

Multiple approaches to evaluating multi-modal collaborative systems

Duncan Stevenson, Matthew Hutchins, Chris Gunn, Matt Adcock and Alexander Krumpholz
CSIRO ICT Centre
GPO Box 664, ACT, Australia
Duncan.Stevenson@csiro.au

ABSTRACT

We present recent work using multi-modal collaborative virtual environments for complex task training. We discuss the importance of multiple approaches to evaluation, including theoretically and empirically based approaches, and suggest how these approaches could transfer to collaborative human-computer situations.

Categories and Subject Descriptors

H5.3 [Group and Organization Interfaces]: Collaborative computing – *collaborative virtual environments, evaluation methodology.*

General Terms

Design, Human Factors

Keywords

Collaborative virtual environments, interaction evaluation

1. INTRODUCTION

This position paper explores some approaches to evaluating multi-modal collaborative systems for complex tasks. We look at recent work in collaborative human-to-human interaction mediated by a networked multi-modal virtual environment and at complex tasks performed by two people across this mediation. We envisage a time when the computer side of human-computer collaboration systems will be more of an active participant in tasks like these, and we draw some conclusions about the nature of the multi-modal dialogue interfaces that will be employed.

Widespread use of multi-modal interaction modes for human-computer dialogue will change the nature of human-computer collaboration. As the Institute for Human and Machine Cognition (IHMC) observes [1], much current human-computer interaction is command-based, with the computer responding to commands from the user. We expect that future collaborative human-computer interaction will have a strong focus on collaboratively

completing some task and that multi-modal interfaces between the human and the computer will enable the dialogue between them to happen in the domain of the task. For example, where the task calls for manipulation of objects then an appropriate interface is likely to use a combination of three-dimensional visual information, spatialised audio and haptic feedback.

Experience in evaluating collaborative activities in networked virtual environments is relevant to the question of evaluating multi-modal dialogue between humans and computers. We propose that a single approach to evaluation will rarely be sufficient for complex systems, and that multiple evaluations from different viewpoints will provide a more complete picture of the system that can be used to guide future improvements. In this paper we discuss our evaluation experiences using both theoretical and empirical strategies. The theoretical approach uses a design framework to analyse how the design of the system reflects the task requirements. Three empirical approaches are discussed, each looking at the system from a different point of view, and providing different types of information on the effectiveness of communication.

2. SURGICAL TRAINING CASE STUDIES

This paper draws on experience with two specific case studies which use collaborative virtual environments for surgical training. These studies are based on a scenario of an instructor and a trainee in different locations, where the instructor is teaching the surgical approach for a particular surgical task. The instructor and the trainee share a virtual space in which the three-dimensional anatomical models are present. Each person interacts via the virtual environment with the other person to progress the learning task and a simulation engine mimics the physical and logical behaviour of the objects in the virtual scene. Interactions are supported by a multi-modal interface using 2D and 3D visual data, voice and simulation audio and haptic interaction with the objects in the scene.

The collaborative environment on which the case studies are built is described by Gunn et al [2]. Each person sits at a “haptic workbench” which uses a mirrored monitor display and active stereo glasses to create a desktop-sized 3D virtual space under the mirror into which the person can reach with their haptic tools to perform pointing, gestural and haptic interactions, as shown in Figure 1. For collaborative use, a miniature broadband videoconferencing system within the workbench links the two participants.

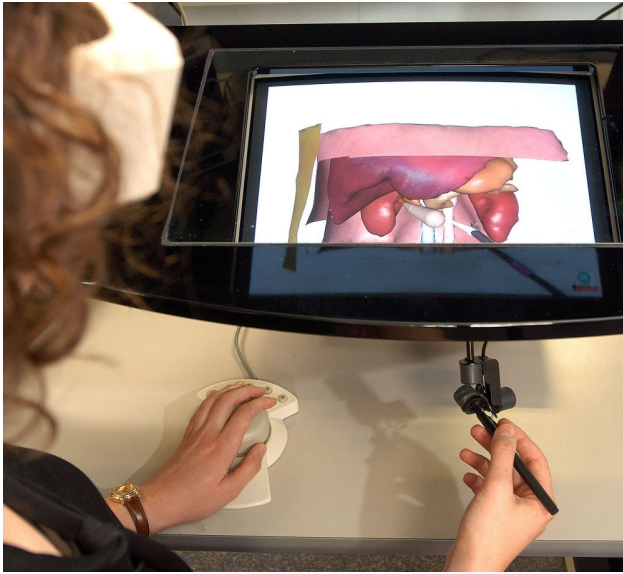


Figure 1: Multi-modal user interface for surgical training

The interaction between the instructor and trainee contains voice, gesture, 3D sketching, demonstrated surgical actions, hand-on-hand guidance and simultaneous shared actions (one grasps and retracts while the other cuts, for example). This is supported by interactive access to embedded ancillary information (patient scans, video clips and interactively drawn diagrams).

The surgical training tasks presented in these case studies require a high level of interactivity between instructor and student. In contrast to training systems for surgical dexterity skills only, these tasks focus on procedures based on reasoning about the underlying anatomy of the patient. Teaching is done by demonstration supported with spoken and gestural dialogue. The first training task, described by Gunn et al. [2] was built as a concept demonstrator and uses the removal of the gall-bladder as its content. The second training task, described by Hutchins et al. [3] is a much more complex activity involving surgical access to the organs of hearing in the temporal bone.

3. A THEORETICAL APPROACH TO EVALUATION BASED ON CHARACTERISING DESIGN DECISIONS

The design of interaction and interface components in a collaborative system has a strong influence on the effectiveness of communication within the system. Hutchins et al. [4] describe AFRADERVITE, a framework for describing and characterizing design decisions about representations in virtual environments for training. The framework allows decisions about the representation of objects, events, and actions from the training domain, and the participants in the environment, to be discussed in terms of several dimensions. The dimensions are: symmetry and heterogeneity; multiplicity and augmentation; decomposition and multimodality; role and significance; and directness. In summary, the way an object is represented in the environment should be determined by its role in the training domain. Objects that are central to the teaching methodology, such as models of anatomy in the case of surgical training, will need to be

represented with a relatively high fidelity, and may need to be represented in multiple ways, such as with 3D models, videos, photos and diagrams. Each type of representation provides different information, and supports different types of interaction and discussion. Objects that are peripheral to the discussion might be represented in a greatly simplified form, saving computational and attentional resources for the more important features.

To use a design framework as a tool in evaluation, we would revisit each of the representations in the system and characterize them against the dimensions of the framework. Then, each can be analysed to determine whether there is a good match for the task requirements, which may have changed since the original design through observations of the system in use. This would give a theoretical basis for discussing the likely effectiveness of the system, which would then be confirmed through empirical observations.

4. EMPIRICAL APPROACHES TO EVALUATING COLLABORATIVE TRAINING ENVIRONMENTS

4.1 Subjective evaluation through questionnaires

Where collaborative systems are to be used voluntarily, one of the important drivers of success will be how the users feel about their experience with the system. Did they enjoy it? Did it deliver what they were expecting? Were they able to communicate naturally and freely with the other participants? Would they recommend it to their peers? We have taken a questionnaire-based approach to answering these types of questions. In trials and demonstrations of both of the surgical training systems described above, we have asked groups of participants to fill out basic questionnaires to capture their responses to the use of the system (for an example, see [2]). These included five-level subjective evaluations, and some free-form text questions. Answers to questions like “what was the most valuable feature of the system?” and “what feature needs the most improvement?” have driven successive iterations in our design process.

4.2 Performance evaluation through transfer trials

In evidence-based medicine, the randomized, double-blind, controlled trial is the standard of evaluation that is required to show that an intervention is effective. This standard has been used for surgical training systems [5], although it requires significant resources. For a training system the trial must show that subjects perform better after the training being tested than if they had no training, or an alternative form of training. Where simulation is used, training must *transfer* from the simulation to the real world. For example, to demonstrate the effectiveness of the temporal bone surgery training system, we sought to show an improvement in subjects’ knowledge of anatomy and surgical planning through pre-training and post-training examinations (the results are to be published in a forthcoming paper).

The generalization of this concept is evaluation through total task performance. We would consider the whole system as being designed to facilitate some task, and demonstrate through

measured task performance (perhaps in terms of speed, quality, number of errors, etc.) that the system under evaluation allows the task to be performed as well or better than some alternative approach.

4.3 Dialogue analysis

Another approach to evaluating a collaborative system is to focus less on task performance, and more on the nature of the communication that users of the system adopt. Daly-Jones et al. [6] demonstrated that video conferencing systems could be studied in terms of features like fluency of the conversation. In [3] we used transcripts of videos taken during training trials of the temporal bone surgery training system to highlight qualitative features of the conversation that demonstrate how the use of a shared 3D model supports a particular style of dialogue. This form of analysis, described by Hutchby & Woofitt [7], is particularly valuable in showing how users take advantage of the multi-modal features of the system to enhance verbal communication with gestures and annotation, and how the conversation is grounded by the shared view of the model.

5. TRANSFER FROM VIRTUAL ENVIRONMENTS TO HUMAN-COMPUTER SYSTEMS

5.1 Transfer from the design framework approach

Our design approach focuses on producing representations in a virtual environment that match the requirements of the task being supported by the system. Correspondingly, the approach to evaluation looks at the designed representations and judges their appropriateness. If we are to consider human-computer communication systems that use perceptual representations, then similar reasoning must apply. That is, the number, type and fidelity of representations for objects, events and actions that appear in the environment must be chosen to support the task of the system, and also to support communication between the user and the computer. An additional representation problem arises in such a system, and that is how to represent the computer as a participant in the dialogue. For example, if the computer is to use gesture or pointing, there must be a natural representation of what is being pointed at.

5.2 Transfer from the subjective and objective measurement approaches

Subjective measures of user satisfaction and quantitative measurements of task performance are clearly directly applicable to human-computer dialogue systems. Some examples of metrics are time to completion, quality of outcomes, and error rates.

5.3 Transfer from the dialogue analysis approach

What we learn from dialogue analysis in collaborative systems is the degree to which recognisable patterns of conversation that occur within face-to-face communication also occur in the shared environment. This will be a useful feature to look out for in human-computer dialogue systems, if the goal is to have the computer mimic human conversation. For instance, the extent to which the human and computer can mutually ground their

conversation, manage turn taking and error correction, and easily refer to objects in the environment, could be established by examining transcripts of dialogues that have taken place, and this will provide valuable indicators as to the “naturalness” of the conversation.

6. CONCLUSION

Our position in this paper is that approaches to evaluating human-human collaborative systems may be generalised to study multi-modal human-computer dialogues. Multiple approaches to evaluation will provide a more complete picture of the effectiveness of the communication than any single technique will. Successful systems will be designed by careful analysis of the task being supported, and through a choice of representations that maximize the opportunity for natural conversational styles.

7. ACKNOWLEDGMENTS

The authors acknowledge the contributions of Peter Cosman and Leroy Heinrichs in developing the gall-bladder-removal case study and Stephen O’Leary and Brian Pyman in developing the temporal bone surgery training system.

This work was performed under the CeNTIE program, supported by the Australian Government through the Advanced Networks Program of the Department of Communications, Information Technology and the Arts, and through the CSIRO ICT Centre.

8. REFERENCES

- [1] Collaborative Human-Machine Interaction, Institute for Human and Machine Cognition, Florida, USA. Online at <http://www.ihmc.us/research/CollaborativeHumanMachineInteraction/>, Accessed 4th November 2005
- [2] Gunn, C., Stevenson, D., Krumm-Heller, A., Srivastava, S., Youngblood, P., Heinrichs, L. and Dev, P. An Interactive Master Class in Remote Surgery, *Proc. OZCHI 2004, Human Factors and Ergonomic Society of Australia, Wollongong, 21-24 November 2004*,
- [3] Hutchins, M., Stevenson, D., Gunn, C., Krumpholz, A., Adriaansen, T, Pyman, B. and O’Leary, S., Communication in a networked haptic virtual environment for temporal bone surgery training, *To appear in Virtual Reality Journal special issue on Haptic Interfaces and Applications*.
- [4] Hutchins, M., Adcock, A., Stevenson, D., Gunn, C. and Krumpholz, A.. The design of perceptual representations for practical networked multimodal virtual training environments. *Proc. HCI International 2005: 11th International Conference on Human-Computer Interaction, Las Vegas, July 22-27, 2005*.
- [5] Seymour N, Gallagher A, Roman S, O’Brien M, Bansai V, Andersen D and Satava R. (2002) Virtual Reality Training Improves Operating Room Performance. *Annals of Surgery, Vol 236 (4) pp 458-464*
- [6] Daly-Jones, O, Monk, A. and Watts, L. Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of attentional focus, *Int. J. Human-Computer Studies (1998), 49, pp21-58*
- [7] Hutchby I and Woofitt W. Conversation Analysis. Polity Press, Cambridge, UK, 1998