A Demonstration of Distributed Pointing and Referencing for Multimodal Collaboration Over Sketched Diagrams

Ed Kaiser[1,2] Paulo Barthelmess[1,2]

[1] Natural Interaction Systems. LLC .
10260 SW Greenburg Road
Portland, OR 97223, USA
+1 503 293 8414
(ed,paulo)@naturalinteraction.com

Xiao Huang[2]

[2] Oregon Health and Science University.
OGI School of Science & Eng .
20000 NW Walker Road
Beaverton, OR 97006, USA
+1 503 748 7803
huangx@cse.ogi.edu

David Demirdjian[3]

[3] MIT, Computer.Science. and Artificial.Intelligence Laboratory (CSAIL).
32 Vassar Street
Cambridge, MA. 02139, USA
+1 617 253 6218
demirdji@ai.mit.edu

ABSTRACT

Groups who are not co-located but still need to collaborate on a common task face the problem of reduced access to the rich multimodal communicative context that they would have if they were collaborating face-to-face. We present a demonstration system that allows participants at remote sites to collaborate in building a project schedule via sketching on multiple distributed whiteboards. We show how participants can be made aware of naturally occurring pointing gestures that reference diagram constituents as those gestures are performed by remote participants. Our system fuses multimodal inputs from pen, speech and 3D gestures, coupled with the dynamic construction of a semantic representation of the interaction, anchored on the sketched diagram, to provide feedback that overcomes some of the intrinsic ambiguities of distant pointing gestures. We also demonstrate how a distributed awareness of the meaning of shared symbols on the whiteboard, like handwritten abbreviations, can be both learned dynamically as the meeting progresses and communicated in unobtrusive ways to remote participants.

Categories and Subject Descriptors

H.5.3 [Group and Organization Interfaces]: Collaborative Computing; Synchronous interaction. H.5.2 [User Interfaces]: Natural language; Input devices and strategies. I.2.10 [Vision and Scene Understanding]: 3D/stereo scene analysis. I.2.6 [Learning]: Language Acquisition.

General Terms

Collaboration; Multimodal; Context-aware.

Keywords

Collaborative Interaction; Multimodal processing; Intelligent interfaces; Gesture; Vocabulary learning.

1. INTRODUCTION

Collaboration has traditionally been associated with participants' ability to orient their actions according to a rich multimodal communicative context. Collaborative communication technology allows participants to interact remotely. However, while providing valuable support for collaboration, such current tools do not take advantage of or sometimes even accommodate natural work practices.

We are interested in exploring applications capable of capturing and interpreting a wide range of group multimodal communicative acts by incorporating recognition of speech, sketches, gestures, and capabilities for modeling dialogues, detecting topic and phase shifts, and identifying user-action patterns. In the present work, we leverage the capabilities of the Charter Suite [1] – a multimodal intelligent system we have extended to support both co-located and distributed collaborative design of project schedules, as well as integrative understanding of multimodal input combinations like redundant speech and handwriting for labeling or referring to chart elements. The distributed version (depicted in Figure 1) promotes cross-site awareness of naturally occurring pointing gestures referencing elements of the shared sketched diagram.

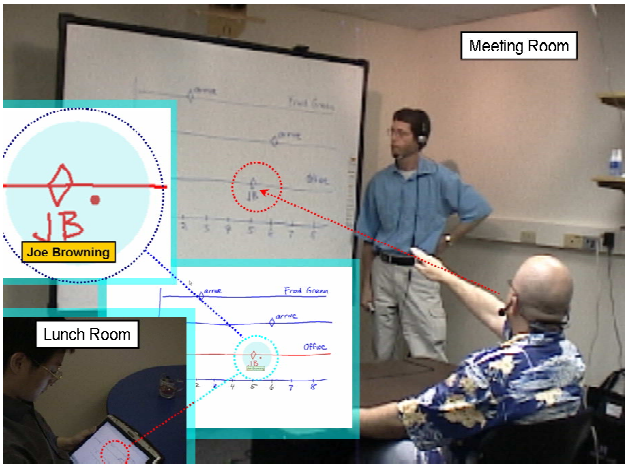

Figure 1. A distributed Three-person scheduling meeting, with speech, sketch, handwriting and 3D gesture recognition. A semantic rendering of the schedule chart, the pointing reference (small red and large blue dots), and the dynamically discovered abbreviation semantics (*JB = Joe Browning*) are available to the remote participant using a tablet PC.

Our demonstration of the Charter suite also highlights its ability to learn new words dynamically as the system is running, like the out-of-vocabulary task name, *Joe Browning*, attached to the middle taskline in Fig. 1. Once such a new term is learned it can be subsequently recognized in various individual input modes like speech recognition. From speech recognition its semantics can be transferred to new symbols in other modes, like the handwritten abbreviation, *JB*, shown in Fig. 1.

2. SYSTEM OVERVIEW

Our distributed Charter targets collaborative sessions in which multiple participants, usually stakeholders, get together to define a schedule of a project. Schedule design is usually a high-stake, highly collaborative activity. The outcome of these planning sessions is expected to shape future work in important ways, in many cases representing commitments of a group of people with

respect to their own individual actions and expectations with respect to the actions of others.

2.1 Multimodal Sketch-recognition: Charter

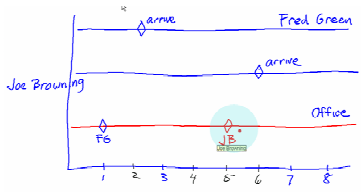

Figure 2: Gantt Chart in Charter.

Charter adopts Gantt Charts as the formalism used to represent project schedules. A Gantt chart is a diagrammatic, two-dimensional representation of a planned series of tasks and important milestones, and incorporates in a compact and graphical form constraints related to time, tasks, budget and other dependencies (Figure 2). Lines represent tasks; milestones are represented as diamonds. Temporal information is associated with the placement of milestones and task lines within a temporal grid defined as an *xy*-axis. Task durations are represented by the length of the lines according to this same temporal framework.

Charter is able to distinguish individual elements of a sketch, including their names given by labels, due to the online parsing and analysis that associates pen strokes combined with speech to items of a semantic representation. Based on this semantic representation, Charter is able to determine the individual diagram constituents that are most likely to be the targets given a gestural region of focus. This is made visible though coloring of the sketch constituents according to the likelihood that individual elements are the targets of a pointing gesture (Figure 1 and 2).

2.2 Vision-based Body-Tracking

In order to model human bodies, we use a 3D cylindrical model of articulated appearance: limbs (head, torso, arms, forearms) are modeled as rigid bodies connected by spherical joints. We have designed an algorithm for estimating articulated motion based on rigid motion estimates of the articulated models constituent parts [2]. In brief, the tracking algorithm maps a 3D cylindrical model to the tri-dimensional reconstruction of the scene provided by the stereo camera. More exactly, the well-known ICP algorithm [3] is used to coarsely align two clouds of 3D points and estimate an initial rigid motion between body parts. The ICP algorithm iteratively estimates the rigid transformation between two clouds of 3D points. First, a set of elementary displacements is estimated by finding, for each point in the first cloud (model) the closest point in the second cloud (scene). Then the global rigid motion is computed by integrating all the elementary displacements and applied to the first cloud. The process is repeated until convergence is achieved.

2.3 Speech/Writing Recognition: SHACER

In our previous Multimodal New Vocabulary Recognition technique (MNVR) [4] we allowed new words to occur only within grammar-defined *carrier phrases* – a predefined, constrained set of standardized phrases. Recently we have generalized our approach, so the system no longer requires the use of *carrier phrases*. We refer to our new system as Multimodal Out-Of-Vocabulary Recognition (MOOVR) using a Speech and HAndwriting reCognizER (SHACER). MOOVR/SHACER keeps the basic MNVR architecture of grammar-based language modeling implemented as a Recursive Transition Network within a continuous speech recognizer – Carnegie Mellon University's (CMUs) Sphinx 2 recognizer [5]. However, instead of having only two grammars we now use an ensemble of four syllable/phone grammars plus a new Word/Phrase-Spotting Recognizer (WPSR) grammar. New terms are enrolled into the WPSR and then used to both improve subsequent recognition and serve as foundation for Mutual Semantic Acquistion (MSA). We also employ a large vocabulary continuous speech recognizer, CMU's Sphinx 3.5 engine, in an implementation called Speechalyzer.

MSA occurs when a user labels a chart constituent with a new handwritten abbreviation for an already enrolled term, while at the same time speaking that term. WPSR recognizes the term and SHACER performs various alignments across the speech phones, letter-to-sound generated phones from the handwriting, and any transcript or lattice available from Speechalyzer. In this way the association between an enrolled spoken term (with known spelling/semantics/pronunciation) and an unknown abbreviation can be made. The popup below the blue gestural focus area on the remote tablet's charter display (Fig. 1) alerts the remote user to the fact that the abbreviation, *JB*, refers to *Joe Browning* – the label of the chart's middle taskline (Fig. 2) that was enrolled by SHACER into the system dictionaries only minutes before it is used to ground and distribute the meaning of this abbreviation.

3. CONCLUSION

The Charter Suite demonstrates a policy of minimal intervention in support of unencumbered work practices. The system builds a semantic interpretation of an interaction by observing a multi-user multimodal dialogue. This detailed interpretation built by observation, recognition and fusion of speech, sketch, handwriting and 3D gesture is the basis for the system's services.

4. ACKNOWLEDGMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or the Department of Interior National Business Center (DOINBC).

5. REFERENCES

[1] Kaiser, E., D. Demirdjian, A. Gruenstein, X. Li, J. Niekrasz, M. Wesson, and S. Kumar. *Demo: A Multimodal Learning Interface for Sketch, Speak and Point Creation of a Schedule Chart*. in *International Conference on Multimodal Interfaces (ICMI '04)*. 2004. State College, PA.

[2] Demirdjian, D., T. Ko, and T. Darrell. *Constraining Human Body Tracking*. in *Proceedings of the International Conference on Computer Vision*. 2003. Nice, France.

[3] Besl, P. and N. MacKay, *A method for registration of 3-d shapes*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992. **14**: p. 239-256.

[4] Kaiser, E.C. *Multimodal New Vocabulary Recognition through Speech and Handwriting in a Whiteboard Scheduling Application*. in *Proceedings of the International Conference on Intelligent User Interfaces*. 2005. San Diego, CA.

[5] Singh, R., *The Sphinx Speech Recognition Systems*, in *Encyclopaedia of Human Computer Interaction*, W. Bainbridge, Editor. 2004, Berkshire Publishing Group.