

The Human Collaboration Control module receives commands from all connected clients and processes and coordinates them in order to find the most meaningful set of commands that represent the intentions of the users. Currently the DAVE_G prototype performs a temporal synchronization of incoming user commands. Future developments will avoid possible ambiguous and conflicting GIS queries through more complete dialog management.

The filtered user commands are processed in the Display and Information Control module, which forms GIS statements and queries the data sets. Immediate visual feedback on gesture and speech input is provided to users through a “personalized” cursor on the screen. The following sections review the recognition of user interactions, multiple users and data retrieval in greater detail.

4.1. Human Reception Control

In order to recognize users interactions and to form meaningful queries to the GIS database, gesture and speech recognition is used for a natural interface design. The following sections describe the techniques used and explain selected implementation issues.

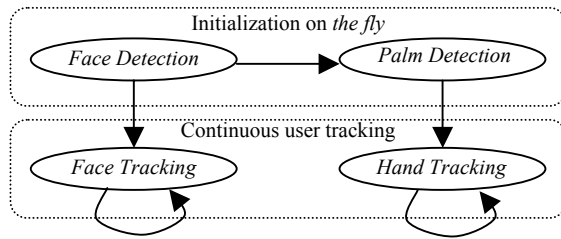


Figure 3. Automatic initialization and continuous tracking allows a “step in and use” interface system

4.1.1. User tracking and hand gesture recognition

One important design pattern for DAVE_G is the so-called “step in and use” functionality, allowing potential users to approach the system and use it instantaneously without prior initialization procedures. “Step in and use” is a prerequisite for collaborative emergency management situations where participants are often not determined ahead of time.

Figure 3 shows a simple sketch of initialization and tracking flow, based on the HCI framework of Sharma et al [7]. A neural network based face detector [22] is applied to the camera image until a face is detected, indicating a potential user is ready for interacting with the system. Since face detection requires high processing power, the actual tracking of the face is performed using a probabilistic framework on skin color and user motion. The color model obtained from the detected face is used to perform automatic palm detection. Using motion energy as another cue, the hand’s location can be tracked continuously.

Finally, continuous gestures are recognized using Hidden-Markov-Models (HMMs), which are already successfully applied in speech recognition software. Supported gestures in DAVE_G are those for pointing, indicating an area and outlining contours.

4.1.2. Speech recognition and command extraction

To recognize spoken words and understand intended user queries, a combination of commercially available speech recognition software [1] and flexible mapping procedures is used to generate queries that can be understood by the prototype (see overview in Figure 4).

A Grammar File based on BNF (Backus Naur Form), defines the abstract structure of phrases and sentences that constrains the speech recognition process and leads to a more accurate and robust recognition performance. This grammar is build such that a very limited subset of natural language can be understood and processed which is complex enough to capture all essential query pronunciations the user might give in a certain task environment.

Words and phrases that are recognized by the grammar are annotated and used in the interpretation of the user’s utterance. An easy to modify Command-Query file specifies the mapping of annotated phrases to actual GIS-query commands with identified attributes and parameters given by the speech input. Thus the prototype is scalable and new user- and GIS-queries are easily integrated.

4.1.3. Fusing speech and gesture to GIS commands

Context and timeline analysis is performed to integrate both interaction cues, speech and hand gestures. Depending on the given command, gesture input might be necessary to provide certain

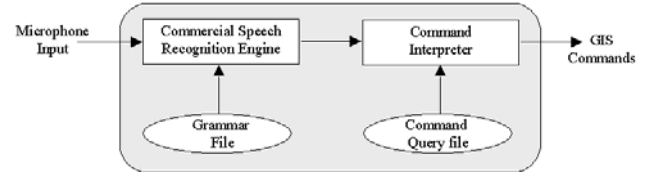


Figure 4. Mapping of Microphone input to GIS Command

spatial information as one or more of the input parameters. Such queries could be for example “zoom here” or “show all cities within this area”. Thus, keywords such as “here”, “there”, “this area” and others indicate information input from the gesture cue.

Since the recognition of both interaction cues is not free of error, neither presence of gesture or phrase annotation by itself is evidence enough for an actual speech-gesture interaction by the user. Instead a certain timely alignment between speech and gesture input has to be found that maximizes the co-occurrence probability. Unfortunately studies have found that this alignment differs significantly between each user and therefore is almost impossible to learn. To account for this high variance we define the probability for co-occurrence on a small time window around the recognized keyword in which such a multimodal interaction input is most likely to occur.

4.2. Human Collaboration Control

The current prototype allows multiple users to interact simultaneously with the GIS. The interpretation process of user queries can be hierarchically divided into two levels. The lower level handles the fusion of all input streams and generates individual user requests as described in the previous section. Controlling a personalized cursor, each user is able to individually interact with the GIS. This freedom might cause some problems with ambiguous or conflicting user commands. In order to tackle the problem, the initial prototype allows a temporal synchronization of incoming user requests. In the future a more intelligent dialog between humans and computers will be realized. This upper level will make use of task specific context and domain knowledge to generate complete and meaningful queries to the GIS-database, incorporating commands from all users. Thus, acting as a dialog manager to resolve ambiguous and conflicting requests.

4.2.1. Dialog management and group collaboration

An agent-based approach will be taken to enable complex communications during problem-solving processes between users and a GIS [9, 20]. The distinctive feature of an agent-based system is that it offers cooperative and helpful assistance to humans in accomplishing their intended tasks. A so-called “GI-agent” needs to reason and support each user’s intentions during collaborative work. A database of previous dialogue interaction must be maintained in order to use it for the interpretation of subsequent user inputs.

The prototype should be more flexible than traditional master-slave GIS operator interaction in several ways. First, the future prototype will allow users to provide partial interaction information without rejecting the command entirely. Second, the system will accept ambiguous commands. In case of unintelligible or incorrect commands, the system will be able to question the user

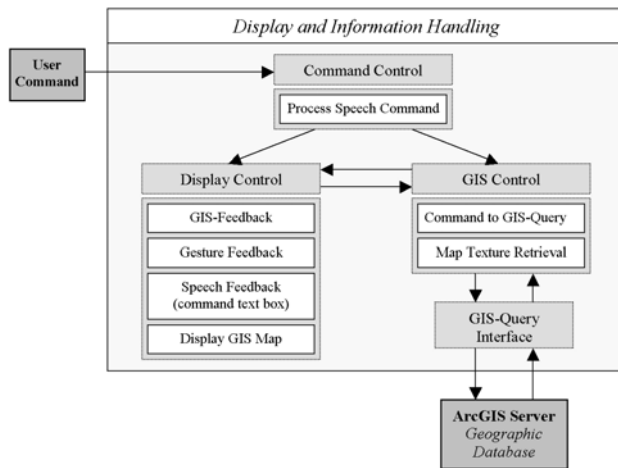


Figure 5. Interface between Display and GIS Server

for further clarification. Third, the GI agent will provide visual feedback to users, maximizing the chance for a successful GIS query.

In order to create a realistic dialogue, the collaborative GI-agent needs to be domain specific. The GI-agent needs to understand the domain users' goals, task structures, and general problem-solving strategies. This information is being acquired through iterative, cognitive task analyses with domain experts and potential users. Additionally, the agent needs to know about spatial data availability and which procedures for data processing and display are valid. Beyond static knowledge processing, knowledge generation during problem solving processes need to be integrated and applied dynamically. In the next project stage, GI agents will be designed, that combine these different types of knowledge, thus, providing cooperative and realistic behavior between the system and the operator during collaborative management situations.

4.3. Display and information Control

An important role for dialog management is the ability to communicate with the user. Therefore, user feedback is an integral part of the Display and Information Control of DAVE_G (see Figure 5).

The Display and Information Control can be divided into three sections: Command Control, GIS Control, and Display Control. The Command Control module filters and passes incoming control and user commands to the respective sub modules. The GIS Control module issues GIS queries to the GIS-database, maintains respective feedback information about currently visible, active, non-visible, and non-active layers and receives the rendered map from the GIS. Queries are issued to the GIS server through the GIS Query Interface Library, which is build upon ArcObjects [5]. The Display Control module renders the selected geospatial datasets, visual interaction information and moving hand cursors on a large screen display. Recognized speech and gesture commands are also displayed, in order to prevent user confusions in the case where the system's perception and recognition is not correct. Future prototype developments of DAVE_G will provide dialogue-based feedback to the user. The system will be enabled to ask clarifying questions and guide the user using the dialog manager in the Human Collaboration Module.

5. CONCLUSION AND FUTURE WORK

Current GIS provide a range of geospatial tools and analysis methods but the rather complex functionalities are usually available only to one user at a time. In collaborative emergency management, current GIS restrict the range of users and narrow chances for efficient task solving.

Therefore, a multimodal, multi-user GIS interface was developed, which supports collaborative work on large screen displays. This research involved tackling research problems in three different fields:

- 1) The development of collaborative and natural gesture speech recognition,
- 2) The design of a human centered interface by applying cognitive system engineering methods to domain analysis in the emergency management domain, and
- 3) The development of an intelligent database that can respond and interact with a group of users.

The developed prototype replaces the traditional keyboard and mouse with free-hand gestures and natural language recognition. The majority of GIS commands are generated by spoken phrases and the systems reliability is highly dependent on the robustness of the underlying speech recognition engine. Since verbal descriptions of geographic phenomena are sometimes ambiguous and not precise enough with spatial relations, a purely natural language based interaction approach to GIS seemed too weak. Thus, gesture recognition was incorporated into the multimodal GIS interface.

At the current state the prototype recognizes and supports multiple users, enabling them to interact directly with a GIS. The fusion of speech and gesture recognition is based on the time analysis of incoming signals. In a second step we will implement a more robust approach that extracts features from the speech and gesture signals and then obtains a measure for co-occurrence, which will provide more insight into multi-modal interactions. A comparable problem exists in fusing the signals from different members of a collaborative working group. Thus, speaker detection and fusing interaction inputs are important steps toward the robustness as well as the envisioned extensions of the next prototype.

The current DAVE_G prototype will be further improved in functionality and usability. In the coming project phases the results of a domain analysis and usability tests will be incorporated. In addition, an agent-based approach will be applied to enable complex communications between users and GIS during collaborative work, refining and filtering out conflicting or confusing input situations.

Further on, our working group will develop more robust vision tools in order to provide stable gesture and speech recognition in extreme situations as in emergency response. Techniques used will be speech and gesture co-analysis, active speaker detection and model-based user tracking. In addition the implementation of a collaborative dialog manager will be given a strong focus in the future stages of the project.

6. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No.BCS-0113030, PI: Alan M. MacEachren, CoPIs: Rajeev Sharma and Guoray Cai. ESRI's software contribution to the project is gratefully acknowledged.

7. REFERENCES

- [1] IBM ViaVoice Speech SDK, IBM, Inc., <http://www-3.ibm.com/software/speech>.
- [2] Brewer, I., MacEachren, A.M., Abdo, H., Gundrum, J. and Otto, G., Collaborative Geographic Visualization: Enabling shared understanding of environmental processes. in *IEEE Information Visualization Symposium*, (Salt Lake City, Utah, 2000).
- [3] Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L. and Clow, J., QuickSet: Multimodal interaction for distributed applications. in *Proceedings of the Fifth International Multimedia Conference (Multimedia '97)*, (1997), ACM Press, 31-40.
- [4] Egenhofer, M. Query Processing in Spatial-Query-by-Sketch. *Journal of Visual Languages and Computing*, 8, 403--424.
- [5] ESRI. ArcObjects Development Tool for GIS, ESRI, Inc., <http://arconline.esri.com/arcobjectsonline/>.
- [6] Kettebekov, S., Krahnstoeve, N., Leas, M., Polat, E., Raju, H., Schapira, E. and Sharma, R. i2Map: Crisis Management using a Multimodal Interface *ARL Federate Laboratory 4th Annual Symposium*, College Park, MD, 2000.
- [7] Krahnstoeve, N., Kettebekov, S., Yeasin, M. and Sharma, R., A Real-Time Framework for Natural Multimodal Interaction with Large Screen Displays. in *Fourth IEEE International Conference on Multimodal Interfaces (ICMI 2002)*, (Pittsburgh, USA, 2002).
- [8] Lamel, L., Bennacef, S., Gauvain, J.L., H.Dartiguest and Temem, J.N. User Evaluation of the MASK Kiosk, ICSLP '98, Sydney, 1998.
- [9] Lesh, N., Rich, C. and Sidner, C.L. Using Plan Recognition in Human-Computer Collaboration.
- [10] Lokuge, I. and Ishizaki, S., Geospace: An interactive visualization system for exploring complex information spaces. in *CHI'95 Proceedings*, (1995).
- [11] Mark, D. NSF Workshop Report -- Geographic Information Science: Critical Issues in an Emerging Cross-disciplinary Research Domain, NSF, Washington, DC, 1999.
- [12] McGee, D.R., Cohen, P.R. and Wu, L., Something from Nothing: Augmenting a Paper Based Work Practice Via Multimodal Interaction. in *Proceedings of the ACM Designing Augmented Reality Environments DARE*, (Helsinki, Denmark, 2000), 71-80.
- [13] Oviatt, S. Ten myths of multimodal interaction. *Communications of the ACM*, 42 (11). 74--81.
- [14] Oviatt, S. and Cohen, P. Multimodal interfaces that process what comes naturally. *Communications of the ACM*, 43 (3). 45-53.
- [15] Oviatt, S.L. Multimodal interfaces for dynamic interactive maps. in *Proceedings of the Conference on Human Factors in Computing Systems (CHI'96)*, ACM Press, New York, 1996, 95-102.
- [16] Rasmussen, J., Pejtersen, A.M. and Goodstein, L.P. *Cognitive systems engineering*. Wiley, New York, 1994.
- [17] Schapira, E. and Sharma, R. Experimental evaluation of vision and speech-based multimodal interfaces. in *Workshop on Perceptive User Interfaces*, 2001.
- [18] Schlaisich, I. and Egenhofer, M., Multimodal Spatial Querying: What People Sketch and Talk About. in *1st International Conference on Universal Access in Human-Computer Interaction*, (New Orleans, LA, 2001), 732-736.
- [19] Sharma, R., Pavlovic, V.I. and Huang, T.S. Toward a multimodal human-computer interface. in *In Proceedings of the IEEE*, 1998, 853--869.
- [20] Stock, O., Strapparava, C. and Zancanaro, M. Augmenting and Executing SharedPlans for Multimodal Communication. in Beun, R.-J. ed. *Multimodal Cooperative Communication*, Springer-Verlag, Berlin Heidelberg, 2001, 89-112.
- [21] Vicente, K.J. *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 1999.
- [22] Yeasin, M. and Kuniyoshi, Y., Detecting and tracking human face using a space-varying sensor and an active head. in *proceedings of the IEEE computer society conference on Computer Vision and Pattern Recognition (CVPR'00)*, (NC, USA, 2000).