The Role of Context in Head Gesture Recognition

Louis-Philippe Morency[⋆] **Candace Sidner**[†] **Christopher Lee**[†] **Trevor Darrell**[⋆]

[⋆]Computer Sciences and Artificial Intelligence Laboratory
Massachussetts Institute of Technology
Cambridge, MA 02139, USA
{lmorency,trevor}@csail.mit.edu

[†]Mitsubishi Electric Research Laboratories (MERL)
Cambridge, MA 02139, USA
{sidner,lee}@merl.com

Abstract

Head pose and gesture offer several key conversational grounding cues and are used extensively in face-to-face interaction among people. We investigate how dialog context from an embodied conversational agent (ECA) can improve visual recognition of user gestures. We present a recognition framework which (1) extracts contextual features from an ECA's dialog manager, (2) computes a prediction of head nod and head shakes, and (3) integrates the contextual predictions with the visual observation of a vision-based head gesture recognizer. We found a subset of lexical, punctuation and timing features that are easily available in most ECA architectures and can be used to learn how to predict user feedback. Using a discriminative approach to contextual prediction and multi-modal integration, we were able to improve the performance of head gesture detection even when the topic of the test set was significantly different than the training set.

Introduction

During face-to-face conversation, people use visual feedback to communicate relevant information and to synchronize rhythm between participants. A good example of nonverbal feedback is head nodding and its use for visual grounding, turn-taking and answering yes/no questions. When recognizing visual feedback, people use more than their visual perception. Knowledge about the current topic and expectations from previous utterances help guide our visual perception in recognizing nonverbal cues. Our goal is to equip an embodied conversational agent (ECA) with the ability to use contextual information for performing visual feedback recognition much in the same way people do.

In the last decade, many ECAs have been developed for face-to-face interaction (Bickmore & Cassell 2004; Sidner *et al.* 2005). A key component of these systems is the dialogue manager, usually consisting of a history of the past events, the current state, and an agenda of future actions. The dialogue manager uses contextual information to decide which verbal or nonverbal action the agent should perform next. This is called context-based synthesis.

Contextual information has proven useful for aiding speech recognition (Lemon, Gruenstein, & Stanley 2002). In (Lemon, Gruenstein, & Stanley 2002), the grammar of the speech recognizer dynamically changes depending on the agent's previous action or utterance. In a similar fashion, we want to develop a context-based visual recognition module that builds upon the contextual information available in the dialogue manager to improve performance.

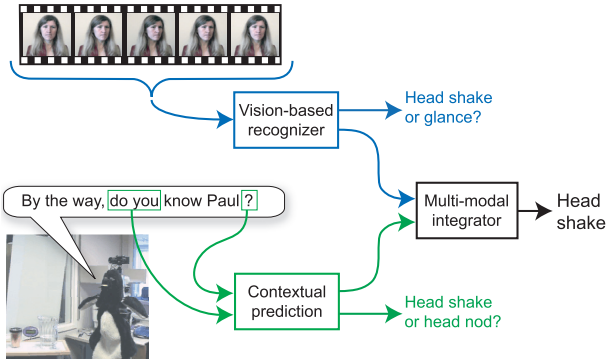

Figure 1: Contextual recognition of head gestures during face-to-face interaction with an embodied agent.

The use of dialogue context for visual gesture recognition has, to our knowledge, not been explored before for conversational interaction. In this paper we present a prediction framework for incorporating dialogue context with vision-based head gesture recognition. The contextual features are derived from the utterances of the ECA, which is readily available from the dialogue manager. We highlight three types of contextual features: lexical, punctuation, and timing, and selected a subset for our experiment that were topic independent. In contrast to previous work based on HMMs (Fujie *et al.* 2004), we use a discriminative approach to predict head nods and head shakes from a small set of recorded interactions. We then combine the contextual predictions with a vision-based recognition algorithm based on the frequency pattern of the user's head motion. Our context-based recognition framework allows us to predict, for example, that in certain contexts a glance is not likely whereas a head shake or nod is (as in Figure 1), or that a head nod is not likely and a head nod misperceived by the vision system can be ignored. This work was published in longer form at the Seventh International Conference on Multimodal Interfaces (Morency *et al.* 2005).

Dialog Context in ECA Architectures

During face-to-face interactions, people use knowledge about the current dialog to anticipate visual feedback from their interlocutor. As depicted in Figure 1, knowledge of the ECA's spoken utterance can help predict which visual feedback is most likely.

The idea of this paper is to use this existing information to predict when visual feedback gestures from the user are likely. Since the dialog manager is already merging information from the input devices with the history and the discourse model, the output of the dialog manager will contain useful contextual information. We highlight three types of contextual features easily available in the dialog manager:

LEXICAL FEATURES Lexical features are computed from the words said by the embodied agent. By analyzing the word content of the current or next utterance, one should be able to anticipate certain visual feedback. For example, if the current spoken utterance started with "Do you", the interlocutor will most likely answer using affirmation or negation. In this case, it is also likely to see visual feedback like a head nod or a head shake. On the other hand, if the current spoken utterance started with "What", then it's unlikely to see the listener head shake or head nod–other visual feedback gestures (e.g., pointing) are more likely in this case.

PUNCTUATION FEATURES Punctuation features modify the way the text-to-speech engine will pronounce an utterance. Punctuation features can be seen as a substitute for more complex prosodic processing that are not yet available from most speech synthesizers. A comma in the middle of a sentence will produce a short pause, which will most likely trigger some feedback from the listener. A question mark at the end of the sentence represents a question that should be answered by the listener. When merged with lexical features, the punctuation features can help recognize situations (e.g., yes/no questions) where the listener will most likely use head gestures to answer.

TIMING Timing is an important part of spoken language and information about when a specific word is spoken or when a sentence ends is critical. This information can aid the ECA to anticipate visual grounding feedback. People naturally give visual feedback (e.g., head nods) during speaker pauses as well as just before the pause occurs. In natural language processing (NLP), lexical and syntactic features are predominant but for face-to-face interaction with an ECA, timing is also an important feature.

The following section describes how we can automatically extract lexical, punctuation and timing features from the dialog system.

Contextual Features

We want to automatically extract contextual information from the dialog manager rather than directly access the ECA internal state. Our proposed method extracts contextual features from the messages sent to the audio and gesture synthesizers. This strategy allows us to extract a summarized version of the dialog context while reducing the cost of extracting contextual cues. Since it does not presume any internal representation, our idea can be applied to most ECA architectures.

In our framework, the dialog manager sends a minimal set of information to the visual analysis module: the next spoken utterance, a time stamp and an approximated duration. The next spoken utterance contains the words, punctuation, and gesture information used to generate the ECA's actions. The utterance information is processed to extract the lexical, punctuation, timing, and gesture features described below. Approximate duration of utterances is generally computed by speech synthesizers and made available in the synthesizer API.

The top three graphs of Figure 2 show how two sample utterances will be coded for the bigram "do you", the question mark and the timing feature. The contextual features are evaluated for every frame acquired by the visual analysis module (about 18Hz). The lexical and punctuation features are evaluated based on the current spoken utterance. The effect of an utterance starts when it starts to be spoken and ends after the pause following the utterance.

We selected our features so that they are topic independent. This means that we should be able to learn how to predict head gesture from a small set of interactions and then use this knowledge on a new set of interactions with a a different topic discussed by the human participant and the robot. However, different classes of dialogs might have different key features, and ultimately these should be learned using a feature selection algorithm (this is a topic of future work).

Contextual Prediction

In this section, we first describe our discriminative approach to learning the influence of contextual features on visual feedback. We learn automatically a likelihood measure of certain visual gestures given a subset of contextual features. Then, we present an experiment where we predict head nods and head shakes just from linguistic data.

Our prediction algorithm takes as input the contextual features and outputs a margin for each visual gesture. The margin is a scalar value representing the confidence that a specific gesture happened. In our experiments, we focus on two head gestures: head nods and head shakes.

We are using a multi-class Support Vector Machine (SVM) to estimate the prediction of each visual gesture.After training the multi-class SVM, we can easily compute a margin for each class and use this scalar value as a prediction for each visual gesture.

We trained the contextual predictor using a data set of seven video sequences where human participants conversed with a humanoid robot. The robot's spoken utterances were automatically processed, as described in previous section, to compute the contextual features. A total of 236 utterances were used to train the multi-class SVM of our contextual predictor. Positive and negative samples were selected from the same data set based on manual transcription of head nods and head shakes.

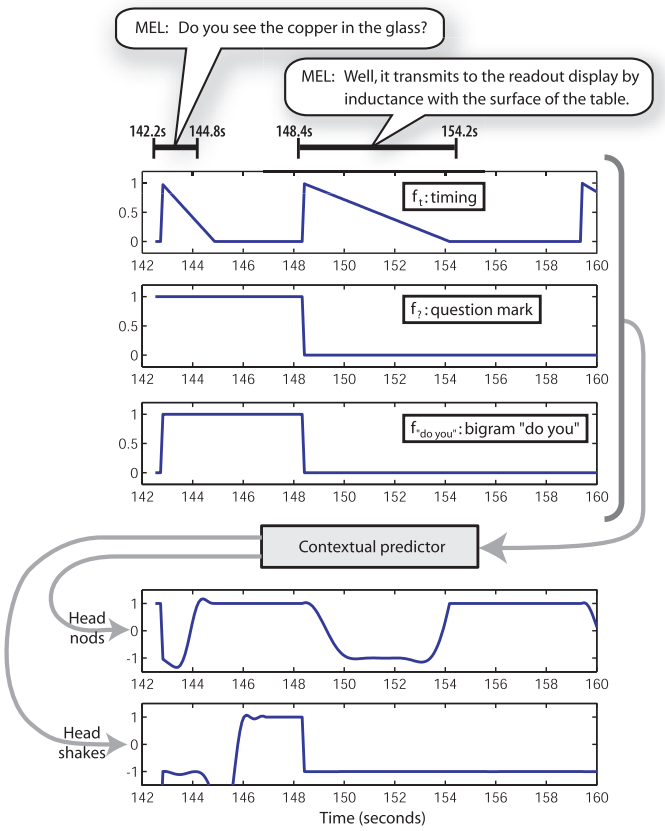

Figure 2: Prediction of head nods and head shakes based on contextual features: (1) distance to end-of-utterance when ECA is speaking, (2) type of utterance and (3) lexical bigram feature. We can see that the contextual predictor learned that head nods should happen near or at the end of an utterance or during a pause while head shakes are most likely at the end of a question.

Figure 2 displays the output of each class of our contextual predictor for a sample dialogue segment between the robot and a human participant held out from the training data. Positive margins represent a high likelihood for the gesture. It is interesting to observe that the contextual predictor automatically learned that head nods are more likely to occur around the end of an utterance or during a pause, while head shakes are most likely to occur after the completion of an utterance. More interestingly, it also learned that head shakes are directly correlated with the type of utterance (a head shake will most likely follow a question), and that head nods can happen at the end of a question to represent an affirmative answer and can also happen at the end of a normal statement to ground the spoken utterance.

Multi-Modal Integration and Recognition

Given a listener's visual feedback based on contextual information from an ECA, we now integrate these predictions with observations from a vision-based head gesture recognizer. We will first describe the visual recognizer used during our experiments and then describe integration of contextual predictions.

We use a two-step process to recognize head gestures: we first track head position and rotation, and then use a computed head velocity feature vector to recognize head gestures. We use a head tracking framework that merges differential tracking with view-based tracking based on the system described by (Morency, Rahimi, & Darrell 2003). We found this tracker was able to track subtle movements of the head for a long period of time. While the tracker recovers the full 3-D position and velocity of the head, features based on angular velocities were found sufficient for gesture recognition.

For vision-based gesture recognition (without dialog context), we trained a multi-class SVM with two different classes: head nods and head shakes. The head pose tracker outputs a head rotation velocity vector at each time step (sampled at approximately 18Hz). We transform the velocity signal into a frequency-based feature by applying a windowed FFT to each dimension of the velocity independently. We resample the velocity vector to have 32 samples per second. This transforms the time-based signal into an instantaneous frequency feature vector more appropriate for discriminative training. The multi-class SVM was trained using the same method described in the previous section.

While we chose to adopt the SVM approach for visual head gesture recognition, other classification schemes could also fit into our context-based recognition framework; all that we require for the multi-modal context fusion described below is that the vision-based head gesture recognizer return a single detection per head gesture. These detections are margins computed directly from the output of the multi-class SVM.

To recognize visual gestures in the context of the current dialog state, we fuse the output of the context predictor with the output of visual head gesture recognizer.

Our integration component takes as input the margins from the contextual predictor and the visual observations from the vision-based head gesture recognizer, and recognizes if a head gesture has been expressed by the human participant. The output from the integrator is further sent to the dialog manager so it can be used to decide the next action of the ECA.

We use a multi-class SVM for the integrator since experimentally it gave us better performance than a linear classifier or simple thresholding. As mentioned earlier, the integrator could be trained on a smaller data set than the contextual predictor. However in our experiments, we trained the integrator on the same data set as the contextual predictor since our training data set included results from the head pose tracker. (Test data was withheld from both during evaluation.)

Experiments

The following experiment demonstrates how contextual features inferred from an agent's spoken dialogue can improve head nod and head shake recognition. The experiment compares the performance of the vision-only recognizer with the context-only prediction and with multi-modal integration.

For this experiment, a first data set was used to train the contextual predictor and the multi-modal integrator, while a

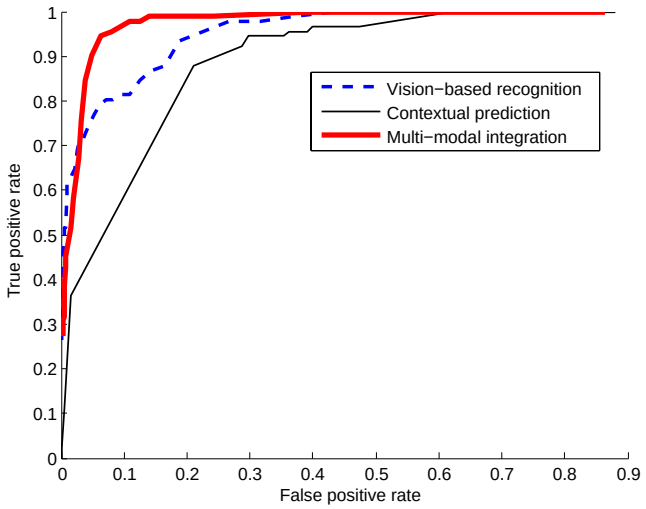

Figure 3: Head nod recognition curves when varying the detection threshold.

second data set with a different topic was used to evaluate the head gesture recognition performance. In the training data set, the robot interacted with the participant by demonstrating its own abilities and characteristics. This data set, called Self, contains 7 interactions. The test data set, called iGlass, consists of nine interactions of the robot describing the iGlassware invention (∼340 utterances).

For both data sets, human participants were video recorded while interacting with the robot. The vision-based head tracking and head gesture recognition were run online (∼18Hz). The robot's conversational model, based on COLLAGEN (Rich, Sidner, & Lesh 2001), determines the next activity on the agenda using a predefined set of engagement rules, originally based on human–human interaction (Sidner *et al.* 2005). Each interaction lasted between 2 and 5 minutes.

During each interaction, we also recorded the results of the vision-based head gesture recognizer as well as the contextual cues (spoken utterances with start time and duration) from the dialog manager. These contextual cues were later automatically processed to create the contextual features necessary for the contextual predictor.

For ground truth, we hand labeled each video sequence to determine exactly when the participant nodded or shook his/her head. A total of 274 head nods and 14 head shakes were naturally performed by the participants while interacting with the robot.

Figure 3 shows head nod detection results for all 9 subjects used during testing. The ROC curves present the detection performance each recognition algorithm when varying the detection threshold. The areas under the curve for each techniques are 0.9543 for the vision only, 0.7923 for the predictor and 0.9722 for the integrator.

Table 1 summarizes the results head nods and head shakes recognition by computing the true positive rates for the fixed negative rate of 0.05. Using a standard analysis of variance (ANOVA) on all the subjects, results on the head nod detection task showed a significant difference among the means of the 3 methods of detection: $F(2, 8) = 20.22$, $p = 0.002$, $d = 0.97$. Pairwise comparisons show a significant difference between all pairs, with $p = 0.006$, $p = 0.039$, and $p < 0.001$ for vision-predictor, vision-integrator, and predictor-integrator respectively. A larger number of samples maybe necessary to see the same significance in head shakes.

	Vision	Predictor	Integrator
Head nods	75%	42%	**90%**
Head shakes	84%	67%	**100%**

Table 1: True detection rates(fixed false positive rate 0.05).

Conclusion and Future Work

Our results show that contextual information can improve user gesture recognition for interactions with embodied conversational agents. We presented a prediction framework that extracts knowledge from the spoken dialogue of an embodied agent to predict which head gesture is most likely. By using simple lexical, punctuation, and timing context features, we were able to improve the recognition rate of the vision-only head gesture recognizer from 75% to 90% for head nods and from 84% to 100% for head shakes. As future work, we plan to experiment with a richer set of contextual cues including those based on gesture display, and to incorporate general feature selection to our prediction framework so that a wide range of potential context features can be considered and the optimal set determined from a training corpus.

References

Bickmore, T., and Cassell, J. 2004. *J. van Kuppevelt, L. Dybkjaer, and N. Bernsen (eds.), Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems*. Kluwer Academic. chapter Social Dialogue with Embodied Conversational Agents.

Fujie, S.; Ejiri, Y.; Nakajima, K.; Matsusaka, Y.; and Kobayashi, T. 2004. 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*, 159–164.

Lemon; Gruenstein; and Stanley, P. 2002. Collaborative activities and multi-tasking in dialogue systems. *Traitement Automatique des Langues (TAL)* 43(2):131–154.

Morency, L.-P.; Sidner, C.; Lee, C.; and Darrell, T. 2005. Contextual recognition of head gestures. In *International Conference on Multi-modal Interfaces*, 18–24.

Morency, L.-P.; Rahimi, A.; and Darrell, T. 2003. Adaptive view-based appearance model. In *Conf. on Computer Vision and Pattern Recognition*, volume 1, 803–810.

Rich, C.; Sidner, C.; and Lesh, N. 2001. Collagen: Applying collaborative discourse theory to human–computer interaction. *AI Magazine* 22(4):15–25.

Sidner, C.; Lee, C.; Kidd, C.; Lesh, N.; and Rich, C. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence* 166(1–2):140–164.