Robust Eye Contact Detection in Natural Multi-Person Interactions Using Gaze and Speaking Behaviour

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

Michael Xuelin Huang
Max Planck Institute for Informatics
Saarland Informatics Campus, Germany
mhuang@mpi-inf.mpg.de

Xucong Zhang
Max Planck Institute for Informatics
Saarland Informatics Campus, Germany
xczhang@mpi-inf.mpg.de

Andreas Bulling
Max Planck Institute for Informatics
Saarland Informatics Campus, Germany
bulling@mpi-inf.mpg.de

ABSTRACT

Eye contact is one of the most important non-verbal social cues and fundamental to human interactions. However, detecting eye contact without specialised eye tracking equipment poses significant challenges, particularly for multiple people in real-world settings. We present a novel method to robustly detect eye contact in natural three- and four-person interactions using off-the-shelf ambient cameras. Our method exploits that, during conversations, people tend to look at the person who is currently speaking. Harnessing the correlation between people's gaze and speaking behaviour therefore allows our method to automatically acquire training data during deployment and adaptively train eye contact detectors for each target user. We empirically evaluate the performance of our method on a recent dataset of natural group interactions and demonstrate that it achieves a relative improvement over the state-of-the-art method of more than 60%, and also improves over a head pose based baseline.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Human-centered computing** → *Collaborative and social computing*;

KEYWORDS

Social Gaze; Gaze Signaling; Gaze Estimation; Group Interactions; Social Signal Processing

ACM Reference Format:
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 *ETRA '18: 2018 Symposium on Eye Tracking Research and Applications, June 14–17, 2018, Warsaw, Poland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3204493.3204549

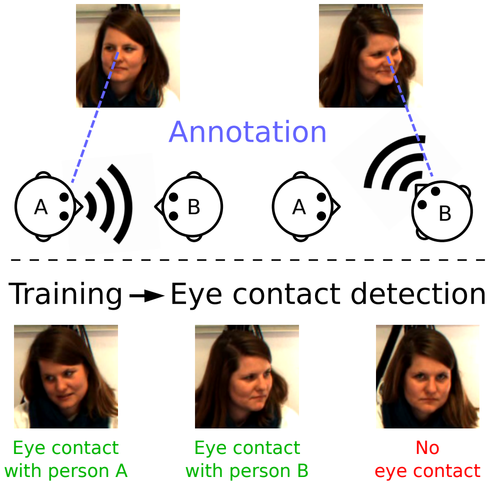

Figure 1: Our method exploits the correlation between gaze and speaking behaviour naturally occurring during multi-person interactions to weakly annotate images (top) that are, in turn, used to train a robust eye contact detector (bottom).

1 INTRODUCTION

Eye contact is fundamental to human social interactions and, as such, a key non-verbal behavioural cue [Kleinke 1986]. Eye contact detection has consequently emerged as an important tool for better understanding human social behaviour and cognition [Farroni et al. 2002]. Eye contact detection is typically understood as the task of automatically detecting whether a person's gaze is directed at another person's eyes or face [Chong et al. 2017a], an object of interest [Shell et al. 2003; Smith et al. 2013, 2005] or both [Zhang et al. 2017a]. Eye contact detection has numerous applications, for example as a key component in attentive user interfaces [Smith et al. 2005] or to analyse turn-taking, social roles, and engagement during multi-person interactions [Oertel and Salvi 2013].

Despite recent advances in appearance-based gaze estimation [Zhang et al. 2015, 2017b, 2018], eye contact detection using off-the-shelf cameras, i.e. without special-purpose eye tracking equipment, remains profoundly challenging. This is because eye contact detection not only requires accurate gaze estimation but also information

ETRA '18, June 14–17, 2018, Warsaw, Poland

ACM ISBN 978-1-4503-5706-7/18/06…$15.00
https://doi.org/10.1145/3204493.3204549

on the 3D position and size of the eye contact target, which is typically unknown in real-world settings. Previous works on automatic analysis of social interactions thus often fell back to using head orientation as a proxy for gaze direction and, in turn, eye contact [Beyan et al. 2017b; Gatica-Perez et al. 2005]. However, while head orientation and gaze are correlated, this correlation is far from perfect during multi-person interactions [Vrzakova et al. 2016].

Hence, more recent works focused on developing methods specifically geared towards eye contact detection. Smith et al. used a classification approach to determine eye contact with a camera, but their method required prior knowledge about the size and location of the target [Smith et al. 2013]. Zhang et al. presented a method for eye contact detection during dyadic (two-person) interactions [Zhang et al. 2017a]. Their method achieved significant performance improvements but only worked for a single eye contact target that had to be closest to the camera. This assumption does not hold for multi-person interactions in which multiple conversation partners need to be differentiated.

To address both limitations, inspired by [Siegfried et al. 2017], we present a novel method to robustly detect eye contact in natural three- and four-person interactions using off-the-shelf ambient cameras. Our method exploits the fact that, during conversations, people tend to look at the person who is currently speaking [Vertegaal et al. 2001]. Analysing the correlation between people's gaze and speaking behaviour therefore allows our method to automatically acquire training data during deployment and adaptively train eye contact detectors for each target user. More specifically, our method first detects speaking behaviour of people based on their mouth movements extracted from several ambient cameras. The speaking behaviour is then associated with gaze estimates obtained using a state-of-the-art convolutional neural network (CNN) gaze estimator [Zhang et al. 2017b] applied on a frontal view on the person whose eye contact with others is to be estimated. Finally, our method weakly labels images to train an eye contact detector on the corresponding CNN face feature representations.

The specific contributions of our work are two-fold. First, we propose the first method for eye contact detection in natural multi-person interactions using RGB cameras. Second, we demonstrate the effectiveness of our method through a detailed performance evaluation on a recent dataset of natural multi-person interactions [Müller et al. 2018], showing that our method outperforms the state-of-the-art method [Zhang et al. 2017a] with more than 60% relative improvement. We further show that our method benefits from ground truth speaking information, and can outperform the state-of-the-art method trained on the whole 20-minute-long interactions after only observing the first four minutes of an interaction.

2 RELATED WORK

Our method is related to previous works on 1) exploring the link between gaze and speech, 2) estimating gaze during social interactions, and 3) computational methods for eye contact detection.

2.1 Link between Gaze and Speech

Research on the link between gaze and speech has a long history. Studies have indicated that gaze can be a cue for turn-taking [Kendon 1967], as well as a collaborative signal to coordinate the insertion of responses [Bavelas et al. 2002]. Recent research confirmed these findings by employing head-mounted eye trackers and cross-correlation analysis to show that speakers tend to end their turns gazing at their interlocutor, while listeners begin speaking with averted gaze [Ho et al. 2015]. Moreover, Hirvenkari et al. found that even uninvolved observers of dyadic interactions followed the interactants' speaking turns with their gaze [Hirvenkari et al. 2013].

Although the roles in multi-person interactions can be more complex than those of dyadic interactions, a strong link between gaze and speech remains. Similar to the dyadic case, research has shown that gaze is an important signal in turn-taking [Ishii et al. 2016; Jokinen et al. 2013]. Most importantly, however, Vertegaal et al. reported a very high chance (88%) that a person looks at the speaker in four-party conversations [Vertegaal et al. 2001]. All of these findings underline the strong link between gaze and speech and, as such, lay the foundation for our method and the idea of using speech to weakly annotate gaze in an automatic fashion.

2.2 Gaze Estimation During Social Interactions

Gaze estimation has been of great interest for researchers in psychology [Bavelas et al. 2002; Kendon 1967] as well as affective computing [Andrist et al. 2014; Huang et al. 2016; Picard 1995]. Previous studies followed two different ways to address the challenges of gaze estimation. Most of them relied on stationary [Jokinen et al. 2013; Vertegaal et al. 2001] or head-mounted [Ho et al. 2015] eye trackers. However, the need for special-purpose equipment represents a significant constraint on the recording setup and can result in unnatural behaviour by participants [Risko and Kingstone 2011].

A second line of work consequently focused on estimating gaze during social interactions using off-the-shelf cameras. Most methods approximated gaze by head pose, for instance, to implement plausible gaze aversion mechanisms on robots [Andrist et al. 2014], track the attentional focus of meeting participants [Stiefelhagen 2002], or to detect a group's interest level [Gatica-Perez et al. 2005]. Most recently, Beyan et al. estimated the visual focus of attention among multiple persons based on head pose in order to detect emergent leaders [Beyan et al. 2016a, 2017a] and predict leadership styles [Beyan et al. 2017b]. Müller et al. used head orientation to detect low rapport in small group interactions [Müller et al. 2018]. While all of these works assumed that head pose can serve as a good proxy for gaze in diverse social interaction tasks, recent research showed that several characteristics of gaze and head orientation are not well correlated in group interactions [Vrzakova et al. 2016].

2.3 Eye Contact Detection

Unlike the general gaze estimation task that attempts to estimate the precise gaze direction in a continuous space [Zhang et al. 2018], eye contact detection is concerned with a binary decision on whether gaze falls onto a target (e.g. a face or a screen) or not. A number of studies have approached this task by either relying on a head-mounted [Chong et al. 2017b; Smith et al. 2005; Ye et al. 2015] or glasses-mounted device [Selker et al. 2001], or requiring LEDs attached to the target [Shell et al. 2004, 2003; Smith et al. 2005].

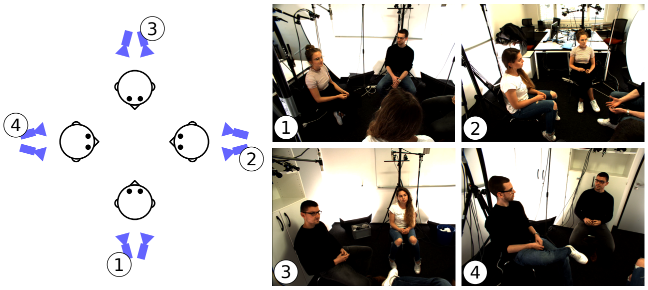

Figure 2: Camera setup used for the dataset recording in [Müller et al. 2018]. Please note that the cameras were placed slightly above the participants to avoid occlusions.

More recent works focused on the significantly more challenging task of using off-the-shelf cameras for eye contact detection [Recasens et al. 2015; Smith et al. 2013]. To overcome limitations of cumbersome and time-consuming data annotation, and to allow for arbitrary geometric relationships between camera and target, Zhang et al. recently proposed an unsupervised method for eye contact detection [Zhang et al. 2017a] built on top of a learning-based gaze estimation method [Zhang et al. 2017b]. A key assumption, and limitation, of their method is that it assumes the gaze target to be the closest to the camera. While this assumption held in the investigated settings, it does not in many other real-world situations, in particular multi-person interactions. Siegfried et al. proposed a method to detect eye contact in dyadic interactions [Siegfried et al. 2017]. However, their method required calibrated depth cameras, a microphone array to detect the beginning and end of utterances of each person, and knowledge of each person's position.

2.4 Summary

Previous works on eye contact detection either required specialised equipment or were limited to dyadic interactions. In contrast, we present the first method for eye contact detection during natural multi-person interactions that requires only an uncalibrated setup of off-the-shelf cameras placed in the environment. We further show that speaking behaviour inferred from mouth movements can be leveraged to weakly annotate gaze estimates in such a setting.

3 DATASET

All experimental evaluations were performed on a subset of a recent dataset of three- and four-person interactions [Müller et al. 2018]. We choose this dataset because, unlike others [Beyan et al. 2016b; Oertel and Salvi 2013], it features two cameras behind each participant providing a view on every other participant. This camera placement makes it particularly well-suited for applying the eye contact detection method by Zhang et al. [Zhang et al. 2017a], as their method requires the target participant to be the closest to the camera. In the following, we first provide an overview of the dataset and then describe the additional eye contact annotations that we collected for the purposes of the current work.

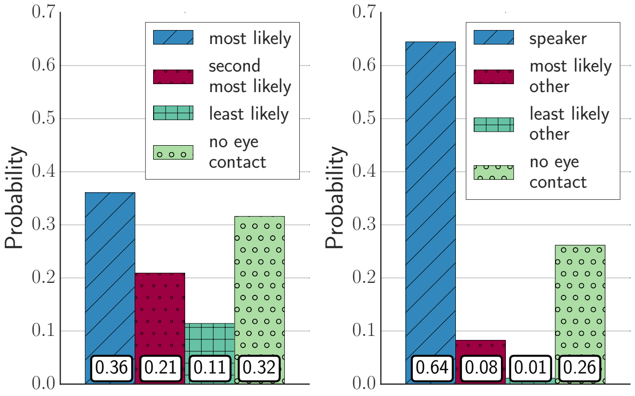

Figure 3: *Left*: Probability of looking to the most often, second most often, and least often looked-at person, along with looking at no face. *Right*: Probability of eye contact with the person who is currently speaking in comparison to the second and third most often looked-at person, along with looking at no face.

3.1 Recording Setup

The dataset [Müller et al. 2018] had originally been recorded to study rapport during multi-person interactions. It consists of 78 participants studying at a German university (43 female, aged between 18 and 38 years), split into 12 four-person and 10 three-person interactions. Participants in each group were instructed to choose and discuss the most controversial topic from a list of possible topics.

The recording was performed in a quiet office room equipped with a 4DV camera system consisting of eight frame-synchronised cameras. As shown in Figure 2, two cameras were placed behind each participant at a slightly elevated position above the head, providing a near frontal view of the faces of all participants even if they turned their head during the conversation. After each recording session, participants provided ratings for felt rapport with the interactants, perceived leadership, dominance, competence and liking, and a five-factor personality assessment (not used here). Furthermore, the authors provided speaking activity annotations for the whole dataset, indicating who was speaking at each moment.

3.2 Gaze Annotations

Given that the dataset by [Müller et al. 2018] did not contain any annotations of participants' gaze behaviour, we asked three annotators to label a subset of 14 recordings with eye contact ground-truth, five of which we used as a dataset for developing our method ("development set"), and nine of which we used for testing ("test set"). This subset was chosen randomly after excluding recordings which suffered from data loss in one camera, as comparing to the method of [Zhang et al. 2017a] on these recordings would have given an unfair advantage to our method. Each of the annotators labelled a different part of the data while being supervised by the lead author to ensure a constant quality of annotations. The annotations consisted of the identifier of the participant whose face is being looked at at a particular moment. Specifically, similar to Zhang et al. [2017a], we defined eye contact as gaze landing within the face region. We also asked them to annotate an additional class containing all non-eye-contact cases, such as looking at the body, walls, or floor, or when participants closed their eyes. Annotations were performed on a per-frame basis at 15-second intervals to strike a balance between annotation effort and coverage. This resulted in eye contact annotations for 3,995 frames from 50 participants, spanning more than 16.5 hours of video recordings.

The annotations revealed that eye contact occurred pervasively during interactions. In Figure 3, we show statistics of eye contact in four-person interactions. The basic pattern is the same in three-person interactions. Although on average one person receives a very large part of the overall eye contact (see Figure 3, left), other people receive significant amounts as well. However, conditioning on the currently speaking person reveals that the current speaker is by far the most likely eye contact target (see Figure 3, right). This pattern lays the foundation for our method.

4 METHOD

Our method improves over the weak labelling and subsequent training of the eye contact detection method proposed in [Zhang et al. 2017a]. Thus, we first briefly summarise that method before we discuss the improvements introduced in our work. Throughout the discussion, we refer to the person whose gaze we analyse as *gazer*, and the person whom the gazer looks at as gaze *target person*.

4.1 Eye Contact Detection Framework

Here we briefly introduce the unsupervised eye contact pipeline in [Zhang et al. 2017a]. Their method took camera images as input and applied facial landmark detection [Baltrušaitis et al. 2016] to extract six key points, including eye and mouth corners. These key points were used to estimate the 3D head pose by fitting them to a generic 3D face model. Then the face images were cropped according to the head pose and data normalisation discussed in [Zhang et al. 2018]. Subsequently, a user-independent CNN model [Zhang et al. 2017b] estimated gaze points in the camera plane, whose origin represents the camera location. All samples of gaze estimates extracted over a time period were clustered by the density-based OPTICS clustering algorithm [Ankerst et al. 1999]. By assuming that *the target object is the closest salient object to the camera*, samples within the cluster closest to the origin were labelled as "eye contact" and the rest as "no eye contact". Afterwards, a binary support vector machine classifier was trained on these annotations with the 4096-dimensional face features extracted from the first fully-connected layer of the CNN model. Compared with the two-dimensional gaze location, this CNN feature representation contains richer information and thus a higher potential of achieving better performance.

Despite the success of this method in unsupervised eye contact detection, the underlying assumption of the gaze target object being the salient object closest to the camera constrains its extension to the multi-person interaction scenario. This is because eye contact with the target person can only be detected on a camera positioned closely to the target person, which restricts the placement of cameras to locations that might not have an optimal view on the gazer. To address this challenge, we propose a novel annotation mechanism that exploits the gaze and speaking behaviour to allow for eye contact detection with multiple target persons from a single frontal view on the gazer.

4.2 Weak Labelling Using Speaking Behaviour

In contrast to the binary classification problem considered in [Zhang et al. 2017a], we have to address a multi-class classification problem, for which we propose a new automatic annotation method. Similar to the work of [Siegfried et al. 2017], we leverage social conventions to perform weak labelling of gaze estimates. Whereas [Siegfried et al. 2017] used speech-based weak labels only to correct for constant shifts in gaze estimates, our approach accommodates nonlinear transformations in the gaze estimate space and provides automatic annotations for the subsequent training of an eye contact detector. We make two assumptions about gaze and speaking behaviour during social interactions:

(1) *People tend to look at the speaker during the interaction.*
(2) *Probability of eye contact with a target person is higher if (s)he speaks more often.*

These assumptions allow our method to 1) locate the face centres and 2) determine the face boundary in the space of gaze estimates.

Figure 4 shows an overview of our method. Our method takes the video stream of a multi-person interaction as input. From a frontal view on the gazer, it extracts gaze estimates using a state-of-the-art CNN-based gaze estimation model [Zhang et al. 2017b]. From the gaze estimates obtained from the whole interaction, we compute the gaze probability distribution. Afterward, we identify the speaking behaviour of different individuals in the interaction based on their mouth movements and associate them with the corresponding gaze estimates across time. We further estimate the gaze probability distributions given a specific gaze target person is speaking. Given this information, we can locate the faces of the gaze targets in the gaze estimate space by comparing the conditional distributions with the general gaze distribution. Our approach subsequently grows regions around the gaze target locations and marks samples falling into those regions as "eye contact with person *j*". Samples not falling into any gaze target region are labelled as "no eye contact". Finally, we use these annotated samples to train an eye contact detector based on the high-dimensional CNN-features as in [Zhang et al. 2017a]. In the following, we discuss each step in detail.

4.2.1 Estimating Distributions of Raw Gaze Estimates. We apply Gaussian Kernel Density Estimation (KDE) to approximate probability density functions of gaze estimates. KDE replaces each sample with a Gaussian distribution, aggregates them, and then outputs the normalised result as a density function. As we use Scott's Rule to estimate the kernel bandwidth [Scott 2015], KDE is completely parameter free. By applying KDE to the 2D gaze estimates of different participants, we derive a gaze density estimate g_i for every gazer i, as well as the conditional gaze density estimate $g_{i|j}$ for gazer i given the potential gaze target person j is speaking.

4.2.2 Locating Face Centres from Gaze Density Estimates. While participants in general are likely to look at the current speaker, there could be a *personal gaze bias* due to individual preferences or external distractions. For example, one participant might frequently look to the floor, while another might often look at a particular person. Such personal gaze bias to some object or person should

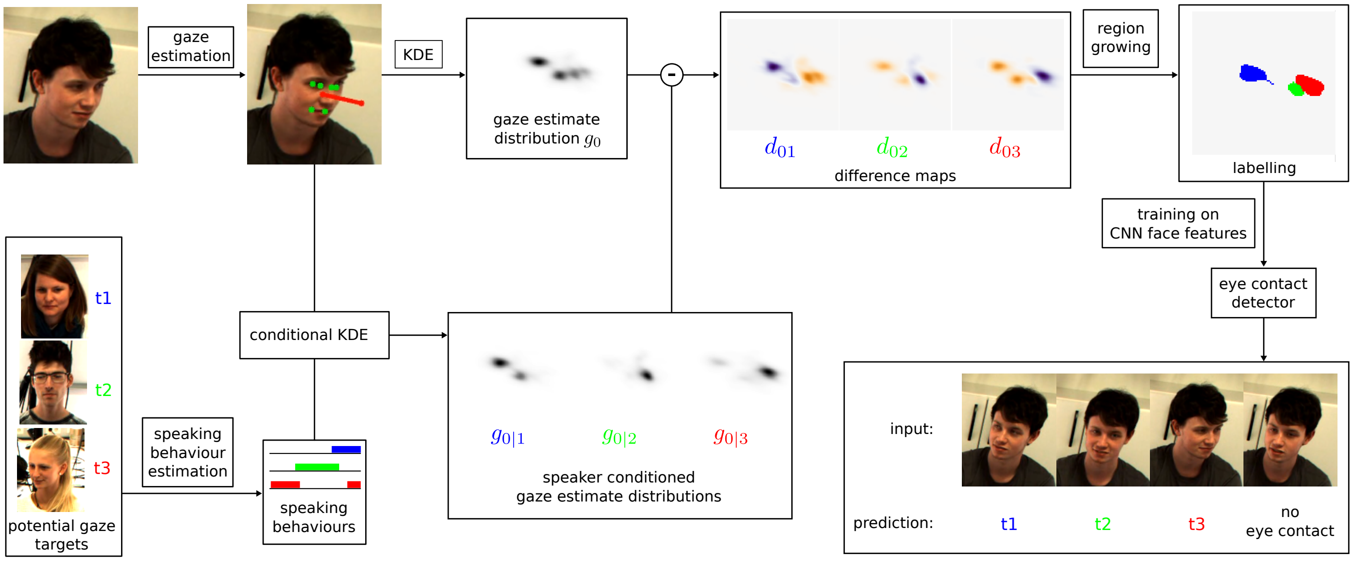

Figure 4: Our method takes images from multiple ambient cameras pointing at the gazer and potential gaze target persons as input. The images are the basis for coarse gaze estimates obtained using a full-face gaze estimation method [Zhang et al. 2017b]. Kernel density estimation (KDE) yields the distribution of gaze estimates. This distribution is contrasted with distributions of gaze estimates for which a fixed gaze target person is speaking. The resulting difference distributions are used to extract locations of the gaze target persons' heads in the space of gaze estimates and to grow corresponding labelled regions around them. Using these labels an eye contact detector based on CNN face features is trained that is able to classify new input images.

be relatively irrespective of who is speaking and be encoded both in the general gaze estimate density g_i and the conditional gaze density estimate $g_{i|j}$. To compensate for this bias in analysing gazer i while participant j is speaking, we compute the difference map between the conditional density estimate and the general gaze density estimate, i.e. $d_{ij} = g_{i|j} - g_i$.

To locate the face centre of participant j in the gaze estimate space of gazer i, we exploit our first assumption that people in general tend to look at the speaker in the interaction. As a result, we retrieve the location with maximum value in the difference map, i.e. $l_{ij} = \arg\max d_{ij}$, where l_{ij} is the face location of participant j in the gaze estimate space of gazer i.

4.2.3 Labelling Frames for Eye Contact Detection. After locating the face centre location l_{ij}, we annotate samples in its vicinity with the participant id j. These samples cover the area corresponding to the face region of the target person in the gaze estimate space. Specifically, starting from a location l_{ij}, we grow a *gaze target region* by following the level sets of g_i. This region is grown subject to two conditions. First, we grow the region until a *probability mass threshold* t_{accept} is covered in g_i. Second, we constrain the region to not grow into areas with negative values in d_{ij}, so as to ensure the samples we annotate only correspond to gaze locations where the gazer is more likely to look if j is speaking.

The probability mass threshold should ideally be determined by the probability $p(ec_{ij})$ that gazer i has eye contact with target person j. Although there is a high chance that the gazer looks at the speaker, to estimate $p(ec_{ij})$ we also need to consider the situation when the gazer is speaking. Therefore, we use the second assumption that the probability of eye contact with a target person is higher if s/he speaks more often. Based on this assumption, we estimate $p(ec_{ij})$ by multiplying the probability of target person j speaking with the probability of the occurrence of eye contact (as opposite to looking at the body of a person or a non-person object, etc.), $p(ec)$, across all participants:

$$\hat{p}(e_{ij}) = p(speak_j)p(ec) \tag{1}$$

We use this estimate $\hat{p}(e_{ij})$ as our probability mass threshold t_{accept} in weak eye contact labelling. We estimate $p(ec)$ only from the recordings in the development set. Given that our method does not use ground truth speaking annotations or audio information, we cannot calculate $p(speak_j)$ directly. Thus, we heuristically set it to $\frac{1}{n}$, where n denotes the number of interactants. Unlike [Zhang et al. 2017a], our method does not rely on an unlabelled "safe margin" to exclude ambiguous samples between "eye contact" and "no eye contact" from training. Instead, we obtained a higher performance by weakly annotating every sample and using a strongly regularised classifier to learn the eye contact model. This is probably because our heuristic results in a sufficiently precise guess as to the extent of "eye contact" regions in gaze estimate space.

4.3 Extracting Speaking Behaviour

To achieve a fully automatic system, we develop a visual speaking indicator based on the sum of the standard deviations of facial action units 25 (lips part) and 26 (jaw drop). We choose to extract this quantity in four-second time intervals around each frame, as this time window maximised the correlation of the speaking indicator with ground truth speaking activity on the development set. To obtain robust estimates of facial action units, we obtain predictions from OpenFace [Baltrušaitis et al. 2016] on all available views on a given person and select the best view for each frame according to the provided confidence scores. As this visual speaking indicator suffers from noise caused by different facial expressions related to action units 25 and 26, we do not precisely know the amount of speaking for each participant. To address this, we use a heuristic threshold to detect speaking behaviours by assuming all participants speak equally often. Specifically, we extract the $1 - \frac{1}{n}$ percentile of the values of the visual speaking indicator of each person, where n is the number of participants (3 or 4) in the recording. Frames above this threshold are classified as "speaking", those below as "not speaking".

4.4 Training the Eye Contact Detector

Our eye contact detector relies on the feature representation extracted from the second last layer of the gaze CNN model [Zhang et al. 2017b]. To make our approach more easily comparable to that of [Zhang et al. 2017a], we also chose to train a support vector classifier (SVC) on this representation. Specifically, we train a one-versus-one multi-class SVC with a radial basis function kernel on the annotated samples, which include classes of gaze on different persons' faces as well as "no eye contact". We use the default value for γ (1 / number of features), and construct a balanced training set by subsampling classes that are overrepresented. On the development set we observed that strong regularisation is important. We therefore set the misclassification penalty parameter C to 0.01 for training our eye contact SVCs.

5 EVALUATION

We compared the performance of our method against the state of the art and investigated its robustness to the quality of speaking behaviour (human-annotated vs. auto-detected). We then examined early predictions, evaluating a progressively increasing amount of training data, followed by the impact of the eye contact prior, $\hat{p}(ec)$ as well as a underlying cause of performance limitation.

5.1 Eye Contact Detection Performance

We compared our method (*ours*) against the following baselines:

(1) Unsupervised eye contact detection [Zhang et al. 2017a]: This is the state-of-the-art eye contact detection method. As this approach assumes that the potential target object is the closest to the camera, we ran it on each camera separately to detect eye contact with the person next to the camera. We used the development set to find the optimal C parameter for the SVC.

(2) Head pose as a proxy to gaze (*ours - head pose*): This is an alternative method that replaces the annotation by gaze estimates with head orientation in our pipeline. This baseline method is motivated by studies that used head orientation as a proxy for gaze direction [Beyan et al. 2017b; Gatica-Perez et al. 2005].

(3) Detection without training (*ours - no train*): This method replaces the eye contact detection model (i.e. SVC) training in our pipeline with a component that predicts eye contact directly by the labelling region in which the raw gaze estimates fall.

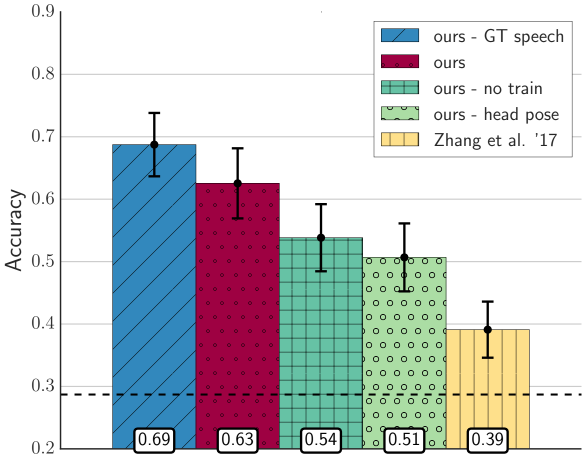

Figure 5: Accuracy of the different eye contact detection methods. Error bars indicate 95% confidence intervals. Random chance level is indicated with the black dashed line.

(4) Labelling with ground truth speaking behaviour (*ours - GT speech*): This method replaces the vision-based speaking behaviour extraction with manual speaking annotation. It thus represents an upper performance bound and, as such, simulates the case when close-to-ground-truth speaking detection is available via specialised audio recording equipment. This method does not need camera views on potential gaze targets.

(5) Random baseline: Eye contact detection using random guessing. For a given participant, this can either be $\frac{1}{4}$ (for four-person-interaction), or $\frac{1}{3}$ (three-person-interaction).

Except for the baselines (1) and (5), the above methods replace different major components in our pipeline, thus shedding light on the contribution of each component to overall performance. Please note that hyper-parameter tuning was done solely on our five-recording development set; final performance numbers were computed at the very end on our nine-recording evaluation set.

Figure 5 shows the performance comparison, with confidence intervals based on the Student's t-distribution indicating the range in which the mean accuracy of the population of subjects will fall with a chance of 95%. The overall results are very encouraging. Our method (0.63) can outperform the no-training counterpart (0.54), and more interestingly, considerably outperform the head pose counterpart (0.51) as well as the state of the art [Zhang et al. 2017a] (0.39) and random guessing (0.29). Furthermore, our method is close to the performance with ground truth speech information (0.69).

The large performance drop (12% absolute decrease) when replacing gaze with head pose estimates is in line with a previous study questioning the reliability of the head as a proxy for gaze in multi-person interactions [Vrzakova et al. 2016]. Moreover, removing the eye contact classifier training in our pipeline also caused a clear decrease in accuracy (9% absolute decrease), indicating that the SVC can effectively leverage the information encoded in the high-dimensional CNN feature space. The moderate gap (6% absolute decrease) between our method and the alternative with ground truth speaking annotation indicates that our fully-automatic method for

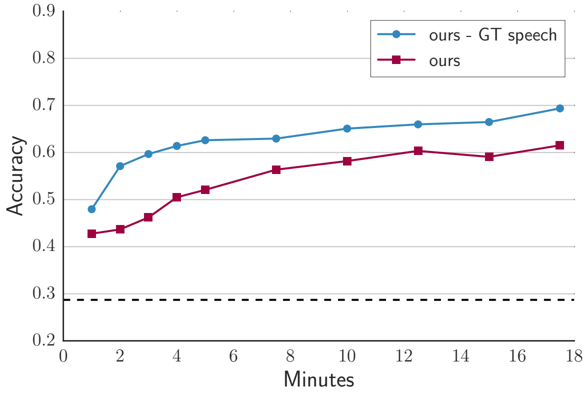

Figure 6: Accuracy of selected methods when only using the first x minutes of an interaction for training. The black dashed line indicates performance of a random predictor.

vision-based speaking detection is quite accurate. Although this vision-based method suffers a slight drop in performance, it enables more flexibility in the recording setup, as the specialised equipment necessary for close-to-ground-truth speaking detection (e.g. lapel microphones or a microphone array) is not always available.

5.2 Online Prediction

As some applications may require eye contact detections at early stages of an interaction, we evaluated the performance of our method for an increasing amount of training data. Figure 6 shows the accuracy of our method (in red) and our upper bound (in blue). As expected, accuracy increases for both methods as the amount of training data increases. More importantly, we see that after training for 12.5 minutes our method performs close (0.60 accuracy) to that of training on the full interactions (0.63 accuracy). It is interesting to note that the performance gap between our method and the upper bound is slightly larger in the range of two to five minutes of training. Furthermore, after three minutes the upper bound already achieves the performance of the fully-automatic method being trained on 12.5 minutes of an interaction. This speaks for specialised audio equipment providing close-to-ground-truth speaking detection like lapel microphones or a microphone array in cases where early prediction is desired. Apart from the online prediction case, these results indicate that annotating speaking status can be helpful if the duration of recordings is limited.

5.3 Influence of the Eye Contact Prior

In this section we evaluate the impact of the prior on the probability of eye contact $p(ec)$, which is used as a parameter for automatic annotation in our method.

Figure 7 shows the performance of our method (in red), the method using speaking ground truth (in blue), and the method without training (in green), given different estimates of $p(ec)$ between 0.25 and 1.0. Probably due to the strongly regularised SVC, $p(ec)$ does not have a significant influence on the training-based methods. Regularisation allows the SVC to leverage the facial appearance information and better tolerate the potential erroneous labelling caused by the nonoptimal $p(ec)$ during learning. However, in the no-training method, $p(ec)$ does have a clear influence on performance, since it directly determines the area of each face region and thus the likelihood of eye contact. Specifically, the accuracy of the no-training method grows with the increase of $p(ec)$.

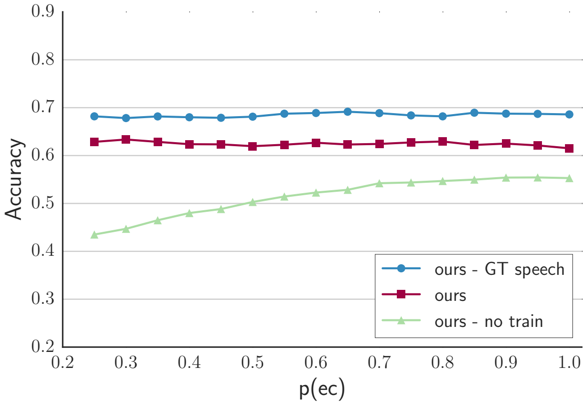

Figure 7: Accuracy of selected methods depending on the eye contact prior $p(ec)$.

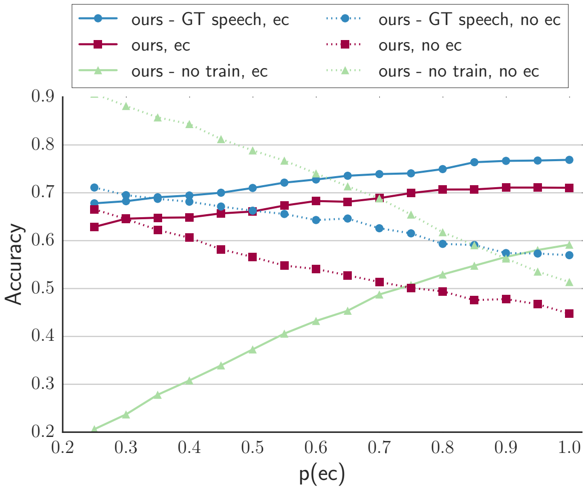

Figure 8: Accuracy of selected methods depending on the eye contact prior $p(ec)$ for ground truth "eye contact" (solid lines) and "no eye contact" samples (dotted lines).

Figure 8 separates the performance of these three methods for ground truth "eye contact" and "no eye contact" samples. Although the overall accuracies of the learning based methods are robust to $p(ec)$, the individual accuracies for "eye contact" and "no eye contact" behave differently. In general, a larger $p(ec)$ increases the accuracy for "eye contact", while it decreases the accuracy for "no eye contact". Thus $p(ec)$ trades off accuracy on "eye contact" samples against the accuracy on "no eye contact" samples. This can be useful if a high accuracy for ground truth "no eye contact" is desired, such as for studies about gaze aversion or autistic behaviours.

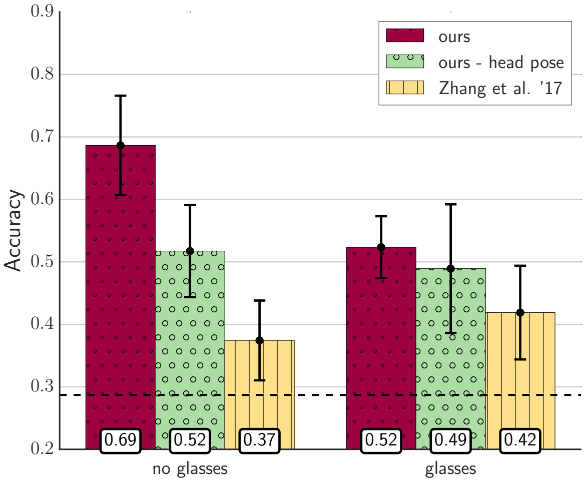

Figure 9: Accuracy of our method as well as the head pose proxy for participants with and without glasses. Error bars indicate 95% confidence intervals. Performance of a random predictor is indicated by the black dashed line.

5.4 Performance With and Without Glasses

Given that eyes can be partially occluded by glasses, we analysed how wearing glasses affected our performance in contrast to the method relying on head orientation and the state-of-the-art method [Zhang et al. 2017a] (see Figure 9). We see that our method reaches an accuracy of almost 0.7 for the no-glasses cases, while it yields only 0.52 for the glasses cases, which is not significantly better than relying on head pose (0.49). The lower accuracy is a direct consequence of the low performance of the underlying gaze estimation method for these cases [Zhang et al. 2017b]. However, our method clearly outperforms the state-of-the-art method no matter if people are wearing glasses or not (0.69 vs. 0.37 and 0.52 vs. 0.42, respectively). It is surprising that the state-of-the-art method reaches a higher accuracy for people wearing glasses than for people without glasses. However, as the confidence intervals for these cases are largely overlapping, this might be a result of chance.

6 DISCUSSION

In this work we proposed a novel method for eye contact detection during multi-person interactions which exploits speaking behaviour as weak supervision to train the eye contact detector. Our method addresses two key limitations of state-of-the-art methods for eye contact detection [Zhang et al. 2017a]: First, it allows detection of eye contact for an arbitrary number of targets. This is important for the meeting scenario studied here, but even more so considering future application scenarios with a larger number of users, such as in a classroom. Second, these targets can be positioned at arbitrary distances from the camera. This is equally important as it significantly reduces constraints on the recording setup, allowing for further studies on optimal camera placements and more seamless integration of the setup into natural environments.

Through evaluations on a recent multi-person dataset, we showed that our method significantly improves over the current state of the art in eye contact detection (see Figure 5). This is encouraging for automatic analysis of group behaviour, for which previous works often had to fall back to using only weakly correlated head orientations as a proxy for gaze and eye contact [Beyan et al. 2017b; Vrzakova et al. 2016]. As a consequence, our approach may lead to new insights into non-verbal group behaviour and to improved prediction performance on diverse social signal processing tasks, such as leadership, interest level, and low rapport detection [Beyan et al. 2017b; Gatica-Perez et al. 2005; Müller et al. 2018].

While post-hoc analysis of eye contact is sufficient for many applications, real-time eye contact detection for multiple users could, for example, be used for future systems that detect low rapport [Müller et al. 2018] and directly execute interventions, e.g. via different kinds of displays [Balaam et al. 2011; Damian et al. 2015; Schiavo et al. 2014]. As shown in our work, our method is capable of online prediction after observing the interaction for a short amount of time (see Figure 6). Using only four minutes of data, our method can outperform the state of the art on eye contact detection being trained on the whole 20-minute-long interactions.

Our evaluations also showed that our method can still benefit from ground truth speaking annotations (see Figure 5). These results are a simulation of a setup including lapel microphones (small microphones e.g. attached to the collar) or microphone arrays, as they can provide close-to-ground-truth speaking detection. If such equipment is available, our method even does not require camera views on the gaze target persons, but only a single view on the person whose gaze we desire to estimate.

While these results are promising, some limitations remain that we intend to address in future work. Our method currently assumes people to be stationary. While this assumption holds for many scenarios, such as the group meetings we investigated, eye contact detection of moving people is an important problem. An improved version of our method could enable studying free-standing conversational groups [Alameda-Pineda et al. 2016] or emotion recognition in free-moving settings [Müller et al. 2015]. Another limitation of our current method is that it can only detect eye contact to people, as it relies on speaking information.

7 CONCLUSION

In this work we proposed a novel method to robustly detect eye contact in natural multi-person interactions recorded using off-the-shelf ambient cameras. We evaluated our method on a recent dataset of natural group interactions, which we annotated with eye contact ground truth, and showed that it outperforms the state-of-the-art in eye contact detection by a large margin. Given the prevalence of cameras in private and public spaces, these results are promising and point towards eye contact detection methods that allow for unobtrusive analysis of social gaze in natural environments, thereby paving the way for new applications in the social and behavioural sciences, social signal processing, and intelligent user interfaces.

ACKNOWLEDGMENTS

This work was funded, in part, by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) at Saarland University, Germany, as well as by a JST CREST research grant under Grant No.: JPMJCR14E1, Japan.

REFERENCES

Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. 2016. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 8 (2016), 1707–1720. https://doi.org/10.1109/TPAMI.2015.2496269

Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational Gaze Aversion for Humanlike Robots. In *Proc. of the 2014 ACM/IEEE International Conference on Human-robot Interaction (HRI '14)*. ACM, New York, NY, USA, 25–32. https://doi.org/10.1145/2559636.2559666

Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. OPTICS: ordering points to identify the clustering structure. In *ACM Sigmod record*, Vol. 28. ACM, 49–60. https://doi.org/10.1145/304182.304187

Madeline Balaam, Geraldine Fitzpatrick, Judith Good, and Eric Harris. 2011. Enhancing Interactional Synchrony with an Ambient Display. In *Proc. of the ACM Conference on Human Factors in Computing Systems*. 867–876. https://doi.org/10.1145/1978942.1979070

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: an open source facial behavior analysis toolkit. In *Proc. of the IEEE Winter Conference on Applications of Computer Vision*. 1–10. https://doi.org/10.1109/WACV.2016.7477553

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, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. 2016a. Identification of Emergent Leaders in a Meeting Scenario Using Multiple Kernel Learning. In *Proc. of the Workshop on Advancements in Social Signal Processing for Multimodal Interaction (ASSP4MI '16)*. ACM, New York, NY, USA, 3–10. https://doi.org/10.1145/3005467.3005469

Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. 2017a. Multi-task Learning of Social Psychology Assessments and Nonverbal Features for Automatic Leadership Identification. In *Proc. of the ACM International Conference on Multimodal Interaction (ICMI 2017)*. ACM, New York, NY, USA, 451–455. https://doi.org/10.1145/3136755.3136812

Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. 2017b. Prediction of the Leadership Style of an Emergent Leader Using Audio and Visual Nonverbal Features. *IEEE Transactions on Multimedia* (2017). https://doi.org/10.1109/TMM.2017.2740062

Cigdem Beyan, Nicolò Carissimi, Francesca Capozzi, Sebastiano Vascon, Matteo Bustreo, Antonio Pierro, Cristina Becchio, and Vittorio Murino. 2016b. Detecting emergent leader in a meeting environment using nonverbal visual features only. In *Proc. of the ACM International Conference on Multimodal Interaction*. ACM, 317–324. https://doi.org/10.1145/2993148.2993175

Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M Jones, Agata Rozga, and James M Rehg. 2017a. Detecting Gaze Towards Eyes in Natural Social Interactions and Its Use in Child Assessment. *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 43. https://doi.org/10.1145/3131902

Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M. Jones, Agata Rozga, and James M. Rehg. 2017b. Detecting Gaze Towards Eyes in Natural Social Interactions and Its Use in Child Assessment. *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3, Article 43 (Sept. 2017), 20 pages. https://doi.org/10.1145/3131902

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

Teresa Farroni, Gergely Csibra, Francesca Simion, and Mark H Johnson. 2002. Eye contact detection in humans from birth. *Proc. of the National Academy of Sciences* 99, 14 (2002), 9602–9605. https://doi.org/10.1073/pnas.152159999

Daniel Gatica-Perez, L McCowan, Dong Zhang, and Samy Bengio. 2005. Detecting Group Interest-Level in Meetings. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. 489–492. https://doi.org/10.1109/ICASSP.2005.1415157

Lotta Hirvenkari, Johanna Ruusuvuori, Veli-Matti Saarinen, Maari Kivioja, Anssi Peräkylä, and Riitta Hari. 2013. Influence of turn-taking in a two-person conversation on the gaze of a viewer. *PloS one* 8, 8 (2013), e71569. https://doi.org/10.1371/journal.pone.0071569

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

Michael Xuelin Huang, Jiajia Li, Grace Ngai, and Hong Va Leong. 2016. StressClick: Sensing Stress from Gaze-Click Patterns. In *Proc. of the ACM Conference on Multimedia*. ACM, 1395–1404. https://doi.org/10.1145/2964284.2964318

Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016. Prediction of Who Will Be the Next Speaker and When Using Gaze Behavior in Multiparty Meetings. *ACM Transactions on Interactive Intelligent Systems* 6, 1, Article 4 (May 2016), 31 pages. https://doi.org/10.1145/2757284

Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and Turn-taking Behavior in Casual Conversational Interactions. *ACM Transactions on Interactive Intelligent Systems* 3, 2, Article 12 (Aug. 2013), 30 pages. https://doi.org/10.1145/2499474.2499481

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

Chris L Kleinke. 1986. Gaze and eye contact: a research review. *Psychological bulletin* 100, 1 (1986), 78. https://doi.org/10.1037/0033-2909.100.1.78

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

Philipp M Müller, Sikandar Amin, Prateek Verma, Mykhaylo Andriluka, and Andreas Bulling. 2015. Emotion recognition from embedded bodily expressions and speech during dyadic interactions. In *Proc. of the International Conference on Affective Computing and Intelligent Interaction*. 663–669. https://doi.org/10.1109/ACII.2015.7344640

Catharine Oertel and Giampiero Salvi. 2013. A Gaze-based Method for Relating Group Involvement to Individual Engagement in Multimodal Multiparty Dialogue. In *Proc. of the ACM International Conference on Multimodal Interaction*. 99–106. https://doi.org/10.1145/2522848.2522865

Rosalind W Picard. 1995. Affective computing. (1995).

Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking?. In *Advances in Neural Information Processing Systems*. 199–207.

Evan F Risko and Alan Kingstone. 2011. Eyes wide shut: implied social presence, eye tracking and attention. *Attention, Perception, & Psychophysics* 73, 2 (2011), 291–296. https://doi.org/10.3758/s13414-010-0042-1

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

David W Scott. 2015. *Multivariate density estimation: theory, practice, and visualization.* John Wiley & Sons.

Ted Selker, Andrea Lockerd, and Jorge Martinez. 2001. Eye-R, a glasses-mounted eye motion detection interface. In *Proc. of the ACM Extended Abstracts on Human Factors in Computing Systems*. ACM, 179–180. https://doi.org/10.1145/634067.634176

Jeffrey S Shell, Roel Vertegaal, Daniel Cheng, Alexander W Skaburskis, Changuk Sohn, A James Stewart, Omar Aoudeh, and Connor Dickie. 2004. ECSGlasses and EyePliances: using attention to open sociable windows of interaction. In *Proc. of the ACM Symposium on Eye Tracking Research & Applications*. ACM, 93–100. https://doi.org/10.1145/968363.968384

Jeffrey S Shell, Roel Vertegaal, and Alexander W Skaburskis. 2003. EyePliances: attention-seeking devices that respond to visual attention. In *Proc. of the ACM Extended Abstracts on Human Factors in Computing Systems*. ACM, 770–771. https://doi.org/10.1145/765891.765981

Rémy Siegfried, Yu Yu, and Jean-Marc Odobez. 2017. Towards the Use of Social Interaction Conventions as Prior for Gaze Model Adaptation. In *Proc. of ACM International Conference on Multimodal Interaction*. ACM. https://doi.org/10.1145/3136755.3136793

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

John D Smith, Roel Vertegaal, and Changuk Sohn. 2005. ViewPointer: lightweight calibration-free eye tracking for ubiquitous handsfree deixis. In *Proc. of the ACM Symposium on User Interface Software and Technology*. ACM, 53–61. https://doi.org/10.1145/1095034.1095043

Rainer Stiefelhagen. 2002. Tracking Focus of Attention in Meetings. In *Proc. of the IEEE International Conference on Multimodal Interfaces (ICMI '02)*. IEEE Computer Society, Washington, DC, USA, 273–. https://doi.org/10.1109/ICMI.2002.1167006

Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. 2001. Eye Gaze Patterns in Conversations: There is More to Conversational Agents Than Meets the Eyes. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 301–308. https://doi.org/10.1145/365024.365119

Hana Vrzakova, Roman Bednarik, Yukiko I. Nakano, and Fumio Nihei. 2016. Speakers' Head and Gaze Dynamics Weakly Correlate in Group Conversation. In *Proc. of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA '16)*. ACM, New York, NY, USA, 77–84. https://doi.org/10.1145/2857491.2857522

Zhefan Ye, Yin Li, Yun Liu, Chanel Bridges, Agata Rozga, and James M Rehg. 2015. Detecting bids for eye contact using a wearable camera. In *Proc. of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Vol. 1. IEEE, 1–8. https://doi.org/10.1109/FG.2015.7163095

Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2017a. 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. 2015. Appearance-based Gaze Estimation in the Wild. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 4511–4520. https://doi.org/10.1109/CVPR.2015.7299081

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017b. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2299–2308. https://doi.org/10.1109/CVPRW.2017.284

Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2018. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018). https://doi.org/10.1109/TPAMI.2017.2778103