Anticipating Averted Gaze in Dyadic Interactions

Philipp Müller
Max Planck Institute for Informatics
Saarland Informatics Campus, Germany
pmueller@mpi-inf.mpg.de

Ekta Sood
Institute for Visualisation and Interactive Systems
University of Stuttgart, Germany
ekta.sood@vis.uni-stuttgart.de

Andreas Bulling
Institute for Visualisation and Interactive Systems
University of Stuttgart, Germany
andreas.bulling@vis.uni-stuttgart.de

ABSTRACT

We present the first method to anticipate averted gaze in natural dyadic interactions. The task of anticipating averted gaze, i.e. that a person will not make eye contact in the near future, remains unsolved despite its importance for human social encounters as well as a number of applications, including human-robot interaction or conversational agents. Our multimodal method is based on a long short-term memory (LSTM) network that analyses non-verbal facial cues and speaking behaviour. We empirically evaluate our method for different future time horizons on a novel dataset of 121 YouTube videos of dyadic video conferences (74 hours in total). We investigate person-specific and person-independent performance and demonstrate that our method clearly outperforms baselines in both settings. As such, our work sheds light on the tight interplay between eye contact and other non-verbal signals and underlines the potential of computational modelling and anticipation of averted gaze for interactive applications.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**.

KEYWORDS

averted gaze, social interactions, anticipatory human-computer interaction

ACM Reference Format:
Philipp Müller, Ekta Sood, and Andreas Bulling. 2020. Anticipating Averted Gaze in Dyadic Interactions. In *Symposium on Eye Tracking Research and Applications (ETRA '20 Full Papers), June 2–5, 2020, Stuttgart, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3379155.3391332

1 INTRODUCTION

Gaze is a central non-verbal cue in social interactions, being connected to many fundamental aspects in conversations, including turn-taking [Kendon 1967], perception of affective state [Adams Jr and Kleck 2003], attraction [Kellerman et al. 1989] and leadership [Capozzi et al. 2019; Müller and Bulling 2019]. One particularly important aspect of gaze in conversations is the presence of averted gaze, which has been shown to be connected to cognitive load [Glenberg et al. 1998], intimacy-modulation [Abele 1986] and floor management [Kendon 1967].

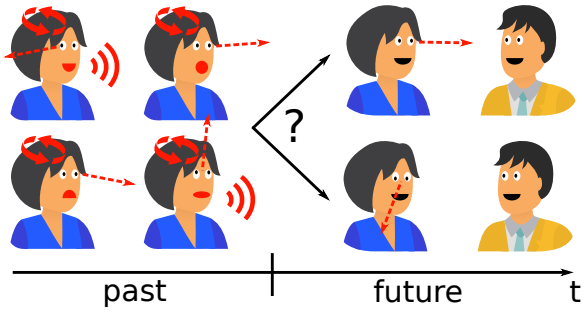

Figure 1: We study the challenging task of averted gaze anticipation in conversations: Given past observation of a person's gaze, head pose, facial expressions and speaking behaviour, we predict averted gaze in the near future.

Recent advances in gaze estimation and eye contact detection make it possible to automatically detect averted gaze, providing valuable input to a number of potential applications in human-robot interaction and assistive systems [Müller et al. 2018; Zhang et al. 2017b]. While these current methods focus on predicting gaze behaviour in the present, the ability to anticipate future states of gaze in conversations is essential to enable systems to proactively manage user attention. For example, if a robot detects that its interlocutors' gaze is going to be averted when it is about to initiate an important action, it can either catch the users attention by an expressive gesture, or delay the onset of the action in order to be less obtrusive. Furthermore, new possibilities for assisting human-human interactions open up by the ability to forecast eye contact. For example proactive feedback could help people having difficulty to maintain socially accepted eye contact behaviour (e.g. people with autism spectrum disorder [Senju and Johnson 2009]).

While first works explored anticipation of visual behaviour in egocentric video [Zhang et al. 2017c] and mobile device interactions [Steil et al. 2018], gaze anticipation in human-human interactions remains completely unexplored. We fill this gap by proposing the first method to anticipate averted gaze in natural dyadic conversations, i.e. to predict whether gaze will be averted in the near future (see Figure 1 for an illustration of the prediction task). Our method consists of a long short-term memory (LSTM) network [Hochreiter and Schmidhuber 1997] which takes as input a slice of prior conversation and outputs whether the interactants' gaze will be mostly averted or not during a subsequent future time interval. We exploit the dependence of subsequent states of eye contact on previous

ETRA '20 Full Papers, June 2–5, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-7133-9/20/06…$15.00
https://doi.org/10.1145/3379155.3391332

eye contact, gaze, head pose, and facial expressions as well as the well-known link between speaking and eye contact [Kendon 1967].

The specific contributions of our work are two-fold. First, we propose the first method to forecast eye contact in dyadic conversations based on the observation of preceding visual and speaking behaviour. Second, we evaluate our method on a newly collected dataset of natural interactions over video conferencing, annotated with eye contact information on 23,131 frames. We show consistent improvements of our method over several baselines. The dataset is publicly available at https://www.perceptualui.org/datasets/DEyeAdicContact and we hope it can become a valuable resource for research on eye contact detection and anticipation.

2 RELATED WORK

Our work is related to previous research on 1) the importance of gaze in social conversations, 2) computational methods for learning-based gaze estimation and eye contact detection, as well as 3) methods for gaze behaviour prediction and anticipation.

2.1 Gaze in Social Conversations

A large body of work has demonstrated the fundamental importance of gaze in conversations. Early work showed that gaze is an important cue in turn-taking [Kendon 1967; Rossano 2013] and coordinates the insertion of responses [Bavelas et al. 2002]. More recently, [Ho et al. 2015] found that while speakers likely gaze at their interlocutor at the end of speaking turns, listeners begin speaking with averted gaze. Furthermore, Glenberg et al. [1998] found a connection between averted gaze and cognitive load by giving participants questions of varying difficulty. The results showed that the frequency of averted gaze was higher with larger cognitive load, and averted gaze also led to better task performance.

Social gaze has also been studied extensively together with other social signals, such as facial expressions [Adams Jr and Kleck 2003; Ekman 1992; Zuckerman et al. 1981]. A key finding is that coordinated gaze behaviour and facial cues can denote affective states, such as avoidance-oriented emotions (e.g., fear and sadness)" [Adams Jr and Kleck 2003]. Another line of work has explored the intimate relationship between gaze and speech [Argyle and Cook 1976; Jokinen et al. 2010; Maglio et al. 2000; Streeck 1993]. For example, the tone of prosodic features and gaze direction was shown to denote emotional states (e.g if someone is angry they might raise their voice and look in the direction of a target) [Hamilton 2016]. Jokinen et al. [2010] leveraged gaze information to better predict turn taking, particularly the time windows for alignment in conversational/naturalistic speech while Müller et al. [2018] combined gaze and speech to improve eye contact detection in group interactions. Finally, recent work combined features based on peoples' visual focus of attention with facial expressions or body pose features to detect leadership [Beyan et al. 2017; Müller and Bulling 2019] or rapport [Müller et al. 2018] in group interactions.

2.2 Gaze Estimation and Eye Contact Detection

Analysing social gaze in conversations either requires specialised mobile eye tracking equipment [Kassner et al. 2014; Tonsen et al. 2017] or computational methods for gaze estimation and eye contact detection from off-the-shelf RGB cameras – the latter research area in computer vision has received particular attention in recent years. Gaze estimation methods can be roughly divided in model-based and appearance-based [Hansen and Ji 2009]: While model-based approaches use a geometric model of the human eye to perform gaze estimation [Valenti et al. 2011; Wood et al. 2015; Yamazoe et al. 2008], appearance based methods directly regress the gaze from the image input [Lu et al. 2012; Zhang et al. 2019]. In contrast to gaze estimation, eye contact detection is the task of predicting a binary label of whether gaze is on a specific target (person, object) or not. While Smith et al. [2013] detected eye contact with the camera an image was taken from, Zhang et al. [2017b] were the first to propose a more general method that was able to detect eye contact with a salient object close to the camera. This method was subsequently generalised to discriminate multiple eye contact targets during group interactions [Müller et al. 2018].

Our work is fundamentally different from these approaches given that their aim is to detect eye contact only *in the present moment* while we present the first method to *anticipate future eye contact* in conversations, particularly averted gaze.

2.3 Gaze Behaviour Prediction and Anticipation

While the previously discussed methods require an image of the target person to estimate gaze and predict eye contact, a parallel line of research explores methods to predict gaze behaviour without such information. One of the most common tasks is to predict *saliency maps*, that is person-independent, two-dimensional heatmaps indicating at which locations in an image people are most likely to look [Harel et al. 2007; Itti et al. 1998; Kümmerer et al. 2016] or user interface [Xu et al. 2016]. In contrast to saliency prediction, *scanpath prediction* attempts to predict sequences of plausible fixations for a given image [Assens Reina et al. 2017; Liu et al. 2013]. Both tasks, however, assume a fixed input image as stimulus as well as a free-viewing task, disregarding the effects of context and top-down influences on gaze behaviour. Borji et al. [2013] introduced such top-down effects by modelling gaze behaviour during driving in a computer game. More recently, there has also been interest in predicting gaze on egocentric videos – a task that requires the system to integrate bottom-up as well as top-down factors across time [Huang et al. 2018; Li et al. 2018; Zhang et al. 2018].

Only few previous works explored the even more challenging task of anticipating future gaze behaviour. Zhang et al. predicted future gaze in egocentric videos by generating future video frames and predicting temporal saliency on these [Zhang et al. 2017c]. Conceptually most similar, albeit in a different setting and using fundamentally different information, is recent work by Steil et al. on attention forecasting [Steil et al. 2018]. There, the authors focused on attention anticipation during everyday mobile interactions by combining visual scene information from a head-mounted camera with information on app usage and device-integrated mobile phone sensors. They demonstrated that imminent shifts of attention to and away from the phone, as well as the future primary attentional focus could be robustly predicted in a wide variety of mobile settings.

To the best of our knowledge, our work is the first to study attention forecasting, particularly anticipating averted gaze behaviour, in everyday conversations from multimodal social signals.

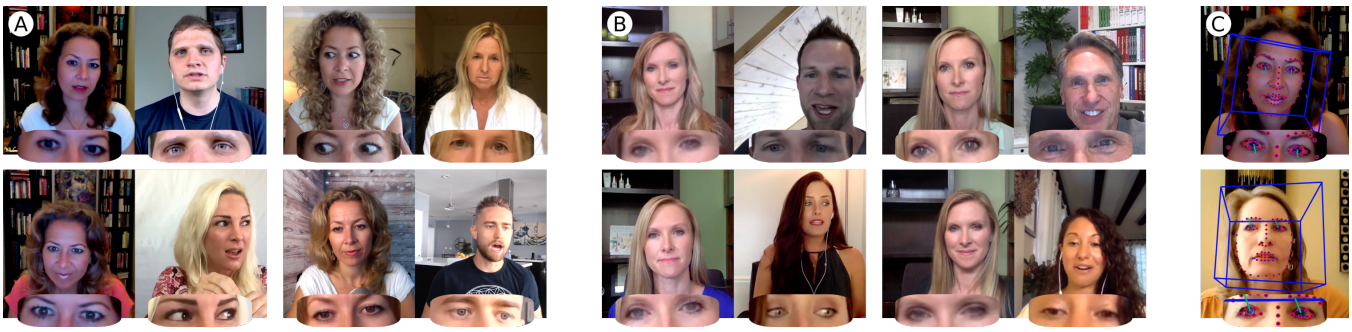

Figure 2: Example images from the dataset with enlarged eye regions for better visibility. A: Images from the Youtube channel "Wisdom From North", B: Images from the Youtube channel "The Spa Dr." For each image, the host is shown on the left and the guest on the right. C: Examples of head pose estimates, keypoint detections and gaze estimates obtained from OpenFace. Image courtesy Jannecke Øinæs and Dr. Trevor Cates.

3 DATASET

To the best of our knowledge, currently no dataset of natural dyadic interactions with fine-grained eye contact annotations exists. We therefore created our own dataset using videos of dyadic interviews published on YouTube. Especially compared to lab-based recordings, these Youtube interviews allow us to analyse behaviour in a natural situation. All interviews were conducted via video conferencing and provide frontal views of interviewer and interviewee side-by-side. Specifically, we downloaded videos from the YouTube channels "Wisdom From North" and "The Spa Dr." that both provide a large number of interviews, often with a high video quality. Videos from the channel "Wisdom From North" have already been utilised in research on facial expression generation [Feng et al. 2017]. While "Wisdom From North" is concerned with spiritual topics, "The Spa Dr." focuses on health and beauty. Each channel features a single host interviewing different guests in each session. We manually selected videos with high video quality, resulting in 60 videos for "The Spa Dr." and 61 videos for "Wisdom From North". All videos are recorded at a frame rate between 24 and 30 fps and vary in length from 17 minutes to 58 minutes (average: 37 minutes). In total the videos contain 74 hours of conversations, amounting to 7,817,821 video frames. Figure 2 shows example images from both Youtube channels. The natural and unconstrained behaviour of interactants comes hand-in-hand with challenges for obtaining accurate eye contact ground truth. In particular, the geometric relation between interactant, camera and screen on which the interlocutor is visible in the interactants' view changes between videos. For example, while both guests in the top two images in Figure 2 B have eye contact with their interlocutors, different camera- and screen positions lead to different gaze directions. In the following, we discuss how we tackle this challenge by semi-automatic gaze annotation.

3.1 Gaze Annotation

We instructed five human annotators to classify the gaze of interviewer and interviewee (in the following referred to as "subjects"). Even though in this study we were only interested in a binary classification of averted gaze versus eye contact, a more fine-grained distinction of averted gaze might prove beneficial for future research. To this end we used in total 11 mutually exclusive classes during annotation. Annotators were asked to select the class "eye contact" if the subject was looking at the location of the other person on her screen or the camera from which she was recorded. We found that annotators were able to reliably determine the placements of camera and screen by skimming through the video prior to starting the annotation. If there was no eye contact, annotators classified whether the subject gazed "up","down", "left", "right", or to the "upper left", "lower left", "upper right" or "lower right". In the following, we refer to the union of these classes as the "no eye contact class". A separate class was dedicated to blinks, while yet another class indicated instances in which annotators were unsure about how to decide, e.g. as a result of low image quality. As annotators worked on disjoint sets of videos, one of the authors was present throughout the first sessions in order to ensure consistency.

To strike a good balance between sufficient coverage and annotation effort, we collected these annotations on a frame-by-frame basis every 30 seconds for the Wisdom From North interviews, and every 15 seconds for The Spa Dr. interviews. We collected annotations for The Spa Dr. on a finer timescale given that the host of that channel almost always keeps eye contact with her interviewees. A coarser time scale would have increased the risk of missing the no eye contact classes in the annotation. In total, we collected 23,131 annotated video frames of which 83% were labelled as "eye contact".

3.2 Semi-automatic Eye Contact Annotation

Annotating such a large dataset on a frame-by-frame basis completely manually is impractical. We therefore designed a semi-automatic method to annotate every frame in the videos by combining the sparse human annotations with eye contact labels calculated using gaze estimates from OpenFace [Baltrusaitis et al. 2018] (see Figure 2 C for an illustration of OpenFace output).

3.2.1 Preprocessing of the gaze estimates. We observed that blinks create artifacts in the OpenFace gaze estimates, as gaze estimates rapidly switch to "looking down" and back to the original position. To remove these artifacts, we first apply a median filter with a width of 0.4 seconds. We chose 0.4 seconds because this represents the typical duration of a blink and it effectively removes the artifacts. Afterwards, we project the gaze estimates on the 2D camera plane.

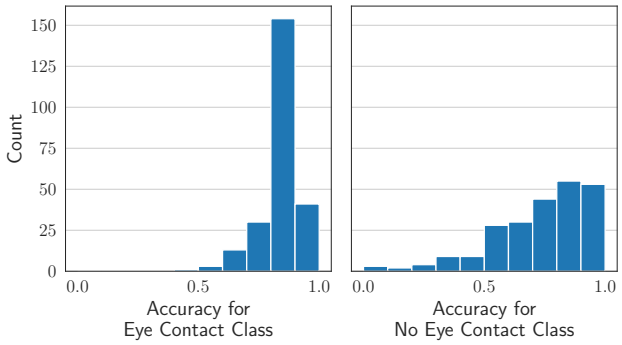

Figure 3: Left: Histogram of accuracies of our semi-automatic eye contact detection approach obtained on ground truth eye contact samples. Right: Corresponding accuracies obtained on ground truth no eye contact samples.

3.2.2 Eye contact classification. The core idea of our method is to extract regions of "eye contact" and "no eye contact" in the space of gaze estimates described before. To this end, our method first computes the convex hull C of all gaze estimates corresponding to "eye contact" annotations. Due to noise in the gaze estimation, C can be too large and encompass regions that correspond to "no eye contact" annotations. To address this issue, we incorporate "no eye contact" annotations in a second step. Specifically, we use kernel density estimation to approximate the distribution of gaze estimates during eye contact p_e as well as the distribution of gaze estimates when there is no eye contact $p_{\neg e}$. Areas within C for which $p_{\neg e} > p_e$, that is, for which there is more probability mass in the "no eye contact" distribution than in the "eye contact" distribution are re-labelled as "no eye contact".

3.2.3 Evaluating the semi-automatic annotations. We evaluated this eye contact annotation approach using leave-one-annotation-out cross-validation for each video and interactant separately. That is, for a given interaction for which we recorded n annotations for interactant i, we used one annotation as test annotation and computed eye contact annotations from the remaining $n-1$ annotations as discussed before. We cycle through all possible test annotations to compute the accuracy of the semi-automatic eye contact annotations on that particular interactant. As the classes are highly imbalanced, we compute accuracies for the eye contact class and the no eye contact class separately.

Using this approach and after averaging the accuracies obtained for each interactant in each interaction, we obtain an overall accuracy of 0.84 for ground truth "eye contact" frames, and an accuracy of 0.74 for ground truth "no eye contact" frames. Figure 3 shows the overall distribution of the accuracies obtained for each interactant in each interaction. As can be seen from the figure, while most accuracies fall into the higher regions, there is a number of very low accuracies. When using our semi-automatic eye contact annotations for analyses or evaluations on the dataset, it is therefore important to exclude these interactions that achieved only low accuracy in the cross-validation evaluation.

4 METHOD

Figure 4 provides an overview of our proposed method to anticipate averted gaze. At its core is a recurrent neural network with long short-term memory (LSTM) units [Hochreiter and Schmidhuber 1997]. Inputs to the network are provided at each timestep for a feature window w_f. At the last timestep, the network outputs a classification score for gaze aversion on the target window w_t. In the following, we describe the extraction of features and provide details on the prediction method.

4.1 Feature Extraction

We extract visual features from the person for which we want to predict averted gaze (in the following also referred to as "target person"), including eye contact, raw gaze, head pose and facial expressions, and speaking status. We do not extract features from the interactant, as they were not effective in preliminary experiments.

4.1.1 Visual Features. We use OpenFace 2.0 [Baltrusaitis et al. 2018] to extract features from the interactants' facial behaviour. In detail, we extract the following sets of features:

- *AUs*: intensities of all 17 facial action units available in OpenFace (17 dimensions)
- *HeadPose*: location and orientation of the head in camera coordinates (2×6 dimensions)
- *Gaze*: gaze estimates obtained by OpenFace projected on the camera plane (2 dimensions)
- *EyeCont*: eye contact detections obtained as described in Section 3.2 (one-hot encoding, 2 dimensions).

4.1.2 Multimodal Speaker Diarisation. We further include a one-dimensional feature that indicates whether the target person is speaking at a particular moment in time (*SpeakDiar*). To this end, we perform speaker diarisation using the pyAudioAnalysis toolkit [Giannakopoulos 2015] and subsequently employ facial action unit information to increase its robustness. The approach taken by pyAudioAnalysis uses latent discriminant analysis (LDA) to reduce the dimensionality of speech features. The method first clusters speech data into a user-defined number of classes (in our case 2) and finally uses a hidden Markov Model (HMM) for smoothing. While this approach worked well on our data, some instances remained in which the speaker prediction erroneously switched away from the current speaker for a small number of seconds, only to switch back afterwards. We address this issue by incorporating visual information to check for the plausibility of short speaker switches. In detail, we make use of a visual speaking indicator based on the sum of the standard deviations of facial action units 25 (lips part) and 26 (jaw drop) as described in [Müller et al. 2018]. Given this speaking indicator, we check all switches in speaker diarisation lasting less than five seconds. The idea is that if the switch from a speaker i to a speaker j in the speaker diarisation class is correct, it should also correspond to a switch in the visual speaking activity indicator in such a way that the visual speaking indicator for i is lower during the switch as compared to before and after the switch, and the visual speaking indicator for j is higher during the switch as compared to before/after. If this is not the case, we ignore the switch in the speaker diarisation, assuming i to be the speaker throughout.

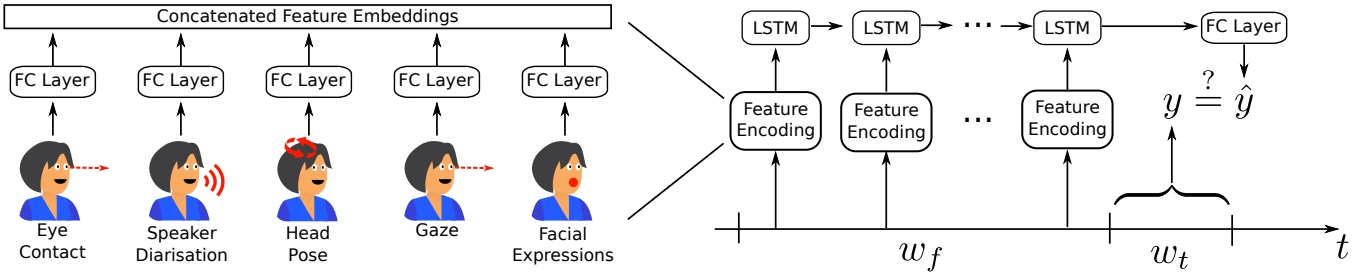

Figure 4: Overview of our eye contact anticipation method. Left: In the feature encoding network, each feature modality is fed through a fully connected layer (FC Layer) separately and the resulting representations are concatenated. Right: features are extracted on a feature window w_f and fed through an embedding network consisting of a fully connected layer for each timestep separately, before they are fed to a LSTM network. At the last timestep of the feature window the LSTM outputs a classification score which is compared to ground truth extracted from the target window w_t.

4.2 Prediction Method

As a first step in our LSTM-based method, each feature channel is embedded into a 16-dimensional space for each timestep separately using a fully-connected layer with ReLU nonlinearities. Subsequently, these embedding vectors are concatenated and fed into a LSTM layer with 32 hidden units and ReLU activation functions. At the final timestep, a dense layer with softmax activation functions is applied to obtain a classification score. We train our models using categorical cross entropy between softmax output and ground truth and add a $l2$-regulariser of 0.001. The learning rate is adjusted dynamically by Adam [Kingma and Ba 2014].

We evaluate our models using 10-fold cross validation. In each iteration, 10 percent of the data are used as test data, another 10 percent as validation data and the rest as training data. For splitting the data into training, validation and testing sets, we make sure that data from one interaction only appears in one of the three sets. For a given train/val/test split, we train the model for 100 epochs and select the model weights achieving best performance on the validation data for evaluation on the test data.

5 EVALUATION

The task of anticipating averted gaze from multimodal non-verbal cues involves extracting features on a feature window w_f and predicting whether gaze is mostly averted on a subsequent target window w_t. For our LSTM network, we discretised time into segments of 200ms given that this is approximately the length of short fixations [Salthouse and Ellis 1980]. As different application scenarios may require anticipation of averted gaze on different time horizons, we evaluated a range of different sizes of the target window w_t, including 0.2, 0.6, 1, 2, 3, 4 and 5 seconds. The gaze aversion ground truth is obtained by thresholding the probability of eye contact according to our semi-automatic eye contact annotations on w_t. In case this probability is larger than 0.5, the sample belongs to the "gaze aversion" class, and to the background class otherwise. We use a length of the feature window w_f of 6.4 seconds, consisting of 32 timesteps of 200ms each, as this feature window length led to the best performance in preliminary experiments.

We investigated performance of models trained and tested on a single person ("person-specific" evaluation) as well as when trained on several persons and tested on a disjoint set of other persons ("person-independent" evaluation). For person-specific evaluation, we exploited that the same "Wisdom From North" host appears in 61 videos but interviewed different guests each time. For the person-independent evaluation, we anticipated averted gaze of the guests of both YouTube channels because they differ in every video. Given that classes are highly imbalanced on both prediction tasks, with averted gaze being the minority, we chose to evaluate our method using average precision. Average precision evaluates a ranking of test examples obtained from the classifier by computing the average of the precisions obtained at all recall levels. While a classifier outputting the negative class only would be able to achieve high accuracy on such an imbalanced class distribution as ours, its average precision would be very low.

5.1 Data Selection

In order to train our and evaluate our models with accurate ground truth, we selected subsets of the whole dataset for which our semi-automatic eye contact annotation method achieved at least an accuracy of 0.7 both on the eye contact and no eye contact class for the person for which we want to anticipate averted gaze. For the person-specific evaluation this resulted in 51 out of 61 videos (32 hours) from "Wisdom From North" and an average accuracy for eye contact detection of 0.87 on the eye contact class and 0.90 on the no eye contact class. We did not conduct a person-specific evaluation for "The Spa Dr." because only 21 of 60 videos would have been included with our accuracy-based selection criterion. For the person-independent case this resulted in 76 of 121 videos (46 hours) from both channels, reaching an average accuracy 0.85 on the eye contact class and 0.83 on the no eye contact class.

5.2 Baselines

The first baseline we evaluated against is one that outputs a random permutation of test examples. That is, the performance of this *random baseline* in terms of average precision is equal to the rate of positive examples, i.e. the probability of averted gaze on the target time window. To be able to judge the performance of our method more thoroughly, we used the eye contact information on the feature window to design two baselines which are significantly

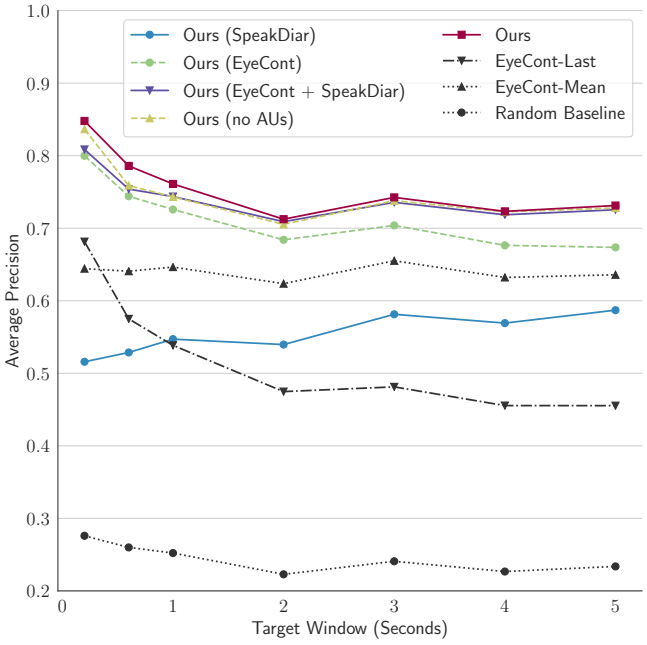

Figure 5: Average precision achieved in the person-specific evaluation for different feature channel ablations of our method and baselines across different target window sizes.

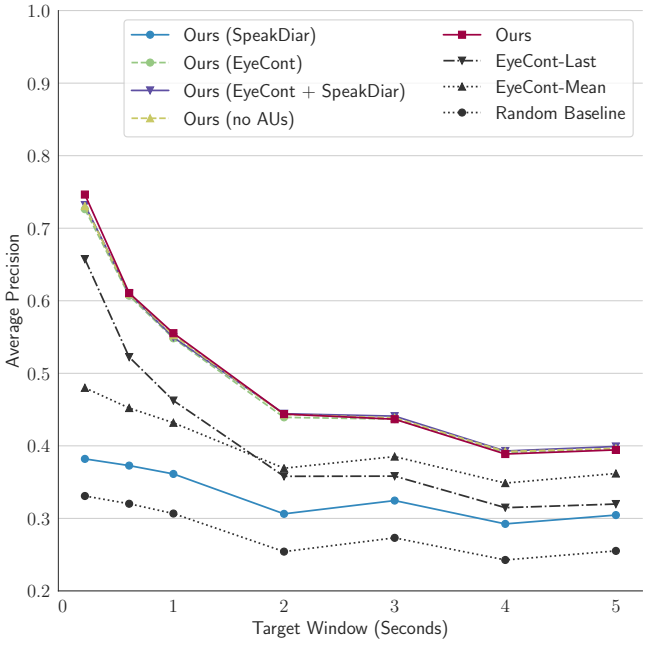

Figure 6: Average precision achieved in the person-independent evaluation for different feature channel ablations of our method and baselines across different target window sizes.

stronger. Specifically, the baseline *EyeCont-Last* classifies a person to have averted gaze on the target window, if she had averted gaze (i.e. no eye contact) at the last timestep of the feature window. In this way, the baseline exploits the assumption of a certain degree of temporal smoothness of gaze behaviour. This baseline is optimal for cases of constant gaze. The ranking used for computing average precision is obtained by ordering examples according to the classification decision. As relying only on one timestep for prediction might be subject to noise, we also designed a second baseline *EyeCont-Mean* which orders test examples according to the probability of averted gaze observed on the feature window. We also assume this baseline to be stronger than the random baseline, as the probability of averted gaze on a time window right before the target window should be closer to the probability of averted gaze on the target window than the general probability of averted gaze computed on the whole training set. Finally, we implemented a baseline based on the assumption of constant gaze velocity. In detail, we computed the velocity of OpenFace gaze estimates by taking the difference of the two last gaze points in the feature window. We extrapolated the future gaze location using this velocity, and checked whether it falls into the eye contact region at the middle of the target window. We omit this baseline in the results, as it only performed close to the random baseline, due to the tendency of gaze extrapolations to overshoot beyond the eye contact region.

5.3 Person-specific Evaluation

Our person-specific evaluation simulates the case in which an eye contact detection system is adapted to a specific person. As different application scenarios require gaze anticipation for different time horizons, we evaluated the performance of our model for different target window lengths. Figure 5 shows the resulting average precision in averted gaze anticipation for different future time windows and different input features to our method. As can be seen from the figure, our method obtains a performance that is consistently better or on-par with all other methods and baselines across all target window sizes. The largest average precision is achieved for a target window size of 0.2 seconds (0.85 AP for our method). As expected, predictive performance decreases with larger target time windows but remains in the range of 0.71 to 0.74 for target time windows between 2 and 5 seconds length. In contrast, the baselines using eye contact detections from the feature window consistently remain below 0.7 AP with *EyeCont − Mean* achieving higher AP than *EyeCont − Last* except for the 0.2 second target window.

We also compare our method to ablations with removed input channels (e.g. *Ours (SpeakDiar)* uses only the speaker diarisation channel as input). Here, the advantage of incorporating facial action units is primarily evident for the target window sizes of 0.2 to 1 second. The largest gap between our method and the method with the facial action unit channel removed (*Ours (no AUs)*) is at a target window size of 0.6 seconds (0.79 vs. 0.76 AP). Starting from target window sizes of 2 seconds, our method is only marginally better than the method without facial action unit input (e.g. 0.742 compared to 0.738 for target window size 3 seconds). Ablating further, we observed that while eye contact input alone (*Ours (EyeCont)*) is able to yield above-baseline performances for all target windows, it is important to combine eye contact with speaker diarisation input

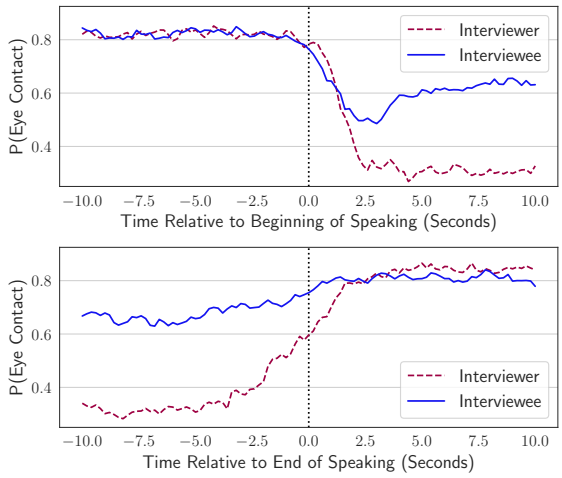

Figure 7: Temporal evolution of the probability of interviewer or interviewee having eye contact at the start (top), or end (bottom) of speaking turns.

(*Ours (EyeCont + SpeakDiar)*) to obtain a strong boost in performance. On the other hand, speaker diarisation input alone (*Ours (SpeakDiar)*) is not sufficient to outperform the baselines.

5.4 Person-independent Evaluation

In the person-independent evaluation we investigated whether it is possible to train an anticipation system for averted gaze that generalises across people. This is significantly more challenging as it adds the variability in behavioural patterns across people, along with variability in the geometric configuration of recording camera, screen and head location as well as in video quality.

The results of this evaluation, performed otherwise analogously to the person-specific case, are summarised in Figure 6. Overall, the differences between our method and the baselines are lower than in the person-specific case, which reflects that exploiting behavioral patterns is more challenging given the higher variability in this person-independent evaluation. Again, our method reaches its highest performance (0.75 AP) for the smallest target window size (0.2 seconds). As could be expected, for larger target window sizes, performance drops more quickly than in the person-specific evaluation (0.44 AP at 2 seconds and 0.39 AP at 5 seconds). However, our method stays consistently above the highest performing baseline for each target window, e.g. outperforming *EyeCont − Last* with 0.75 AP compared to 0.66 AP for a 0.2 second feature window, and outperforming *EyeCont − Mean* with 0.44 AP compared to 0.39 AP for a 3 second target window. In contrast to the person-specific evaluation, our ablation analysis reveals that the ablation of our method using eye contact input only (*Ours (EyeCont)*) performs on par with our method on almost all target window sizes. Only for a target window size of 0.2 our method is slightly better than this comparison approach (0.75 AP versus 0.73 AP).

5.5 Eye Contact at Speaker Changes

The comparably low performances in our person-independent evaluation point at the difficulty of generalising averted gaze anticipation across people. To obtain further insights into the variability of averted gaze depending on person-specific and situational factors we analysed the temporal evolution of eye contact around speaking turn transitions. More specifically, we compared the average eye contact behaviour of guests (interviewees) with the average eye contact behaviour of the host of "Wisdom From North" (example of an interviewer) at speaking turn transitions (see Figure 7). In detail, we first computed for each person and interaction separately the probability of having eye contact at an offset of Δ seconds relative to a speaking turn transition. Subsequently, we averaged these probabilities across all interactions of the host or all interactions of guests, respectively. We performed this analysis separately for speaking turn transitions at which the target person starts speaking, and for speaking turn transitions at which the target person stops speaking. We varied Δ from -10 to 10, obtaining a 20 second time window centered on speaker turn transitions. In this analysis, we only considered speaking turn transitions for which there was no second speaking turn transition 15 seconds before or after. In this way, no effects of other speaking turn boundaries are introduced.

The results of this analysis (see Figure 7) show that for both interviewer and interviewees the probability of eye contact during listening (before starting to speak or after stopping to speak) is higher than during speaking (after starting to speak or before stopping to speak). This is a well known effect [Rossano 2013] that has shown to even be robust enough to be exploited as a means for weak annotation in the context of training multi-person eye contact detection systems [Müller et al. 2018]. While the probabilities of eye contact are similar (around 0.8) for both interviewer and interviewee during listening, during speaking the probability of eye contact is lower for the interviewer (below 0.4) than for the interviewee (above 0.6). While it is difficult to attribute this difference specifically to interpersonal- or situational causes, it underlines the difficulty of person-independent averted gaze anticipation as experienced in our earlier analyses.

A second interesting difference between interviewer and interviewees observable in our analysis is a gaze aversion effect for interviewees at around 2.5 seconds after beginning to speak. While the probability of eye contact of the interviewer decreases steadily before settling on a plateau, the interviewees probability of eye contact decreases, reaches a local minimum at about 3 seconds after starting to speak and eventually increases again. While the data available to us does not grant a definite conclusion, one plausible explanation is that interviewees show gaze aversion due to cognitive load [Glenberg et al. 1998] when starting to speak. In the interview situations in our dataset the interviewees often respond to questions of the interviewer. It is likely that interviewee cognitive load is high during the first seconds of their response as recollection processes and planning of the response might be especially resource-demanding at the beginning and level off only later. In contrast, the interviewer is not confronted with questions frequently and consequently does not show a gaze aversion effect when starting to speak.

6 DISCUSSION

6.1 On Performance

Our method achieved above-baseline performance consistently across all evaluation scenarios. Especially in the person-specific evaluation, it improved on already strong baselines by a clear margin. For a small target window size, performances were in the region of 0.76 to 0.85, which may already be reliable enough for some applications. For example, as a result of the large inherent variability in social behaviour, a visual chatbot adapting its behaviour based on anticipated user gaze might not be perceived too negatively when it selects behaviour based on a false anticipation from time to time. However, our evaluations for the person-independent case also showed that the problem of averted gaze anticipation is far from being solved. While we achieved high performance for small target windows in this case as well, average precision dropped to below 0.5 for target windows of 2 seconds and larger. Furthermore, in the subject-independent case our method was not yet able to harness the combination of different input features effectively.

It is surprising that the LSTM with only eye contact and speaker diarisation input channels achieves a performance close to the full model in many cases. As the performance of this reduced feature set is still clearly better than the baselines we tried, it appears that the LSTM is able to exploit the temporal patterns present in eye contact and speaker diarisation channels in order to anticipate averted gaze.

Our analysis of eye contact at speaker changes showed significant differences between interviewee and interviewer behaviour, further emphasizing the challenge to create systems that can reliably anticipate averted gaze in a subject-independent manner.

6.2 On Potential Applications

Automatic anticipation of averted gaze during interactions opens up multiple possibilities for exciting new applications. In human-agent interactions, a visual chatbot could use knowledge about users' future eye contact behaviour to adapt its behaviour. If for example an agent wants to show something to the user, and at the same time anticipates that the user will avert her gaze in the near future and overlook the agent's action, the agent could generate an utterance to catch the user's attention. Alternatively, if an agent wants to be unobtrusive, it might wait with the initiation of its action until it anticipates that the user will have eye contact again.

Anticipating averted gaze also enables new applications supporting human-human interactions. Current research on real-time feedback in social interactions is limited to intervening after a target behaviour has been observed [Damian et al. 2015; Schiavo et al. 2014]. With averted gaze anticipation, feedback systems could intervene earlier, not allowing the undesired behaviour to occur in the first place. Feedback could be given explicitly, e.g. by a symbol appearing on the screen or presented via an augmented reality device. Another promising possibility are subtle ways of changing visual behaviour, e.g. by presenting cues that are not consciously perceived but still influence gaze behaviour [Bailey et al. 2009].

Another exciting potential future application of averted gaze anticipation is to investigate whether it can be used to train people to exert stronger conscious control over their gaze behaviour. In a fashion similar to biofeedback [Schwartz and Andrasik 2017], people could be informed by e.g. a sound if averted gaze is anticipated.

6.3 On Possible Improvements and Extensions

While our work represents an important step towards automatic averted gaze anticipation, several possibilities for future improvements and extensions remain. First of all, performance of averted gaze anticipation could probably be improved by increasing the accuracy of eye contact detections used as input (currently between 80% and 90%). Furthermore, a highly accurate, fully automatic eye contact detection approach would eliminate the eye contact labelling step and could be a building block of a system that adapts itself to a target user during deployment. This is especially important because our evaluations have shown that person-independent prediction is particularly challenging. While latest methods for eye contact detection have improved significantly, in terms of performance in challenging everyday settings [Müller et al. 2018; Zhang et al. 2017b], additional improvements are needed to provide close to gold-standard predictions. Further performance improvements in gaze anticipation might be gained by additional input features. For example, the link between the difficulty of a question and gaze aversion [Glenberg et al. 1998] could be exploited by verbal analysis.

Beyond performance improvements, our approach could also be extended to novel settings. Appropriate eye contact behaviour of robots was shown to be beneficial for feelings of social connectedness between robots and users [Zhang et al. 2017a] and robots can make use of gaze aversion mechanisms to make a more thoughtful impression and effectively manage the conversational floor [Andrist et al. 2014]. Anticipating averted gaze in interactions situated in physical spaces, potentially including complex tasks, can help robots to initiate such appropriate gaze behaviour proactively in response to users' anticipated gaze, achieving seamless interaction.

Further possible extensions include outputting more fine-grained predictions, going beyond a binary classification of averted gaze vs. eye contact towards a richer set of predictions similar to mobile attention forecasting [Steil et al. 2018]. It might also be helpful for applications to anticipate the spatial location or the object towards which gaze averted from the interactant will be directed.

7 CONCLUSION

Averted gaze is of fundamental importance in human social encounters and, as such, also is the ability to automatically predict averted gaze for applications in human-machine interaction. We proposed the first method to anticipate averted gaze in natural interactions and evaluated it for different future time horizons on a novel dataset of dyadic video conferences. Our analyses showed that our method significantly outperforms baselines for both person-specific and person-independent evaluation settings. While averted gaze anticipation remains challenging, our work marks an important step towards accurate and robust methods for anticipatory human-computer interaction.

ACKNOWLEDGMENTS

This work was funded by the European Research Council (ERC; grant agreement 801708). P. Müller was funded by a JST CREST research grant under Grant No.: JPMJCR14E1, Japan, and E. Sood by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 - 390740016.

REFERENCES

Andrea Abele. 1986. Functions of gaze in social interaction: Communication and monitoring. *Journal of Nonverbal Behavior* 10, 2 (1986), 83–101. https://doi.org/10.1007/BF01000006

Reginald B Adams Jr and Robert E Kleck. 2003. Perceived gaze direction and the processing of facial displays of emotion. *Psychological science* 14, 6 (2003), 644–647. https://doi.org/10.1046/j.0956-7976.2003.psci_1479.x

Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *Proc. of the ACM/IEEE International Conference on Human-robot Interaction*. 25–32. https://doi.org/10.1145/2559636.2559666

Michael Argyle and Mark Cook. 1976. Gaze and mutual gaze. (1976). https://doi.org/10.1017/S0007125000073980

Marc Assens Reina, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. 2017. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proc.of the ICCV Workshop on Egocentric Perception, Interaction and Computing*. 2331–2338. https://doi.org/10.1109/ICCVW.2017.275

Reynold Bailey, Ann McNamara, Nisha Sudarsanam, and Cindy Grimm. 2009. Subtle Gaze Direction. *ACM Transactions on Graphics* 28, 4, Article 100 (Sept. 2009), 14 pages. https://doi.org/10.1145/1559755.1559757

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *Proc. of the IEEE International Conference on Automatic Face & Gesture Recognition*. 59–66. https://doi.org/10.1109/FG.2018.00019

Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. 2002. Listener responses as a collaborative process: The role of gaze. *Journal of Communication* 52, 3 (2002), 566–580. https://doi.org/10.1111/j.1460-2466.2002.tb02562.x

Cigdem Beyan, Vasiliki-Maria Katsageorgiou, and Vittorio Murino. 2017. Moving as a Leader: Detecting Emergent Leadership in Small Groups using Body Pose. In *Proc. of the ACM International Conference on Multimedia*. 1425–1433. https://doi.org/10.1145/3123266.3123404

Ali Borji, Dicky N Sihite, and Laurent Itti. 2013. What/where to look next? Modeling top-down visual attention in complex interactive environments. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44, 5 (2013), 523–538. https://doi.org/10.1109/TSMC.2013.2279715

Francesca Capozzi, Cigdem Beyan, Antonio Pierro, Atesh Koul, Vittorio Murino, Stefano Livi, Andrew P Bayliss, Jelena Ristic, and Cristina Becchio. 2019. Tracking the Leader: Gaze Behavior in Group Interactions. *iScience* 16 (2019), 242. https://doi.org/10.1016/j.isci.2019.05.035

Ionut Damian, Chiew Seng Sean Tan, Tobias Baur, Johannes Schöning, Kris Luyten, and Elisabeth André. 2015. Augmenting Social Interactions: Realtime Behavioural Feedback using Social Signal Processing Techniques. In *Proc. of the ACM Conference on Human Factors in Computing Systems*. 565–574. https://doi.org/10.1145/2702123.2702314

Paul Ekman. 1992. Facial expressions of emotion: an old controversy and new findings. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 335, 1273 (1992), 63–69.

Will Feng, Anitha Kannan, Georgia Gkioxari, and C Lawrence Zitnick. 2017. Learn2smile: Learning non-verbal interaction through observation. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. 4131–4138. https://doi.org/10.1109/IROS.2017.8206272

Theodoros Giannakopoulos. 2015. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PloS one* 10, 12 (2015). https://doi.org/10.1371/journal.pone.0144610

Arthur M Glenberg, Jennifer L Schroeder, and David A Robertson. 1998. Averting the gaze disengages the environment and facilitates remembering. *Memory & cognition* 26, 4 (1998), 651–658. https://doi.org/10.3758/BF03211385

Antonia F. de C. Hamilton. 2016. Gazing at me: the importance of social meaning in understanding direct-gaze cues. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, 1686 (2016), 20150080. https://doi.org/10.1098/rstb.2015.0080

Dan Witzner Hansen and Qiang Ji. 2009. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 3 (2009), 478–500. https://doi.org/10.1109/TPAMI.2009.30

Jonathan Harel, Christof Koch, and Pietro Perona. 2007. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*. 545–552.

Simon Ho, Tom Foulsham, and Alan Kingstone. 2015. Speaking and listening with the eyes: gaze signaling during dyadic interactions. *PloS one* 10, 8 (2015), e0136905. https://doi.org/10.1371/journal.pone.0136905

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. 2018. Predicting Gaze in Egocentric Video by Learning Task-dependent Attention Transition. In *Proc. of the European Conference on Computer Vision*. 754–769.

Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 11 (1998), 1254–1259. https://doi.org/10.1109/34.730558

Kristiina Jokinen, Kazuaki Harada, Masafumi Nishida, and Seiichi Yamamoto. 2010. Turn-alignment using eye-gaze and speech in conversational interaction. In *Proc. of the Annual Conference of the International Speech Communication Association*.

Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Adj. Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 1151–1160. https://doi.org/10.1145/2638728.2641695

Joan Kellerman, James Lewis, and James D Laird. 1989. Looking and loving: The effects of mutual gaze on feelings of romantic love. *Journal of research in personality* 23, 2 (1989), 145–161. https://doi.org/10.1016/0092-6566(89)90020-2

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta psychologica* 26 (1967), 22–63. https://doi.org/10.1016/0001-6918(67)90005-4

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. 2016. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563* (2016).

Yin Li, Miao Liu, and James M Rehg. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proc. of the European Conference on Computer Vision*. 619–635.

Huiying Liu, Dong Xu, Qingming Huang, Wen Li, Min Xu, and Stephen Lin. 2013. Semantically-Based Human Scanpath Estimation with HMMs. In *Proc. of the IEEE International Conference on Computer Vision*. 3232–3239. https://doi.org/10.1109/ICCV.2013.401

Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. 2012. Head pose-free appearance-based gaze sensing via eye image synthesis. In *Proc. of the International Conference on Pattern Recognition*. IEEE, 1008–1011.

Paul P Maglio, Teenie Matlock, Christopher S Campbell, Shumin Zhai, and Barton A Smith. 2000. Gaze and Speech in Attentive User Interfaces. In *Proc. of the International Conference on Multimodal Interfaces*. Springer, 1–7.

Philipp Müller and Andreas Bulling. 2019. Emergent Leadership Detection Across Datasets. In *Proc. of the International Conference on Multimodal Interaction*. https://doi.org/10.1145/3340555.3353721

Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behavior. In *Proc. ACM International Conference on Intelligent User Interfaces*. 153–164. https://doi.org/10.1145/3172944.3172969

Philipp Müller, Michael Xuelin Huang, Xucong Zhang, and Andreas Bulling. 2018. Robust Eye Contact Detection in Natural Multi-Person Interactions Using Gaze and Speaking Behaviour. In *Proc. of the International Symposium on Eye Tracking Research and Applications*. 31:1–31:10. https://doi.org/10.1145/3204493.3204549

Federico Rossano. 2013. Gaze in Conversation. *The handbook of conversation analysis* (2013), 308.

Timothy A Salthouse and Cecil L Ellis. 1980. Determinants of Eye-Fixation Duration. *The American journal of psychology* (1980), 207–234. https://doi.org/10.2307/1422228

Gianluca Schiavo, Alessandro Cappelletti, Eleonora Mencarini, Oliviero Stock, and Massimo Zancanaro. 2014. Overt or Subtle? Supporting Group Conversations with Automatically Targeted Directives. In *Proc. of the ACM International Conference on Intelligent User Interfaces*. 225–234. https://doi.org/10.1145/2557500.2557507

Mark S Schwartz and Frank Andrasik. 2017. *Biofeedback: A practitioner's guide*. Guilford Publications.

Atsushi Senju and Mark H Johnson. 2009. The eye contact effect: mechanisms and development. *Trends in cognitive sciences* 13, 3 (2009), 127–134. https://doi.org/10.1016/j.tics.2008.11.009

Brian A Smith, Qi Yin, Steven K Feiner, and Shree K Nayar. 2013. Gaze locking: passive eye contact detection for human-object interaction. In *Proc. of the ACM Symposium on User Interface Software and Technology*. ACM, 271–280. https://doi.org/i10.1145/2501988.2501994

Julian Steil, Philipp Müller, Yusuke Sugano, and Andreas Bulling. 2018. Forecasting user attention during everyday mobile interactions using device-integrated and wearable sensors. In *Proc. of the International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 1. https://doi.org/10.1145/3229434.3229439

Jürgen Streeck. 1993. Gesture as communication I: Its coordination with gaze and speech. *Communications Monographs* 60, 4 (1993), 275–299. https://doi.org/10.1080/03637759309376314

Marc Tonsen, Julian Steil, Yusuke Sugano, and Andreas Bulling. 2017. InvisibleEye: Mobile Eye Tracking Using Multiple Low-Resolution Cameras and Learning-Based Gaze Estimation. *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 106:1–106:21. https://doi.org/10.1145/3130971

Roberto Valenti, Nicu Sebe, and Theo Gevers. 2011. Combining Head Pose and Eye Location Information for Gaze Estimation. *IEEE Transactions on Image Processing* 21, 2 (2011), 802–815. https://doi.org/10.1109/TIP.2011.2162740

Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of eyes for eye-shape registration and gaze estimation. In *Proc. of the IEEE International Conference on Computer Vision*. 3756–3764. https://doi.org/10.1109/ICCV.2015.428

Pingmei Xu, Yusuke Sugano, and Andreas Bulling. 2016. Spatio-Temporal Modeling and Prediction of Visual Attention in Graphical User Interfaces. In *Proc. of the ACM Conference on Human Factors in Computing Systems*. 3299–3310. https://doi.org/10.1145/2858036.2858479

Hirotake Yamazoe, Akira Utsumi, Tomoko Yonezawa, and Shinji Abe. 2008. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions. In *Proc. of the Symposium on Eye Tracking Research & Applications*. ACM, 245–250. https://doi.org/10.1145/1344471.1344527

Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. 2017c. Deep Future Gaze: Gaze Anticipation on Egocentric Videos Using Adversarial Networks. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 4372–4381. https://doi.org/10.1109/CVPR.2017.377

Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2017b. Everyday Eye Contact Detection Using Unsupervised Gaze Target Discovery. In *Proc. of the ACM Symposium on User Interface Software and Technology*. 193–203. https://doi.org/10.1145/3126594.3126614

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2019. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 1 (2019), 162–175. https://doi.org/10.1109/TPAMI.2017.2778103

Yanxia Zhang, Jonas Beskow, and Hedvig Kjellström. 2017a. Look but Don't Stare: Mutual Gaze Interaction in Social Robots. In *Proc. of the International Conference on Social Robotics*. Springer, 556–566. https://doi.org/10.1007/978-3-319-70022-9_55

Zehua Zhang, David J Crandall, Chen Yu, and Sven Bambach. 2018. From Coarse Attention to Fine-Grained Gaze: A Two-stage 3D Fully Convolutional Network for Predicting Eye Gaze in First Person Video. In *Proc. of the British Machine Vision Conference*.

Miron Zuckerman, Bella M DePaulo, and Robert Rosenthal. 1981. Verbal and Nonverbal Communication of Deception. In *Advances in experimental social psychology*. Vol. 14. Elsevier, 1–59. https://doi.org/10.1016/S0065-2601(08)60369-X