From Conversational Tooltips to Grounded Discourse: Head Pose Tracking in Interactive Dialog Systems

Louis-Philippe Morency
Computer Science and Artificial Intelligence
Laboratory at MIT
Cambridge, MA 02139
USA
lmorency@csail.mit.edu

Trevor Darrell
Computer Science and Artificial Intelligence
Laboratory at MIT
Cambridge, MA 02139
USA
trevor@csail.mit.edu

ABSTRACT

Head pose and gesture offer several key conversational grounding cues and are used extensively in face-to-face interaction among people. While the machine interpretation of these cues has previously been limited to output modalities, recent advances in face-pose tracking allow for systems which are robust and accurate enough to sense natural grounding gestures. We present the design of a module that detects these cues and show examples of its integration in three different conversational agents with varying degrees of discourse model complexity. Using a scripted discourse model and off-the-shelf animation and speech-recognition components, we demonstrate the use of this module in a novel "conversational tooltip" task, where additional information is spontaneously provided by an animated character when users attend to various physical objects or characters in the environment. We further describe the integration of our module in two systems where animated and robotic characters interact with users based on rich discourse and semantic models.

Categories and Subject Descriptors

I.4.8 [**Scene Analysis**]: Image Processing and Computer Vision—*Tracking, Motion*; H.5.2 [**User Interfaces**]: Information Interfaces and Presentation

General Terms

Algorithms

Keywords

Head pose tracking, Head gesture recognition, Interactive dialog system, Grounding, Conversational tooltips, Human-computer interaction

ICMI'04, October 13–15, 2004, State College, Pennsylvania, USA.
Copyright 2004 ACM 1-58113-954-3/04/0010 ...$5.00.

1. INTRODUCTION

Multimodal interfaces have begun to become practical as multimedia sensor streams become prevalent in everyday interaction with machines. These new interfaces integrate information from different sources such as speech, eye gaze and body gestures. Head (as well as body) pose serves as a critical cue in most human-to-human conversational interaction; we use our face pose to signal conversational turn-taking intent, offer explicit and implicit acknowledgement, and refer to specific objects of interest in the environment. These cues ought to be equally if not more valuable in human–machine interaction.

Previous work has demonstrated the utility of generating agreement gestures or deictic references in the output modalities of animated interface agents [6, 24]. However, input processing has largely been limited to sensing face pose for basic agent turn-taking [26]; advanced interpretation has required offline processing. Until recently, the task of robustly and accurately sensing head pose using computer vision proved too challenging for perception of grounding cues in real time. Many face detectors and motion estimators are available, but detectors generally have not demonstrated sufficient accuracy, and motion analysis has often been too brittle for reliable, long-term use.

We have developed a face-processing system designed to serve as a conversational grounding module in a conversational dialog system. Our system is based on motion stereo methods, can automatically initialize to new users, and builds a user-specific model on the fly to perform stable tracking. Below, we detail the design and algorithmic choices which lead to our present tracking system, and the methods used to appropriately train recognizers to detect grounding gestures from the tracked pose data. We have developed our module in toolkit form so that it can be quickly integrated with existing interactive conversational systems.[1] We describe the use and evaluation of our module in three deployed systems for conversational interaction.

We first present the use of our module in a scripted, off-the-shelf animated conversational character. Using animation software from Haptek [10], speech synthesis from AT&T [1], and speech recognition from Nuance [20], we create a baseline animated character which offers information about a number of objects in the environment. Without perception of grounding cues, spoken commands are used to select a topic. With our module, we

[1] Our toolkit is available for download by interested parties at http://www.ai.mit.edu/projects/vip/watson/.

show how *conversational tooltips* can be provided, which spontaneously offer additional information about objects of apparent visual interest to a user. A quantitative user study showed that users were able to effectively use conversational tooltips to quickly select an object of interest.

We then describe the integration of our module with two interactive conversation agents based on natural language discourse models augmented with multimodal gesture representation. These systems have been used as interactive hosts for guiding visitors through a building or a set of technology exhibits. In use with both animated and robotic agents of this form, our system allowed users to successfully interact with passively sensed head pose grounding gestures.

2. PREVIOUS WORK

Many techniques have been proposed for tracking a user's head based on passive visual observation. To be useful for interactive environments, tracking performance must be accurate enough to localize a desired region, robust enough to ignore illumination and scene variation, and fast enough to serve as an interactive controller. Examples of 2-D approaches to face tracking include color-based [31], template-based [12] and eigenface-based [9] techniques.

Techniques using 3-D models have greater potential for accurate tracking but require knowledge of the shape of the face. Early work presumed simple shape models (e.g., planar [3], cylindrical [13], or ellipsoidal [2]). Tracking can also be performed with a 3-D face texture mesh [23] or 3-D face feature mesh [30].

Very accurate shape models are possible using the active appearance model methodology [7], such as was applied to 3-D head data in [4]. However, tracking 3-D active appearance models with monocular intensity images is currently a time-consuming process, and requires that the trained model be general enough to include the class of tracked users.

In contrast to these head-tracking systems, our system is robust to strong illumination changes, automatically initializes without user intervention, and can re-initialize automatically if tracking is lost (which is rare). In addition, it can track head pose under large rotations and does not suffer from drift.

Several systems have exploited head-pose cues or eye gaze cues in interactive and conversational systems. Stiefelhagen developed several successful systems for tracking face pose in meeting rooms and has shown that face pose is very useful for predicting turn-taking [27]. Takemae et al. also examined face pose in conversation, and showed that if face pose could be tracked accurately it was useful in creating a video summary of a meeting [28]. Siracusa et al. developed a kiosk front end that used head pose tracking to interpret who was talking to who in a conversational setting [26].

Justine Cassell [5, 6] and Candace Sidner [24, 25, 22], have developed rich models of multimodal output in the context of embodied natural language conversation, including multimodal representations of agreement and grounding gestures, but used mostly simple visual-based inputs like face detection. Here we extend our face-tracking system to enable the recognition of such gestures, describe a novel interaction paradigm based on face-responsive *tool tips*, and report on the results of using our recognition system within the embodied NLP frameworks of MACK, an embodied conversational agent, and Mel, a multimodal robot.

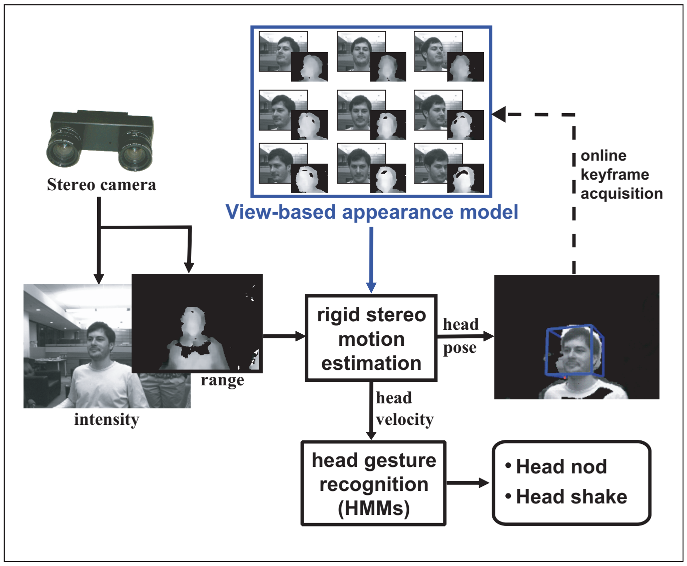

Figure 1: Visual grounding module.

3. A VISUAL GROUNDING MODULE

Our goal is to design a vision module that can detect important visual cues that occur when interacting with a multimodal conversational agent. Previous literature on turn-taking, grounding and engagement suggest that head gaze and gesture are important visual cues and can improve human-computer interaction. To integrate a vision module with an interactive dialog system, this module must meet certain requirements:

- Automatic initialization
- User independence
- Robustness to different environment (lighting, moving background, etc.)
- Sufficient sensitivity to recognize natural (subtle) gestures
- Real-time processing
- Stability over a long period of time

These requirements guide the development of our head-pose tracker and head-gesture recognizer. Figure 1 presents an overview of our visual grounding module.

3.1 Head-Pose Tracking

Our head-pose tracker takes advantage of depth information available from a stereo camera [8], which makes it less sensitive to lighting variations and a moving background and simplifies the segmentation process. Using a fast frontal face detector [29], we automatically initialize the tracking by merging the region of interest of the detected face with the segmentation from the depth image. After initialization, the estimates of head movement are used to update the region of interest. This simple segmentation technique makes it possible to track any user even if they have a beard or wear glasses.

Since important visual cues are often subtle (such as a head nod for acknowledgment), we decided to use a motion-based approach instead of simply using face detection at each frame. We compute the transformation between two frames using a hybrid error function which combines the robustness of ICP (Iterative Closest Point) and the precision of the normal flow constraint [15]. Our motion-based tracking algorithm can detect small movements quite accurately. Since it uses the real estimate of the shape of the object from the depth information, it can differentiate between translation and rotation accurately.

Human interactions with an embodied conversational agent are often prolonged so the tracking algorithm needs to be robust enough to not drift over time. To solve this problem yet still be user independent, we created a tracking framework that merges differential tracking with view-based tracking. In this framework, called the Adaptive View-Based Appearance Model [17], key frames are acquired online during tracking and used later to bound the drift. When the head pose trajectory crosses itself, the view-based model can track objects undergoing large motion for long periods of time with bounded drift.

An adaptive view-based appearance model consists of pose-annotated key frames acquired during tracking and a covariance matrix of all random variables representing each pose with a Gaussian distribution. Pose estimation of the new frame and pose adjustments of the view-based model are performed simultaneously using a Kalman filter tracking framework (see [17] for more details). The state vector of the normal differential Kalman filter tracker is extended to include the pose variables of the view-based model. The observation vector consists of pose-change measurements between the new frames and each relevant key frame (including the previous frame for differential tracking). Each pose-change measurement is then used to update all poses via the Kalman Filter update.

When merged with our stereo-based registration algorithm, the adaptive view-based model makes it possible to track the position, orientation and velocity of the head with good accuracy over a long period of time [17]. The position and orientation of the head can be used to estimate head gaze which is a good estimate of the person's attention. When compared with eye gaze, head gaze is more accurate when dealing with low resolution images and can be estimated over a larger range than eye gaze [16].

When compared with an inertial sensor (*Inertia Cube*^{2}), our head pose tracking system has a rotational RMS error smaller than the 3° accuracy of the inertial sensor[17]. We performed the comparison using video sequences recorded at 6 Hz with average length of 801 frames (∼133sec). During recording, subjects underwent rotations of about 125 degrees and translations of about 90cm, including translation along the Z axis. As described in the next subsection, the velocity information provided by the tracking system can be used to estimate head gestures such as head nods and shakes.

3.2 Head Gesture Recognition

Head gesture is often used in human conversation to communicate some feedback or emphasize an answer. Creating a visual module able to recognize head gesture during natural conversation is challenging, since most head gestures are fast, subtle movements. Using the output velocity of our head-pose tracker as input for our gesture detector, we can detect even subtle movements of the head. Since some gestures are performed at different speeds depending on the situation and the user, we decided to train our detector using Hidden Markov Models (HMMs).

To ensure that our training data was a good sampling of natural gestures, we acquired two data sets for positive examples. As a first data set, we used recorded sequences of 11 subjects interacting with a simple character displayed on the screen. In this case, the subjects were asked to answer each question with a head nod or a head shake. As a second data set, we used tracking results from 10 subjects interacting with a robot (Mel from MERL [14]). In this case, subjects were interacting naturally with the robot and performed many nonverbal gestures to acknowledge and ground information. The rotational velocity estimated by our head tracker was segmented manually to identify head nods and head shakes. Both data sets were used during training so that we could detect command-style gestures as well as natural gestures.

The head pose tracker returns the rotational and translational velocity at each frame. Since head nods and head shakes are performed by rotating the head, we used only the rotational component of the velocity for training. After analyzing the training set, we determined that most head nods and head shakes were performed in a time window between 1/2 and 1 second. Since the frame rate of the recorded sequences varied between 25-30Hz, we decided to use a window size of 30 frames for our training and testing. If the gesture duration was shorter then 1 second, then we zero-padded the sequence.

We trained two continuous Hidden Markov Models (extension of [11]) to recognize head nods and head shakes. The HMMs were trained using the Bayes Net Toolbox for Matlab[18]. During testing, we run each HMM independently and recognize the head gesture based on both likelihoods. The thresholds were learned experimentally during a pre-user study.

The complete visual grounding module described in this section can robustly estimate a user's head position and orientation as well as detect head nods and head shakes.

4. GROUNDING WITH SCRIPTED DIALOG: CONVERSATIONAL TOOLTIPS

Visual tooltips are an extension of the concept of mouse-based tooltips where the user's attention is estimated from the head-gaze estimate. We define visual tooltips as a three-step process: deictic gesture, tooltip and answer. During the first step, the system analyzes the user's gaze to determine if a specific object or region is under observation. Then the system informs the user about this object or region and offers to give more information. During the final step, if the user answers positively, the system gives more information about the object.

There are many applications for visual tooltips. Most museum exhibitions now have an audio guide to help visitors understand the different parts of the exhibition. These audio guides use proxy sensors to know where the visitor is or need input on a keypad to start the prerecorded information. Visual tooltips are a more intuitive interface.

To work properly, the system that offers visual tooltips needs to know where the user is focused and if the user wants more information. A natural way to estimate the user's focus is to look at the user's head orientation. If a user is interested in a specific object, he or she will usually move his or her head in the direction of that object [27]. Another interesting observation is that people often nod or shake their head when answering a question. To test this hypothesis, we designed a multimodal experiment that accepts

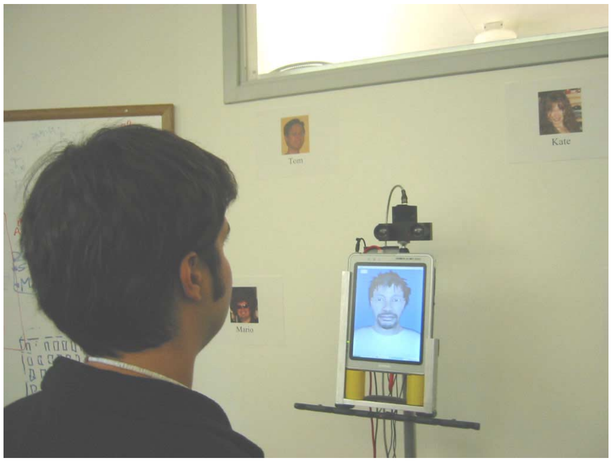

Figure 2: Multimodal kiosk built to experiment with *Conversational tooltip*. A stereo camera is mounted on top of the avatar to track the head position and recognize head gestures. When the subject look at a picture, the avatar offers to give more information about the picture. The subject can accept, decline or ignore the offer for extra information

speech as well as vision input from the user. The following section describes the experimental setup and our analysis of the results.

4.1 Experimental Setup

We designed this experiment with three tasks in mind: exploring the idea of visual tooltips, observing the relationship between head gestures and speech, and testing our head-tracking system. We built a multimodal kiosk that could provide information about some graduate students in our research group (see Figure 2). The kiosk consisted of a Tablet PC surrounded by pictures of the group members. A stereo camera [8] and a microphone array were attached to the Tablet PC.

The central software component of our kiosk consists of a simple event-based dialogue manager that gets input from the vision toolbox (Section 3) and the speech recognition tools [20] and can produce output via the text-to-speech routines [1] and the avatar [10].

When the user approaches the kiosk, the head tracker starts sending pose information and head nod detection results to the dialogue manager. The avatar then recites a short greeting message that informs the user of the pictures surrounding the kiosk and asks the user to say a name or look at a specific picture for more information. After the welcome message, the kiosk switches to listening mode (the passive interface) and waits for one of two events: the user saying the name of one of the members or the user looking at one of the pictures for more than n milliseconds. When the vocal command is used, the kiosk automatically gives more information about the targeted member. If the user looks at a picture, the kiosk provides a short description and offers to give more information. In this case, the user can answer using voice (yes, no) or a gesture (head nods and head shakes). If the answer is positive, the kiosk describes the picture, otherwise the kiosk returns to listening mode.

For our user study, we asked 10 people (between 24 and 30 years old) to interact with the kiosk. Their goal was to collect information about each member. They were informed about both ways to interact: voice (name tags and yes/no) and gesture (head gaze and head nods). There were no constraints on the way the user should interact with the kiosk.

4.2 Results

10 people participated in our user study. The average duration of each interaction was approximately 3 minutes. At the end of each interaction, the participant was asked some subjective questions about the kiosk and the different types of interaction (voice and gesture).

A log of the events from each interaction allowed us to perform a quantitative evaluation of the type of interaction preferred. The avatar gave a total of 48 explanations during the 10 interactions. Of these 48 explanations, 16 were initiated with voice commands and 32 were initiated with conversational tooltips (the user looked at a picture). During the interactions, the avatar offered 61 tooltips, of which 32 were accepted, 6 refused and 23 ignored. Of the 32 accepted tooltips, 16 were accepted with a head nod and 16 with a verbal response. Our results suggest that head gesture and pose can be useful cues when interacting with a kiosk.

The comments recorded after each interaction show a general appreciation of the conversational tooltips. Eight of the ten participants said they prefer the tooltips compared to the voice commands. One of the participants who preferred the voice commands suggested an on-demand tooltip version where the user asked for more information and the head gaze is used to determine the current object observed. Two participants suggested that the kiosk should merge the information coming from the audio (the yes/no answer) with the video (the head nods and head shakes).

5. INTEGRATION WITH DISCOURSE MODELS

Our head-tracking module has been successfully integrated with two different discourse models: MACK, an embodied conversational agent (ECA) designed to study verbal and nonverbal signals for face-to-face grounding, and Mel, a robot that can collaborate with a person in hosting an activity. Both projects integrate multimodal input signals: speech recognition, head-pose tracking and head-gesture recognition.

5.1 Face-to-Face Grounding

MACK (Media lab Autonomous Conversational Kiosk) is an embodied conversational agent (ECA) that relies on both verbal and nonverbal signals to establish common ground in computer–human interactions [19]. Using a map placed in front of the kiosk and an overhead projector, MACK can give directions to different research projects of the MIT Media Lab. Figure 3 shows a user interacting with MACK.

The MACK system tokenizes input signals into utterance units (UU) [21] corresponding to single intonational phrases. After each UU, the dialog manager decides the next action based on the log of verbal and nonverbal events. The dialogue manager's main challenge is to determine if the agent's last UU is grounded (the information was understood by the listener) or is still ungrounded (a sign of miscommunication).

As described in [19], a grounding model has been developed based on the verbal and nonverbal signals happening during human–human interactions. The two main nonverbal patterns observed in the grounding model are gaze and head nods. In the final ver-

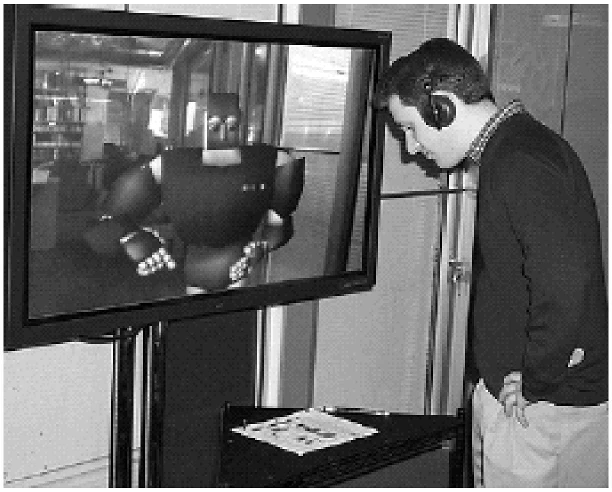

Figure 3: MACK was designed to study face-to-face grounding [19]. Directions are given by the avatar using a common map placed on the table which is highlighted using an over-head projector. The head pose tracker is used to determine if the subject is looking at the common map.

sion of MACK, our head-tracking module was used to estimate the gaze of the user and detect head nods. Nonverbal patterns are used by MACK to decide whether to proceed to the next step(UU) or elaborate on the current step. Positive evidence of grounding is recognized by MACK if the user looks at the map or nods his or her head. In this case, the agent goes ahead with the next step 70% of the time. Negative evidence of grounding is recognized if the user looks continuously at the agent. In this case, MACK will elaborate on the current step 73% of the time. These percentages are based on the analysis of human–human interactions.

5.2 Human–Robot Engagement

Mel is a robot developed at Mitsubishi Electric Research Labs (MERL) that mimics human conversational gaze behavior in collaborative conversation [24]. One important goal of this project is to study engagement during conversation. The robot performs a demonstration of an invention created at MERL in collaboration with the user (see Figure 4).

Mel's conversation model, based on COLLAGEN [22], determines the next move on the agenda using a predefined set of engagement rules, originally based on human–human interaction [25]. The conversation model also assesses engagement information about the human conversational partner from the Sensor Fusion Module, which keeps track of verbal (speech recognition) and nonverbal cues (multiview face detection[29]).

A recent experiment using the Mel system suggested that users respond to changes in head direction and gaze by changing their own gaze or head direction[24]. Another interesting observation is that people tend to nod their head at the robot during explanation. These kind of positive responses from the listener could be used to improve the engagement between a human and robot.

Mel has been augmented with our head-tracking module so that it can estimate head gaze more accurately and detect head nods [14]. The original conversation model of Mel was modified to include head nods as an additional engagement cue. When the robot is speaking, head nods can be detected and used by the system to know that the listener is engaged in the conversation. This is a more natural interface when compared to the original version where the robot had to ask a question to get the same feedback.

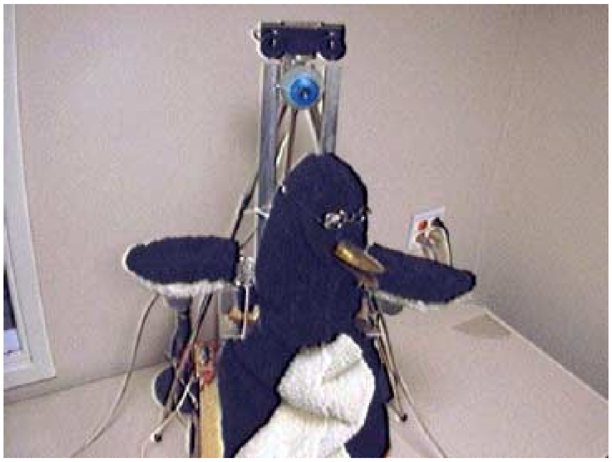

Figure 4: Mel has been developed to study engagement in collaborative conversation[14]. The robot uses information from the stereo camera to estimate head pose and recognize head gesture.

This augmented version of Mel has been tested by multiple subjects and seems to give more engaging conversation. As shown during MACK experiments, nonverbal grounding cues like head nods are performed by human subjects when interacting with an embodied conversational agent. The visual grounding module enriches the input sensor information of the embodied conversational agent and improves the user experience.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented the design concepts necessary to build a visual grounding module for interactive dialog systems. This module can track head pose and detect head gestures with the accuracy needed for human–robot interaction. We presented a new user interface concept called conversational tooltips and showed that head gestures and head pose can be useful cues when interacting with a kiosk. Finally, we showed how our visual module was integrated with two different discourse models: an embodied conversational agent and a robot. In both cases, the visual grounding module enriched the input sensor information and the user experience. As future work, we would like to integrate the visual module more closely with the discourse model and include context information inside the vision processing.

7. REFERENCES

[1] AT&T. *Natural Voices*. http://www.naturalvoices.att.com.

[2] S. Basu, I. Essa, and A. Pentland. Motion regularization for model-based head tracking. In *Proceedings. International Conference on Pattern Recognition*, 1996.

[3] M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *ICCV*, pages 374–381, 1995.

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH99*, pages 187–194, 1999.

[5] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjalmsson, and H. Yan. Embodiment in conversational interfaces: Rea. In *Proceedings of the CHI'99 Conference*, pages 520–527, Pittsburgh, PA, 1999.

[6] J. Cassell, T. Bickmore, H. Vilhjalmsson, and H. Yan. A relational agent: A model and implementation of building user trust. In *Proceedings of the CHI'01 Conference*, pages 396–403, Seattle, WA, 2001.

[7] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *PAMI*, 23(6):681–684, June 2001.

[8] V. Design. *MEGA-D Megapixel Digital Stereo Head*. http://www.ai.sri.com/ konolige/svs/, 2000.

[9] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *PAMI*, 20(10):1025–1039, October 1998.

[10] Haptek. *Haptek Player*. http://www.haptek.com.

[11] A. Kapoor and R. Picard. A real-time head nod and shake detector. In *Proceedings from the Workshop on Perspective User Interfaces*, November 2001.

[12] R. Kjeldsen. Head gestures for computer control. In *Proc. Second International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems*, pages 62–67, 2001.

[13] M. La Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of textured-mapped 3D models. *PAMI*, 22(4):322–336, April 2000.

[14] C. Lee, N. Lesh, C. Sidner, L.-P. Morency, A. Kapoor, and T. Darrell. Nodding in conversations with a robot. In *Extended Abstract of CHI'04*, April 2004.

[15] L.-P. Morency and T. Darrell. Stereo tracking using ICP and normal flow. In *Proceedings International Conference on Pattern Recognition*, 2002.

[16] L.-P. Morency, A. Rahimi, N. Checka, and T. Darrell. Fast stereo-based head tracking for interactive environment. In *Proceedings of the Int. Conference on Automatic Face and Gesture Recognition*, 2002.

[17] L.-P. Morency, A. Rahimi, and T. Darrell. Adaptive view-based appearance model. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 2003.

[18] K. Murphy. *Bayes Net Toolbox for Matlab*. http://www.ai.mit.edu/ murphyk/Software/BNT/bnt.html.

[19] Y. Nakano, G. Reinstein, T. Stocky, and J. Cassell. Towards a model of face-to-face grounding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, July 2003.

[20] Nuance. *Nuance*. http://www.nuance.com.

[21] J. Pierrehumbert. *The phonology and phonetic of English intonation*. Massachusetts Institute of Technology, 1980.

[22] C. Rich, C. Sidner, and N. Lesh. Collagen: Applying collaborative discourse theory to human–computer interaction. *AI Magazine, Special Issue on Intelligent User Interfaces*, 22(4):15–25, 2001.

[23] A. Schodl, A. Haro, and I. Essa. Head tracking using a textured polygonal model. In *PUI98*, 1998.

[24] C. Sidner, C. D. Kidd, C. Lee, and N. Lesh. Where to look: A study of human–robot engagement. In *Proceedings of Intelligent User Interfaces*, Portugal, 2004.

[25] C. Sidner, C. Lee, and N. Lesh. Engagement when looking: Behaviors for robots when collaborating with people. In *Diabruck: Proceedings of the 7th workshop on the Semantic and Pragmatics of Dialogue*, pages 123–130, University of Saarland, 2003. I. Kruiff-Korbayova and C. Kosny (eds.).

[26] M. Siracusa, L.-P. Morency, K. Wilson, J. Fisher, and T. Darrell. Haptics and biometrics: A multimodal approach for determining speaker location and focus. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, November 2003.

[27] R. Stiefelhagen. Tracking focus of attention in meetings. In *Proceedings of International Conference on Multimodal Interfaces*, 2002.

[28] Y. Takemae, K. Otsuka, and N. Mukaua. Impact of video editing based on participants' gaze in multiparty conversation. In *Extended Abstract of CHI'04*, April 2004.

[29] P. Viola and M. Jones. Robust real-time face detection. In *ICCV*, page II: 747, 2001.

[30] L. Wiskott, J. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *PAMI*, 19(7):775–779, July 1997.

[31] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *PAMI*, 19(7):780–785, July 1997.