Towards Context-Based Visual Feedback Recognition for Embodied Agents

Louis-Philippe Morency* Candace Sidner† 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@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 contextual information can improve visual recognition of feedback gestures during interactions with embodied conversational agents. We present a visual recognition model that integrates cues from the spoken dialogue of an embodied agent with direct observation of a user's head pose. In preliminary experiments using a discriminative framework, contextual information improved the performance of head nod detection.

1 Introduction

During face-to-face conversation, people use visual nonverbal feedback to communicate relevant information and to synchronize rhythm between participants. A good example of nonverbal feedback is head nodding and its usage for visual grounding, turn-taking and answering yes/no questions. When recognizing visual feedback, people use more then their visual perception. Knowledge about the current topic and expectations from the previous dialog are also included with the visual perception to recognize 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. A key component of these systems is the dialog manager, usually consisting of a history (also called a memory) of the past events and current state, and an agenda of the future actions (see Figure 1). The dialog 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 et al., 2002). In these systems, the grammar of the speech recognizer dynamically changes depending on the agent's previous action or sentence. In a similar fashion, we want to develop a context-based visual recognition module that builds upon the contextual information available in the dialog manager to improve performance.

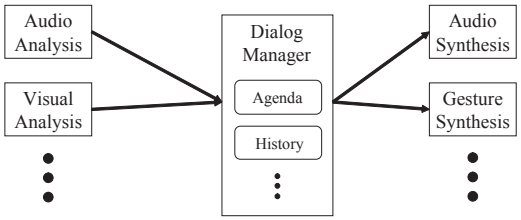

Figure 1: Simplified architecture for embodied conversational agent.

The use of dialog context for visual gesture recognition has, to our knowledge, not been explored before for conversational interaction. Here we present a model for incorporating text-based dialog cues into head-nod recognition. We exploit discriminative classifiers in our work, but other classification schemes could also fit into our general approach.

We have designed a visual analysis module that can recognize head nods based on both visual and dialog context cues. The contextual information is based on the spoken sentences of the ECA, which are readily available from the dialog manager. We use a sum-based technique to detect the head nods based on the spoken sentences and the frequency pattern of the head motion. The experiments were based on 30 video recordings of human participants interacting with an interactive robot.

There has been considerable previous work on gestures with ECA. Bickmore and Cassell (2004) de-

veloped an ECA that exhibited many gestural capabilities to accompany its spoken conversation, and could interpret spoken linguistic utterances from human users. Sidner et al. (2004) have experimented with people interacting with a humanoid robot. They found that more than half their participants naturally nodded at the robot's conversational contributions even though the robot could not interpret head nods. Nakano et al. (2003) analyzed eye gaze and head nods in a computer–human conversation and found that humans monitored the lack of negative feedback. They incorporated their results in an ECA that updated dialogue state. Numerous other ECAs (Traum and Rickel, 2002; Carolis et al., 2001) are exploring aspects of gestural behavior in human-ECA interactions. Physically embodied ECAs—for example, ARMAR II (Dillman et al., 2004, 2002) and Leo (Breazeal et al., 2004)–have also begun to incorporate the ability to recognize human gestures and track human counterparts.

Few of these systems have incorporated tracking of fine motion actions, or visual gesture, and none have included top-down dialog context in the visual recognition process. This paper describes our system for contextual-based visual feedback recognition.

2 Context-Based Visual Analysis

In general, our goal is to efficiently integrate dialog context information from an embodied agent with a visual analysis module. We define a visual analysis module as a software component that can analyze images (or video sequences) and recognize visual feedback of a human participant during interaction with an embodied agent.

Figure 1 is a general view of the architecture for an embodied conversational agent. In this architecture, the dialog manager contains two main subcomponents, an agenda and a history.

The agenda keeps a list of all the possible actions the agent and the user (i.e. human participant) can do next. This list is updated by the dialog manager based on its discourse model (prior knowledge) and on the history. Some interesting contextual cues can be estimated from the agenda:

- What will be the next spoken sentence of our embodied agent?
- Are we expecting some specific answers from the user?
- Is the user expected to look at some common space?

The history keeps a log of all the previous events that happened during the conversation. This information can be used to learn some interesting contextual cues:

- How did the user answer previous questions (speech or gesture)?
- Does the user seem to understand the last explanation?

Based on the history, we can build a prior model about the type of visual feedback shown by the user. Based on the agenda, we can predict the type of visual feedback that will be shown by the user.

Following the definitions of Cassell and Thorisson (1999) for nonverbal feedback synthesis, we outline three categories for visual feedback analysis: (1) content-related feedback, (2) envelope feedback, and (3) emotional feedback. Contextual information can be used to improve recognition in each category.

Content-related feedback is concerned with the content of the conversation. For example, a person uses head nods or pointing gestures to supplement or replace a spoken sentence. For this type of feedback, contextual information inferred from speech can greatly improve the performance of the visual recognition system. For instance, to know that the embodied agent just asked a yes/no question should indicate to the visual analysis module a high probability of a head nod or a head shake.

Grounding visual cues that occur during conversation fall into the category of envelope feedback. Such visual cues include eye gaze contact, head nods for visual grounding, and manual beat gestures. Envelope feedback cues accompany the dialog of a conversation much in the same way audio cues like pitch, volume and tone envelope spoken words. Contextual information can improve the recognition of envelope visual feedback cues. For example, knowledge about when the embodied agent pauses can help to recognize visual feedback related to face-to-face grounding.

Emotional feedback visual cues indicate the emotional state of a person. Facial expression is an emotional feedback cue used to show one of the 6 basic emotions (Ekman, 1992) such as happiness or anger. For this kind of feedback, contextual information can be used to anticipate a person's facial expression. For example, a person smiles after receiving a compliment.

We are developing a framework to integrate the contextual information of a dialog manager with the visual cues recognized by a computer-vision module. To efficiently integrate contextual information,

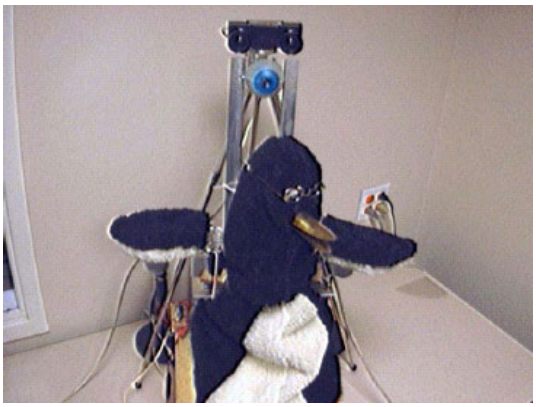

Figure 2: Mel, interactive robot used during our experiment.

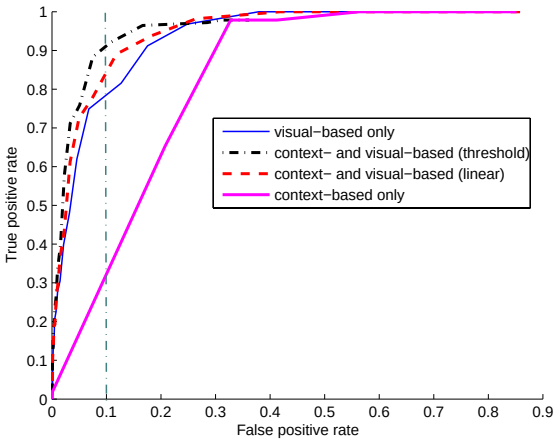

Figure 3: Recognition curves for different head nod detection algorithms.

we need to have a flexible visual recognition algorithm that can deal with multiple sources of information. In this paper, we use a simple cascade of discriminative classifiers to differentiate gestures and to learn contextual events. We show preliminary results indicating that even relatively impoverished dialog cues can have a significant impact on recognition performance. By observing the intra-sentence word-position and whether the sentence was a question, we can significantly improve recognition over visual observation alone.

3 Experiment

Our experiments demonstrate the use of contextual information inferred from an agent's spoken dialogue to improve head-nod recognition. We tested our prototype on 30 video recordings of human participants interacting with an interactive robot, Mel, developed at Mistubishi Electronic Research Lab (MERL) (see Figure 2). Mel interacted with the subject by demonstrating an invention created at MERL. Each interaction lasted between 2 and 5 minutes. Mel's conversational model, based on COLLAGEN (Rich et al., 2001), determines the next item on the agenda using a predefined set of engagement rules, originally based on human–human interaction (Sidner et al., 2003). The conversational model also uses a Sensor Fusion Model (Sidner et al., 2004) to asses engagement information about the user. This module keeps track of verbal—speech recognition—and nonverbal—head-pose estimation and head gesture recognition (Morency and Darrell, 2004)—cues.

For each subject, we had a video sequence of the complete interaction as well as the head pose and velocity for each frame. We labeled each video sequence to determine exactly when the participant noded his head, when the robot spoke, and which type of sentence was spoken by the robot (statement or question).

We trained a first discriminative classifier, a support vector machine (SVM), with data that was a good sampling of natural head gestures (Lee et al., 2004). This SVM was trained using a frequency representation of the head motion over a time window of 1 second. A second SVM was trained for each frame using input features from the spoken dialogue. In this experiment we used two contextual features from the ECA's agenda: the word position inside the spoken sentence and whether the sentence was a question or a statement. The word position is coded between 0 and 1, where 1 represents the middle of a sentence and 0 represents an extremity of the sentence or a pause (between sentences).

Our hypothesis was that including contextual information inside the head-nod detector would increase the number of recognized head nods or, equivalently, would reduce the number of false detections. We tested three different configurations: (1) using only the visual-based approach, (2) using only the contextual information as input, and (3) using the visual approach with the contextual information. Figure 3 shows results for each recognition algorithm when varying the detection threshold. We show the recognition results for two different techniques of combining the discriminative classifiers. For the first technique (black curve in Figure 3), we fixed the context-based threshold to its optimal value and varied the velocity-based threshold. For the second technique (red curve in Figure 3), we linearly combined the output of the two discriminative classifiers.

For a fixed false positive rate of 0.1, 92% of the head nods were detected by the combined approach while only 78% were detected by the visual-based approach and 32% were detected by the context-based method. These results show that we can achieve better performance when integrating contextual information in the visual feedback recognition algorithm.

4 Conclusion and Future Work

Our results show that contextual information can improve visual feedback recognition for interactions with embodied conversational agents. We presented a visual recognition model that integrates knowledge from the spoken dialogue of an embodied agent. By using simple contextual features like word positioning and question/statement differentiation, we were able to improve the performance of our head nod detector from 78% to 92% recognition rate. As future work, we would like to experiment with a richer set of contextual cues and apply our model to different type of visual feedback.

Acknowledgements

Thanks to C. Mario Christoudias.

References

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

Breazeal, Hoffman, and A. Lockerd. Teaching and working with robots as a collaboration. In *The Third International Conference on Autonomous Agents and Multi-Agent Systems AAMAS 2004*, pages 1028–1035. ACM Press, July 2004.

De Carolis, Pelachaud, Poggi, and F. de Rosis. Behavior planning for a reflexive agent. In *Proceedings of IJCAI*, Seattle, September 2001.

Justine Cassell and Kristinn R. Thorisson. The poser of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 1999.

Dillman, Becher, and P. Steinhaus. ARMAR II – a learning and cooperative multimodal humanoid robot system. *International Journal of Humanoid Robotics*, 1(1):143–155, 2004.

Dillman, Ehrenmann, Steinhaus, Rogalla, and R. Zoellner. Human friendly programming of humanoid robots–the German Collaborative Research Center. In *The Third IARP Intenational Workshop on Humanoid and Human-Friendly Robotics*, Tsukuba Research Centre, Japan, December 2002.

Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3–4):169–200, 1992.

Lee, Lesh, Sidner, Morency, Kapoor, and Trevor Darrell. Nodding in conversations with a robot. In *Extended Abstract of CHI'04*, April 2004.

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

Louis-Philippe Morency and Trevor Darrell. From conversational tooltips to grounded discourse: Head pose tracking in interactive dialog systems. In *Proceedings of the International Conference on Multi-modal Interfaces*, College State, PA, October 2004.

Nakano, Reinstein, Stocky, and Justine 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.

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

Sidner, Kidd, Lee, and Neal Lesh. Where to look: A study of human–robot engagement. In *Proceedings of Intelligent User Interfaces*, Portugal, 2004.

Sidner, Lee, and Neal 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.).

D. Traum and J. Rickel. Embodied agents for multiparty dialogue in immersive virtual world. In *Proceedings of the International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2002)*, pages 766–773, July 2002.