Demo: A Multimodal Learning Interface for Sketch, Speak and Point Creation of a Schedule Chart

Ed Kaiser[1] David Demirdjian[2] Alexander Gruenstein[3] Xiaoguang Li[1] John Niekrasz[3] Matt Wesson[1] Sanjeev Kumar[1]

[1]Oregon Health and Science University
OGI School of Science & Eng.
20000 NW Walker Road
Beaverton, OR 97006, USA
+1 503 748 7803
{kaiser,xiaoli,wesson,skumar}@cse.ogi.edu

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

[3] Stanford University, Center for the Study of
Language and Information (CSLI)
220 Panama Street
Stanford, CA. 94305, USA
1 650 723 3084
{alexgru,niekrasz}@csli.stanford.edu

ABSTRACT

We present a video demonstration of an agent-based test bed application for ongoing research into multi-user, multimodal, computer-assisted meetings. The system tracks a two person scheduling meeting: one person standing at a touch sensitive whiteboard creating a Gantt chart, while another person looks on in view of a calibrated stereo camera. The stereo camera performs real-time, untethered, vision-based tracking of the onlooker's head, torso and limb movements, which in turn are routed to a 3D-gesture recognition agent. Using speech, 3D deictic gesture and 2D object de-referencing the system is able to track the onlooker's suggestion to move a specific milestone. The system also has a speech recognition agent capable of recognizing out-of-vocabulary (OOV) words as phonetic sequences. Thus when a user at the whiteboard speaks an OOV label name for a chart constituent while also writing it, the OOV speech is combined with letter sequences hypothesized by the handwriting recognizer to yield an orthography, pronunciation and semantics for the new label. These are then learned dynamically by the system and become immediately available for future recognition.

Categories and Subject Descriptors

I.2.11 (Distributed Artificial Intelligence): Multiagent systems; H.5.2 (User Interfaces): Graphical user interfaces, natural language, voice I/O; I.2.6 (Learning): Language Acquisition.

General Terms

Design, Experimentation, Human Factors.

Keywords

Multimodal interaction, vision-based body-tracking, vocabulary learning.

1. INTRODUCTION

The agent architecture of our system significantly extends the multimodal QuickSet and MAVEN architectures [1] that have preceded it. Tracked users are now untethered, and the system now has significant capabilities for dynamic learning of new vocabulary, contextualized by handwriting, speech, and sketch.

2. SYSTEM OVERVIEW

Our test bed scenario posits two meeting participants at work before a whiteboard (Fig. 1) creating a project schedule for the creation and delivery of a demonstration system. They first lay out a schedule grid (using sketch-based recognition informed by an ontologically derived spoken interface — Sect. 2.1, 2.2), then sketch in task-lines and label them (using speech and handwriting for dynamic enrollment of new vocabulary — Sect. 2.3), and finally they designate milestones. The system recognizes the onlooker's speech and 3D gesture (using untethered, real-time, vision-based body-tracking — Sect. 2.4), tracking his gesture-aided suggestion to move a milestone over (Fig. 3). The system also synchronously creates an MS Project schedule (Fig. 2, lower right), in real-time via a Python AAA agent capable of controlling applications that expose a COM interface.

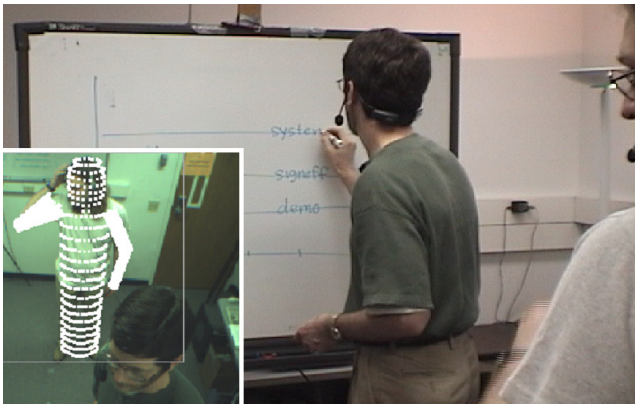

Figure 1. Two-person GANTT chart scheduling meeting, with tracking.

2.1 Multimodal Sketch-recognition: Charter

Pen gestures on the whiteboard are captured by Charter, an AAA based multimodal user interface (Fig. 2), and then routed to 2D sketch and handwriting recognition agents[1]. Charter extends the basic parsing architecture of temporally constrained multimodal integration by maintaining a single chart feature-structure representation that persists over time, growing and shrinking as necessary. Charter is also the display mechanism for the beautified Gantt chart produced by multimodal integration.

2.2 Understanding Conversational Speech

The conversational understanding capabilities of the system build on and extend CSLI's Conversational Intelligence Architecture (CIA) [4]. The dialogue manager (DM) — a generic dialogue management toolkit employed in applications spanning a variety of domains — is used to process logical forms produced by Gemini [3], which robustly parses the list of results from speech

ICMI'04, October 13–15, 2004, State College, Pennsylvania, USA.
ACM 1-58113-890-3/04/0010.

[1] From Natural Interaction Systems: http://www.naturalinteraction.com

recognition. In this multimodal setting, the toolkit was enhanced to produce multiple hypothesized interpretations of an utterance in context (expressed as knowledge-base updates), which are passed to the multimodal integrator, facilitating mutual disambiguation of input modes [1]. Figure 2 shows several hypotheses produced by a single utterance (upper right). An utterance may attach to multiple nodes on the tree of dialogue moves [4], which is pruned once one of the hypotheses is confirmed — as well as potentially augmented — by integration with other modalities (middle right).

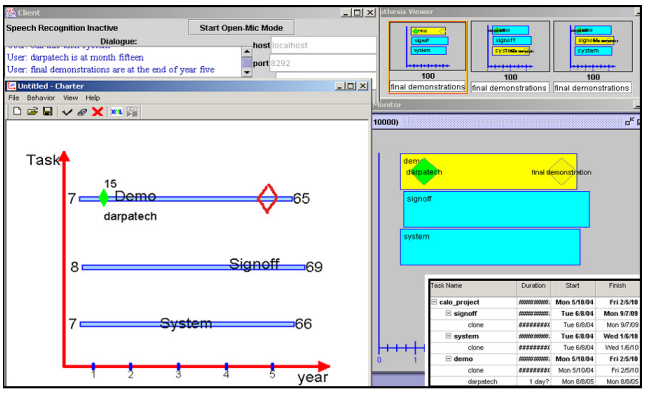

Figure 2. Charter sketch input/display (left), hypotheses (top left & right) from speech processing, confirmation (right), MS Project (lower right).

The toolkit also features a conversational history browser that shows both the meeting's transcription and the history of the chart, thus visualizing the causal relationship between multimodal actions and chart modifications. Search capabilities make it possible to answer queries such as: "Who moved the March deadline to May?"

2.3 Dynamic New Vocabulary Enrollment

We have augmented CMU's Sphinx2 speech recognizer both to function as an agent in our AAA architecture and to optionally use an embedded Recursive Transition Network (RTN) grammar. The grammar writer can specify the contextual location of licensed out-of-vocabulary (OOV) word occurrences. At run-time, when those spoken contexts arise, OOV words are recognized as sequences of phones constrained by the use of a syllabic sub-grammar. These phone sequences are then mapped to orthographies using a sound-to-letter module [5]. If semantically interpretable handwriting occurs co-temporally then the letter string hypotheses from the handwriting recognizer (along with their associated letter-to-sound strings) are paired with the OOV-based orthographies using an edit distance measure. The highest scoring orthography, pronunciation and semantic-hypothesis tuple is then dynamically enrolled in the system becoming immediately available for future utterance recognition at author-specified points in the grammar.

2.4 Vision-based Body-Tracking

The location and pose of the persons in the meeting is estimated using a vision system, which uses a stereo camera calibrated with respect to the other devices in the scene (e.g. whiteboard). The stereo camera provides input for a real-time, untethered vision-based tracking algorithm that estimates the body pose of the user in the field of view (Fig. 3).

The vision-based tracking technique is described in [6]. In brief, the approach consists of fitting a 3D articulated CAD model to the 3D reconstruction of the scene provided by the stereo camera. The 3D CAD model only consists of the upper body parts (head, torso, arms, forearms), Fig. 3.

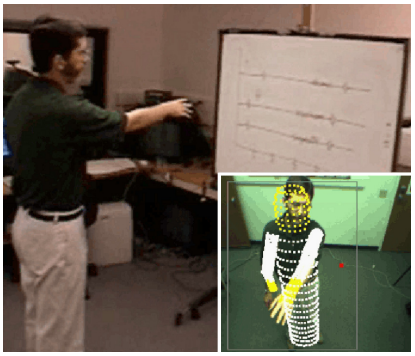

Fig. 3. Vision-based body-tracking in support of 3D deictic gesture recognition.

More precisely, a fitting error function is defined as the Euclidean distance between the 3D CAD model and scene reconstruction (this framework allows the use of flesh color models for tracking head and hands). For further robustness, the 3D CAD model is constrained to only certain configurations (e.g., person standing and facing approximately towards the white board). The fitting error function and the model constraints are put into a Quadratic Programming problem that is solved in an efficient way, thereby allowing real-time tracking.

To determine what the user is pointing at, the object state of the system is maintained in a Feature Structure Database (FSDB), which can resolve queries across types very quickly. When the tracked user says, "Move that milestone to the end of year one," and co-temporally the system recognizes a deictic 3D pointing gesture, then a proximity query on the point of intersection with the whiteboard yields a ranked list of possible object references. A cross product of that object list and ranked speech interpretations yields an object de-referenced command that is tracked and implemented by the system.

3. ACKNOWLEDGMENTS

This work was funded by DARPA grant NBCH-D-03-0010(1).

4. REFERENCES

[1] E. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen, and S. Feiner, "Mutual Disambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality," ICMI '03, 12-19.

[2] P.R. Cohen, M. Johnston, D.R. McGee, S.L. Oviatt, J.A. Pittman, I. Smith, L. Chen, and J. Clow, "QuickSet: Multimodal Interaction for Distributed Applications," Intl. Multimedia Conference, '97, 31-40.

[3] J. Dowding, J.M. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore and D. Moran, "Gemini: A Natural Language System for Spoken-Language Understanding," ACL '93, 54-61.

[4] O. Lemon, A. Gruenstein and S. Peters, "Collaborative Activities and Multi-tasking in Dialogue Systems," *Traitment Automatique desLangues,* 2002, 43(2), 131-154.

[5] A.W. Black, K.A. Lenzo, "Flite: a small fast run-time synthesis engine," 4th ISCA Workshop on Speech Synthesis, 2001.

[6] D. Demirdjian, T. Ko and T. Darrell, "Constraining Human Body Tracking," Proc. of Int'l Conf. on Computer Vision, Nice, France, 1071-1078, Oct. 2003.