Multimodal Co-located Collaboration

Edward Tse
University of Calgary
2500 University Dr. N.W., Calgary, Alberta, Canada T3H 2V1
Tel: 403-210-9502
tsee@cs.ucalgary.ca[1]

ABSTRACT
People naturally perform multimodal interactions in everyday real world settings, as they collaborate over large visual surfaces such as paper maps on walls and tables. While a new generation of research technologies now supports co-located collaboration, they do not yet directly leverage such rich multimodal interaction. Even though a rich behavioural foundation is emerging that informs the design of co-located collaborative technologies, most systems still typically limit input to a single finger or pointer. This problem is partly caused by the fact that single point interaction is the only input easily accessible to researchers through existing toolkits and input APIs. In this research, I distil existing theories, models and ethnographic studies on co-present collaboration into behavioural foundations that describe the individual and group benefits for using gesture and speech multimodal input in a co-located large display setting. Next, I will develop a toolkit that facilitates rapid prototyping of multimodal co-located applications over large digital displays. Finally, using these applications, I will conduct a number of studies exploring design in a multimodal setting.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces.
General terms: Design, Human Factors, Experimentation

Keywords: Multimodal Interaction, Co-located Interaction, Tabletop Interaction, Whole handed gestures, Verbal Alouds

INTRODUCTION
Consider everyday life. Co-located collaborators often work over artefacts placed atop physical tabletops, such as maps containing rich geospatial information. Their work is very nuanced, where people use gestures and speech in subtle ways as they interact with artefacts on the table and communicate with one another.

With the advent of large multi-touch surfaces, researchers are now applying knowledge of co-located tabletop interaction to create appropriate technical innovations in digital table design. My research focus is on advancing our understanding of multimodal co-located interaction, specifically on the feasibility and potential benefits and problems of multimodal co-located input. My motivation for this thesis can be summarized as follows:

1. *Co-located collaborators can leverage the power of digital displays for saving and distributing annotations, for receiving real-time updates, and for exploring and updating large amounts of data in real time.*
2. *Multimodal interaction allows people to interact with a digital surface using the same hand gestures and speech utterances that they use in the physical environment*
3. *An important side effect of multimodal interaction is that it provides improved awareness to people working together in a co-located environment.*

To investigate this thesis, my research will first examine the work practices and findings reported in various ethnographic studies of safety critical environments; e.g., air traffic control, military command and control, underground subway traffic management and hospital emergency rooms. I will examine what types of speech and gesture interactions are used, how it is performed (e.g., simultaneously vs sequentially), how conflicts are handled (e.g., turn taking protocols), and how these natural activities can be supported by technology. Ultimately, this research will involve investigating and bridging a range of perspectives: human computer interaction, human factors, social factors/psychology, cognitive psychology and technological applications.

My research addresses the following fundamental limitations now found in the co-located setting:

1. *Traditional desktop computers are unsatisfying for highly collaborative situations involving multiple co-located people exploring and problem-solving over rich digital spatial information.*
2. *While existing research toolkits provide support for multiple people, current systems fail to support the rich hand gestures and speech acts that people do in everyday collaborative situations. While many toolkits are designed to replicate existing Windows, Icons and Pointer metaphors seen in the desktop environment ethnographic studies of collaborative settings shows a very different type of interaction.*

UIST'06, October 15–18, 2006, Montreux, Switzerland.

[1] This work was done in collaboration with Mitsubishi Electric Research Laboratories in Cambridge, Massachusetts

3. *Ethnographic studies illustrate how the 'single user' assumptions inherent in current large display input devices limit collaborators who are accustomed to using multiple fingers and two-handed gestures often in concert with speech.*

BACKGROUND

People naturally use speech and gestures in their everyday communications over artefacts. Consequently, researchers are now becoming interested in exploiting speech and gestures in computer supported cooperative work systems. In this section, I provide a brief background to some of the ethnographic studies, mostly drawn from observations of safety critical environments, which form the motivations and foundations for my work in multimodal co-located interaction. Next, I extract implications for design for multimodal co-located system development. Finally, I will explore technological explorations of multimodal and co-located systems.

Ethnographic and Empirical Studies

Ethnographic studies of mission critical environments such as military command posts, air traffic control centers and hospital emergency rooms have shown that paper media such as maps and flight strips are preferred even when digital counterparts are available [Cohen, 2002, Cohen 1997, Chin, 2003, Hutchins, 2000]. For example, Cohen et. al.'s ethnographic studies illustrate why paper maps on a tabletop were preferred over electronic displays by Brigadier Generals in military command and control situations [Cohen, 2002]. The 'single user' assumptions inherent in the electronic display's input device and its software limited commanders, as they were accustomed to using multiple fingers and two-handed gestures to mark (or pin) points and areas of interest with their fingers and hands, often in concert with speech [Cohen, 2002, McGee, 2001]. Several ethnographic researchers have focused on how gesture and speech provide improved awareness to group members in a co-located environment.

These empirical and ethnographic studies provide motivation for multimodal support in co-located environments and have consequently led to specific design implications.

Implications to Design

This section begins with the low level mechanics of gesture and speech input and then moves into high level theories influencing group work.

Deixis: Speech Refined by Gestures: Deictic references are speech terms ('this', 'that', etc.) whose meanings are disambiguated by spatial gestures (e.g., pointing to a location). A typical deictic utterance is "Put that…" (points to item) "there…" (points to location) [Bolt, 1980]. Deixis often makes communication more efficient since complex locations and object descriptions can be replaced in speech by a simple gesture. For example, contrast the ease of understanding a person pointing to this sentence while saying 'this sentence here' to the utterance 'the 5th sentence in the paragraph starting with the word deixis located in the middle of page 3'. Furthermore, when speech and gestures are used as multimodal input to a computer, Bolt states [1980] and Oviatt confirms [1997] that such input provides individuals with a briefer, syntactically simpler and more fluent means of input than speech alone.

Complementary Modes: Speech and gestures are strikingly distinct in the information each transmits. For example, studies show that speech is less useful for describing locations and objects that are perceptually accessible to the user, with other modes such as pointing and gesturing being far more appropriate [Bolt, 1980, Oviatt, 1999]. Similarly, speech is more useful than gestures for specifying abstract or discrete actions (e.g., Fly to Boston).

Simplicity, Efficiency, and Errors: Empirical studies of speech/gestures *vs.* speech-only interaction by individuals performing map-based tasks showed that parallel speech/gestural input yields a higher likelihood of correct interpretation than recognition based on a single input mode [Oviatt, 1999]. This includes more efficient use of speech (23% fewer spoken words), 35% less disfluencies (content self corrections, false starts, verbatim repetitions, spoken pauses, etc.), 36% fewer task performance errors, and 10% faster task performance [Oviatt, 1997].

Speech, Gesture and Cognition: McNeil explains how cognitive science proves that gesture and speech originate from the same cognitive system in the human mind and that there are various different types of gestures: deictic, iconic, cohesive, beat and metaphoric [McNeil, 1992]. This shows how the deictic pointing gestures of current point and click interfaces encompass a very small portion of the gestures that people use in everyday conversation. Consequently, a system needs to understand how rich gestures are used in accordance with speech so that the gesture type can be determined.

Consequential Communication: Gutwin describes how speech and gestural acts provide awareness to group members through consequential communication [Gutwin, 2004, Segal, 1994]. For example, researchers have noticed that people will often verbalize their current actions aloud (e.g., "I am moving this box") for a variety of reasons [Hutchins, 1997, Heath, 1991, Segal, 1994]:

- to make others aware of actions that may be missed,
- to forewarn others about the action they are about to take,
- to serve as an implicit request for assistance,
- to allow others to coordinate their actions,
- to reveal the course of reasoning,
- to contribute to a history of the decision making process.

Distributed Cognition: While much of human computer interaction is focused on understanding cognition and factors within an individual, both Clark and Hollan emphasize a need to understand distributed cognition in a team setting [Hollan, 2000]. This means that researchers should consider the whole team as a cognitive system and

use their communicative acts to understand patterns of information flow within the system [Hutchins, 2000].

RESEARCH OBJECTIVES

My thesis consists of four main research objectives:

1. Distill existing theories and ethnographic studies into a set of behavioural foundations that inform the design of multimodal co-located systems and list individual and group benefits: I have performed a survey of existing theories of team work and ethnographic research in safety critical environments. I examined interaction in real world situations, paying particular attention to the speech and gesture acts used to produce a list of individual and group benefits of multimodal co-located interaction. This summary will outline some of the benefits and provide motivation for adding multimodal interaction to co-located environments. This forms the basis of the design of multimodal co-located applications in this thesis and will be used in the evaluation of multimodal co-located systems.

2. Develop a toolkit to allow rapid prototyping of multimodal co-located interactive systems: Using some of the experience gained from my Master's Thesis [Tse, 2005] on building toolkits to support application development using multiple mice and keyboards, I will develop a software toolkit that will facilitate the rapid prototyping of responsive and demonstrative multimodal gesture and speech applications in a co-located environment. This toolkit will be designed to support gestures on existing table top input devices (e.g., Diamond Touch, Smart DViT), however, there may be plans to extend this infrastructure to support other multimodal input devices at a later time. To evaluate the toolkit I will build the applications described in Objectives 3 and 4, and I will get others to develop multimodal co-located systems using my toolkit.

3. Develop and evaluate multimodal co-located wrappers over existing commercial applications to further my understanding and inform the design of true multi-user multimodal interactive systems: Using the design implications and behavioural foundations and the prototyping toolkit I will develop several multi user, multimodal co-located interface wrappers atop of existing commercial applications on a interactive table top display. By studying existing commercial applications I will be able to rapidly prototype rich multimodal applications that would otherwise be impossible for me to develop from the ground up. User studies of these systems will also provide an opportunity to observe how people naturally mitigate interference and turn taking when interacting with single user applications over a multimodal co-located table top display. They will also be used to evaluate how the design implications work when moved out of the physical world in the realm of a digital tabletop. All of these observations will be used to inform the design of a true multi user multimodal system.

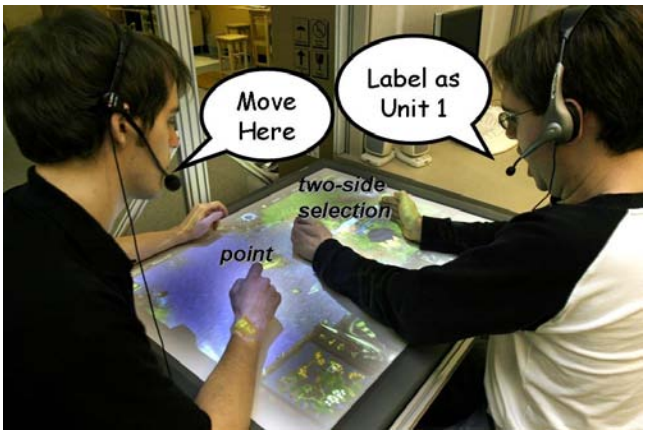

Figure 1. Two people interacting with speech and gesture over Blizzard's Warcraft III.

4. Develop true multi user multimodal co-located systems and evaluate the technical and behavioural nuances of multimodal co-located systems development: Again, using the toolkit developed in Objective two, I will create several applications that explore new interaction possibilities available exclusively in a multi user multimodal co-located environment. I will get groups of people to perform collaborative tasks in this environment paying particular attention to the inter-person behaviours and technical nuances of multimodal co-located application development and possibly develop new techniques to mitigate these issues in the co-located environment.

CURRENT STATUS

Two papers have been published describing work that I have done towards the completion of Objectives 1, 2 and 3. The AVI paper [2006] describes the behavioural foundations of my research along with a set of design implications for designers. This technique was applied in the design of multi user multimodal wrappers over existing single user applications such as Google Earth, and Warcraft III (Figure 1). The Pervasive Games paper [2006] describes multimodal tabletop interaction specifically within the context of home gaming. Digital tables provide a balance between the rich gestural affordances provided by manipulating physical objects (e.g., a racing wheel) and the versatility of playing many different games seen in home console gaming. The rich hand gestures (e.g., using five fingers to pick up a digital hot tub) and face to face affordances of digital tables provides a gaming experience beyond the affordances of a simple game controller.

Currently I am working on a system to allow speech and gesture to be trained over any existing single user application by demonstration. For example, saying "computer, when I do [one finger drag on table] you do [left mouse drag over same area]" would teach the computer that a single finger should be mapped to a left mouse drag. Similarly saying "Computer, when I say [fly to Boston] you do [keyboard and mouse macro]" should allow a sequence of keyboard and mouse commands to be mapped to a single voice utterance. Multiple users could interact using built in floor control policies (e.g., last

gesture and speech wins). This would allow end users to focus on the design of their gesture and speech wrappers rather than dealing with the complexity of adding speech and gesture recognizers to existing applications.

My future work consists of two steps: first, I will evaluate how multiple people interact over existing single user applications using speech and gestures through a series of informal observational studies. During these pilot studies I will look closely at the issues that arise when multi user speech and gesture interaction is provided. The lessons learned from the observational studies will be used to filter out one or two fundamental issues that will be studied in the development of true multi user multimodal applications built from the ground up (Research Objective 4).

The empirical results obtained from studies of true multi user multimodal applications will be used to further refine the behavioural foundations of Research Objective 1. This refinement will detail how the behavioural foundations were put into practice in a true multi user multimodal application and describe some of the benefits and shortcomings of our design.

CONCLUSION

This thesis argues that single point touch interaction on a large display reduces the expressive capabilities of people's hands and arms to simple deictic pointers. Richer multimodal interaction that is aware of the hand postures, movements and speech acts that people naturally perform in a co-located environment will not only provide improved group awareness, but it will improve the accuracy and effectiveness of the collaborations in co-located environments.

The related work has shown that there is a wealth of ethnographic, theoretical and technical research that has investigated and argued for the benefits of multimodal interaction in a co-located environment. This proposal has identified a largely unexploited area in Human-Computer Interaction: multimodal co-located collaboration. The research I propose in this document aims to ground the individual and group benefits of multimodal interaction in the co-located setting. The contributions offered by this research are: an improved understanding of the benefits and tradeoffs of multimodal input in a co-located setting, a toolkit that allows rapid prototyping of multimodal interactive systems, the exploration of multi-user multimodal co-located wrappers around existing single user applications, and the design of evaluation of true multimodal co-located systems with the goal of understanding the nuances and design implications of effective multimodal co-located interaction.

REFERENCES

1. Bolt, R.A., Put-that-there: Voice and gesture at the graphics interface. *Proc ACM Conf. Computer Graphics and Interactive Techniques Seattle*, 1980, 262-270.
2. Clark, H. *Using language.* Cambridge Univ. Press, 1996.
3. Cohen, P.R., Coulston, R. and Krout, K., Multimodal interaction during multiparty dialogues: Initial results. Proc IEEE Int'l Conf. Multimodal Interfaces, 2002, 448-452.
4. Chin, T., Doctors Pull Plug on Paperless System, American Medical News, Feb 17, 2003, http://ama-assn.org/amednews/2003/02/17/bil20217.htm
5. Gutwin, C., and Greenberg, S. The importance of awareness for team cognition in distributed collaboration. In E. Salas, S. Fiore (Eds) *Team Cognition: Understanding the Factors that Drive Process and Performance,* APA Press, 2004, 177-201.
6. Heath, C.C. and Luff, P. Collaborative activity and technological design: Task coordination in London Underground control rooms. *Proc ECSCW*, 1991, 65-80
7. Hollan, J., Hutchins, E., & Kirsh, D. Distributed Cognition: Toward a New Foundation for Human Computer Interaction. Proceedings of ACM TOCHI Vol 7 No 2 Jun 2000 pp. 174-196
8. Hutchins, E., and Palen, L. Constructing Meaning from Space, Gesture, and Speech. *Discourse, tools, and reasoning: Essays on situated cognition*. Heidelberg, Germany: Springer-Verlag, 1997 Pp. 23-40.
9. Hutchins, E. (2000) The Cognitive Consequences of Patterns of Information Flow. Proc. Intellectica 2000/1, 30, pp. 53-74.
10. McGee, D.R. and Cohen, P.R., Creating tangible interfaces by augmenting physical objects with multimodal language. Proc ACM Conf. Intelligent User Interfaces, 2001, 113-119.
11. McNeill, D. 1992. Hand and Mind: What Gestures Reveal About Thought. University of Chicago Press, Chicago.
12. Oviatt, S. Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction* 12, 1997.
13. Oviatt, S. L. Ten myths of multimodal interaction, *Comm. ACM*, 42(11), 1999, 74-81.
14. Segal, L. Effects of checklist interface on non-verbal crew communications, NASA Ames Research Center, Contractor Report 177639. 1994
15. Tse, E. (2004) The Single Display Groupware Toolkit. MSc Thesis, Department of Computer Science, University of Calgary, Calgary, Alberta, Canada, November.
16. Tse, E., Greenberg, S., Shen, C. and Forlines, C. (2006) Multimodal Multiplayer Tabletop Gaming. PerGames '06, May 7th Dublin, Ireland, pp. 139-148.
17. Tse, E., Shen, C., Greenberg, S. and Forlines, C. (2006) Enabling Interaction with Single User Applications through Speech and Gestures on a Multi-User Tabletop. Proc. AVI '06, May 23-26, Venezia, Italy, pp. 336-343.