Co-occurrence Graphs: Contextual Representation for Head Gesture Recognition during Multi-Party Interactions

Louis-Philippe Morency
Institute for Creative Technologies
University of Southern California
Marina del Rey, CA 90292
morency@ict.usc.edu

ABSTRACT

Head pose and gesture offer several conversational grounding cues and are used extensively in face-to-face interaction among people. To accurately recognize visual feedback, humans often use contextual knowledge from previous and current events to anticipate when feedback is most likely to occur. In this paper we describe how contextual information from other participants can be used to predict visual feedback and improve recognition of head gestures in multi-party interactions (e.g., meetings). An important contribution of this paper is our data-driven representation, called co-occurrence graphs, which models co-occurrence between contextual cues such as spoken words and pauses, and visual head gestures. By analyzing these co-occurrence patterns we can automatically select relevant contextual features and predict when visual gestures are more likely. Using a discriminative approach to multi-modal integration, our contextual representation using co-occurrence graph improves head gesture recognition performance on a publicly available dataset of multi-party interactions.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Motion*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Discourse*

General Terms

Algorithms

Keywords

Co-occurrence graphs, Contextual information, visual gesture recognition, human-human interaction

1. INTRODUCTION

During multi-party interactions such as in meetings, information is exchanged between participants using both audio and visual channels. Visual feedback can range from a simple eye glance to a large arm gesture or posture change. One important visual cue is head nod during conversation. Head nods are used for displaying agreement, grounding information or during turn-taking [6, 7].

People do not provide feedback at random. Rather they react to the current topic, previous utterances and the speaker's current verbal and nonverbal behavior [1]. More generally, speakers and listeners co-produce a range of lexical, prosodic, and nonverbal patterns. Such feedback is an essential and predictable aspect of natural conversation and its absence can significantly disrupt participants ability to communicate [3, 29]. Recognizing these visual gestures is important for understanding all the information exchanged during a meeting or conversation, and can be particularly crucial for identifying more subtle factors such as the effectiveness of communication [24], points of confusion, status relationships between participants [25], or the diagnosis social disorders [20]. Our goal is to automatically discover these patterns using only easily observable features of human face-to-face interaction (e.g. prosodic features and eye gaze), and exploit them to improve recognition accuracy.

In this paper we describe how contextual information from other participants can be used to predict visual feedback and improve recognition of head gestures in multi-party interactions (e.g., meetings). An important contribution of this paper is our contextual representation based on co-occurrence graphs which models co-occurrence between contextual cues such as spoken words and pauses, and visual head gestures. By analyzing these co-occurrence patterns, we show how to automatically select relevant contextual features and predict when visual gestures are most likely.

The following section describes previous work in visual gesture recognition and explains the differences between our context-based approach and other recognition models. Section 3 discusses the contextual information available during multi-party interactions. Section 4 introduces co-occurrence graphs and Section 5 describes how to use them to encode contextual information. Section 6 presents the way we collected the data used for training and evaluating our model as well as the methodology used to evaluate the performance of our approach.

2. PREVIOUS WORK

Head pose and gesture offer several key conversational grounding cues and are used extensively in face-to-face interaction among people. Stiefelhagen developed several systems for tracking face pose in meeting rooms and has shown

UCVP '09 Boston, Massachusetts USA

that face pose is very useful for predicting turn-taking [22]. Takemae *et al.* also examined face pose in conversation and showed that if tracked accurately, face pose is useful in creating a video summary of a meeting [23]. Siracusa *et al.* developed a system that uses head pose tracking to interpret who was talking to who in conversational setting [21]. The position and orientation of the head can be used to estimate head gaze which is a good estimate of a person's attention.

Recognition of head gestures has been demonstrated by tracking eye position over time. Kapoor and Picard presented a technique to recognize head nods and head shakes based on two Hidden Markov Models (HMMs) trained and tested using 2D coordinate results from an eye gaze tracker [11]. Kawato and Ohya suggested a technique for head gesture recognition using between eye templates [12]. Fujie *et al.* also used HMMs to perform head nod recognition [8]. In their paper, they combined head gesture detection with prosodic low-level features computed from Japanese spoken utterances to determine strongly positive, weak positive and negative responses to yes/no type utterances.

Several researchers have developed models to predict when backchannel should happen based mostly on unimodal inputs. Ward and Tsukahara [28] propose a unimodal approach where backchannels are associated with a region of low pitch lasting 110ms during speech. Models were produced manually through an analysis of English and Japanese conversational data. Nishimura et al. [19] present a unimodal decision-tree approach for producing backchannels based on prosodic features. The system analyzes speech in 100ms intervals and generates backchannels as well as other paralinguistic cues (e.g., turn taking) as a function of pitch and power contours. Cathcart et al. [4] propose a unimodal model based on pause duration and trigram part-of-speech frequency. Lee and Marsella [13] mixed a trigram representation and HMMs to predict speaker's head nods. The model was constructed by identifying, from the HCRC Map Task Corpus [2], trigrams ending with a backchannel. In contrast to these gesture generation systems, our approach uses the contextual information from other participants to improve gestures recognition.

Context has been previously used in computer vision to disambiguate recognition of individual objects given the current overall scene category [26]. In contrast to the idea of fusing multiple modalities from the human participant to improve recognition (e.g., Kaiser *et al.* work on multi-modal interaction in augmented and virtual reality [10]), our approach takes its contextual information directly from the other human participants. More closely related, Morency et al. [15] encoded dialogue context using a static set of encoding templates to improve head nod recognition during dyadic interactions. Our work extends this approach in two main aspects: (1) contextual representation using co-occurrence graphs, and (2) context from multi-party interactions.

3. DIALOG CONTEXT DURING MULTI-PARTY INTERACTIONS

Our goal is to quantify the relationship between contextual information and visual gestures by looking at the time distribution of visual gestures given a contextual event. In our case, a contextual event can be a spoken word, a pause

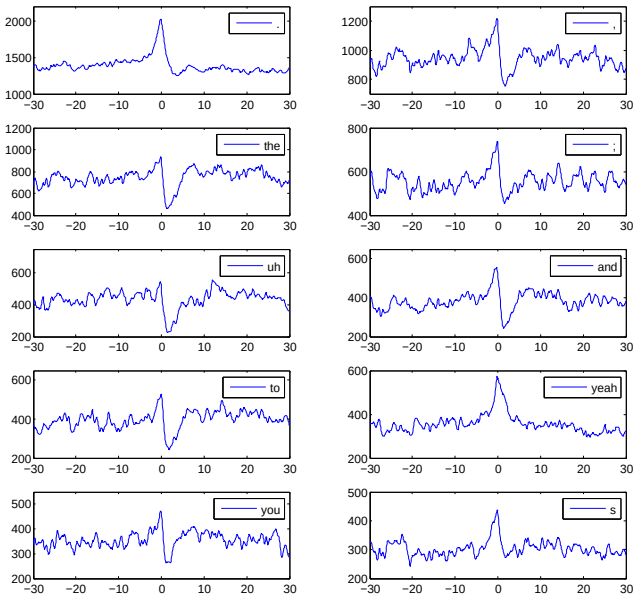

Figure 1: Examples of co-occurrence graphs. When analyzing this relationship between head nods and contextual events, three temporal patterns appear: ignition, transition and negation.

or the end of a sentence. If a relationship exists between a contextual event and a specific visual gesture (e.g., head nod) then we will expect to see a structure in the relative distribution. If no relationship exist, the relative distribution should be random.

We define context as the set of events happening from other sources than the person of interest. For example, in a multi-party conversation between four people, we define context for one participant as the set of events coming from the three other participants. Since our goal in this paper is to recognize visual gestures, we focus on context events related to spoken utterances:

- **Prosodic cues** Prosody refers to the rhythm, pitch and intonation of speech. Several studies have demonstrated that listener feedback is correlated with a speaker's prosody [19]. For example, Ward and Tsukahara [28] show that short listener backchannels (listener utterances like "ok" or "uh-huh" given during a speaker's utterance) are associated with a lowering of pitch over some interval. As an approximation to prosody, the punctuation of the transcription can be used.

- **Pauses** Listener feedback often follows speaker pauses or filled pauses such as "um" (see [4]). Speakers will often use pauses to get backchannel feedback from the other participants.

- **Lexical cues** These type of contextual events include all spoken words by other participants. Some conjunctive words such as "and" will bring head nod since one idea just ended but right after the conjunction, head nods are less likely since a new sub-sentence (i.e. phrases, constituent) will follow, during which visual feedback is less likely.

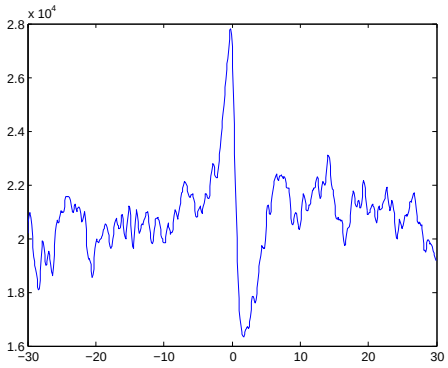

Figure 2: Cumulative number of head nods (Y axis) in function of the time alignment with all contextual events (X axis): spoken words, prosodic and timing. We can observe a relationship between contextual events and head nods between -5 and 5 seconds.

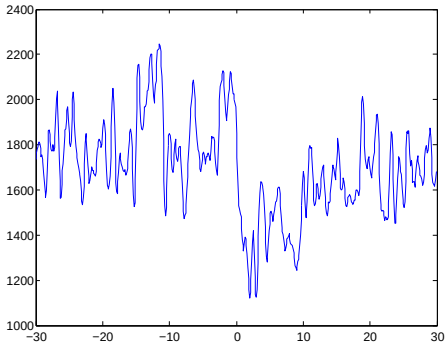

Figure 3: Cumulative number of head shakes (Y axis) in function of the time alignment with all contextual events (X axis). We can still observe a relationship between -5 and 5 seconds, but not as clear as for head nods (most likely due to the smaller number of head shakes in our dataset).

3.1 AMI Meeting Corpus

To study this relationship between context and visual gesture we looked at the annotations from the AMI meeting corpus [5]. This corpus contains 46 meetings with annotated head gestures and spoken words of all four participants[1]. Each meeting varies between 20-40 minutes. The corpus contains follow-up meetings with the same participants. These series usually contain 3 or 4 meetings.

Participants were video recorded using a frontal camera and a close-talking microphone. The video sequences were manually annotated with spoken words, punctuation and head gestures (head nods and head shakes). The dataset contains 9745 head nods and 1279 head shakes. In our analysis, we used a total of 184 sequences (some meetings had only 3 participants annotated with head gestures).

Following the context discussion of Section 3, the contextual events were extracted from the spoken words and punctuation. The lexical event were represented by specific words and pair of words. The prosodic event were approximated by the punctuation cues annotated in each sequence. The timing and pauses were also deduced from the punctuation cues. The contextual events for a specific participant consisted of all the lexical, prosodic and pauses events from all 3 other participants.

[1]The corpus contains a larger number of meetings but we used only the meetings that had both head gestures and spoken words annotated

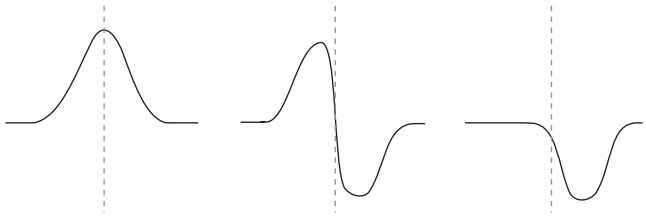

Figure 4: Schematic representation of the three patterns observed when analyzing co-occurrence of head nods and contextual events: (left) ignition pattern, (middle) transition pattern and (right) negative pattern.

4. CO-OCCURRENCE GRAPHS

Our goal is to analyze the relationship between contextual events and visual gestures. Our approach is to create a co-occurrence graph for each contextual event and each possible type of visual gesture. The co-occurrence graph, centered at the contextual event, represents how many visual gesture instances happened around that event. The co-occurrence graphs can be seen as temporal generalization of the co-occurrence matrices introduced by Haralick *et al.* [9].

For each instance of a contextual event, we slide a window of 0.1 second from -30 second before the event to 30 seconds after the event. If a visual gesture happens during a specific time window, the corresponding bin in the co-occurrence graph is incremented. By doing this for each instance of a specific contextual event, we get a time distribution of visual gesture given the contextual event. Figures 1 shows examples of co-occurrence graphs for different contextual events.

Figures 2 and 3 show cumulative co-occurrence graphs for head nods and head shakes. The cumulative co-occurrence graph for head nods shows an interesting point: most of the relationship between head nods and contextual event seems to happen between -5 and 5 seconds. Past this time, the relationship seems mostly random. The relationship between head shakes and contextual events is not as clear, mostly due to the smaller set of head shakes.

4.1 Patterns in Co-occurrence Graphs

By observing the co-occurrence graphs of Figure 1, three patterns appear: ignition, transition and negation. These patterns are illustrated in Figure 4.

- **Ignition pattern** The first pattern is the ignition pattern (left) where a contextual event positively influence visual gesture. This type of relationship means that a visual gesture is more likely to happen around the contextual event. This is true for the period which

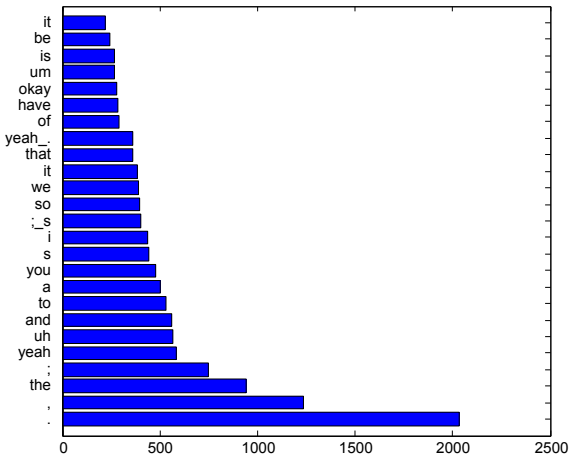

Figure 5: Top 25 contextual features. Horizontal axis: maximum number of time a head nod happened in a window or +/-5 seconds around the contextual feature.

represents the end of a sentence. This is also true for positive feed such as the word "yeah".

- **Transition pattern** The second pattern is the transition pattern (middle) where a contextual event represents a mid-point between two phrases. This type of relations will bring a high likelihood around or before the event but right after the event this likelihood will be lower. Two good examples of this type of events are the comma and the word "and". These events will usually occur in the middle of a sentence, between two constituents.

- **Negative pattern** The last pattern is the "negative" pattern (right) where a contextual event negatively influence a visual gesture. This type of relations means that a visual gesture is unlikely to happen after this event. The words "the" and "to" are two good examples of this type of patterns. These words do not bring visual feedback and usually following one of these words will be a large number of other spoken words.

The analysis of the co-occurrence graphs shown in Figure 1 confirm our intuition that the context is related to visual feedback (e.g., head nods). Also, these co-occurrence graphs contains patterns that can potentially help to recognize when a specific gesture is more likely.

4.2 Co-occurrence Ranking of Contextual Features

A good contextual feature is an event (1) that happens on a regular basis so that there is a good chance to see this same event in a new meeting, and (2) that is related to visual feedback. One criterium that includes both advantages is the maximum number of co-occurrence between the contextual event and the visual gesture. This criteria is equal to the maximum peak of each co-occurrence graphs.

Figure 5 shows the top 25 contextual features. The top feature is the period, which usually represent the end of a sentence. This goes with our intuition that people usually

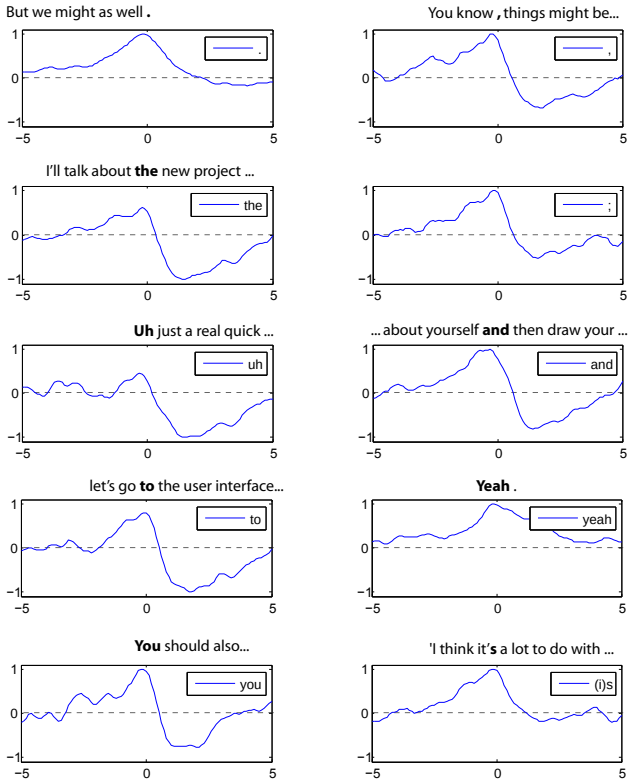

Figure 6: Examples of contextual representation using co-occurrence graphs. Taken from the AMI dataset, each plot shows a spoken utterance (from one of the meeting participant) and underneath how context was encoded. The sentences are centered around a specific contextual event shown in bold (same as in Figure 1).

do grounding gesture at the end of a sentence. Also the second feature is the comma which represents a pause in a sentence. Pauses are also good timing for grounding gesture. The other top contextual features are more interesting since they are lexical features and bring interesting questions as why they are related with visual gestures. The following section shows how using co-occurrence graphs to represent contextual features brings an intuitive feature representation.

5. CONTEXTUAL REPRESENTATION USING CO-OCCURRENCE GRAPHS

Based on the observed relationship between contextual events and visual gestures, in this section we present our contextual representation using co-occurrence graphs. Our goal is to encode contextual events so they keep their relationship with the visual gesture only when the relationship is strong and useful.

As we observed earlier in the co-occurrence graphs, most of the variation happens between -5 and 5 seconds. For this reason, we define an inlier region (between -5 and 5 seconds) and an outlier region (outside -5 and +5 seconds). The outlier region represents the number of visual gestures

randomly happening when the contextual event does not have influence. The mean value in the outlier region can be used as an estimate of this randomness.

The first step for computing the contextual representation of an event is to re-center the co-occurrence graph by subtracting the mean from the outlier region. By doing so, the contextual feature will be set to zero if no (or random) influence. The final step is to re-scale the inlier co-occurrence graph to be contained between 1 and -1. Figure 6 shows the final representation of the top 10 contextual events.

During encoding of a new sequence, the value of a contextual feature will be computed from the time between the current frame and the contextual event. If more than one contextual event happens in short time, the highest value is kept. If no contextual event happened in the last 5 seconds, then the value for this contextual feature is zero.

Figure 6 shows examples of sentences from the AMI meeting dataset. The sentences are placed so that the contextual event is centered at zero. This figure gives concrete examples of the relationship between context and its feature representation. Also, we can see how the three feature patterns described earlier apply to different sentences.

6. EXPERIMENTS

We performed experiments to compare our context-based recognition approach with a vision-only approach and context-only approach.

6.1 Dataset

For our experiments, we used the AMI meeting corpus [5] introduced in Section 3.1 which contains 46 different meetings. The first four meetings were used to train and test our visual gesture recognizer while the last 42 meetings were used for computing the co-occurrence graphs as described in Section 5. The total number of head nods in the first four meetings was 1176 while only 103 head shakes occurred. The results presented in Section 6.4 are for head nod recognition.

The video sequences from the first four meetings were processed using the Watson software [17] to obtain the head position and orientation of each participant in real-time. Watson tracks the 6 degrees-of-freedom head pose using a using a framework called Adaptive View-Based Appearance Model with an average accuracy of 3.5° and 0.75in. The library also recognizes two head gestures using a support vector machines (SVMs): head nods and head shakes. The head tracker is automatically initialized using frontal face detection [27]. Two participants had to be left out because of tracking problems (e.g., occlusion).

6.2 Context Integration

Following [18], we adopt a late fusion approach for context-based recognition where first contextual prediction and vision-only recognition are done independently and then their results are combined by a third module called multimodal integrator. The contextual predictor outputs a likelihood measurement at the same frame rate as the vision-only recognizer so that the multi-modal integrator can merge both measurements. This late fusion approach has the advantage that contextual predictor can be trained on a different dataset then the multimodal integrator.

For the contextual predictor and the multimodal integrator, we use Latent-Dynamic Conditional Random Field models as it was shown in [14] to be well suited for context-based recognition. The LDCRF classifiers were trained using the objective function described in [16]. During evaluation, we compute ROC curves using the maximal marginal probabilities. During training, the number of hidden states per label (from 2 to 6 states per label) and the regularization term (with values $10^k, k = -3..3$) were selected automatically using the validation set.

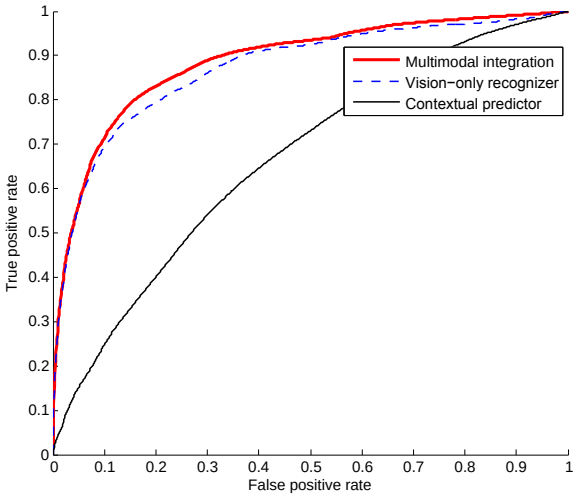

Figure 7: ROC curves of head nod recognition comparing our context-based model to a vision-only technique. The contextual events were encoded using co-occurrence graphs (described in Section 5).

6.3 Methodology

The experiments were performed using a leave-one-out testing approach. For validation, we did holdout cross-validation where a sequence is randomly picked from the training set and kept for validation. The optimal validation parameters (number of hidden states and regularization factor) were picked automatically based on the equal error rate on the validation set.

The dataset contained an unbalanced number of gesture frames compared to background frames. To have a balanced training set and to reduce the training time, the training set was preprocessed to create a smaller dataset containing an equal number of gesture and background examples. The training set contained subsequences from either the background class, or from the gesture class with a buffer of background frames before and after the gesture. The size of the buffer before and after the gesture randomly varied between 2 and 50 frames. Background subsequences were randomly extracted from the original sequences with length varying between 30-60 frames.

6.4 Results

We compared our context-based recognition approach to a vision-only algorithm and context-only approach. Figure 7 shows the ROC curves comparing our context-based approach with a vision-only technique. The ROC curves present the detection performance for both recognition algorithms when varying the detection threshold. Pairwise one-tailed t-test comparison show a marginally significant difference between the two approaches, with $p = 0.05$ for

the equal error rate. We should note that while 42 interactions were used to create our co-occurrence graphs, only 4 interactions were used for the testing. We can expect a better statistical significance as we increase the number of test sequences.

Our experiments show that by using a contextual representation based on co-occurrence graphs, contextual information from other participants can improve the performance of vision-based gesture recognition.

7. CONCLUSION

In this paper we described how contextual information from other participants can be used to predict visual feedback and improve recognition of head gestures in multi-party interactions (e.g., meetings). An important contribution of this paper was our contextual representation based on co-occurrence graphs which models co-occurrence between contextual cues such as spoken words and pauses, and visual head gestures. By analyzing these co-occurrence patterns we automatically selected relevant contextual features and predicted when visual gestures was most likely. Using a discriminative approach to multi-modal integration, our data-driven context representation improved head gesture recognition performance.

8. REFERENCES

[1] J. Allwood, J. Nivre, and E. Ahlsn. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, pages 1–26, 1992.

[2] H. Anderson, M. Bader, E.G. Bard, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert. The mcrc map task corpus. *Language and Speech*, 34(4):351–366, 1991.

[3] Janet B. Bavelas, Linda Coates, and Trudy Johnson. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952, 2000.

[4] N. Cathcart, Jean Carletta, and Ewan Klein. A shallow model of backchannel continuers in spoken dialogue. In *European ACL*, pages 51–58, 2003.

[5] AMI consortium. *AMI meeting corpus*. http://corpus.amiproject.org/.

[6] A. Dittmann and L. Llewellyn. Relationship between vocalizations and head nods as listener responses. *Journal of Personality and Social Psychology*, 9:79–84, 1968.

[7] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292, 1972.

[8] Shinya Fujie, Yasuhi Ejiri, Kei Nakajima, Yosuke Matsusaka, and Tetsunori Kobayashi. A conversation robot using head gesture recognition as para-linguistic information. In *Proceedings of 13th IEEE International Workshop on Robot and Human Communication, RO-MAN 2004*, pages 159–164, September 2004.

[9] R. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6):610–621, 1973.

[10] 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. In *Proceedings of the 5th international conference on Multimodal interfaces (ICMI 2003)*, pages 12–19, Vancouver, B.C., Canada, November 2003. ACM press.

[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] S. Kawato and J. Ohya. Real-time detection of nodding and head-shaking by directly detecting and tracking the between-eyes. In *Proceedings. Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 40–45, 2000.

[13] J. Lee and S. Marsella. Learning a model of speaker head nods using gesture corpora. In *The 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, Budapest, Hungary, 2009.

[14] L.-P. Morency and T. Darrell. Conditional sequence model for context-based recognition of gaze aversion. In *Proceedings of the 4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 2007.

[15] L.-P. Morency, I. de Kok, and J. Gratch. Context-based recognition during human interactions: Automatic feature selection and encoding dictionary. In *Proceedings of the International Conference on Multimodal interfaces (ICMI 2008)*, 2008.

[16] L.-P. Morency, A. Quattoni, and T. Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[17] Louis-Philippe Morency. *Watson: Head tracking and gesture recognition library*. http://groups.csail.mit.edu/vision/vip/watson/.

[18] Louis-Philippe Morency, Candace Sidner, Christopher Lee, and Trevor Darrell. Contextual recognition of head gestures. In *Proceedings of the International Conference on Multi-modal Interfaces*, October 2005.

[19] Ryota Nishimura, Norihide Kitaoka, and Seiichi Nakagawa. A spoken dialog system for chat-like conversations considering response timing. *LNCS*, 4629:599–606, 2007.

[20] A.A. Rizzo, D. Klimchuk, R. Mitura, T. Bowerly, J.G. Buckwalter, and T. Parsons. A virtual reality scenario for all seasons: The virtual classroom. *CNS Spectrums*, 11(1):35–44, 2006.

[21] 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.

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

[23] 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.

[24] Linda Tickle-Degnen and Robert Rosenthal. The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1(4):285–293, 1990.

[25] Larissa Z. Tiedens and Alison R. Fragale. Power moves: Complementarity in dominant and submissive nonverbal behavior. *Journal of Personality and Social Psychology*, 84(3):558–568, 2003.

[26] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *IEEE Intl. Conference on Computer Vision (ICCV)*, Nice, France, October 2003.

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

[28] N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 23:1177–1207, 2000.

[29] V. H Yngve. On getting a word in edgewise. In *Sixth regional Meeting of the Chicago Linguistic Society*, pages 567–577, 1970.