Recognizing Gaze Aversion Gestures in Embodied Conversational Discourse

Louis-Philippe Morency
MIT CSAIL
Cambridge, MA 02141
lmorency@csail.mit.edu

C. Mario Christoudias
MIT CSAIL
Cambridge, MA 02141
cmch@csail.mit.edu

Trevor Darrell
MIT CSAIL
Cambridge, MA 02141
trevor@csail.mit.edu

ABSTRACT

Eye gaze offers several key cues regarding conversational discourse during face-to-face interaction between people. While a large body of research results exist to document the use of gaze in human-to-human interaction, and in animating realistic embodied avatars, recognition of conversational eye gestures—distinct eye movement patterns relevant to discourse—has received less attention. We analyze eye gestures during interaction with an animated embodied agent and propose a non-intrusive vision-based approach to estimate eye gaze and recognize eye gestures. In our user study, human participants avert their gaze (i.e. with "look-away" or "thinking" gestures) during periods of cognitive load. Using our approach, an agent can visually differentiate whether a user is thinking about a response or is waiting for the agent or robot to take its turn.

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

Eye gaze tracking, Eye gestures, Embodied conversational agent, Aversion gestures, Turn-taking, Human-computer interaction

1. INTRODUCTION

In face to face interaction, eye gaze is known to be an important aspect of discourse and turn-taking. To create effective conversational human-computer interfaces, it is desirable to have computers which can sense a users' gaze and infer appropriate conversational cues. Embodied conversational agents, either in robotic form or implemented as virtual avatars, have the ability to demonstrate conversational gestures through eye gaze and body gesture, and should also be able to perceive similar displays as expressed by a human user.

It has long been known that eye gaze is a direct way to measure a users attention and engagement, and/or interest in a particular display or person [1]. Gaze has also been demonstrated to be an effective cue to indicate explicit turn-taking in multi-party conversation; users typically attend to the speaker who has the floor.

Even in two-way interaction without any physical objects under discussion, however, gaze is also a useful discourse cue. As pointed out by Kendon [16], humans can use gaze to mediate the length of a conversational turn. When finished with a turn and willing to give up the floor, users tend to look to their conversational partner. Conversely, when they wish to hold the floor even as they pause their speech, they often look away. This often appears as a "non-deictic" or *gaze-averting* eye gesture while they pause their speech to momentarily consider a response [12, 30].

In the user study presented in this paper, we observed that human users made similar gaze aversion gestures while considering their reply when interacting with a virtual embodied conversational agent. Currently, conversational speech systems rely primarily on audio cues to determine utterance end-points, and thus may have difficulty knowing in such cases whether a user is in fact finished speaking. Also, a human participant may have to think about their answer when asked a question, and this delay could be misinterpreted by the system if no visual feedback like eye gaze aversion is recognized.

It is disruptive for a user to have an agent not realize the user is still thinking about a response and interrupt prematurely, or to wait inappropriately thinking the user is still going to speak when the user has in fact finished their turn. To overcome this limitation of existing conversational systems, we have developed an automatic system to recognize eye movement patterns that indicate a user is still holding their conversational turn.

The remainder of this paper is organized as follows. In Section 2 we review relevant related work, and in Section 3 we discuss how conversational gaze gestures can help embodied agents. The result of our user study on the use of eye aversion gestures by subjects interacting with an ECA is presented in Section 4. Our automatic gaze tracking algorithm is described in Section 5 including the experimental results. Finally, a summary and discussion of future work are provided in Section 6.

2. PREVIOUS WORK

Eye gaze tracking has been a research topic for many years [41, 19]. Stiefelhagen *et al.* suggested an approach for eye gaze estimation from frontal images based on neural networks [34]. Some recent systems can estimate the eye gaze with an accuracy of less than a few degrees; these video-based systems require high resolu-

ICMI'06, November 2–4, 2006, Banff, Alberta, Canada.

tion images and usually are constrained to small fields of view (4x4 cm) [25, 7]. Many systems require an infra-red illumination source and filtered camera [42, 15]. In this paper we develop a passive vision-based eye gaze estimator sufficient for inferring conversational gaze aversion cues; as suggested in [40], we can improve tracking accuracy by integration with head pose tracking.

A considerable body of work has been carried out regarding eye gaze and eye motion patterns for perceptive user interfaces. Velichkovsky suggested the use of eye motion to replace the mouse as a pointing device [37]. Qvarfordt and Zhai used eye-gaze patterns to sense the user interest with a map-based interactive system [28]. Li and Selker developed the InVision system which responded to a user's eye fixation patterns in a kitchen environment [18].

There has been considerable work on gestures with embodied conversational agents. Bickmore and Cassell developed an embodied conversational agent (ECA) that exhibited many gestural capabilities to accompany its spoken conversation and could interpret spoken utterances from human users [4]. Sidner *et al.* have investigated how people interact with a humanoid robot [31]. Nakano *et al.* analyzed eye gaze and head nods in computer–human conversation and found that their subjects were aware of the lack of conversational feedback from the ECA [23]. Numerous other projects (e.g. [35, 6]) explore aspects of gestural behavior in human-ECA interactions. Physically embodied ECAs—for example, ARMAR II [9, 10] and Leo [5]—incorporate the ability to perform articulated body tracking and recognize human gestures.

Recently, many researchers have worked on modeling eye gaze behavior for the purpose of synthesizing a realistic ECA. Colburn *et al.* use hierarchical state machines to model eye gaze patterns in the context of real-time verbal communication [8]. Fukayama *et al.* use a two-state Markov model based on three gaze parameters (amount of gaze, mean duration of gaze and gaze points while averted) [11]. Lee *et al.* use an eye movement model based on empirical studies of saccade and statistical models of eye-tracking data [17]. Pelachaud and Bilvi proposed a model that embeds information on communicative functions as well as on statistical information of gaze patterns [26].

The goal of our paper is to observe the kind of eye gestures people make when interacting with an ECA, and evaluate how well we can recognize these gestures.

3. CONVERSATIONAL GAZE CUES

Eye gaze plays an important role in face-to-face interactions. Kendon proposed that eye gaze in two-person conversation offers different functions: monitor visual feedback, express emotion and information, regulate the flow of the conversation (turn-taking), and improve concentration by restricting visual input [16]. Many of these functions have been studied for creating more realistic ECAs [36, 38, 11], but they have tended to explore only gaze directed towards individual conversational partners or objects.

We define three types of distinctive eye motion patterns, or "eye gestures": eye contact, deictic gestures, and non-deictic gestures. Eye contact implies one participant looking at the other participant: during typical interactions, the listener usually maintains fairly long gazes at the speaker while the speaker tends to look at the listener as he or she is about to finish the utterance [16, 24]. Deictic gestures are eye gestures with a specific reference which can be a person not currently involved in the discussion, or an object. Griffin and Bock showed in their user studies that speakers look at an object approximately 900ms before referencing it vocally [13]. Non-deictic gestures are eye movements to empty or uninformative regions of space. This gesture is also referred to as a *gaze-averting*

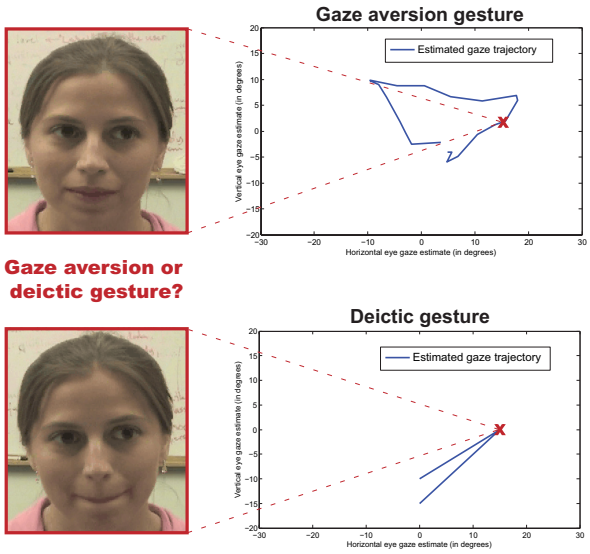

Figure 1: Comparison of a typical gaze aversion gesture (top) with a "deictic" eye movement (bottom). Each eye gesture is indistinguishable from a single image (see left images), however the eye motion patterns of each gesture are clearly different (see right plots).

gesture [12] and the eye movement of a thinker [30]. Researchers have shown that people will make gaze-averting gestures to retrieve information from memory [29] or while listening to a story [32]. Gaze aversion during conversation has been shown to be a function of cognitive load [12].

These studies of human-to-human interaction give us insight regarding the kind of gestures that could be useful for ECAs. Humans do seem to make similar gestures when interacting with an animated agent. Colburn *et al.* looked at eye contact with ECAs and found a correlation between the time people spend looking at an avatar versus the time they spend looking at another human during conversation [8].

We have observed that eye motions that attend to a specific person or object tend to involve direct saccades, while gaze aversion gestures tend to include more of a "wandering" eye motion. Looking at still images may be inconclusive in terms of deciding whether it is a gaze aversion gesture or a deictic eye movement, while looking at the dynamics of motion tends to be more discriminative (Figure 1). We therefore investigate the use of eye motion trajectory features to estimate gaze aversion gestures.

To our knowledge, no work has been done to study gaze aversion by human participants when interacting with ECAs. Recognizing such eye gestures would be useful for an ECA. A gaze aversion gesture while a person is thinking may indicate the person is not finished with their conversational turn. If the ECA senses the aversion gesture, it can correctly wait for mutual gaze to be re-established before taking its turn. In this paper, we show how to visually differentiate gaze aversion gestures from eye contact and deictic eye movements.

4. USER STUDY

Our user study was designed with two tasks in mind: (1) to observe the kind of eye gestures people make when interacting with an ECA, and (2) to evaluate how well we can recognize these ges-

tures. We built a multimodal kiosk with an interactive avatar that can perform a survey of 100 questions (see Figure 2). Sample questions asked during the user study include:

1. Are you a student?
2. Is your age an even number?
3. Do you live on campus?
4. Do you like Coke better than Pepsi?
5. Is one of the official languages in Canada Spanish?
6. Does Canada have a president?
7. Is fifteen minus five equal to nine?
8. Is five a prime number?

Our user study was composed of 6 participants: 2 men and 4 women, aged between 25-35 years old. Each interaction lasted approximately 10-12 minutes. At the beginning of the experiment, participants were informed that they would interact with an avatar who would ask them 100 questions. Participants were asked to answer every question with a positive answer (by saying "yes" and/or head nodding) or a negative answer (by saying "no" and/or head shaking) and were not aware that their eye gestures were being monitored.

The kiosk consisted of a 15.4" screen and a monocular camera with an integrated microphone placed on top of the screen. Participants sat in a chair placed 1.3 meters in front of the screen. The screen was elevated so that the eyes of the avatar were approximately at the same height as the eyes of the participant. The central software component of our kiosk consisted of a simple event-based dialogue manager that produced output using the AT&T text-to-speech engine [2] and the Haptek virtual avatar [14]. The experimenter, sitting to the right of each participant, used a remote keyboard to trigger the dialogue manager after each answer from each participant.

4.1 Results and Discussion

Since eye gaze gestures can be subtle and sometimes hard to differentiate even for a human, we asked three people to annotate the aversion gestures in the video sequences corresponding to each subject. The following definition was given to each coder for gaze-aversion gestures: eye movements to empty or uninformative regions of space, reflecting "look-away" or "thinking".

Even though the user study didn't include any explicit deictic reference to the environment around the user, human participants naturally made deictic eye gestures during the interactions. Most of these deictic gestures were targeted to the video camera or the experimenter sitting on the right side of the participant. Coders were instructed to label these deictic eye gestures as "non-gaze-aversion".

During the process of labeling and segmentation of the 6 video sequences from our user study, the 3 coders labeled 114, 72 and 125 gaze-aversion gestures, respectively. The variation between coders is not that surprising since gaze gestures are sometime subtle. We decided to define our ground truth as the intersection of all three coders, for a total of 72 gaze-aversion gestures.

The average length of gaze aversion gestures was 1.3 seconds. Since all verbal answers from the users were short "yes" or "no", waiting an extra 2 seconds for an answer may be a long time for the embodied agent. Without any visual feedback recognition, the dialog manager could potentially identify silence as a sign for misunderstanding and would repeat the question or ask the user if he/she understood the question.

Figure 2: Multimodal interactive kiosk used during our user study.

On average, our 6 participants made gaze aversion gestures 12 times per interaction with a standard deviation of 6.8. Since 100 questions were asked during each interaction, on average 12% of the time, people made a gaze aversion gesture that was labeled by all 3 coders. In our experiment, most gaze-aversions were gestures where the participant was thinking about their answer. Since our dialog manager was relatively simple, few gaze-aversion gestures had the purpose of floor-holding. We anticipate that with a more complex dialog manager and a better speech recognizer, human participants would express an even greater amount of gaze-aversion gestures.

Our results suggest that people do make gaze aversion gestures while interacting with an ECA and that it would be useful for an avatar to recognize these patterns. The question now is to know if we can recognize these eye gaze gestures. In the following section we present our eye gesture recognition framework and discuss experimental results on eye gaze aversion gesture recognition

5. EYE GESTURE RECOGNITION

Our goal is to recognize eye gestures during multimodal conversation with an embodied agent. To ensure a natural interaction, we want a recognition framework with the following capabilities:

- User-independent
- Non-intrusive
- Automatic initialization
- Robust to eye glasses
- Works with monocular cameras
- Takes advantage of other cues (e.g., head tracking)

As discussed earlier, we wish to recognize eye gestures that can help an ECA differentiate when a user is thinking from when a user is waiting for more information from the ECA. Since our goal is eye gesture and not precise eye gaze estimation, we built an eye

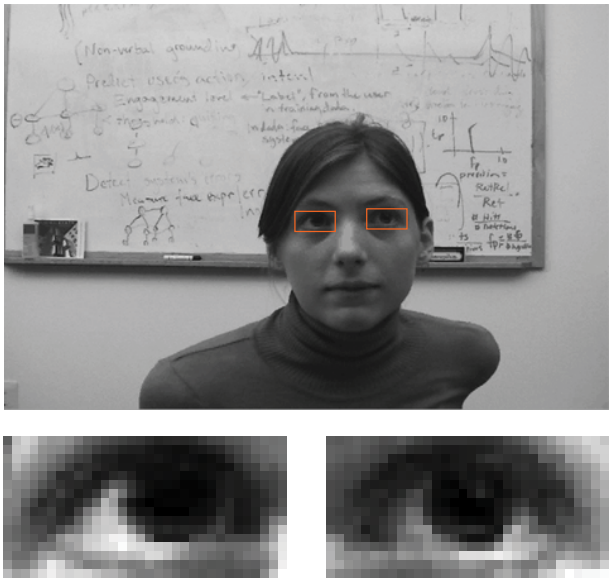

Figure 3: Example image resolution used by our eye gaze estimator. The size of the eye samples are 16x32 pixels.

gaze estimator that produces sufficient precision for gesture recognition but works with a low-resolution camera. Figure 3 shows an example of the resolution used during training and testing.

Our approach for eye gesture recognition is a four-step process: (1) detect the location of each eye in the image using a cascade of boosted classifiers, (2) track each eye location over time using a head pose tracker, (3) estimate the eye gaze based on a view-based appearance model, and (4) recognize eye gestures using a sliding time-window of eye motion estimates and a discriminative classifier.

5.1 Eye Detection

For eye detection, we first detect faces inside the entire image and then search inside the top-left and top-right quarters for the right and left eyes, respectively. Face and eye detectors were trained using a cascaded version of Adaboost [39]. For face detection, we used the pre-trained detector from Intel OpenCV.

To train our left and right eye detectors, we collected a database of 16 subjects looking at targets on a whiteboard. This dataset was also used to train the eye gaze estimator described in Section 5.3. A tripod was placed 1 meter in front of the whiteboard. Targets were arranged on a 7x5 grid so that the spacing between each target was 10 degrees (see Figure 4). The top left target represented a eye direction of -30 degrees horizontally and +20 degrees vertically. Two cameras were used to image each subject: one located in front of the target (0,0) and another in front of the target (+20,0).

Participants were asked to place their head on the tripod and then look sequentially at the 35 targets on the whiteboard. A keyboard was placed next to the participant so that he/she could press the space bar after looking at a target. The experiment was repeated under 3 different lighting conditions (see Figure 5). The location and size of both eyes were manually specified to create the training set. Negative samples were selected from the non-eye regions inside each image.

5.2 Eye Tracking

The results of the eye detector are sometime noisy due to missed detections, false-positives and jitter in the detected eye location.

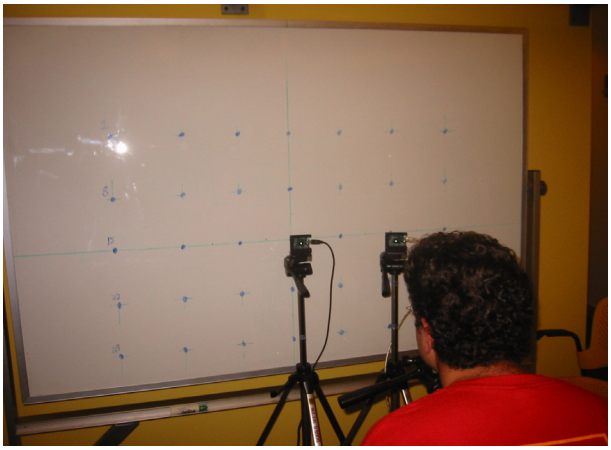

Figure 4: Experimental setup used to acquire eye images from 16 subjects with ground truth eye gaze. This dataset was used to train our eye detector and our gaze estimator.

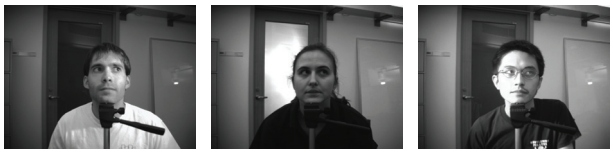

Figure 5: Samples from the dataset used to train the eye detector and gaze estimator. The dataset had 3 different lighting conditions.

For these reasons we need a way to smooth the estimated eye locations and keep a reasonable estimate of the eye location even if the eye detector doesn't trigger.

Our approach integrates eye detection results with a monocular 3D head pose tracker to achieve smooth and robust eye tracking, that computes the 3D position and orientation of the head at each frame. We initialize our head tracker using the detected face in the first frame. A 3D ellipsoid model is then fit to the face based on the width of the detected face and the camera focal length. The position and orientation of the model are updated at each frame after tracking is performed.

Our approach for head pose tracking is based on the Adaptive view-based appearance model [20] and differs from previously published ellipsoid-based head tracking techniques [3] in the fact that we acquire extra key-frames during tracking and adjust the key-frame poses over time. This approach makes it possible to track head pose over a larger range of motion and over a long period of time with bounded drift. The view registration is done using an iterative version of the Normal Flow Constraint [33].

Given the new head pose estimate for the current frame, the region of interest (ROI) around both eyes is updated so that the center of the ROI reflects the observed head motion. The eye tracker will return two ROIs per eye: one from the eye detector (if the eye was detected) and the other from the updated ROI based on the head velocity.

5.3 Gaze Estimation

To estimate the eye gaze given a region of interest inside the image, we created two view-based appearance models [27, 22], one

model for each eye. We trained the models using the dataset described in Section 5.1, which contains eye images of 16 subjects looking at 35 different orientations, ranging [-30,30] horizontally and [-20,20] vertically.

We define our view-based eigenspaces models $\mathcal{P}_l$ and $\mathcal{P}_r$, for the left and right eye respectively, as:

$$\mathcal{P} = \{\bar{I}_i, \mathcal{V}_i, \varepsilon_i\}$$

where $\bar{I}_i$ is the mean intensity image for one of the 35 views i, ε_i is the eye gaze of that view and $\mathcal{V}_i$ is the eigenspace matrix. The eye gaze is represented as $\varepsilon = [\ R^x \quad R^y\ ]$, a 2-dimensional vector consisting of the horizontal and vertical rotation.

To create the view-based eigenspace models, we first store every segmented eye image in a one-dimensional vector. We can then compute the average vectors $\bar{I}_i = \frac{1}{n}\sum_{j=1}^{n} I_i^j$ and stack all the normalized intensity vectors into a matrix:

$$\mathcal{I}_i = \left[\ \left(I_i^1 - \bar{I}_i\right)\left(I_i^2 - \bar{I}_i\right)\cdots\ \right]^T$$

To compute the eigenspaces $\mathcal{V}_i$ for each view, we find the SVD decomposition $\mathcal{I}_i = U_i D_i \mathcal{V}_i^T$.

At recognition time, given a seed region of interest (ROI), our algorithm will search around the seed position for the optimal pose with the lowest reconstruction error e_i^*. For each region of interest and each view of the appearance model i, the reconstruction error is computed:

$$e_i = |I_t' - \bar{I}_i - \vec{w}_i \cdot \mathcal{V}_{I_i}|^2, \tag{1}$$

The lowest reconstruction error e_i^* will be selected and the eye gaze ε_i associated with the optimal view i will be returned. In our implementation, the search region was [+4,-4] pixels along the X axis and [+2,-2] pixels along the Y axis. Also, different scales for the search region were tested during this process, ranging from 0.7 to 1.3 times the original size of the ROI.

Gaze estimation was done independently for each seed ROI returned by the eye tracker described in the previous section. ROIs associated with the left eye are processed using the left view-based appearance model and similarly for the right eye. If more then one seed ROI was used, then the eye gaze with the lowest reconstruction error is selected. The final eye gaze is approximated based on a simple average of the left and right eye gaze estimates.

To test the accuracy of our eye gaze estimator we ran a set of experiments using the dataset described earlier in Section 5.1. In these experiments, we randomly selected 200 images, then retrained the eye gaze estimator and compared the estimated eye gaze with the ground truth estimate.

We tested two aspects of our estimator: its sensitivity to noise and its performance using different merging techniques. To test our estimator's sensitivity to noise, we added varying amounts of noise to the initial region of interest. Figure 6 shows the average error on the eye gaze for varying levels of noise. Our eye gaze estimator is relatively insensitive to noise in the initialized region of interest, maintaining an average error of under 8 degrees for as much as 6 pixel noise.

We also tested two techniques to merge the left and right eye gaze estimates: (1) picking the eye gaze estimate from the eye with the lowest reconstruction error and (2) averaging the eye gaze estimates from both eyes. Figure 6 also summarizes the result of this experiment. We can see that the averaging technique consistently works better than picking the lowest reconstruction error. This result confirms our choice of using the average to compute eye gaze estimates.

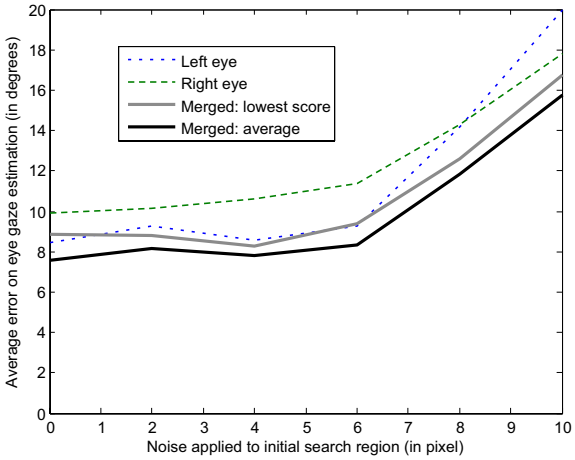

Figure 6: Average error of our eye gaze estimator when varying the noise added to the initial region of interest.

5.4 Gesture Recognition

Our goal is to recognize gaze aversion gestures given a video sequence. We have found that dynamic cues can differentiate such eye gestures from eye contact and deictic gestures (Figure 1). For this reason, our input features for the gesture recognition algorithm are based on temporal windows of eye gaze observations.

In our implementation we use a supervised discriminative classification approach: the Support Vector Machine (SVM). Our choice of a discriminative approach over a generative approach like Hidden Markov Models (HMMs) was influenced by our previous work on head gesture recognition, where SVMs were shown to outperform HMMs when trying to differentiate subtle head nods from other natural gestures of a human participant interacting with an ECA [21]. Gaze aversion gestures can be very subtle to differentiate from other natural eye gestures, as confirmed by the large variability in the human coder results in Section 4.1.

In order to create training data, we first ran the eye gaze estimator described in the previous sections and obtained an estimate of the eye gaze for each frame. The data was labeled with the start and end points of each gesture. We concatenate eye gaze estimates over a fixed window size to create a feature vector x, and shift the window through all of the video sequences to create training samples. In our implementation the window size was set to 20 frames ($\sim$0.7 seconds for a frame rate of 30Hz). Training samples were labeled as positive if the start and end times were inside a labeled gaze aversion gesture. Otherwise, the training sample was labeled as negative. The training set had the same number of negative and positive samples.

After training of the SVM, the margin $m(x)$ of the feature vector x can easily be computed given the learned set of support vectors x_i, the associated set of labels y_i and weights w_i, and the bias b:

$$m(x) = \sum_{i=1}^{l} y_i w_i K(x_i, x) + b \tag{2}$$

where l is the number of support vectors and $K(x_i, x)$ is the kernel function. In our experiments, we used a radial basis function (RBF) kernel:

$$K(x_i, x) = e^{-\gamma \|x_i - x\|^2} \tag{3}$$

where γ is the kernel smoothing parameter learned automatically using cross-validation on our training set.

At testing, we create a feature vector x for each frame of the video sequence and a frame is labeled as a gesture if the margin $m(x)$ is larger then a threshold k. The following section shows the performance of our gesture recognizer.

5.5 Results and Discussion

The eye gaze estimator and eye gesture recognizer were applied to unsegmented video sequence of all 6 human participants from the user study described in Section 4. Each video sequence lasted approximately 10-12 minutes, and was recorded at 30 frames/sec, for a total of 105,743 frames. During these interactions, human participants would rotate their head up to +/-70 degrees around the Y axis and +/-20 degrees around the X axis, and would also occasionally translate their head, mostly along the Z axis. The following eye gesture recognition results are on the unsegmented sequences, including extreme head motion.

First, it is interesting to look at the qualitative accuracy of the eye gaze estimator for a sample sequence of images. Figure 8 illustrates a gaze aversion gesture where the eye gaze estimates are depicted by cartoon eyes and the head tracking result is indicated by a white cube around the head. Notice that the estimator works quite well even if the participant is wearing eye glasses.

To analyze the performance of our eye gesture recognizer, we performed a leave-one-out experiment where we retrained the gesture recognizer with 5 out of 6 participants and tested the trained recognizer with the sixth participant. Figure 7 shows the ROC curves for all 6 participants in gray and the average ROC curve in black.

We computed the true positive rate by dividing the number of recognized gestures by the total number of ground truth gestures. An aversion eye gesture is tagged as recognized if the recognizer triggered at least once during a time window around the aversion eye gesture. The false positive rate is computed by dividing the number of falsely recognized frames by the total number of non-gesture frames. A frame is tagged as falsely recognized if the eye gesture recognizer triggers and if this frame is outside any time window of a ground truth aversion eye gesture. The denominator is the total number of frames outside any time window.

With a small false positive rate of 0.02, our eye gesture recognizer correctly labeled more then 87% of all aversion eye gestures. This result shows that we can accurately recognize gaze-aversion eye gestures during interactions with an embodied conversational agent.

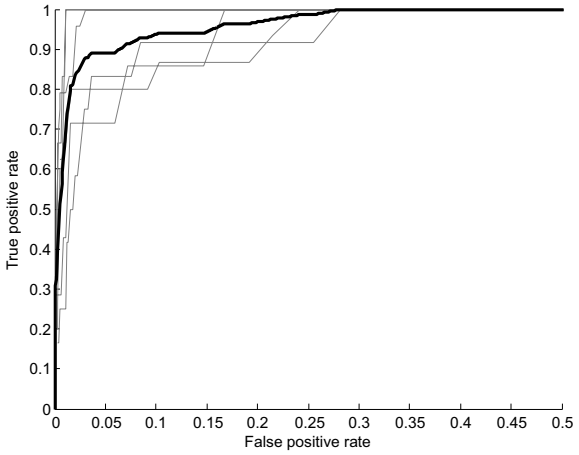

Figure 7: Accuracy of our eye gesture recognizer. This plot shows the leave-one-out ROC curves for all 6 participants (gray) and the average ROC curve (black).

6. CONCLUSION

We introduced an automatic system that allows conversational agents to detect gaze aversion gestures in human conversational partners. We found that human participants avert their gaze (i.e. perform "look-away" or "thinking" gestures) to hold the conversational floor even while answering relatively simple questions. Using our approach, an agent could visually differentiate whether a user is thinking about a response or is waiting for the agent or robot to take its turn. Interesting avenues of future work include extending our user study to interactions with a more complex dialog manager and the use of dialogue context for improving eye gesture recognition.

7. REFERENCES

[1] M. Argyle and M. Cook. *Gaze and mutual gaze*. Cambridge University Press, 1976.

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

[3] S. Basu, I. Essa, and A. Pentland. Motion regularization for model-based head tracking. In *In Intl. Conf. on Pattern Recognition (ICPR '96)*, 1996.

[4] 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.

[5] 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.

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

[7] M.R.M. Mimica C.H. Morimoto. Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, 98:52–82, 2005.

[8] R. Alex Colburn, Michael F. Cohen, and Steven M. Ducker. The role or eye gaze in avatar mediated conversational interfaces. Technical Report MSR-TR-2000-81, Microsoft Research, July 2000.

[9] 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.

[10] 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.

[11] Atsushi Fukayama, Takehiko Ohno, Naoki Mukawa, Minako Sawaki, and Norihiro Hagita. Messages embedded in gaze of interface agents — impression management with agent's gaze. In *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 41–48, 2002.

[12] A. M. Glenberg, J. L. Schroeder, and D.A. Robertson. Averting the gaze disengages the environment and facilitates remembering. *Memory and cognition*, 26(4):651–658, July 1998.

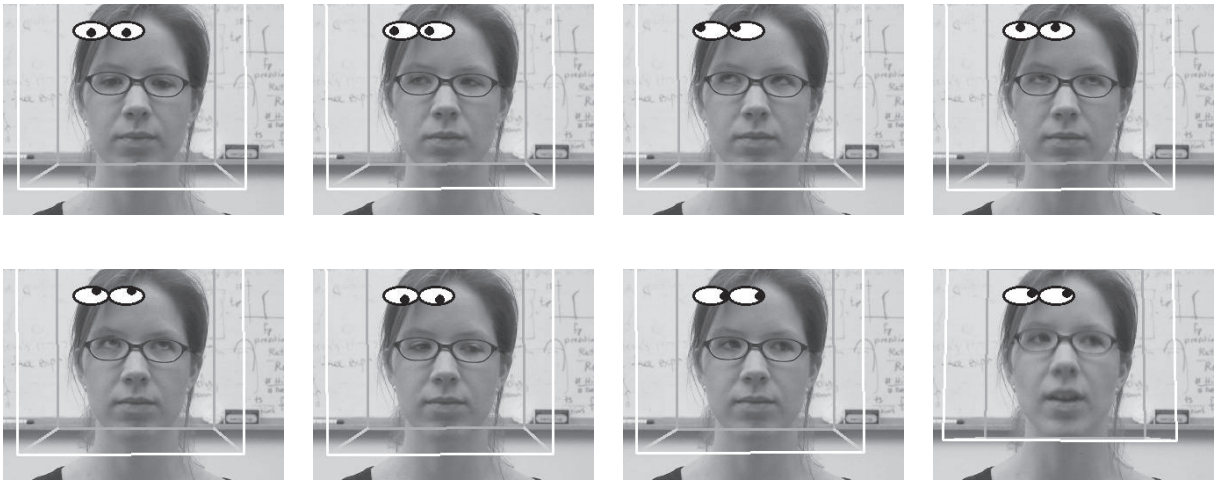

Figure 8: Eye gaze estimation for a sample image sequence from our user study. The cartoon eyes depict the estimated eye gaze. The cube represents the head pose computed by the head tracker.

[13] Z. M. Griffin and K. Bock. What the eyes say about speaking. *Psychological Science*, 11(4):274–279, 2000.

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

[15] Craig Hennessey, Borna Noureddin, and Peter Lawrence. A single camera eye-gaze tracking system with free head motion. In *ETRA '06: Proceedings of the 2006 symposium on Eye tracking research & applications*, pages 87–94, 2006.

[16] A. Kendon. Some functions of gaze direction in social interaction. *Acta Psyghologica*, 26:22–63, 1967.

[17] Sooha Park Lee, Jeremy B. Badler, and Norman I. Badler. Eyes alive. In *SIGGRAPH '02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 637–644, 2002.

[18] Michael Li and Ted Selker. Eye pattern analysis in intelligent virtual agents. In *Conference on Intelligent Virutal Agents(IVA02)*, pages 23–35, 2001.

[19] P. Majaranta and K. J. Raiha. Twenty years of eye typing: systems and design issues. In *ETRA '02: Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 15–22, 2002.

[20] Louis-Philippe Morency, Ali Rahimi, and Trevor Darrell. Adaptive view-based appearance model. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 803–810, 2003.

[21] 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.

[22] Louis-Philippe Morency, Patrik Sundberg, and Trevor Darrell. Pose estimation using 3d view-based eigenspaces. In *ICCV Workshop on Analysis and Modeling of Faces and Gestures*, pages 45–52, Nice, France, 2003.

[23] 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.

[24] D.G. Novick, B. Hansen, and K. Ward. Coordinating turn-taking with gaze. In *Proceedings of the Fourth International Conference on Spoken Language*, volume 3, pages 1888–1891, 1996.

[25] Takehiko Ohno and Naoki Mukawa. A free-head, simple calibration, gaze tracking system that enables gaze-based interaction. In *ETRA '04: Proceedings of the 2004 symposium on Eye tracking research & applications*, pages 115–122, 2004.

[26] C. Pelachaud and M. Bilvi. Modelling gaze behavior for conversational agents. In *International Working Conference on Intelligent Virtual Agents*, pages 15–17, September 2003.

[27] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, June 1994.

[28] Pernilla Qvarfordt and Shumin Zhai. Conversing with the user based on eye-gaze patterns. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 221–230, 2005.

[29] Daniel C. Richardson and Michael J. Spivey. Representation, space and hollywood squares: looking at things that aren't there anymore. *Cognition*, 76:269–295, 2000.

[30] D.C. Richardson and R. Dale. Looking to understand: The coupling between speakers and listeners eye movements and its relationship to discourse comprehension. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, pages 1143–1148, 2004.

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

[32] M. J. Spivey and J. J. Geng. Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological Research/Psychologische Forschung*, 65(4):235–241, 2001.

[33] G. Stein and A. Shashua. Direct estimation of motion and extended scene structure from moving stereo rig. In *Proc. of CVPR*, June 1998.

[34] R. Stiefelhagen, J. Yang, and A. Waibel. Tracking eyes and monitoring eye gaze. In *Proceedings of the Workshop on*

Perceptual User Interfaces (PUI'97), pages 98–100, Alberta, Canada, 1997.

[35] D. Traum and J. Rickel. Embodied agents for multi-party 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.

[36] David Traum and Jeff Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *AAMAS '02: Proceedings of the first international joint conference on Autonomous agents and multiagent systems*, pages 766–773, 2002.

[37] B. M. Velichkovsky and J. P. Hansen. New technological windows in mind: There is more in eyes and brains for human-computer interaction. In *CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*, 1996.

[38] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 301–308, 2001.

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

[40] X. Xie, R. Sudhakar, and H. Zhuang. A cascaded scheme for eye tracking and head movement compensation. *IEEE Transactions on Systems, Man and Cybernetics*, 28(4):487–490, July 1998.

[41] L. Young and D. Sheena. Survey of eye movement recording methods. *Behavior Research Methods and Instrumentation*, 7:397–429, 1975.

[42] S. Zhai, C. Morimoto, and S. Ihde. Manual and gaze input cascaded (magic) pointing. In *CHI99*, pages 246–253, 1999.