Designing, Developing & Evaluating Multimodal Applications

Adam Cheyer & Luc Julia
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025 USA
{Adam.Cheyer, Luc.Julia}@sri.com

Introduction

SRI's CHIC[1] group has been working with multimodal pen and voice applications since 1994. In this paper, we briefly describe several implemented systems, provide an overview of the infrastructure used for these projects, and then outline a novel methodology that encourages an incremental, integrated approach for simultaneously designing, developing and evaluating multimodal systems. We finish by listing a few research areas we consider worthy of further study.

Pen-Voice Applications

Although we have implemented a few applications of a more textual nature, we have focused primarily on multimodal applications that possess graphical or spatial qualities. In this section, we will briefly describe three applications that incorporate synergistic fusion of pen and voice.

Multimodal Map

Our Multimodal Map application provides an interactive interface on which the user may draw, write, or speak. In a travel planning domain (Figure 1), available information includes data about hotels, restaurants and tourist sites that have been retrieved by software agents from commercial web sites [2]. A typical query might be "Show me all french restaurants within two miles of here."

The primary research focus of this work is on how to generate the most appropriate interpretation for the incoming streams of multimodal input. Our approach employs an agent-based framework (see section on architecture below) to coordinate competition and cooperation among distributed information sources, which work in parallel to resolve the ambiguities arising at every level of the interpretation process:

- *Low-level processing of the data stream:* Pen input may be interpreted as a gesture by one algorithm or as handwriting by a separate recognition process. Multiple hypotheses may be returned by any modality recognition component.
- *Anaphora resolution*: When resolving references, separate information sources contribute to the resolution process. For example, given the utterance "Show photo of the hotel on Main Street", a natural language agent may contribute it's context about what the user was speaking about recently, a gesture recognition agent might provide results from a simultaneous pointing or arrow gesture, the map interface might indicate that only one hotel is visible to the user, the database agent provides information about which hotels have addresses on Main Street, and so forth. New information sources and fusion strategies can be added at runtime, without having to change other code in the system.
- *Cross-modal influences*: When multiple modalities are used together, one modality may reinforce or help disambiguate another. For example, an arrow has different interpretations when accompanied by the command "scroll map" than for "show photo". Two modalities may also submit conflicting information (e.g. "scroll west" with an arrow drawn to the east).

MVIEWS: Tools for the Video Analyst

Full-motion video has inherent advantages over still imagery for characterizing events and movement. Military and intelligence analysts currently view live video imagery from airborne and ground-based video platforms, but few tools exist for efficient exploitation of the video and its accompanying metadata. In pursuit of this goal,

[1] Computer Human Interface Center (CHIC)

SRI developed MVIEWS[2], a system for annotating, indexing, extracting, and disseminating information from video streams for surveillance and intelligence applications [1]. An analyst watching one or more live video feeds is able to use pen and voice to annotate the events taking place (Figure 2). The annotation streams are indexed by speech and gesture recognition technologies for later retrieval, and can be quickly scanned using a timeline interface, then played back during review of the film. Pen and speech can also be used to command various aspects of the system, including image processing functions, with multimodal utterances such as "Track this" or "If any object enters this area, notify me immediately."

Tasking Multiple Robots

Integrating many of the agents from the two previous applications, we developed a prototype interface for controlling and tasking a team of robots and their sensors [5]. In addition to directing robots using a multimodal map-style interface (e.g., "You are here facing this direction. Go pick this up."), and controlling and annotating robot's video input (e.g., "Zoom in on this. Grab this region for the report."), pen and voice were used in a cooperative map-building task (Figure 3). An operator with a general idea of a floor space layout can sketch a rough map and indicate constraints on individual entities. The result is cleaned up and directed to the robots, which attempt to match their local sensors to the global map, updating information as they go. Clarification dialogs may be required between human and mobile machines.

Open Agent Architecture

The Open Agent Architecture™ (OAA)[3] is a general-purpose infrastructure for constructing systems composed of multiple software components written in different programming languages and distributed across multiple platforms [8]. Similar in spirit to distributed object frameworks such as OMG's CORBA or Microsoft's DCOM, OAA provides support for describing more flexible and adaptable interactions than the tightly-bound method calls provided by these architectures. In addition, OAA's facilitation-based approach provides numerous services suitable for developing multimodal applications, including the following:

- Agents communicate using a logic-based tasking language called ICL. Several agent-enabled systems exist that can translate from English to ICL and back to English, enabling users to interact closely with agents in a natural way.
- The infrastructure, through Facilitator agents, supports conflict management, competitive and cooperative parallelism, failure conditions across multiple agents, etc. (See example above on anaphora resolution).
- OAA has built-in support for developing collaborative applications where multiple humans and agents share the same workspace. All three of the applications described above are collaboration-enabled.

In addition to those described above, OAA has been used to implement more than 30 applications in various domains, many of them multimodal in nature [10]. OAA has also been used by organizations outside of SRI. Examples include OGI's QuickSet system [4] and EPFL's telepresent surgical simulations.

Integrated Design, Development & Evaluation

Wizard of Oz (WOZ) simulations have proven an effective technique for discovering how users would interact with systems that are beyond the current state of the art [7]. In [3], we describe a novel extension to the WOZ methodology that we call a WOZZOW[4] simulation. Here's how it works (using Multimodal Map as an example):

1. Instead of constructing a specialized simulation environment whose sole purpose is to collect data from users, we run a real, working OAA application in multi-user collaboration mode so the displays are synchronized. One display is configured in a minimalist way, with no scrollbars, toolbars, or buttons, to allow only pen and voice input; the other is presented with all system dialog boxes and GUI controls visible.

---

[2] Multimodal Video Image Exploitation WorkStation (MVIEWS)

[3] More information can be found on the OAA homepage at http://www.ai.sri.com/~oaa

[4] A WOZZOW simulation is a 2-way Wizard of Oz simulation, where both the naïve user and expert wizard are subjects of the experiment simultaneously.

2. An uninitiated user (the "subject") is told to write, draw, or speak to the system to accomplish a complex task such as planning a weekend in Toronto. In a second room is hidden our Wizard, an experienced user of the application, whose role is to perform the actions requested by the subject as quickly as possible, using any combination of pen, voice, or GUI controls. In this way, the subject is lead to believe that the system is interpreting his input. In the case of the Wizard, the system really is processing her multimodal requests.

In a single experiment, we simultaneously collect data input from both an unconstrained new user (unknowingly) operating a simulated system – providing answers about how pen and voice are combined in the most natural way possible – and from an expert user (under duress) making full use of our best automated system. In analyzing the Wizard's interactions, we can learn how well the real system performs, and investigate the roles of a standard GUI (e.g., buttons, scrollbars) relative to a multimodal interface.

A WOZZOW simulation provides many advantages over a standard WOZ simulation:

- There is a very low cost to turn an OAA application into a WOZZOW simulation thanks to OAA's built-in collaboration, logging and playback facilities.
- Resulting improvements to the end-user system garnered from the experiments are *quantifiable*. Groups of subject input data can be run over the real system before and after findings are incorporated (e.g, enhancing speech grammars, fusion algorithms), and the rate of success can be measured.
- An application develops in an incremental style, where the performance of the real system is tested even as the simulation side of the experiment provides information about future enhancements.

In [6] and [9], we provide initial results of experiments using this approach for the multimodal map application.

Further Research Areas

At SRI, we are particularly interested in studying the use of language and gesture in interacting with a computerized terrain model, particularly in the context of solving spatial problems. Specific issues include:

- Deictic and gestural reference to features of the terrain: How do people refer to and distinguish between features of a terrain model with words and gesture?
- Discourse structure: How does the structure of the interaction enable more economical communication, and how can a computer system utilize this structure in interpreting spoken and gestural input? How is the discourse structured by the structure of the terrain model and of the task or operation being executed in the terrain?
- Spatial language: How does language carve up space, and what is its relation to more geometric representations of space used in terrain models, particularly for perspective-relative relational terms?

We performed initial studies using 2-dimensional multimodal maps and the WOZZOW approach, and are now investigating spatial reference with respect to 3-dimensional, realistic terrain visualizations.

References

1. Cheyer, A. & Julia L. (1998). MVIEWS: Multimodal Tools for the Video Analyst. Conference on Intelligent User Interfaces (IUI'98), San Francisco, January 1998.
2. Cheyer A. & Julia L. (1998). Multimodal Maps: An Agent-based Approach. In Multimodal Human-Computer Communication, Lecture Notes in Artificial Intelligence #1374, Bunt/Beun/Borghuis (Eds.), Springer, pp 111-121.
3. Cheyer, A., Julia, L & Martin, J.C. (1998). A Unified Framework for Constructing Multimodal Experiments and Applications. CMC'98 : January 1998, Tilburg, (The Netherlands).
4. Cohen, P. R., Johnston, M., McGee, D., Smith, I., Oviatt, S., Pittman, J., Chen, L., and Clow, J. (1997). QuickSet: Multimodal interaction for simulation set-up and control. Proceedings of the Fifth Applied Natural Language Processing meeting . Association for Computational Linguistics: Washington, D.C.
5. Julia L. (1998). Tasking Robots through Multimodal Interfaces: the "Coach Metaphor". In Collective Robotics, Lecture Notes in Artificial Intelligence #1456, Drogoul (Ed.), Springer, pp 38-47.
6. Kehler A., Martin J.C., Cheyer A., Julia L., Hobbs J. & Bear J. (1998). On representing Salience and Reference in Multimodal Human-Computer Interaction. AAAI'98 (Representations for Multi-Modal Human-Computer Interaction) : Madison (USA), pp 33-39.
7. Oviatt, S. (1996). Multimodal interfaces for dynamic interactive maps. Proceedings of CHI'96. pp 95-102.

8. Martin, D., Cheyer, A. & Moran, D. (1999). The Open Agent Architecture: A framework for building distributed software systems. Applied Artificial Intelligence: An International Journal. Volume 13, Number 1-2. January-March 1999.
9. Martin, J.C., Julia, L. & Cheyer. (1998). A. A Theoretical Framework for Multimodal User Studies. CMC'98 : January 1998, Tilburg, (The Netherlands).
10. Moran D., Cheyer A., Julia L., Martin D. & Park S. (1998). Multimodal User Interfaces in the Open Agent Architecture. Journal of Knowledge-Based SYSTEMS, #10, pp 295-303.

Figure 1: Multimodal Map

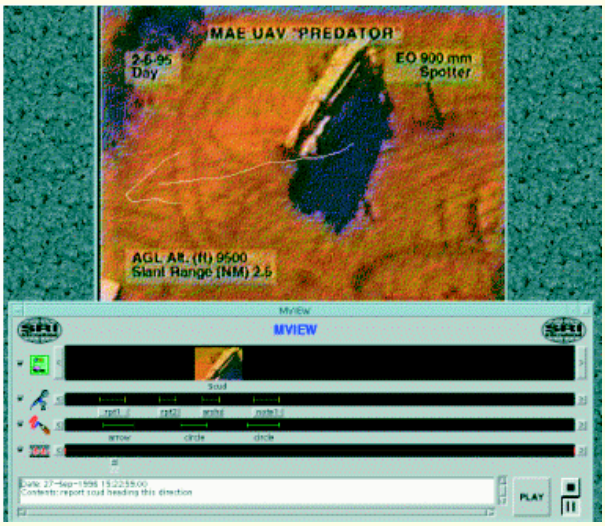

Figure 2: MVIEWS tools for the video analyst

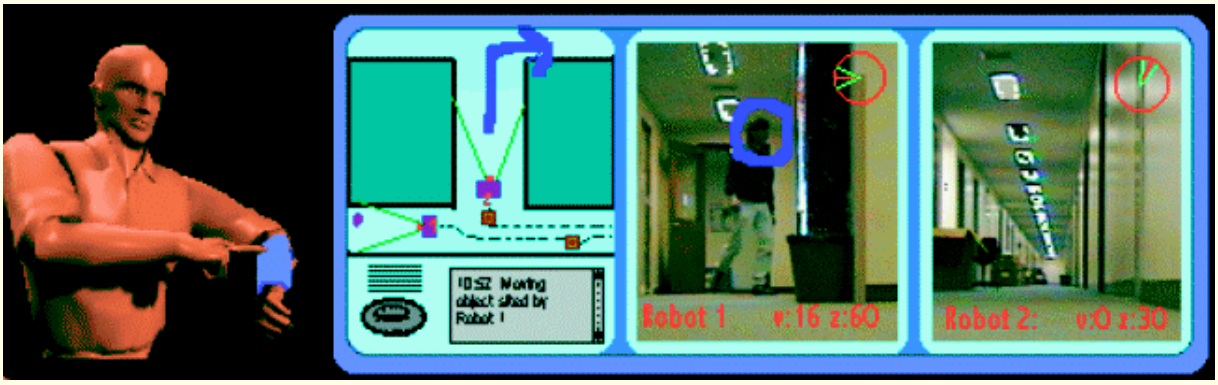

Figure 3. Concept screen for Multmodal Robot Tasking. Prototype implemented on laptop.