QuickSet: Multimodal Interaction for Simulation Set-up and Control

Philip R. Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman,
Ira Smith, Liang Chen and Josh Clow
Center for Human Computer Communication
Oregon Graduate Institute of Science and Technology
P.O.Box 91000
Portland, OR 97291-1000 USA
Tel: 1-503-690-1326
E-mail: pcohen@cse.ogi.edu
http://www.cse.ogi.edu/CHCC

ABSTRACT

This paper presents a novel multimodal system applied to the setup and control of distributed interactive simulations. We have developed the QuickSet prototype, a pen/voice system running on a hand-held PC, communicating through a distributed agent architecture to NRaD's[1] LeatherNet system, a distributed interactive training simulator built for the US Marine Corps (USMC). The paper briefly describes the system and illustrates its use in multimodal simulation setup.

KEYWORDS: multimodal interfaces, agent architecture, gesture recognition, speech recognition, natural language processing, distributed interactive simulation.

1. INTRODUCTION

In order to train personnel more effectively, the US military is developing large-scale distributed simulation capabilities. Begun as SIMNET in the 1980's [23], these distributed, interactive environments attempt to provide a high degree of fidelity in simulating combat, including simulations of the individual combatants, the equipment, entity movements, atmospheric effects, etc. There are four general phases of user interaction with these simulations: Creating entities, supplying their initial behavior, interacting with the entities during a running simulation, and reviewing the results. The present research concentrates on the first two of these stages.

Our contribution to the distributed interactive simulation (DIS) effort is to rethink the nature of the user interaction. As with most modern simulators, DISs are controlled via graphical user interfaces (GUIs). However, the simulation GUI is showing signs of strain, since even for a small-scale scenario, it requires users to choose from hundreds of entities in order to select the desired ones to place on a map. To compound these interface problems, the military is intending to increase the scale of the simulations dramatically, while at the same time, for reasons of mobility and affordability, desiring that simulations should be creatable from small devices (e.g., PDAs). This impending collision of trends for smaller screen size and for more entities requires a different paradigm for human-computer interaction.

We have argued generically that GUI technologies offer advantages in allowing users to manipulate objects that are on the screen, in reminding users of their options, and in minimizing errors [7]. However, GUIs are often weak in supporting interactions with many objects, or objects not on the screen. In contrast, it was argued that linguistically-based interface technologies offer the potential to describe large sets of objects, which may not all be present on a screen, and can be used to create more complex behaviors through specification of rule invocation conditions. Simulation is one type of application for which these limitations of GUIs, as well as the strengths of natural language, especially spoken language, are apparent [6].

It has become clear, however, that speech-only interaction is not optimal for spatial tasks. Using a high-fidelity "Wizard-of-Oz" methodology [20], recent empirical results demonstrate clear language processing and task performance advantages for multimodal (pen/voice) input over speech-only input for map-based systems [17,18].

3. QUICKSET

To address these simulation interface problems, and motivated by the above results, we have developed QuickSet (see Figure 1) a collaborative, handheld, multimodal system for configuring military simulations based on LeatherNet [5], a system used in training platoon leaders and company commanders at the USMC base at 29 Palms, California. LeatherNet simulations are created using the ModSAF simulator [10] and can be visualized in a CAVE-based virtual reality environment [11, 26] called CommandVu (see Figure 2 — QuickSet systems are on the soldiers' tables). In addition to LeatherNet, QuickSet is being used in a second effort called ExInit (Exercise

[1] NRaD = US Navy Command and Control Ocean Systems Center Research Development Test and Evaluation (San Diego).

Initialization), that will enable users to create division-sized exercises. Because of the use of OAA, QuickSet can interoperate with agents from CommandTalk [14], which provides a speech-only interface to ModSAF.

QuickSet runs on both desktop and hand-held PC's, communicating over wired and wireless LAN's, or modem links. The system combines speech and pen-based gesture input on multiple 3-lb hand-held PCs (Fujitsu Stylistic 1000), which communicate via wireless LAN through the Open Agent Architecture (OAA)[2] [8], to ModSAF, and also to CommandVu. With this highly portable device, a user can create entities, establish "control measures" (e.g., objectives, checkpoints, etc.), draw and label various lines and areas, (e.g., landing zones) and give the entities behavior.

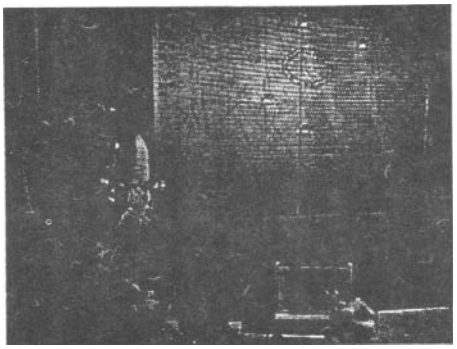


Figure 1: QuickSet running on a wireless handheld PC.

In the remainder of the paper, we illustrate the system briefly, describe its components, and discuss its application.

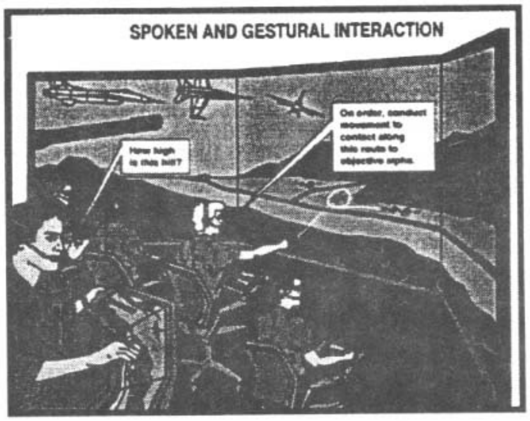

SPOKEN AND GESTURAL INTERACTION
How high is this hill?
On order, conduct movement to contact along this route to objective alpha.

Figure 2: Artist's rendition of QuickSet used with CommandVu virtual display of distributed interactive simulation.

4. SYSTEM ARCHITECTURE

Architecturally, QuickSet uses distributed agent technologies based on the Open Agent Architecture for interoperation, information brokering and distribution. An agent-based architecture was chosen to support this application because it offers easy connection to legacy applications, and the ability to run the same set of software components in a variety of hardware configurations, ranging from stand-alone on the handheld PC, to distributed operation across numerous workstations and PCs. Additionally, the architecture supports mobility in that lighter weight agents can run on the handheld, while more computationally-intensive processing can be migrated elsewhere on the network. The agents may be written in any programming language (here, Quintus Prolog, Visual C++, Visual Basic, and Java), as long as they communicate via an interagent communication language. The configuration of agents used in the Quickset system is illustrated in Figure 3. A brief description of each agent follows:

QuickSet interface: On the handheld PC is a georeferenced map of the region such that entities displayed on the map are registered to their positions on the actual terrain, and thereby to their positions on each of the various user interfaces connected to the simulation. The map interface agent provides the usual pan and zoom capabilities, multiple overlays, icons, etc. The user can draw directly on the map, in order to create points, lines, and areas. The user can create entities, give them behavior, and watch the simulation unfold from the handheld. When the pen is placed on the screen, the speech recognizer is activated, thereby allowing users to speak and gesture simultaneously.

Speech recognition agent: The speech recognition agent used in QuickSet employs either IBM's VoiceType Application Factory or VoiceType 3.0 recognizers. The recognizers use an HMM-based continuous speaker-independent speech recognition technology for PC's under Windows 95/NT. Currently, the system has a vocabulary of 450 words. It produces a single most likely interpretation of an utterance.

Gesture recognition agent: OGI's gesture recognition agent processes all pen input from a PC screen or tablet. The agent weights the results of both HMM and neural net recognizers, producing a combined score for each of the possible recognition results. Currently, 45 gestures can be recognized, resulting in the creation of 21 military symbols, irregular shapes, and various types of lines.

Natural language agent: The natural language agent currently employs a definite clause grammar and produces typed feature structures as a representation of the utterance meaning. Currently, for this task, the language consists of noun phrases that label entities, as well as a variety of imperative constructs for supplying behavior.

Multimodal integration agent: The multimodal interpretation agent accepts typed feature structure meaning representations from the language and gesture recognition agents, and produces a unified multimodal interpretation.

[2] Open Agent Architecture is a trademark of SRI International.

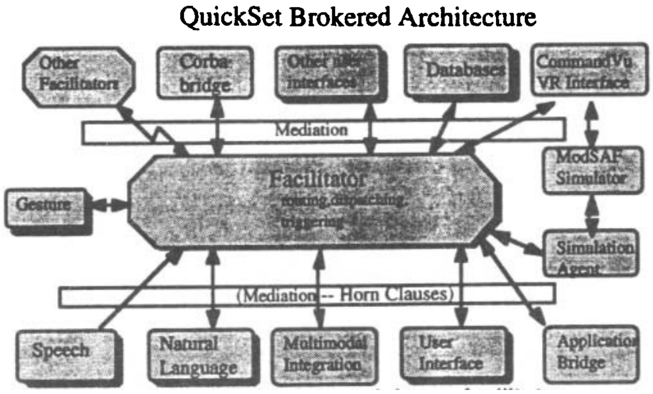


Figure 3: A blackboard is used by a facilitator agent, who routes queries to appropriate agents for solution.

Simulation agent: The simulation agent, developed primarily by SRI International, but modified by us for multimodal interaction, serves as the communication channel between the OAA-brokered agents and the ModSAF simulation system. This agent offers an API for ModSAF that other agents can use.

Web display agent: The Web display agent can be used to create entities, points, lines, and areas. It posts queries for updates to the state of the simulation via Java code that interacts with the blackboard and facilitator. The queries are routed to the running ModSAF simulation, and the available entities can be viewed over a WWW connection using a suitable browser.

Other user interfaces: When another user interface connected to the facilitator subscribes to and produces the same set of events as others, it immediately becomes part of a collaboration. One can view this as human-human collaboration mediated by the agent architecture, or as agent-agent collaboration.

CommandVu agent: Since the CommandVu virtual reality system is an agent, the same multimodal interface on the handheld PC can be used to create entities and to fly the user through the 3-D terrain. For example, the user can ask "CommandVu, fly me to this platoon <gesture on the map>."

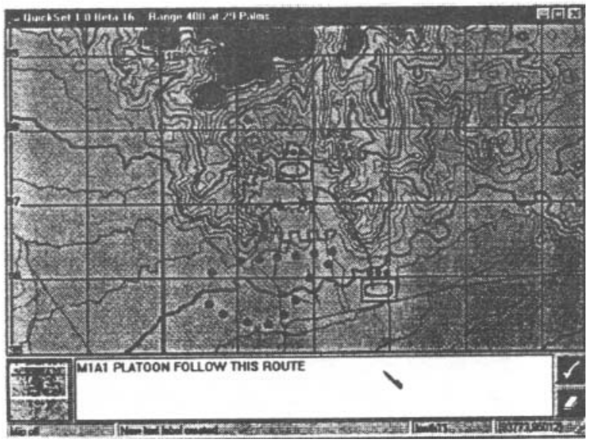
Application bridge agent: The bridge agent generalizes the underlying applications' API to typed feature structures, thereby providing an interface to the various applications such as ModSAF, CommandVu, and Exinit. This allows for a domain-independent integration architecture in which constraints on multimodal interpretation are stated in terms of higher-level constructs such as typed feature structures, greatly facilitating reuse.

CORBA bridge agent: This agent converts OAA messages to CORBA IDL (Interface Definition Language) for the Exercise Initialization project.

More detail on the architecture and the individual agents are provided in [12, 22].

5. EXAMPLE

Holding QuickSet in hand, the user views a map from the ModSAF simulation, and with spoken language coupled with pen gestures, issues commands to ModSAF. In order to create a unit in QuickSet, the user would hold the pen at the desired location and utter (for instance): "red T72 platoon" resulting in a new platoon of the specified type being created at that location.



M1A1 PLATOON FOLLOW THIS ROUTE

Figure 4: The QuickSet interface as the user establishes two platoons, a barbed-wire fence, a breached minefield, and then issues a command to one platoon to follow a traced route.

The user then adds a barbed-wire fence to the simulation by drawing a line at the desired location while uttering "barbed wire." Similarly a fortified line is added. A minefield of an amorphous shape is drawn and is labeled verbally, and finally an M1A1 platoon is created as above. Then the user can assign a task to the new platoon by saying "M1A1 platoon follow this route" while drawing the route with the pen. The results of these commands are visible on the QuickSet screen, as seen in Figure 4, in the ModSAF simulation, and in the CommandVu 3D rendering of the scene. In addition to multimodal input, unimodal spoken language and gestural commands can be given at any time, depending on the user's task and preference.

6. MULTIMODAL INTEGRATION

Since any unimodal recognizer will make mistakes, the output of the gesture recognizer is not accepted as a simple unilateral decision. Instead the recognizer produces a set of probabilities, one for each possible interpretation of the gesture. The recognized entities, as well as their recognition probabilities, are sent to the facilitator, which forwards them to the multimodal interpretation agent. In combining the meanings of the gestural and spoken interpretations, we attempt to satisfy an important design consideration, namely that the communicative modalities should compensate for each other's weaknesses [7, 16]. This is accomplished by selecting the highest scoring *unified interpretation* of speech and gesture. Importantly, the unified interpretation might not include the highest scoring gestural (or spoken language) interpretation because it might not be semantically compatible with the other mode. The key to this interpretation process is the use of a typed feature structure [1, 3] as a meaning representation language that is common to the natural language and gestural interpretation agents. Johnston et al. [12] present the details of multimodal integration of continuous speech and pen-based gesture, guided by research in users' multimodal integration and synchronization strategies [19]. Unlike many previous approaches to multimodal integration (e.g, [2, 9, 12, 15, 25]) speech is not "in charge," in the sense of relegating gesture a secondary and dependent role. This mutually-compensatory interpretation process is capable of analyzing multimodal constructions, as well as speech-only and pen-only constructions when they occur. Vo and Wood's system [24] is similar to the one reported here, though we believe the use of typed feature structures provides a more generally usable and formal integration mechanism than their frame-merging strategy. Cheyer and Julia [4] sketch a system based on Oviatt's [17] results and the OAA [8], but do not discuss the integration strategy nor multimodal compensation.

7. CONCLUDING REMARKS

QuickSet has been delivered to the US Navy (NRaD) and US Marine Corps. for use at 29 Palms, California, where it is primarily used to set up training scenarios and to control the virtual environment. It is also installed at NRaD's Command Center of the Future. The system was used by the US Army's 82nd Airborne Corps. at Ft. Bragg during the Royal Dragon Exercise. There, QuickSet was deployed in a tent, where it was subjected to an extreme noise environment, including explosions, low-flying jet aircraft, generators, and the like. Not surprisingly, spoken interaction with QuickSet was not feasible, although users gestured successfully. Instead, users wanted to gesture. Although we had provided a multimodal interface for use in less hostile conditions, nevertheless we needed to provide,and in fact have provided, a complete overlap in functionality, such that any task can be accomplished just with pen or just with speech when necessary. Finally, QuickSet is now being extended for use in the ExInit simulation initialization system for DARPA's STOW-97 Advanced Concept Demonstration that is intended for creation of division-sized exercises.

Regarding the multimodal interface itself, QuickSet has undergone a "proactive" interface evaluation in that the studies that were performed in advance of building the system predicted the utility of multimodal over unimodal speech as an input to map-based systems [17, 18]. In particular, it was discovered in this research that multimodal interaction generates simpler language than unimodal spoken commands to maps. For example, to create a "phase line" between two three-digit <x,y> grid coordinates, a user would have to say: "create a line from nine four three nine six one to nine five seven nine six eight and call it phase line green" [14]. In contrast, a QuickSet user would say "phase line green" while drawing a line. Creation of area features with unimodal speech would be more complex still, if not infeasible. Given that numerous difficult-to-process linguistic phenomena (such as utterance disfluencies) are known to be elevated in lengthy utterances, and also to be elevated when people speak locative constituents [17, 18], multimodal interaction that permits pen input to specify locations and that results in brevity offers the possibility of more robust recognition.

Further development of QuickSet's spoken, gestural, and multimodal integration capabilites are continuing. Research is also ongoing to examine and quantify the benefits of multimodal interaction in general, and our architecture in particular.

ACKNOWLEDGMENTS

This work is supported in part by the Information Technology and Information Systems offices of DARPA under contract number DABT63-95-C-007, in part by ONR grant number N00014-95-1-1164, and has been done in collaboration with the US Navy's NCCOSC RDT&E Division (NRaD), Ascent Technologies, Mitre Corp., MRJ Corp., and SRI International.

REFERENCES

1. Calder, J. Typed unification for natural language processing. In E. Klein and J. van Benthem (Eds.), Categories, Polymorphisms, and Unification. Centre for Cognitive Science, University of Edinburgh, Edinburgh, 1987, 65-72.

2. Brison, E. and N. Vigouroux. (unpublished ms.). Multimodal references: A generic fusion process. URIT-URA CNRS. Universitė Paul Sabatier, Toulouse, France.

3. Carpenter, R. *The logic of typed feature structures.* Cambridge University Press, Cambridge, 1992.

4. Cheyer, A., and L. Julia. Multimodal maps: An agent-based approach. *International Conference on Cooperative Multimodal Communication (CMC/95),* May 1995. Eindhoven, The Netherlands, 1995, 24-26.

5. Clarkson, J. D., and Yi., J., LeatherNet: A synthetic forces tactical training system for the USMC commander. *Proceedings of the Sixth Conference on Computer Generated Forces and Behavioral Representation.* Institute for simulation and training. Technical Report IST-TR-96-18, 1996, 275-281.

6. Cohen, P. R. Integrated Interfaces for Decision Support with Simulation, *Proceedings of the Winter Simulation Conference*, Nelson, B. and Kelton, W. D. and Clark, G. M., (eds.), ACM, New York, December, 1991, 1066-1072.

7. Cohen, P. R. The Role of Natural Language in a Multimodal Interface. *Proceedings of UIST'92*, ACM Press, New York, 1992, 143-149.

8. Cohen, P.R., Cheyer, A., Wang, M., and Baeg, S.C. An Open Agent Architecture. *Working notes of the AAAI Spring Symposium Series on Software Agents* Stanford Univ., CA, March, 1994, 1-8.

9. Cohen, P. R., Dalrymple, M., Moran, D.B., Pereira, F. C. N., Sullivan, J. W., Gargan, R. A., Schlossberg, J. L., and Tyler, S.W. Synergistic Use of Direct Manipulation and Natural Language, *Human Factors in Computing Systems: CHI'89 Conference Proceedings*, ACM, Addison Wesley Publishing Co New York, 227-234, 1989.

10. Courtemanche, A.J. and Ceranowicz, A. ModSAF Development Status. *Proceedings of the Fifth Conference on Computer Generated Forces and Behavioral Representation*, Univ. Central Florida, Orlando, 1995, 3-13.

11. Cruz-Neira, C. D.J. Sandin, T.A. DeFanti, "Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE," *Computer Graphics* (Proceedings of SIGGRAH'93), ACM SIGGRAPH, August 1993, 135-142.

12. Johnston, M., Cohen, P. R., McGee, D., Oviatt, S. L., Pittman, J., and Smith, I.. Unification-based multimodal integration, in submission.

13. Koons, D.B., C.J. Sparrell and K.R. Thorisson. 1993. Integrating simultaneous input from speech, gaze, and hand gestures. In Mark T. Maybury (ed.) *Intelligent Multimedia Interfaces*. AAAI Press/ MIT Press, Cambridge, MA, 257-276.

14. Moore, R., Dowding, J. Bratt, H. Gawron, J. M., and Cheyer, A., CommandTalk: A Spoken-Language Interface for Battlefield Simulations, 1997, (this volume).

15. Neal, J.G. and Shapiro, S.C. Intelligent multi-media interface technology. In J.W. Sullivan and S.W. Tyler, editors, *Intelligent User Interfaces*, chapter 3, pages 45–68. ACM Press Frontier Series, Addison Wesley Publishing Co., New York, New York, 1991.

16. Oviatt, S. L., Pen/Voice: Complementary multimodal communication, Proceedings of SpeechTech'92, New York, February, 1992, 238-241.

17. Oviatt, S.L. Multimodal interfaces for dynamic interactive maps. *Proceedings of CHI'96 Human Factors in Computing Systems* (April 13-18, Vancouver, Canada), ACM Press, NY, 1996, 95-102.

18. Oviatt, S. L., Multimodal interactive maps: Designing for human performance, *Human-Computer Interaction*, in press.

19. Oviatt, S. L, A. DeAngeli, and K. Kuhn. In press. Integration and synchronization of input modes during multimodal human-computer interaction. *Proceedings of the Conference on Human Factors in Computing Systems (CHI '97)*, ACM Press, New York.

20. Oviatt, S. L., Cohen, P. R, Fong, M. W. and Frank, M. P., A rapid semi-automatic simulation technique for interactive speech and handwriting, *Proceedings of the 1992 International Conference Spoken Language Processing, vol. 2*, University of Alberta, J. Ohala (ed.), October, 1992, 1351-1354.

21. Oviatt, S. L., Cohen, P. R., Wang, M. Q.,Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity, *Speech Communication,* 15 (3-4), 1994.

22. Pittman, J.A., Smith, I.A., Cohen, P.R., Oviatt, S.L., and Yang, T.C. QuickSet: A Multimodal Interface for Military Simulation. in *Proceedings of the Sixth Conference on Computer-Generated Forces and Behavioral Representation*, Orlando, Florida, 1996.

23. Thorpe, J. A., The new technology of large scale simulator networking: Implications for mastering the art of warfighting. *Proceedings of the 9th Interservice/industry Training Systems Conference*, Orlando, Florida, December, 1987, 492-501.

24. Vo, M. T. and C. Wood. Building an application framework for speech and pen input integration in multimodal learning interfaces. *International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, GA, 1996.

25. Wauchope, K. Eucalyptus: Integrating natural language input with a graphical user interface. Naval Research Laboratory, Report NRL/FR/5510--94-9711, 1994.

26. Zyda, M. J., Pratt, D. R., Monahan, J. G., and Wilson, K. P., NPSNET: Constructing a 3-D virtual world, Proceedings of the 1992 Symposium on Interactive 3-D Graphics, March, 1992.