Sensing, Interpreting, and Anticipating Human Social Behaviour in the Real World

A dissertation submitted towards the degree
Doctor of Engineering
(Dr.-Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland University

by

Philipp Müller, M.Sc.

Saarbrücken
2020

Day of Colloquium	13^{th} of November, 2020
Dean of the Faculty	Prof. Dr. Thomas Schuster

Examination Committee

Chair	Prof. Dr. Jürgen Steimle
Reviewer, Advisor	Prof. Dr. Andreas Bulling
Reviewer	Prof. Dr. Elisabeth André
Reviewer	Prof. Dr. Antonio Krüger
Academic Assistant	Dr. Paul Swoboda

Abstract

Low-level nonverbal social signals like glances, utterances, facial expressions and body language are central to human communicative situations and have been shown to be connected to important high-level constructs, such as emotions, turn-taking, rapport, or leadership. A prerequisite for the creation of social machines that are able to support humans in e.g. education, psychotherapy, or human resources is the ability to automatically **sense**, **interpret**, and **anticipate** human nonverbal behaviour. While promising results have been shown in controlled settings, automatically analysing unconstrained situations, e.g. in daily-life settings, remains challenging. Furthermore, anticipation of nonverbal behaviour in social situations is still largely unexplored.

The goal of this thesis is to move closer to the vision of social machines in the real world. It makes fundamental contributions along the three dimensions of sensing, interpreting and anticipating nonverbal behaviour in social interactions. First, robust recognition of low-level nonverbal behaviour lays the groundwork for all further analysis steps. Advancing **human visual behaviour sensing** is especially relevant as the current state of the art is still not satisfactory in many daily-life situations. While many social interactions take place in groups, current methods for unsupervised eye contact detection can only handle dyadic interactions. We propose a novel unsupervised method for multi-person eye contact detection by exploiting the connection between gaze and speaking turns. Furthermore, we make use of mobile device engagement to address the problem of calibration drift that occurs in daily-life usage of mobile eye trackers. Second, we improve the **interpretation of social signals** in terms of higher level social behaviours. In particular, we propose the first dataset and method for emotion recognition from bodily expressions of freely moving, unaugmented dyads. Furthermore, we are the first to study low rapport detection in group interactions, as well as investigating a cross-dataset evaluation setting for the emergent leadership detection task. Third, human visual behaviour is special because it functions as a social signal and also determines what a person is seeing at a given moment in time. Being able to **anticipate human gaze** opens up the possibility for machines to more seamlessly share attention with humans, or to intervene in a timely manner if humans are about to overlook important aspects of the environment. We are the first to propose methods for the anticipation of eye contact in dyadic conversations, as well as in the context of mobile device interactions during daily life, thereby paving the way for interfaces that are able to proactively intervene and support interacting humans.

Zusammenfassung

Blick, Gesichtsausdrücke, Körpersprache, oder Prosodie spielen als nonverbale Signale eine zentrale Rolle in menschlicher Kommunikation. Sie wurden durch vielzählige Studien mit wichtigen Konzepten wie Emotionen, Sprecherwechsel, Führung, oder der Qualität des Verhältnisses zwischen zwei Personen in Verbindung gebracht. Damit Menschen effektiv während ihres täglichen sozialen Lebens von Maschinen unterstützt werden können, sind automatische Methoden zur **Erkennung**, **Interpretation**, und **Antizipation** von nonverbalem Verhalten notwendig. Obwohl die bisherige Forschung in kontrollierten Studien zu ermutigenden Ergebnissen gekommen ist, bleibt die automatische Analyse nonverbalen Verhaltens in weniger kontrollierten Situationen eine Herausforderung. Darüber hinaus existieren kaum Untersuchungen zur Antizipation von nonverbalem Verhalten in sozialen Situationen.

Das Ziel dieser Arbeit ist, die Vision vom automatischen Verstehen sozialer Situationen ein Stück weit mehr Realität werden zu lassen. Diese Arbeit liefert wichtige Beiträge zur autmatischen **Erkennung menschlichen Blickverhaltens** in alltäglichen Situationen. Obwohl viele soziale Interaktionen in Gruppen stattfinden, existieren unüberwachte Methoden zur Augenkontakterkennung bisher lediglich für dyadische Interaktionen. Wir stellen einen neuen Ansatz zur Augenkontakterkennung in Gruppen vor, welcher ohne manuelle Annotationen auskommt, indem er sich den statistischen Zusammenhang zwischen Blick- und Sprechverhalten zu Nutze macht. Tägliche Aktivitäten sind eine Herausforderung für Geräte zur mobile Augenbewegungsmessung, da Verschiebungen dieser Geräte zur Verschlechterung ihrer Kalibrierung führen können. In dieser Arbeit verwenden wir Nutzerverhalten an mobilen Endgeräten, um den Effekt solcher Verschiebungen zu korrigieren. Neben der Erkennung verbessert diese Arbeit auch die **Interpretation sozialer Signale**. Wir veröffentlichen den ersten Datensatz sowie die erste Methode zur Emotionserkennung in dyadischen Interaktionen ohne den Einsatz spezialisierter Ausrüstung. Außerdem stellen wir die erste Studie zur automatischen Erkennung mangelnder Verbundenheit in Gruppeninteraktionen vor, und führen die erste datensatzübergreifende Evaluierung zur Detektion von sich entwickelndem Führungsverhalten durch. Zum Abschluss der Arbeit präsentieren wir die ersten Ansätze zur **Antizipation von Blickverhalten** in sozialen Interaktionen. Blickverhalten hat die besondere Eigenschaft, dass es sowohl als soziales Signal als auch der Ausrichtung der visuellen Wahrnehmung dient. Somit eröffnet die Fähigkeit zur Antizipation von Blickverhalten Maschinen die Möglichkeit, sich sowohl nahtloser in soziale Interaktionen einzufügen, als auch Menschen zu warnen, wenn diese Gefahr laufen wichtige Aspekte der Umgebung zu übersehen. Wir präsentieren Methoden zur Antizipation von Blickverhal-

ten im Kontext der Interaktion mit mobilen Endgeräten während täglicher Aktivitäten, als auch während dyadischer Interaktionen mittels Videotelefonie.

Acknowledgements

First of all I would like to thank my supervisor Prof. Dr. Andreas Bulling for the great support he gave me since I joined his group as a student assistant. Since this time Andreas was a constant source of knowledge, inspiration and optimism. I especially grew to appreciate his ability to stay focussed on the positive side of things in times of difficulties when the prospects of a project look bleak.

I am grateful that I had the chance to collaborate with a number of wonderful people. Without the experience and help I received from Dr. Mykhaylo Andriluka, Dr. Michael Xuelin Huang, Dr. Xucong Zhang, Dr. Daniel Buschek, Prof. Dr. Yusuke Sugano, and Dr. Julian Steil, this thesis would not have been possible.

I thank the Japan Science and Technology Agency (JST) for providing the funding for my PhD (CREST Grant No.: JPMJCR14E1).

I would also like to thank Prof. Dr. Elisabeth André and Prof. Dr. Antonio Krüger for serving as reviewers of my thesis.

The Max Planck Institute for Informatics is a great place for research. I enjoyed a very pleasant time in a community of motivated and supportive people. In particular, I thank Prof. Dr. Bernt Schiele for helping to establish such a great working atmosphere, as well as Connie Balzert for her kindness and support on administrative issues.

Finally, I thank my family and friends for supporting me in all non-academic aspects of life. I am particularly grateful to Larry for her presence in my life.

Table of Contents

1 Introduction

Social interactions permeate our lives. They are ubiquitous both in professional and private contexts and their outcomes can shape our future significantly, be it the start of a romantic relationship, a new job, or the process through which a friendship evolves. Research in human science has shown that how we feel in- and judge social interactions is heavily influenced by nonverbal behaviour (Knapp *et al.*, 2013). In many cases, nonverbal behaviour is trusted more than its verbal counterpart - especially when both disagree (Pentland, 2010; Mehrabian and Ferris, 1967). This is rational, as nonverbal behaviour is under less conscious control than verbal behaviour and thus harder to fake (Burgoon *et al.*, 2016). Furthermore, many studies have shown the close link between nonverbal behaviour and important attributes of interactions. For example, gaze has been shown to be connected to, among others, turn-taking, liking, perceived dominance, and attraction (Kleinke, 1986). Even when only observing a short sequence of an interaction, humans are able to make use of the information present in nonverbal behaviour to for instance predict deception or measurements of the quality of the doctor-patient relation (Ambady and Rosenthal, 1992).

Today, machines are able to fulfil an increasing number of tasks that only humans were capable of in the past, including classification of images, navigation, or autonomously vacuum cleaning an apartment. In certain areas, machines even outperform humans, demonstrated by the success of super-human Chess and Go players (Campbell *et al.*, 2002; Silver *et al.*, 2017). In contrast to these advances, machines still lack behind when it comes to sensing, interpreting and anticipating human social behaviour. This is a severe obstacle to the creation of "social" machines that are able to support humans in fields like education, care, human resources, or psychotherapy. Such machines have the potential to offer new services and reduce the load on systems in society. For example, the average time patients have to wait before starting a psychotherapy financed by public health insurance in Germany is 19.9 weeks with rural regions facing significantly worse supply in therapies than cities (Bundespsychotherapeutenkammer, 2018). Machines that understand social behaviour could provide assistance to patients during this waiting time and be applied in aftercare, potentially leading to better outcomes without raising the demand on human labour. Furthermore, in the field of human resources, systems that are able to analyse group behaviour e.g. in terms of leadership or rapport can give

valuable feedback to employees and managers. This could help to improve collaboration and to increase the fit between people and their roles in organisations.

The vision of social machines has fueled years of research in affective computing and social signal processing, leading to impressive results. For example, machines are now able to extract facial expressions (Baltrusaitis *et al.*, 2018), estimate gaze (Kassner *et al.*, 2014; Zhang *et al.*, 2018c), detect eye contact (Zhang *et al.*, 2017b) and estimate body pose (Cao *et al.*, 2018; Insafutdinov *et al.*, 2017). Furthermore, they are able to infer higher level behaviours like emotions (Metallinou *et al.*, 2012), leadership (Beyan *et al.*, 2016a), rapport (Wang and Gratch, 2009) or group cohesion (Hung and Gatica-Perez, 2010). However, these achievements are still subject to constraints that hinder application in real life. For example, eye contact detection is only possible in dyadic interactions, mobile eye-trackers need to be re-calibrated repeatedly in real life conditions due to headset shifts and emergent leadership detection has only been shown to work when training and testing on the same dataset. Furthermore, while an increasing number of approaches to human behaviour anticipation has been proposed in recent years (Shen *et al.*, 2018; Zhang *et al.*, 2017a; Chiu *et al.*, 2019), gaze anticipation in social interactions has not been investigated. This is despite the potential of gaze anticipation to enable machines to more seamlessly produce adequate social behaviour like gaze following or to pro-actively manage user attention.

The goal of this thesis is to move towards methods that are able to sense, interpret and anticipate human nonverbal behaviour in social interactions happening in real life. In the following we give an introduction to each of these three dimensions of the thesis.

1.1 Gaze Sensing

Accurate gaze sensing is still a challenge in real-world conditions. This is in contrast to the significant importance of gaze in social interactions, where it serves as a building block in turn-taking and is connected to a variety of attributes, including leadership and attraction (Kendon, 1967; Kawase, 2014; Kellerman *et al.*, 1989). The three basic ways of sensing gaze in human-human interactions are stationary eye-trackers, mobile eye trackers and gaze estimation from ambient cameras. While stationary eye-trackers provide high fidelity gaze estimates, they are impractical for real-world social interactions, as they require dedicated hardware and are constrained to desktop settings. Consequently, we focus on mobile eye-trackers and ambient cameras. While mobile eye-trackers can deliver highly accurate gaze estimates when calibrated correctly, the quality of the calibration deteriorates in real-life conditions due to headset shifts, resulting in significantly worse gaze estimates (Sugano and Bulling, 2015a). To overcome this challenge, we propose a novel method to automatically re-calibrate mobile eye-trackers making use of users' tendency to engage with their mobile phones (Chapter 5). While mobile eye tracking provides a solution for many situations, they require user augmentation and can raise privacy concerns (Steil *et al.*, 2019a). An alternative can be gaze sensing from ambient cameras, which is still very challenging. A promising approach in this setting is to predict eye contact instead of continuous gaze estimates. In this way the problem is simplified while at the same time retaining the most important information for social

interaction analysis. However, current methods for eye contact detection from ambient cameras are only capable of detecting eye contact to single target objects (Zhang *et al.*, 2017b). This is in contrast to many social situations in which several other humans are present as potential targets for eye contact. To enable eye contact detection in such scenarios, we design a novel method which makes use of the correlation between gaze behaviour and speaking turns in order to train a personalised, multi-target eye contact detector without the need of manual supervision (Chapter 4).

1.2 Nonverbal Behaviour Interpretation

While encouraging progress in inferring higher-level social behaviour like bodily expressions of emotions, rapport or emergent leadership has been made, current methods are still subject to important constraints. Emotion recognition from video sequences has been intensively studied, but bodily expressions of emotions were only investigated for isolated people and expressions (Bänziger *et al.*, 2012), or within dyadic interactions of people wearing motion capture equipment (Metallinou *et al.*, 2010). Emotional expressions in real life however are often embedded in an interaction and performed by unaugmented people. We record the first dataset of bodily expressions of emotions embedded in dyadic interactions of freely-moving unaugmented people and develop a method to infer emotion classes from video input (Chapter 6). While knowledge about the emotion a person feels at a certain moment in time can be helpful for e.g. adapting the behaviour of a household robot, many application contexts require a more high-level description of a social interaction. Such a high-level measurement of interaction quality is rapport, the close and harmonious relationship in which interaction partners are "in sync" with each other. The failure to build rapport in interactions was shown to be connected to decreased collaboration and worse interpersonal outcomes (Burns, 1984; Kelley *et al.*, 2014; Tsui and Schultz, 1985). Being able to detect when a participant is not able to establish rapport with a group of people can open the door to offer support and improve the interaction. Such a system could be an important tool for group meetings at work or during studying, but no prior work on the detection of low rapport in interactions of more than two people exists to date. In this thesis, we are the first to record a dataset and design prediction algorithms to detect the failure to establish rapport within a group interaction (Chapter 7). Another crucial component of the social structure of a group are emergent leaders. These are people who gain influence in a group through the interaction, without necessarily holding formal authority (Stein and Heller, 1979). Even without formal authority, emergent leaders influence group performance (Druskat and Pescosolido, 2006; Kickul and Neuman, 2000). Therefore, it is important for organisations to know who their emergent leaders are. Automatic approaches to emergent leadership detection from nonverbal behaviour have been increasingly studied in recent years (Beyan *et al.*, 2017c; Sanchez-Cortes *et al.*, 2012). However all these approaches were trained and tested on the same dataset. This is unrealistic in real-world settings, where deploying a trained system in similar, but slightly different environments is desired. We are the first to investigate such a scenario (Chapter 8), showing that it is possible to achieve an emergent leadership detection

accuracy of 0.68 (random baseline: 0.29) on a dataset unseen at training time by using a combination of pose- and visual focus of attention features extracted with the help of our eye contact detection approach presented in Chapter 4.

1.3 Visual Behaviour Anticipation

Machines with the ability to anticipate gaze behaviour in social situations can be beneficial for a number of applications. For example, machines might be able to more seamlessly engage in joint attention and gaze following (Skantze *et al.*, 2014), or to proactively support humans who have difficulty in maintaining eye contact behaviour that is judged as socially appropriate (e.g. in autism spectrum disorder (Senju and Johnson, 2009)). Furthermore, as gaze has a perceptual function alongside functioning as a social signal and cue (Gobel *et al.*, 2015), anticipating gaze informs the machine what the user is going to perceive in the near future. This can, for example, be helpful for increasing the users' safety by issuing a warning in case the user is going to overlook something in the environment due to e.g. being immersed in a chat conversation on the mobile phone. Despite these important application scenarios, no attempt has been made to anticipate gaze behaviour in social situations. In this thesis we are the first to propose a method for gaze anticipation during mobile device interactions in everyday social situations like eating or studying in a library (Chapter 9). This method combines features extracted from the mobile phone with analysis of the video recorded from an egocentric camera. To evaluate our method, we recorded a 90-hour dataset of users interacting with their mobile phone during daily-life activities on campus. Furthermore, we are the first to develop a method to anticipate eye contact in natural dyadic interactions during video conferencing (Chapter 10). Our method analyses an interactants' past gaze behaviour, facial expressions, and head movement to predict eye contact with the conversation partner in the near future. We demonstrate the effectiveness of our method through evaluations on a newly annotated dataset of video conferencing interviews obtained from Youtube.

2 Related Work

This chapter provides an introduction to related work relevant to the aims of this thesis. First, we review the state of the art in gaze sensing methods applicable in face-to-face social interactions. Subsequently, we discuss approaches to the interpretation of social nonverbal behaviour, and conclude the related work with an examination of research on human behaviour anticipation.

In the literature a distinction is made between social *signals* and social *cues*. The exact use of these two terms varies. The ethological perspective of Mehu and Scherer (2012) sees social signals as entities that evolved because of their effect on perceivers. In contrast, while social cues can still be useful for communication, they evolved independent of effects on perceivers. In their seminal survey on social signal processing Vinciarelli *et al.* (2009) follow a different understanding of social cues and signals. Here, social cues like eye contact, posture, vocal behaviour, and others combine to form a social signal like for example disagreement. In the context of this thesis, the precise distinction between social cue and social signal is not important. We will use both terms to describe aspects of human behaviour in social interactions that carry informative value.

2.1 Gaze Sensing

As gaze is recognised as a highly important social signal, gaze sensing in everyday interactions is key to realising the full potential of social machines. Gaze sensing needs to be **accurate**, **robust**, and **unobtrusive** under real life conditions in order to provide valuable input to further processing. To be adopted by consumers, the devices should also be **inexpensive**. The most accurate eye tracking results are achieved by stationary eye trackers like the Tobii Pro Spectrum[1] which is able to record 1200 frames per second and estimate gaze with an error of only 0.3° of visual angle. Such stationary eye trackers are extensively used in laboratory research (Theeuwes *et al.*, 1998; Schwedes and Wentura, 2012; Sattar *et al.*, 2015). However stationary eye trackers are not a good fit for real-life social interactions. They assume users to be located at a fixed position for robust performance and need to be positioned close to the user (e.g. between 55 and 75 cm distance to the user for Tobii Pro Spectrum), which makes them obtrusive. Furthermore, even consumer variants like the Tobii 4C[2] are still significantly more expensive compared to standard RGB cameras. Consequently, this thesis focuses on two approaches to gaze sensing that are better suited for real-life social interactions: mobile eye tracking and eye contact detection using ambient RGB cameras. While mobile eye tracking opens up the possibility to estimate users' gaze throughout their daily life, ambient cameras are a cheap and unobtrusive means of sensing gaze patterns in group interaction like business meetings or study groups. Both approaches come with specific advantages and challenges which will be discussed in the following.

2.1.1 Mobile Eye Tracking, Calibration & Recalibration

Mobile eye trackers usually consist of scene and eye cameras mounted on a dedicated headset (Kassner *et al.*, 2014), with first models integrating the cameras into a regular glasses frame[3], making them unobtrusive and potentially increasing social acceptability. Mobile eye trackers like the Pupil Core[4] can reach an accuracy of up to 0.6° of visual angle, making them a viable option to measure social signals like eye contact or joint attention. As with stationary eye trackers, to achieve such high accuracies, a mobile eye tracker needs to be calibrated. Calibration usually consists of a procedure during which the user is instructed to gaze at defined points. The resulting pairs of gaze points and pupil detections obtained from the eye cameras are then used to calculate a mapping to estimate gaze on new pupil detections (Kasprowski *et al.*, 2014). A major challenge for mobile eye tracking under real-life conditions is that this mapping is compromised by calibration drift, which describes the deterioration of gaze estimation accuracy as a result of headset slippage (Sugano and Bulling, 2015a). Such slippage is hardly avoidable if eye trackers are supposed to be worn during daily activities. Asking the user to

[1] `https://www.tobiipro.com/product-listing/tobii-pro-spectrum/`, date: 10.03.2020
[2] `https://gaming.tobii.com/tobii-eye-tracker-4c/`, date: 10.03.2020
[3] `https://pupil-labs.com/products/invisible/`, date: 10.03.2020
[4] `https://pupil-labs.com/products/core/tech-specs/`, date: 10.03.2020

repeatedly re-calibrate the eye tracker would worsen the user experience and likely be a major hurdle to the acceptance of mobile eye trackers.

To address these issues, approaches to automatic calibration and recalibration of eye trackers have been developed. The first works on automatic eye tracker calibration focussed on the stationary setting, employing eyeball models for calibration (Takegami *et al.*, 2002; Yamazoe *et al.*, 2008). Later work exploited the link between mouse clicks and gaze to automatically calibrate the eye tracker (Sugano *et al.*, 2008). Subsequently, different kinds of user interactions like typing and dragging were studied as inputs to automatic eye tracker calibration in addition to mouse clicks (Huang *et al.*, 2016a). Recent work by Zhang *et al.* (2018b) extended the idea of automatic calibration via user interactions to support several different mobile and handheld devices. More general self-calibration approaches in the stationary setting exploited bottom-up saliency maps extracted on images viewed by the user (Sugano *et al.*, 2010; Chen and Ji, 2015) as well as gaze patterns recorded from other users on the same images (Alnajar *et al.*, 2013).

For mobile eye trackers, fewer automatic calibration approaches have been proposed to date. Corneal images are one possibility that has been explored in the literature (Lander *et al.*, 2017; Takemura *et al.*, 2014a). In Lander *et al.* (2017), an infrared as well as a RGB eye camera were utilised. Fast and reliable pupil tracking was achieved using the infrared camera, while corneal images were extracted using the RGB eye camera in order to construct a connection to the scene. The drawback of such a corneal imaging based approach is the need for an additional RGB eye camera, while common mobile eye trackers are only equipped with an infrared eye camera. Furthermore, corneal imaging approaches can also struggle more with suboptimal lighting conditions (Takemura *et al.*, 2014a). As the RGB camera can be used to decode privacy-sensitive scene information from the eye (Backes *et al.*, 2008), corneal imaging is not a viable option in privacy-sensitive situations. The severity of calibration drift in mobile eye trackers was first demonstrated by Sugano and Bulling (2015a), who proposed a method based on saliency maps that was able to retain the quality of an initial manual calibration. The employed saliency maps combined standard bottom-up saliency approaches with person and face detectors. This approach showed promising results on data recorded in a free-viewing setting. However, such a free-viewing setting does not approximate the complexity of gaze behaviour in real life, which involves top-down components from task-driven behaviour. As a result, it remains unclear to what extent this saliency-based recalibration method is applicable in real-life interactions. This thesis for the first time investigates calibration drift in eye tracking recordings of real-life behaviour in different environments and including social- and mobile device interactions (Chapter 5). Furthermore, it proposes two novel automatic recalibration methods which exploit mobile phone interactions and outperform bottom-up saliency-based recalibration in real-life settings.

At the time of publication of the method presented in Chapter 5, Santini *et al.* (2019) introduced a geometrical approach to obtain a slippage-robust input feature for gaze estimation in mobile eye trackers. While their approach showed promising results on a dataset of museum visits, in contrast to Chapter 5 of this thesis, their method was not evaluated on real-life interactions that include activities like eating, mobile device interaction, studying, and extensive locomotion between different buildings.

2.1.2 Eye Contact Detection from Ambient Cameras

In contrast to the continuous gaze estimation task, eye contact detection uses a discrete output space, indicating the objects or people that a target person is looking at. For social behaviour analysis, such a discretised output is desirable as it allows to directly compute features like the amount of gaze a certain person receives from interactants. Due to the difficulty of estimating the gaze direction from ambient cameras, many works analysing eye contact in group interactions have fallen back to relying on head pose estimates as a proxy to gaze direction (Stiefelhagen, 2002; Gatica-Perez *et al.*, 2005; Beyan *et al.*, 2017b). In order to estimate gaze in spite of lacking direct observations of gaze direction, a number of works employed a Bayesian approach that is treating gaze direction as a hidden variable (Otsuka *et al.*, 2005, 2007; Otsuka and Yamato, 2008; Ba and Odobez, 2010). For example, Ba and Odobez (2010) proposed a multimodal method to detect the focus of attention of meeting participants. Head pose estimates were augmented by a dynamic Bayesian network based context model incorporating participants' locations, speaking proportion and activity on the projection screen in order to model participants' visual focus of attention. However, incorporating information from the eye has the potential to obtain a much more accurate estimate of participants' attention, especially as it was shown that significant differences exist between head pose and actual eye movements in group interactions (Vrzakova *et al.*, 2016).

Even though the human eye may only cover an area of a few pixels if recorded from an ambient camera, progress in eye contact detection from ambient cameras that takes actual eye information in account has been made. The first work to incorporate vision-based gaze detection into an eye contact detection model for group interactions was presented by Gorga and Otsuka (2010). After detecting the eye region, the authors performed wavelet decomposition and extracted radial signals from the pupil-centred wavelet images. Subsequently, support vector machines (SVMs) were trained to classify gaze direction into up to 5 horizontally arranged classes. Hyperparameters have to be tuned according to the concrete geometrical layout of the room. In contrast, this thesis for the first time presents a method to detect eye contact in group interactions from ambient cameras that does not require any form of manual supervision (Chapter 4). Instead, our method makes use of the link between gaze and speaking activity in order to train a dedicated eye contact detector for every participant in the interaction.

After the publication of our work presented in Chapter 4, Otsuka *et al.* (2018) used convolutional neural networks (CNNs) to estimate eye contact in group meetings based on multimodal inputs consisting of head pose, utterance and horizontal eye direction. Subsequently, Zhang *et al.* (2019a) presented an approach to detect eye contact using eye gaze and head pose estimates extracted from OpenFace (Baltrusaitis *et al.*, 2018) as input to a multi-layer perceptron. In both cases, the networks have to be trained in a supervised fashion for the specific seating position for which they are to be applied.

A separate line of work on visual focus estimation investigated the case in which both the participant as well as the attention targets are present in the same image (Soo Park and Shi, 2015; Recasens *et al.*, 2015; Ohshima and Nakazawa, 2019; Guan *et al.*, 2020; Chong *et al.*, 2020). Soo Park and Shi (2015) introduced the "social saliency prediction"

task, in which the joint focus of attention of a group of people present in an image is estimated. In contrast, Recasens *et al.* (2015) estimated the gaze targets of individual people present in images by training a deep neural network that is able to extract head orientation and gaze and choose objects in peoples' line of sight which are likely to be attended. This approach was later extended to follow peoples' gaze in videos across views (Recasens *et al.*, 2017). In a recent study, Ohshima and Nakazawa (2019) studied the specific case of eye contact detection from a third-person view on a dyadic interaction. These works are not applicable to the situation usually present in meetings recorded with ambient cameras, where peoples' gaze targets are not in the same camera view as the people themselves (Gatica-Perez *et al.*, 2005; Beyan *et al.*, 2017b; Müller *et al.*, 2018a).

2.2 Nonverbal Social Behaviour Interpretation

Basic nonverbal cues like eye contact, proximity, facial expressions, or body movements are connected to higher-level aspects of social situations like emotions, leadership, collaboration quality, and rapport in a large variety of different ways. These connections can be rather stable and well-understood, like the close connection between different facial action units and displayed emotions (Wiggers, 1982), or more elusive and context-dependent, like the connection between rapport and nonverbal behaviour (Tickle-Degnen and Rosenthal, 1990). Research on the interpretation of basic nonverbal signals in social situations is diverse both in the settings and modalities that are investigated as well as in the target concepts that are extracted.

Early research studied the use of wearable devices ("sociometric badges") to detect face-to-face proximity and conversations over several days to build a model of interpersonal relations inside a group (Choudhury and Pentland, 2003). This approach was extended with email data (Waber *et al.*, 2007), and was also used in conjunction with topic models to mine social interaction routines during a long-term mission in a confined space (Zhang *et al.*, 2018d). Using sociometric badges in conjunction with ambient cameras, Alameda-Pineda *et al.* (2015) analysed social gatherings to detect the formation of subgroups and respective social attention attractors. A different task that has been extensively studied is the detection of violent behaviour in crowds (Mohammadi *et al.*, 2015, 2016; Marsden *et al.*, 2017), or in smaller groups recorded from surveillance cameras (Bilinski and Bremond, 2016; Fu *et al.*, 2018).

While these perspectives on social behaviour are highly valuable, many interactions in private and professional life take place in small groups with a defined beginning and end. Examples include business meetings, school classes, or study groups. A growing body of research has shown that a detailed analysis of such small group interactions can lead to important insights on various aspects of the interaction (Feese *et al.*, 2011; Avci and Aran, 2016; Nanninga *et al.*, 2017; Beyan *et al.*, 2019b). What is common across many prediction tasks in group analysis is that information about the target can be extracted from several modalities and the combination of different modalities can often lead to improved performance. For example, Gatica-Perez *et al.* (2005) studied the prediction of group interest level in meetings, showing that adding visual features to audio features results in the highest performance. Later work highlighted the importance of gaze in individual and group engagement (Oertel and Salvi, 2013). Further aspects of group interactions inferred from visual and auditive nonverbal behaviour include group cohesion (Hung and Gatica-Perez, 2010; Nanninga *et al.*, 2017), the performance of group decisions (Avci and Aran, 2016; Kubasova *et al.*, 2019), and the prediction of dominant participants (Bai *et al.*, 2019). One of the most active research areas in small group analysis in recent years is the identification of leaders (Feese *et al.*, 2011; Sanchez-Cortes *et al.*, 2012; Beyan *et al.*, 2017b; Okada *et al.*, 2019) as well as the estimation of participant personality (Aran and Gatica-Perez, 2013; Kindiroglu *et al.*, 2017; Celiktutan and Gunes, 2017; Beyan *et al.*, 2019b). A special case of small groups are dyadic interactions consisting of two people only. They have been extensively studied in the context of rapport estimation (Wang and Gratch, 2009; Hagad *et al.*,

2011), emotion recognition (Lee *et al.*, 2011; Metallinou *et al.*, 2013), desire during speed-dating (Veenstra and Hung, 2011), or blaming behaviour of married couples (Black *et al.*, 2010).

In the following we will focus in more detail on the three prediction targets that are most important in the context of this thesis: emotion recognition from body movements, rapport estimation and emergent leadership detection.

2.2.1 Emotion Recognition from Body Movements

Research on emotion recognition from different nonverbal channels like prosody (Schuller *et al.*, 2003; Vogt *et al.*, 2008), facial expressions (Cohen *et al.*, 2000) or physiological signals (Kim and André, 2008) has a long history. An especially interesting channel for emotion recognition are body movements, as they are believed to be particularly hard to fake (Burgoon *et al.*, 2016). Behavioural science research has established a close connection between emotions and body movements (Wallbott, 1998; De Meijer, 1989; Pollick *et al.*, 2001).

A large body of research on the inference of emotions from body movements exists, with a comprehensive albeit slightly outdated survey by Karg *et al.* (2013). Most studies focussed on recognising emotions from non-interacting individuals e.g. by using motion capture equipment to record isolated portraits of sadness, joy, anger and fear (Kapur *et al.*, 2005). Using a support vector machine on top of features extracted from the motion capture data, the authors were able to achieve classification accuracies of above 80%. Subsequently, Bernhardt and Robinson (2007) proposed a method to detect affect in non-stylised motions like knocking, throwing, or walking. Furthermore, the same authors also studied the harder problem of emotion detection from connected action sequences (Bernhardt and Robinson, 2009). More recent work on emotion recognition from body movements of single individuals continued to investigate the recognition of emotions displayed in walking (Stephens-Fripp *et al.*, 2017; Randhavane *et al.*, 2019; Bhattacharya *et al.*, 2019) and the contribution of different pose-based cues on emotion classification during motion captured daily activities (Fourati *et al.*, 2019), as well as the development of real-time systems (Wang *et al.*, 2015b).

While most existing work on emotion recognition from body movements studied isolated individuals, there is a growing body of research focussing on interacting individuals. The IEMOCAP corpus features dyadic interactions of seated people (Busso *et al.*, 2008), with facial expressions and wrist movements recorded using motion capture equipment. The study of emotion recognition based on full-body movements in interactions was made possible by later research by Metallinou *et al.* (2010). In their USC CreativeIT database, the authors recorded improvised dyadic interactions using full-body motion capture equipment. Based on this dataset, subsequent work presented a method to automatically track emotional trends of participants using body language and speech information in a Gaussian mixture model framework (Metallinou *et al.*, 2013). The USC CreativeIT database was further used to study the prediction of interaction attitudes (friendly versus conflictive) from hand movements (Yang *et al.*, 2014a). Work on a different dataset by Wang *et al.* (2014) investigated emotion recognition from bodily

expressions in child-robot interaction with pose skeletons extracted using Kinect sensors. More recently, Bozkurt *et al.* (2017) introduced the JESTKOD database consisting of speech and motion capture recordings of dyadic interactions in agreement or disagreement scenarios, which has been used for emotion recognition from speech and body motion (Fatima and Erzin, 2017).

While these databases and approaches are valuable in advancing the state of the art in emotion recognition from body movements, they rely on specialised recording equipment and thus do not reflect the challenges associated with emotion recognition in dyadic interactions in real-world settings where only common RGB cameras and microphones are available. To bridge this gap, this thesis presents the first database and method for emotion recognition from bodily expressions in dyadic interactions of unaugmented people (Chapter 6).

2.2.2 Rapport Estimation

Rapport, the tendency of interaction partners to feel "in sync" with each other, is arguably one of the most important aspects of social interactions as it lays the foundation of mutual understanding and effective communication. The failure to build rapport can result in decreased collaboration and worse interpersonal outcomes (Burns, 1984; Kelley *et al.*, 2014; Tsui and Schultz, 1985). Early work by Tickle-Degnen and Rosenthal (1990) stressed the fact that rapport is not a property that a single individual possesses but only exists in the interaction between people. Furthermore, the authors identified attention, positivity, and coordination to be important factors related to rapport. The connection between rapport and nonverbal behaviour was for example investigated by Harrigan *et al.* (1985), who found that physicians received higher rapport ratings when they were sitting with arms in symmetrical side-by-side positions and uncrossed arms directly facing the patient.

Motivated by the connections between nonverbal behaviour and rapport, computational approaches to estimate rapport have been developed. In the visual domain, Wang and Gratch (2009) employed selected facial action units (AUs) to predict felt rapport in human-human as well as in human-agent interactions. Their findings indicate that rapport is encoded in the absence of AUs associated with negative emotions rather than in the presence of AUs associated with positive emotions. Apart from facial behaviour, body postures and body posture congruence were used for rapport prediction in dyadic interactions (Hagad *et al.*, 2011). More recent work investigated nonverbal features together with verbal behaviour for estimating rapport. Temporal pattern mining was applied to extract rules for rapport management, comparing dyads consisting of friends with dyads consisting of strangers in a peer-tutoring task (Zhao *et al.*, 2014, 2016). In a human-agent interaction setting, Cerekovic *et al.* (2016) combined facial expressions with linguistic content, nonverbal auditory cues, and body pose and motion cues to predict rapport. Furthermore, they found that using results of a personality test as features can improve rapport prediction.

While these works on rapport prediction are encouraging, they focus exclusively on dyadic interactions. This is in contrast to many daily-life interactions consisting

of multiple people, thereby limiting the applicability of existing rapport prediction algorithms in the real world. Due to the potential negative consequences for interaction quality and interaction outcome it is especially important to be able to identify when individuals fail to establish rapport with others in group interactions. To overcome these limitations, this thesis proposes the first dataset and method to detect low rapport in group interactions from nonverbal behaviour (Chapter 7).

2.2.3 Emergent Leadership Detection

Emergent leaders obtain their leadership position through interaction with a group, and do not necessarily hold formal authority (Stein and Heller, 1979). Even without such formal authority, they are of significant importance for group performance (Druskat and Pescosolido, 2006; Kickul and Neuman, 2000). Connections between emergent leadership and nonverbal behaviour are well established in the literature. Examples include a study by Baird Jr (1977) indicating a relationship between gesticulation with the shoulders and arms and emergent leadership, as well as a recent study by Gerpott *et al.* (2018) who conducted an eye tracking study in which people watched recordings of group meetings. The results revealed that observers gaze at emergent leaders more often and longer than at non-leaders. Further work investigating the behaviour of emergent leaders during turn transitions found that leaders are more likely to show prolonged gaze at the end of utterances, which might act as an offer to take the floor (Kalma, 1992).

The connections between emergent leadership and nonverbal behaviour were exploited by a number of approaches to automatic emergent leadership detection in group interactions, thereby focussing on two datasets. The ELEA dataset, recorded by Sanchez-Cortes *et al.* (2012), consists of groups of three to four people solving the winter survival task. In this task, participants are asked to agree on a list of items that will be useful for survival after a plane crash in a remote location during winter. Approaches to emergent leadership detection on ELEA used audio- and visual as well as multi-modal features (Sanchez-Cortes *et al.*, 2012, 2013). In Sanchez-Cortes *et al.* (2012) these include speaking-turn based features as well as basic prosodic features alongside visual features describing head- and body activity. Fusing features from the audio and visual domain resulted in best performance for emergent leadership detection. Subsequent work analysed visual focus of attention (VFOA) features based on head orientation together with speaking turn features as well as multimodal features that combine both input channels, e.g. being looked at while speaking (Sanchez-Cortes *et al.*, 2013).

A second, more recent line of work on emergent leadership detection in group interactions centres on the PAVIS dataset (Beyan *et al.*, 2016b), which consists of meetings of four people instructed to solve similar survival tasks as in the ELEA dataset. Initial work on the PAVIS dataset focussed on the detection of emergent leaders from nonverbal visual features exclusively (Beyan *et al.*, 2016b). This work used head pose estimation to extract a 15-dimensional featureset describing the visual focus of attention (VFOA) of participants, which was used to train a SVM to detect emergent leaders. Subsequent work combined these VFOA features with head activity and body activity

based features in a multiple kernel learning framework (Beyan *et al.*, 2016a). The highest performance for the emergent leadership detection task was achieved by combining VFOA features with features describing the geometrical configurations of body parts of a person (Beyan *et al.*, 2017c). Later work focussed on detecting both the person who received the highest, as well as the person who received the lowest leadership scores in a group. Performance improvements for this task were gained by using deep visual activity features (Beyan *et al.*, 2018) and by employing sequential analysis (Beyan *et al.*, 2019a). Apart from detecting emergent leaders in group interactions, the PAVIS dataset was also used to classify leadership style (Beyan *et al.*, 2017b, 2018).

While this previous work on emergent leadership prediction is highly valuable for understanding which aspects of nonverbal behaviour can be used to detect emergent leaders, approaches are always trained and tested on the same dataset. This assumption of identical training and testing distributions however does not reflect the challenges associated with real-world application scenarios, where it is required to be able to apply a trained system in similar, but slightly different settings (e.g. a different organisation or meeting situation). To overcome this limitation, this thesis for the first time evaluates emergent leadership detection algorithms across different datasets, showing that a combination of VFOA and body pose features can achieve accuracies of up to 0.68 on a target dataset unseen at training time (Chapter 8).

2.3 Behaviour Anticipation

Humans are capable of anticipating others' behaviour and make use of this capability in order to increase the smoothness of interactions (Duarte *et al.*, 2018; Aglioti *et al.*, 2008). Driven by diverse application scenarios like autonomous driving, human-robot interaction or foveated rendering, a large body of research has addressed the problem of automatic human behaviour anticipation.

To increase driving safety and to enable autonomous driving it is crucial to anticipate the behaviour of other traffic participants. Work in this area has focussed on constrained settings like lane change prediction (Mandalia and Salvucci, 2005; Woo *et al.*, 2017) as well as more general tasks like predicting the trajectories of pedestrians, cars, and bikes in traffic scenes (Bhattacharyya *et al.*, 2018; Ma *et al.*, 2019).

A different context in which human behaviour anticipation is highly relevant is human-robot interaction. To coordinate robot behaviour with human behaviour, e.g. when shaking hands or passing on an object, anticipation of future human motion is important. Anticipation of human motion can take place on different levels. On the lower level, an increasing number of studies have been concerned with prediction future body pose of humans (Chiu *et al.*, 2019; Chao *et al.*, 2017; Walker *et al.*, 2017; Butepage *et al.*, 2017). For example, Chiu *et al.* (2019) proposed a new recurrent neural network architecture to anticipate human pose both on short and long timescales. Their network is able to anticipate human pose for different action classes including walking, eating, smoking and discussion without being provided with action class labels. On a higher level, research has focussed on the anticipation of gestures and actions (Saponaro *et al.*, 2013; Schydlo *et al.*, 2018; Narber *et al.*, 2015; Vamplew and Adams, 1995; Huang and Mutlu, 2016). For example, Huang and Mutlu used human gaze to anticipate which ingredient a human will pick next for a sandwich that is going to be assembled by a robot (Huang and Mutlu, 2016). By making use of the anticipated human actions, the robot was able to complete the sandwich making task faster compared to behaving in a purely reactive way. In recent work, Schydlo *et al.* (2018) employed an encoder-decoder recurrent neural network model to anticipate human action sequences using gaze and body pose cues. The anticipated actions include directed giving and placing actions as well as actions like pouring or drinking.

In addition to these works centred on a human-robot interaction setting, recent years have seen a large number of studies on egocentric action anticipation (Furnari and Farinella, 2019; Shen *et al.*, 2018; Liu *et al.*, 2019; Farha and Gall, 2019; Guan *et al.*, 2019; Furnari *et al.*, 2018). In this line of research, observations obtained from an egocentric camera are used to predict the next action the wearer will be performing. Progress on egocentric action anticipation is to a large part driven by the EPIC-Kitchens Action Anticipation Challenge (Damen *et al.*, 2018), which includes common actions occurring during cooking, food preparation and washing up.

In contrast to the large body of research on anticipating human actions and body movements, relatively little work has been done on gaze anticipation. The existing work can be grouped according to the timescale on which gaze is anticipated. While some works have attempted to anticipate gaze on a very short timescale by predicting the

landing location of the currently executed saccade (Arabadzhiyska *et al.*, 2017; Griffith *et al.*, 2018; Morales *et al.*, 2018; Wang *et al.*, 2017), others have the goal of predicting gaze location for up to several seconds (Xu *et al.*, 2018; Zhang *et al.*, 2017a, 2018a). A major motivation for saccade landing point prediction is gaze-contingent rendering, which has applications in vision science research, for example to simulate loss of central vision (Pidcoe and Wetzel, 2006). Furthermore, gaze contingent rendering can be used to improve the perceived quality of rendered images (Arabadzhiyska *et al.*, 2017). Methods for saccade landing point prediction analyse the beginning of a saccade in order to determine its target location, for example by standard linear regression (Arabadzhiyska *et al.*, 2017) or using long short-term memory networks (Morales *et al.*, 2018). Only few works have explored gaze anticipation for longer time horizons than the duration of a saccade. Zhang *et al.* (2017a, 2018a) were the first to anticipate gaze location in egocentric videos of meal preparation and object search for time horizons of up to 3.2 seconds. Their approach consists of two steps. First, it generates future video frames using a generative adversarial network (GAN). Second, based on these generated frames, future saliency maps are predicted. In a different work, Xu *et al.* (2018) studied gaze anticipation for the next 250 ms in 360° videos during free-viewing. Using image content and scanpath history as input, they predicted the displacement of the future gaze point relative to the current gaze point.

To the best of our knowledge, no prior work attempted gaze anticipation in social interactions. In this thesis, we collect datasets and develop methods to anticipate gaze allocation on a mobile phone embedded in daily life situations (Chapter 9). Furthermore, we are the first to present a method to anticipate eye contact during dyadic conversations (Chapter 10).

Thesis Summary

3

This chapter summarises the thesis by discussing the main contributions (Section 3.2), limitations and future work (Section 3.3), as well as the significance of the thesis (Section 3.4). A visual overview of this thesis is given in Figure 3.1, while Table 3.1 provides a list of chapters with corresponding publications.

3.1 Outline of the Thesis

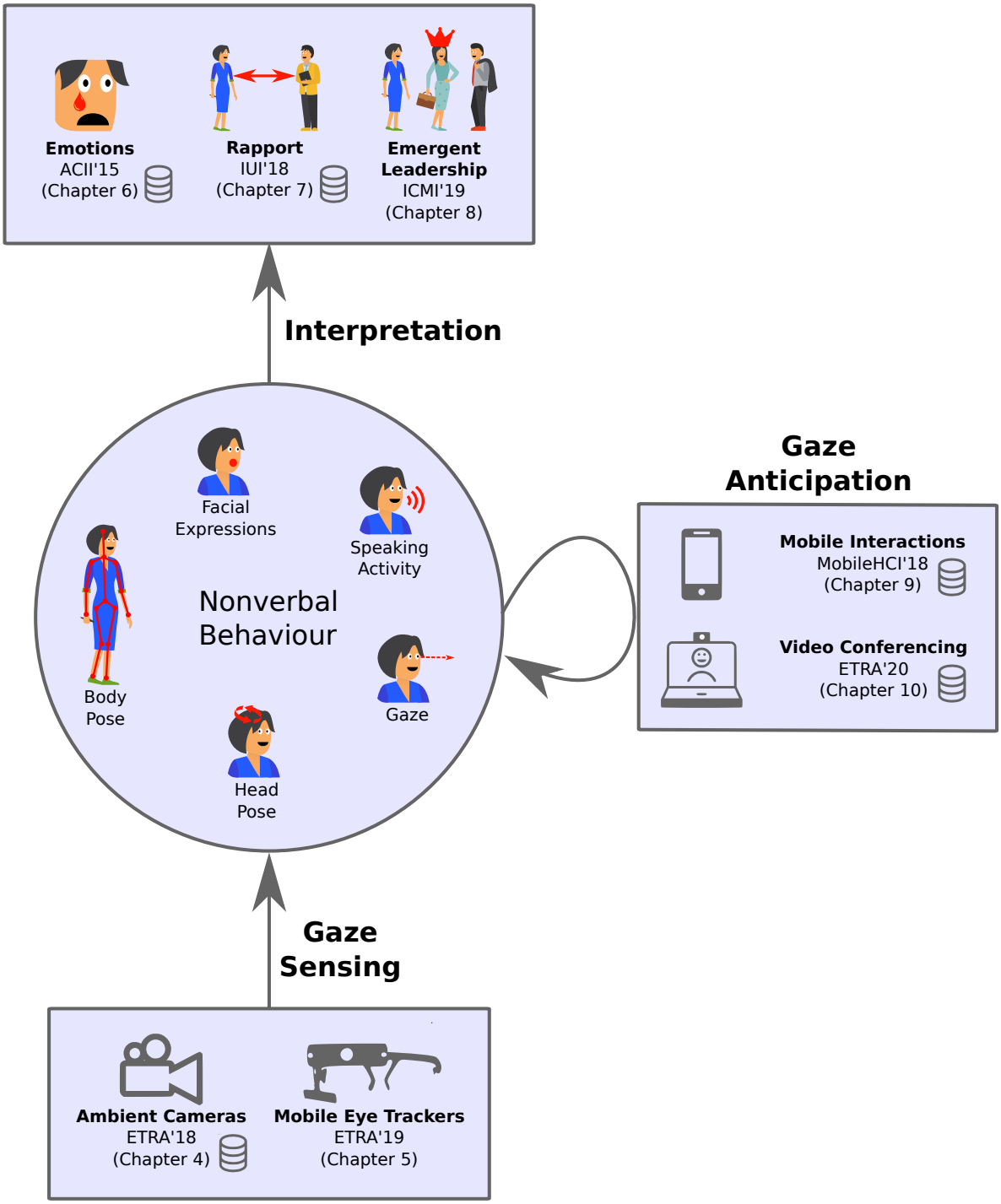


Figure 3.1: Outline of the chapters of this thesis (see Table 3.1 for details on the corresponding publications). Chapters containing a dataset contribution are indicated by a database icon.

Chapter	Publication
4	*Robust Eye Contact Detection in Natural Multi-Person Interactions Using Gaze and Speaking Behaviour* Philipp Müller, Michael Xuelin Huang, Xucong Zhang, and Andreas Bulling; In Proc. of the ACM Symposium on Eye Tracking Research and Applications (**ETRA**), 2018. (Müller *et al.*, 2018b)
5	*Reducing Calibration Drift in Mobile Eye Trackers by Exploiting Mobile Phone Usage* Philipp Müller, Daniel Buschek, Michael Xuelin Huang, and Andreas Bulling; In Proc. of the ACM International Symposium on Eye Tracking Research and Applications (**ETRA**), 2019. (Müller *et al.*, 2019)
6	*Emotion recognition from embedded bodily expressions and speech during dyadic interactions* Philipp Müller, Sikandar Amin, Prateek Verma, Mykhaylo Andriluka, and Andreas Bulling; Proc. of the International Conference on Affective Computing and Intelligent Interaction (**ACII**), 2015. (Müller *et al.*, 2015)
7	*Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behavior* Philipp Müller, Michael Xuelin Huang, and Andreas Bulling; In Proc. of the ACM International Conference on Intelligent User Interfaces (**IUI**), 2018. (Müller *et al.*, 2018a)
8	*Emergent Leadership Detection Across Datasets* Philipp Müller and Andreas Bulling; In Proc. of the International Conference on Multimodal Interaction (**ICMI**), 2019. (Müller and Bulling, 2019)
9	*Forecasting User Attention During Everyday Mobile Interactions Using Device-Integrated and Wearable Sensors* 🏆 **Best Paper Award** Julian Steil, Philipp Müller, Yusuke Sugano, and Andreas Bulling; In Proc. of the ACM International Conference on Human-Computer Interaction with Mobile Devices and Services (**MobileHCI**), 2018. (Steil *et al.*, 2018b)
10	*Anticipating Averted Gaze in Dyadic Interactions* Philipp Müller, Ekta Sood, and Andreas Bulling; In Proc. of the ACM International Symposium on Eye Tracking Research and Applications (**ETRA**), 2020. (Müller *et al.*, 2020)

Table 3.1: Publications included in this thesis with corresponding chapters.

3.2 Summary of Contributions

This chapter summarises the contributions this thesis makes towards the goal of methods that are able to **sense**, **interpret**, and **anticipate** nonverbal behaviour in social interactions.

Gaze Behaviour Sensing. The first goal of this thesis is to improve gaze sensing in real-world scenarios that are relevant to social interaction analysis. To this end, we introduce the first method for unsupervised eye contact detection from ambient cameras in group interactions. Our method makes exploits the link between speaking behaviour and gaze to train a dedicated eye contact detector for every participant (see Section 3.2.1 and Chapter 4). To combat calibration drift in mobile eye trackers, we introduce a novel automatic recalibration approach that makes use of mobile phone interactions and leads to robust improvements across a wide variety of settings (see Section 3.2.2 and Chapter 5).

Nonverbal Behaviour Interpretation. The second goal of this thesis is to bring nonverbal behaviour interpretation to more realistic settings. We propose the first dataset and method for emotion recognition from body movements of freely-moving dyads not wearing dedicated motion capture equipment (see Section 3.2.3 and Chapter 6). Furthermore, we propose the first dataset and method for low rapport detection in small group interactions. In evaluations on our novel dataset we show that features based on facial action units are most effective in detecting low rapport (see Section 3.2.4 and Chapter 7). Finally, we are first to investigate a cross-dataset evaluation setting for the emergent leadership detection task. Using a combination of features based on eye contact and body poses, we are able to detect emergent leaders with an accuracy of 0.68 on a dataset unseen at training time (see Section 3.2.5 and Chapter 8).

Gaze Behaviour Anticipation. The third goal of this thesis is to enable machines to anticipate human visual behaviour during social interactions. We propose novel methods for two different scenarios. In the context of mobile device interactions during daily-life situations, we develop the first approach to detect attention shifts to- and from the device, as well as whether the primary focus of attention will be on the device in the near future. Our approach employs information extracted from egocentric scene cameras, inertial measurement units, and the mobile phone (see Section 3.2.6 and Chapter 9). In the context of dyadic conversations, we propose the first method to anticipate averted gaze of participants. Our method uses gaze, speaking behaviour, facial expressions, and head pose to predict whether gaze will be averted during a time window in the future (see Section 3.2.7 and Chapter 10).

3.2.1 Unsupervised Eye Contact Detection in Multi-Person Interactions

Eye contact is one of the most important nonverbal signals in social interactions. Being able to detect eye contact between multiple people from remote cameras has the potential to open up exciting possibilities for research on group interactions, going beyond the often used head orientation as a proxy for gaze information (Beyan *et al.*, 2017c; Gatica-Perez *et al.*, 2005). However, research on eye contact detection in multi-party interactions has only investigated supervised settings that require labelled training data for the target interaction setting (Otsuka *et al.*, 2018; Zhang *et al.*, 2019a). This represents a fundamental limitation given that such training data may not be available. On the other hand, existing methods for unsupervised eye contact detection have only been developed for situations with a single potential target object (Zhang *et al.*, 2017b). While this can be sufficient for dyadic interactions, multi-party interactions inherently require support for multiple eye contact targets.

Contributions. To address these limitations, we propose the first method for unsupervised eye contact detection in multi-person interactions (Chapter 4). In order to train a person specific eye contact detector for each interactant, we make use of social interaction conventions. Specifically, we exploit the fact that people tend to look at the current speaker. Using a recent method for appearance based gaze estimation (Zhang *et al.*, 2017c) to a frontal view on the face of an interactant, we obtain a large number of raw gaze estimates of this particular person, defining a space of gaze estimates. By exploiting the correlation between speaking behaviour and gaze we locate the other interactants in the space of gaze estimates. After partitioning the gaze estimate space according to these inferred locations, we train a dedicated eye contact detector based on features extracted from the persons face using a convolutional neural network. The resulting detector is able to classify whether and with whom a person has eye contact. We evaluate our method against the state of the art in unsupervised eye contact detection, showing a relative improvement of more than 60%, and provide further in-depth experiments including ablation studies and performance depending on amount of training data. With this novel method for unsupervised eye contact detection it is now possible, for the first time, to analyse the patterns of visual behaviour in group interactions for which no eye contact annotations are available. As a result of this new capability, in Chapter 8 we are able to present the first cross-dataset evaluations for the emergent leadership detection task.

For our evaluations, we annotated the group interaction dataset presented in Chapter 7 of this thesis with eye contact information. The annotations were performed every 15 seconds for every interactant, indicating whether the interactant has eye contact with another person at the current frame, and if so, who is the target of eye contact. In total, eye contact annotations are provided for 3,995 frames from 50 participants, spanning more than 16.5 hours of video recordings. This dataset can be valuable for the development of eye contact detection algorithms and the analysis of gaze in group interactions. The dataset is available upon request from the authors.

3.2.2 Reducing Calibration Drift in Mobile Eye Trackers

Mobile eye trackers are a powerful means to enable gaze sensing in social interactions throughout users' daily lives. However, real-world behaviour of users still poses a significant challenge to robust mobile eye tracking, as involuntarily touching the eye tracking headset, scratching one's nose, or temporarily taking the eye tracker off can lead to severe calibration drift. This results in inaccurate gaze estimates that severely limit the applicability of mobile eye tracking in real-life situations. Repeated manual calibration of the eye tracker to maintain accuracy is not a viable option, as such a cumbersome process would heavily deteriorate user experience. While methods for automatically recalibrating mobile eye-trackers without the need of explicit user input have been proposed (Sugano and Bulling, 2015a), they rely on bottom-up saliency maps which do not reflect daily task-driven behaviour adequately.

Contributions. By analysing real-life recordings of mobile eye-tracking, this thesis reveals that people are likely to look at their phone once it appears in view (Chapter 5). Building on this observation, it proposes two novel automatic recalibration methods for mobile eye trackers. The first method builds saliency maps based on the locations of the mobile phone detected in the egocentric view of the eye tracker. These task-driven saliency maps are then used to adjust the mapping from pupil positions to gaze locations using the visual saliency based recalibration approach from Sugano and Bulling (2015a). In contrast, the second method is able to recalibrate the eye tracker without using images obtained from the egocentric camera. This method uses two assumptions: First, people are likely to look at their phone when touch events occur. Second, the phone is usually at a similar location when touch events occur. This thesis shows that these assumptions are reasonable and enable recalibration without the use of images obtained from the egocentric camera. This is especially beneficial in privacy-sensitive situations when scene camera input can not be provided (Steil *et al.*, 2019b). The proposed methods are evaluated on the novel mobile eye tracking dataset presented in Chapter 9. Both proposed methods significantly outperform the state-of-the-art saliency based recalibration approach by Sugano and Bulling (2015a). Furthermore, evaluations specific to different environments and phone usage conditions show that this pattern of results is robust in real-life conditions. These results are an important step towards the ability to sense gaze in peoples' daily lives without the need of repeated and cumbersome manual recalibration.

3.2.3 Emotion Recognition from Embedded Bodily Expressions during Dyadic Interactions

Emotion recognition from bodily expression has received considerable attention in previous research. These works, however, have focused on recognising emotional expressions in isolation, of non-interacting individuals, or using intrusive motion capture equipment (Metallinou *et al.*, 2013; Wang *et al.*, 2015b). Such constrained settings do not reflect desired application scenarios for emotion recognition as they often consist of

non-isolated expressions of interacting individuals embedded in a physical environment. Furthermore, emotion recognition methods for the real world need to work on the basis of imperfect data extracted from standard consumer sensors and can not rely on expensive and impractical motion capture equipment.

Contributions. In Chapter 6 this thesis proposes the first method for emotion recognition from embedded bodily expressions during dyadic interactions recorded with standard RGB cameras. In a first step, the method detects body parts using a CNN based approach (Sermanet *et al.*, 2014). In a second step, dense trajectory features (Wang *et al.*, 2013a) are extracted around those body parts. Dense trajectory features were originally developed for action recognition and consist of several feature descriptors including histograms of oriented gradients, histograms of optical flow and motion boundary histograms extracted along short tracklets. On top of this features, we train support vector machines to classify time windows into the emotion classes *happiness*, *anger*, *surprise*, *sadness*, and *neutral*. In evaluations on a newly recorded dataset of dyadic interactions, performances well above chance level were achieved.

This novel dataset is the first resource of emotionally charged dyadic interactions taking place in a real kitchen environment without the use of any augmentation like motion capture equipment. Interactions are recorded by eight frame-synchronized cameras and four ambient microphones. In total, eight pairs of actors were recorded, each improvising 28 short scenarios lasting 38 seconds on average. This resulted in 224 video clips with a total length of 143 minutes. Subsequently, the expressed emotions were annotated by observers using both a categorial as well as a dimensional emotion model. This dataset can serve as an important evaluation case for methods bringing emotion recognition from bodily expressions to the real world. The dataset is publicly available at `https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/human-activity-recognition/mpiiemo-dataset` (date: 24.02.2020).

3.2.4 Detecting Low Rapport in Group Interactions

The absence of rapport between interactants has been shown to result in decreased collaboration and worse interpersonal outcomes (Burns, 1984; Kelley *et al.*, 2014; Tsui and Schultz, 1985). Systems that are able to detect the failure to establish rapport are a prerequisite to automatically intervene in interactions to ensure positive outcomes. Previous work on estimating rapport from nonverbal behaviour has focussed on dyadic interaction settings. However, many interactions in daily life involve more than two participants. It remains unclear whether it is possible to detect individuals failing to establish rapport in group interactions, and which feature channels play an important role in this setting.

Contributions. This thesis proposes the first method to detect individuals failing to establish rapport in group interactions (Chapter 7). The method extracts visual and audio features and uses support vector machines (SVMs) to obtain classifications. In

extensive evaluations on a newly recorded group interaction dataset, the best performances are achieved by features based on facial action units reaching an average precision of 0.7. Classifications based on hand motion, speaking activity and prosody achieve lower, but still above-chance accuracy. Adding participants' personality scores measured by a NEO-FFI questionnare (Costa and MacCrae, 1992) to the facial feature set was able to boost performance when only the first third of an interaction was analysed. Further analyses detail the connections between the failure to establish rapport and single facial features as well as the pattern of correlations between rapport and other relevant measures in group interactions including leadership, dominance, liking and cohesion.

Our newly recorded dataset on which we evaluated low rapport detection methods is the first dataset of group interactions with rapport ratings (*MPIIGroupInteraction*). It consists of 78 German-speaking participants having discussions in groups of three to four people. In total, 22 interactions were recorded, each lasting for approximately 20 minutes. Video was recorded using eight frame-synchronised cameras positioned such that a frontal view of each participants' face was always available. To record audio, a microphone was placed in front of every participant. For each group, a discussion topic which was controversial among the participants was chosen from a list of possible topics. After the discussion, participants were presented with several questionnaires, measuring felt rapport with every other participant in the group as well as perceived leadership, dominance, liking, and cohesion. Furthermore a NEO-FFI personality questionnaire was administered (Costa and MacCrae, 1992). This dataset is the first group interaction dataset to combine self-reported rapport ratings with emergent leadership ratings. In contrast to other emergent leadership detection datasets (Sanchez-Cortes *et al.*, 2012; Beyan *et al.*, 2016b) in which participants are instructed to solve survival tasks, the open discussions in the *MPIIGroupInteraction* dataset are less structured, thus serving as a valuable additional perspective on the emergent leadership detection problem. Because of these characteristics, the dataset can become a valuable resource for group interaction research. In this thesis, it is used for cross-dataset evaluations of emergent leadership detection (Chapter 8), as well as for evaluation of a novel eye contact detection approach (Chapter 4). The dataset is available upon request from the authors.

3.2.5 Emergent Leadership Detection Across Datasets

Although they do not possess formal authority, emergent leaders have a significant influence on group performance (Druskat and Pescosolido, 2006; Kickul and Neuman, 2000). Being able to detect such individuals during meetings could be highly valuable for organisations. As a result, emergent leadership detection has been extensively studied, leading to highly effective multi-modal approaches (Sanchez-Cortes *et al.*, 2012; Beyan *et al.*, 2017c). All these approaches were trained and tested on the same dataset. In contrast to that, in most practical use cases in the real world, training and test distributions are not the same. For example, if an organisation decides to apply automatic emergent leadership detection in their group meetings it is impractical to first collect a large number of questionnaire responses from the meeting participants in

order to train the system. In contrast, it is desirable to apply a system that was trained beforehand in a scenario which is similar, but slightly different from the actual test case. Due to the lack of cross-dataset evaluations in emergent leadership detection, it remains unclear to which extent this is possible.

Contributions. In Chapter 8, this thesis for the first time presents a cross-dataset evaluation for emergent leadership detection. Using state-of-the-art features, emergent leadership detectors are trained on a source dataset (PAVIS (Beyan *et al.*, 2016b)) and evaluated on a target dataset (MPIIGroupInteraction (Müller *et al.*, 2018a)). To extract eye contact detections without the need of supervision on the target dataset, the unsupervised eye contact detection method presented in Chapter 4 is applied. Combining pose features with visual focus of attention (VFOA) features based on our eye contact detections, our approach is able to detect emergent leaders on the target dataset with an accuracy of 0.68, demonstrating for the first time the feasibility of cross-dataset emergent leadership detection as well as the practical value of our unsupervised eye contact detection method from Chapter 4. We further provide in-depth analyses highlighting the most important VFOA features in the cross-dataset setting, as well as an investigation of performances achievable with limited observation time of target interactions. As such, this thesis takes an important step towards emergent leadership detection in real-life conditions, encouraging future work to go beyond unrealistic evaluation scenarios on single datasets.

3.2.6 Anticipating Human Attentive Behaviour During Mobile Interactions

Automatic anticipation of attentive behaviour can be beneficial to enable machines to interact more seamlessly with humans but also to support humans in maintaining socially acceptable gaze behaviour in daily life. Due to its pervasiveness (Buschek *et al.*, 2018; Dingler and Pielot, 2015) and co-occurrence with activities like working, commuting, and diverse social interactions, attention allocation during phone usage can have important consequences. For example, inappropriate attention to the mobile phone during social interactions can be perceived as being impolite. On the other hand, frequent attention shifts between phone and physical environment in such situations might result in the user loosing track of tasks that need to be completed on the mobile phone. Current interfaces lack the ability to proactively intervene in such settings, as they can only observe shifts of attention after they occurred. Instead, this thesis envisions a novel type of interface that is able to anticipate such shifts in attention, opening up a new space for techniques supporting interacting humans.

Contributions. This thesis proposes the first method to anticipate human attentive behaviour in the context of mobile device interaction embedded in daily-life activities (Chapter 9). It proposes three novel prediction tasks: (1) anticipation of attention shifts from the environment to the mobile device, (2) anticipation of attention shifts from the mobile device to the environment, and (3) anticipation of the primary focus of attention (on or off the device) in the near future. Our method extracts features from a

head-mounted RGB and depth camera and head mounted inertial measurement unit (IMU) sensors in addition to IMU and mobile device usage features (e.g. touch events or screen on/off) extracted from the mobile phone. Using a random forest classifier, shifts to- and from the environment as well as primary attentional focus can be predicted above chance level. This proof-of-concept method is an important building block towards realising the vision of interfaces that are able to proactively accommodate and manage user attention. Such interfaces can help to integrate mobile device interaction with less friction into our daily social lives.

To evaluate the prediction performance of the proposed method, we collected the novel *MPIIMobileAttention* dataset. It consists of 20 participants being recorded during everyday activities freely roaming a university campus, including having lunch in the cafeteria or studying in the library. Each participant took part in three consecutive recording blocks lasting on average 77 minutes, resulting in a total of 77 hours. Data was recorded from the mobile phone (device interaction and IMU measurements), a mobile eye tracker including front-facing scene camera and an additional stereo camera with integrated inertial measurement unit. The dataset incorporated different conditions of phone use. During chat blocks, the study manager sent questions via an instant messenger application which the participants were instructed to answer. Outside of these chat blocks, the participants were free to use the phone as they liked. As a result of the real-life recording conditions, the variety of sensors and the different phone usage conditions, the dataset is a valuable resource for research on mobile eye tracking and device interaction. For example, in Chapter 5, this thesis makes use of the dataset to evaluate solutions to the problem of calibration drift in mobile eye trackers. The full *MPIIMobileAttention* dataset is made publicly available at `https://www.mpii.mpg.de/MPIIMobileAttention/` (date: 24.02.2020).

3.2.7 Anticipating Averted Gaze in Dyadic Interactions

Averted gaze plays an important role in social interactions, being connected to cognitive load (Glenberg *et al.*, 1998), intimacy-modulation (Abele, 1986) and floor management (Kendon, 1967). Interfaces that are able to anticipate averted gaze in the near future would be able to adapt to, and manage user attention proactively. While first works on gaze anticipation exist for egocentric videos of meal preparation and object search (Zhang *et al.*, 2017a, 2018a) as well as free-viewing of 360°videos (Xu *et al.*, 2018), no approach to averted gaze anticipation in social interactions has been proposed yet. Furthermore, no suitable dataset exists that combine natural interactions with sufficient dataset size and an adequate density of eye contact annotations.

Contributions. This thesis presents the first method to anticipate averted gaze in dyadic interactions (Chapter 10). The method extracts eye contact, speaker diarisation, head pose, raw gaze, and facial expressions on a past time window and predicts averted gaze on a future time window using a short long-term memory (LSTM) network. Evaluations on a newly collected dataset of dyadic video conferencing interactions reveal that this method is able to anticipate whether gaze will be mostly averted in the near

future with a performance of up to 0.85 average precision in a person-specific evaluation, and up to 0.75 average precision in a person-specific evaluation, clearly improving over baselines. Further analyses on the interplay between speaker turns and eye contact reveal differences between interviewer- and interviewee roles. With these initial results on averted gaze anticipation in social interactions, this thesis lays the groundwork for interfaces that are able to use knowledge of future gaze behaviour to support users proactively in their daily-life interactions.

To evaluate averted gaze anticipation approaches, this thesis presents the first dataset of natural dyadic social interactions with fine-grained eye contact annotations. In total, 121 videos of video conferencing interviews on topics including spirituality, health and beauty were collected from two channels on Youtube. All videos were recorded at a frame rate between 24 and 30 fps, with lengths ranging from 17 minutes to 58 minutes (average: 37 minutes). In total, the dataset contains 74 hours of conversation. Frame-based eye contact annotations are provided every 30 seconds for one of the two Youtube channels, and every 15 seconds for the other one. Furthermore, these annotations are combined with gaze estimates obtained using OpenFace (Baltrusaitis *et al.*, 2018) to extrapolate to non-annotated frames. This dataset can serve as an important resource for research on eye contact detection and anticipation as well as to understand the interplay between speech and gaze in dyadic interactions. The dataset is available upon request from the authors.

3.3 Limitations and Future Work

This thesis has made significant contributions towards sensing, interpreting, and anticipating human social behaviour in the real world. This chapter identifies the most important remaining challenges and discusses opportunities for future research to fully realise the vision of social machines.

3.3.1 Datasets

Only by recording data that is natural and truly captures the phenomena that are to be investigated the field is able to develop new methods and quantify progress. This thesis presented several novel datasets that are closer to real life than existing corpora and that can serve as testbeds for future methods. For example, we introduced the first dataset for emotion recognition from body movements in dyadic interactions of unaugmented people (Chapter 6), as well as the first dataset of group discussions with rapport annotations (Chapter 7). Furthermore, we presented the first mobile eye tracking dataset containing multi-modal recordings of mobile device usage in a wide variety of daily life social interaction settings (Chapter 9). Despite these advances, several directions in which datasets can be improved further remain.

Modern machine learning models like deep neural networks (Ciresan *et al.*, 2012; Krizhevsky *et al.*, 2012) profit immensely from large training data. In contrast, the currently available datasets for human social behaviour analysis are rather small. For example, while the Imagenet dataset (Deng *et al.*, 2009), which is commonly used to train image classification models consists of millions of labelled images, common datasets for emergent leadership detection only consist of a small number of group interactions (e.g. 40 meetings in Sanchez-Cortes *et al.* (2012)). While the two tasks are not directly comparable, the drastic difference in dataset sizes illustrates that machine learning methods with larger numbers of trainable parameters are difficult to apply to social behaviour understanding tasks. In order to harness the power of modern machine learning methods, it is therefore key to collect larger datasets than available today for many social behaviour processing tasks.

Apart from size, diversity of available data is another important aspect to develop methods covering the full range of human life. In this thesis, a novel mobile eye tracking dataset featuring a diverse set of environments and activities was presented (Chapter 9). Future research needs to go beyond that by including people, environments, and activities that are currently not well represented. For example, there are distinctive differences in eye contact behaviour during conversations for different ethnic groups (Rossano, 2013). Concerning diversity in environments and activities, many current datasets focus on rather generic tasks and environments, like meetings in office spaces (Sanchez-Cortes *et al.*, 2012; Beyan *et al.*, 2016b). To adequately reflect the variety of real life, future datasets will have to encompass interactions in which participants engage in activities like, among others, cooking, eating, doing sports, or having a party. In addition to the unique challenges in sensing that come with such activities, one key challenge for accurate modelling of social behaviour will be to successfully distinguish

social components of behaviour from behaviour determined by the current activity. Furthermore, with the emergence of augmented- and virtual reality technologies, it will be increasingly important to study in which ways approaches to analyse human nonverbal behaviour have to be adapted to fit such scenarios.

Finally, a major challenge for the creation of datasets on social behaviour are the difficulties present in annotation. While concepts like eye contact or facial expressions that have a direct connection to video content are relatively easy to annotate for observers, more elusive aspects of social interactions including emotions or rapport can pose significant challenges. Observers can not tell whether a person feels a certain emotion or feels in rapport with her interactants. They are only able to tell whether it looks like a person experiences a certain feeling. Research on rapport has shown that there can be significant differences between judgements made by observers and by participants themselves (Bernieri *et al.*, 1996; Cerekovic *et al.*, 2016). Obtaining judgments by participant on a fine-grained timescale is challenging, as such judgments can interfere with the interaction. As a result, studies usually ask participants for global ratings at the end of interactions (Sanchez-Cortes *et al.*, 2012; Beyan *et al.*, 2016b; Müller *et al.*, 2018a). In future research, it can be worthwhile to investigate to what extent techniques like experience sampling (Larson and Csikszentmihalyi, 2014) can be embedded in social interactions to obtain fine-grained annotations without influencing the interaction too heavily.

3.3.2 Additional Modalities

This thesis combines different feature modalities to address several challenging tasks in human social behaviour sensing, interpretation, and anticipation. For example, appearance-based gaze estimates are used in conjunction with speaking behaviour to enable unsupervised eye contact detection in group interactions (Chapter 4). Furthermore, pose and eye contact features are used in emergent leadership detection (Chapter 8), and a combination of gaze, eye contact, speaker diarisation, head pose, and facial expressions is employed to anticipate averted gaze in dyadic conversations (Chapter 10). Despite the existence of such multimodal approaches, future work can still explore many additional modalities for various social signal processing tasks. Relevant modalities can both include measurements taken during the interaction as well as before or after the interaction.

Emergent leadership and low rapport detection in group interactions have focussed on measuring visual and auditive behaviour during interactions (Beyan *et al.*, 2017c; Müller *et al.*, 2018a). Future research could benefit from exploiting information present in physiological sensors. Data from electroencephalograms, electrocardiograms, skin conductivity, and changes in respiration has been successfully used for emotion recognition (Kim and André, 2008; Liu *et al.*, 2011) and might prove equally valuable for emergent leadership and low rapport detection. Furthermore, features obtained from inertial measurement units embedded in smart watches could complement visual pose estimation, or even partly replace it in contexts where video cameras are perceived to be a privacy violation. To better put nonverbal behaviour of participants into context it

might be helpful to analyse it in conjunction with verbal behaviour and measures of task progress.

Apart from such additional modalities extracted while interactions take place, potentially valuable measurements can be taken before or after the interaction. In its work on low rapport detection, this thesis showed that incorporating the results of a personality test performed after interactions into the prediction algorithm could improve performance under certain circumstances (see Chapter 7). This finding encourages future work to investigate more thoroughly how information extracted outside of the interaction can be helpful for different social signal processing tasks. Apart from participants' personalities which can be estimated using e.g. mobile phone interaction behaviour (de Montjoye *et al.*, 2013), online social networks and email communication data could be used to extract information about the relations between participants and, as such, help with interpreting observed interaction behaviour appropriately. This is particularly relevant for rapport prediction given that the behavioural correlates of rapport have been shown to change depending on how developed a relationship is (Ogan *et al.*, 2012).

3.3.3 Learning by Interacting

Current research on nonverbal behaviour interpretation and anticipation usually adopts an approach where machine learning algorithms are trained on static datasets, without the possibility of algorithms to intervene in the interaction. While this is a user-friendly approach from the researcher's point of view, it also comes with drawbacks. Without the possibility to intervene, the ability to infer causal relations from observed interactions is severely limited. For example, while an algorithm can observe that the presence of facial action units associated with negative emotions is connected to low rapport (Chapter 7), without the ability to intervene in the interaction it is hard to tell whether such facial expressions are a result of, or a reason for the development of low rapport. Consequently, future research on problems like low rapport and emergent leadership detection, but also gaze anticipation, will need to put a larger focus on machines that take part in human social interactions to infer causal relationships and be able to shape interactions by intervention. A crucial prerequisite for such machines is the ability to adaptively generate nonverbal behaviour in social interactions. First promising approaches using reinforcement learning to adaptively shape a robot's behaviour in social interactions have been investigated (Hemminahaus and Kopp, 2017; Weber *et al.*, 2018; Ritschel *et al.*, 2019). For reinforcement learning approaches to be effective, it is important to provide rewards at a fine-grained timescale. If future research can find ways to measure the current state of rapport or emergent leadership in group interactions, e.g. via experience sampling as discussed in Chapter 3.3.1, it might be possible for machines to understand the relationships between basic nonverbal behaviour (e.g. eye contact and facial expressions) and higher-level social behaviour (e.g. rapport and leadership) in a causal way. This would open up possibilities to actively manage such higher-level social behaviour in interactions. Providing fine-grained rewards for the gaze anticipation task is comparably easy, as future gaze can be directly observed without requiring user intervention.

3.3.4 Privacy

Recording, processing, and interpreting nonverbal behaviour poses significant risks to user privacy. Nonverbal behaviour contains many different kinds of information users might not agree to share with other parties, including personality (Hoppe *et al.*, 2018; Beyan *et al.*, 2019b), emotions (Metallinou *et al.*, 2013), attraction (Kellerman *et al.*, 1989), or deception (Granhag and Strömwall, 2002). Furthermore, the aim of nonverbal behaviour analysis to support users in their daily lives implies that behaviour needs to be recorded in privacy-sensitive situations like conversations between couples, while studying with peers, or even during psychotherapy. Doing so without compromising user privacy represents a significant challenge. Users' concerns about privacy were shown to depend heavily on sensor types and recording situations (Klasnja *et al.*, 2009; Steil *et al.*, 2019a). It is the task of future research to find practical solutions to these concerns, providing a sweet spot on the trade-off between privacy of users and utility of recorded data. Promising approaches include the design of privacy-preserving features (Wyatt *et al.*, 2007a,b), the application of differential privacy to e.g. hinder user identification (Steil *et al.*, 2019a), or more generally privacy-preserving machine learning (Al-Rubaie and Chang, 2019), and the situation-dependent de-activation of recording (Steil *et al.*, 2019b).

Even if a social behaviour analysis system would be able to guarantee perfect user privacy this does not mean users will automatically trust the system. Digital devices usually do not offer intuitive ways to verify their behaviour and often appear as black boxes to users. Thus, it is important that mechanisms guaranteeing user privacy are communicated transparently and in a way that is comprehensible to a non tech-savvy audience. One example for transparent communication is the manual shutter for egocentric cameras employed by Steil *et al.* (2019b). It is a challenging task for future research to find ways to communicate the workings of privacy-preserving technologies like differential privacy to users in order to win their trust.

Furthermore, it is well-established that the presence of sensing devices can alter peoples' behaviour. A prominent example is the decrease of traffic to privacy-sensitive Wikipedia sites after the revelations on the NSA/PRISM surveillance programs (Penney, 2016). In eye tracking, a study by Risko and Kingstone (2011) has shown that humans change their eye behaviour (looking less at a poster of a woman in a bikini) when they believe their eyes are being tracked. Research on the effect of surveillance cameras shows that the presence of video recording can lead to decreased likelihood of cheating (Jansen *et al.*, 2018), as well as a decreased bystander effect (Van Bommel *et al.*, 2014). Future research will have to elucidate more thoroughly in which ways human behaviour might change as a response to ubiquitous social behaviour sensing, and whether the benefits of such systems outweigh potential negative effects on society.

3.3.5 Applications

Research has proposed numerous applications that can benefit from social behaviour processing technologies, including virtual patients and therapists (Kenny *et al.*, 2007;

Van Vuuren and Cherney, 2014), systems trying to equalise participation levels in group interactions (Schiavo *et al.*, 2014), or socially assistive robots in elderly care (Kachouie *et al.*, 2014). However, until now none of these applications has seen widespread adoption in society. The introduction of widely used applications carries the potential to boost the field of social behaviour processing in several ways. First, the outlook of applications that can be monetised would trigger increased research efforts by private companies. Recent years have seen massive investments of the private sector into artificial intelligence research (OECD, 2018), resulting in many state-of-the-art approaches in deep learning being developed by private companies (Amodei *et al.*, 2016; Szegedy *et al.*, 2017; Silver *et al.*, 2017). Companies can often allocate more resources for large-scale data collection and computing than publicly funded universities are capable of. If companies start to invest heavily in social behaviour processing technologies, more rapid advances could be made, leading to increasingly impressive and useful applications. Second, widespread adoption of social behaviour processing applications could drastically increase the amounts of data available to research - especially if applications are developed by organisations that are willing to share their data with the research community. If users notice the value of applications provided to them, they will be more likely to agree to share their data. For example, an application combining mood tracking (Caldeira *et al.*, 2017) with social behaviour analysis could help users to understand how their mood depends on behaviour observed during interactions - and vice versa. At the same time, the combination of nonverbal behaviour and mood measurements on a large scale would provide highly valuable input to researchers. Finally, widespread adoption of applications involving social behaviour analysis would provide valuable feedback to research. The opinions and wishes of people using applications in their private lives can differ from those that users would raise during scientific studies. This can be crucial to inform research on challenges that have been overlooked in the past.

In the end, novel applications are not only meaningful as a tool to boost research on social behaviour processing, but can improve our lives directly. It is about time for nonverbal behaviour analysis to realise its promise to help users.

3.4 Significance of the Thesis

Automatic social behaviour analysis has the potential to substantially change the way we interact with machines and with each other. This thesis identified key challenges that need to be addressed in order to realise this potential and contributes effective solutions for a variety of problems, thereby moving towards real-world social behaviour understanding.

3.4.1 Robust Gaze Sensing

While gaze behaviour is highly relevant in social interactions, robust gaze sensing is still challenging under real-world conditions. This thesis provided novel approaches to significantly improve gaze sensing in such conditions. We presented the first method for unsupervised eye contact from ambient cameras in group interactions (Chapter 4). Our method exploits the link between speaking activity and eye contact in order to train a dedicated eye contact detector for every group member. This method improved significantly over the state-of-the-art in unsupervised eye contact detection. The high quality of the provided eye contact estimates was demonstrated by successful application in emergent leadership detection (Chapter 8). Furthermore, we proposed novel approaches to reduce calibration drift in mobile eye trackers. Here, we make use of the connection between gaze and engagement with the mobile phone to recalibrate the eye tracker. While the first method uses phone detections obtained from the egocentric camera, the second method employs touch events registered on the mobile phone to identify points in time in which the user is likely to look at the phone. Both methods significantly outperform the previous state-of-the-art approach to automatic mobile eye tracker recalibration based on saliency maps.

In addition to the improvements achieved in accuracy, our methods proposed for robust gaze sensing also underline the fact that knowledge of high-level behaviour can help to improve basic nonverbal behaviour sensing. This is in contrast to the dominant approach that views social signal processing as a linear process starting with the detection of basic nonverbal signals and culminating in the recognition of high-level behaviours (Vinciarelli *et al.*, 2009).

3.4.2 Interpreting Behaviour in Social Interactions

While large progress has been made in the interpretation of nonverbal behaviour in social interactions, the studied settings do not yet cover the full breadth and naturalness of the real world. This thesis advanced the interpretation of nonverbal behaviour by studying emotion recognition and low rapport detection in less constrained settings than previous work. We were the first to design a method to detect emotions from body movements of unaugmented humans interacting in a real kitchen environment (Chapter 6). Compared to previous work which relied on dedicated motion capture equipment to measure peoples' movements, our approach comes significantly closer to real-life interactions where people can not be assumed to wear such equipment. In

addition to emotions, rapport is a valuable concept in understanding social interactions. Although it is potentially highly relevant in group interactions in school or at the workplace, previous work only investigated automatic rapport estimation in dyadic interactions. This thesis for the first time presents a method to detect low rapport in groups of multiple people, showing that features built on facial expressions play a key role in this challenging task (Chapter 7).

3.4.3 Group Analysis across Domains

To bring automatic analysis of group interactions to the real world, it has to be possible to apply analysis systems on data unseen at training time. This thesis made two essential contributions towards this goal. First, it enables eye contact detection from ambient cameras without the need of manual training data collection (Chapter 4). This is crucial for application settings like group analysis in organisations, where it would be impractical to provide manual labels for every different arrangement of seating positions in meeting rooms or even for every individual employee. Building upon this new approach to eye contact detection, we were able to present the first method for cross-dataset emergent leadership detection (Chapter 8). We trained emergent leadership detection algorithms on a source dataset of group interactions and evaluated these algorithms on a different target dataset featuring participants of a different nationality and a different group task. We could show that emergent leadership classification on the target dataset unseen at training time was possible with an accuracy of 0.68 using a combination of eye contact- and pose based features. These findings pave the way for user-friendly emergent leadership detection systems that can be used flexibly in organisations to improve human resource management and to support individuals.

3.4.4 Anticipating Gaze Behaviour

This thesis for the first time proposed methods to anticipate gaze behaviour in social interactions. We did so in two different, but equally relevant scenarios. First, we presented a method to anticipate gaze switches from and to the mobile phone as well as whether the primary focus of gaze will be on the mobile phone in the near future (Chapter 9). Our method makes use of features extracted from egocentric scene cameras, inertial measurement units, and the mobile phone. We evaluated our method on a novel mobile eye tracking dataset of daily life behaviour in different environments and involving various kinds of social interactions happening both on the mobile phone and with people in the same physical space. The ability to anticipate gaze allocation opens up novel ways to reduce friction between human-human and human-device interaction in peoples' daily lives. Second, we developed the first method able to anticipate averted gaze in dyadic conversations (Chapter 5). Our method uses a recurrent neural network that analyses gaze- and speaking behaviour as well as facial expressions and head movements in order to anticipate whether gaze will be averted in the near future. This approach showed promising results on a novel dataset of dyadic video conferencing interviews collected from Youtube. Our method enables novel paradigms in human-agent interaction in

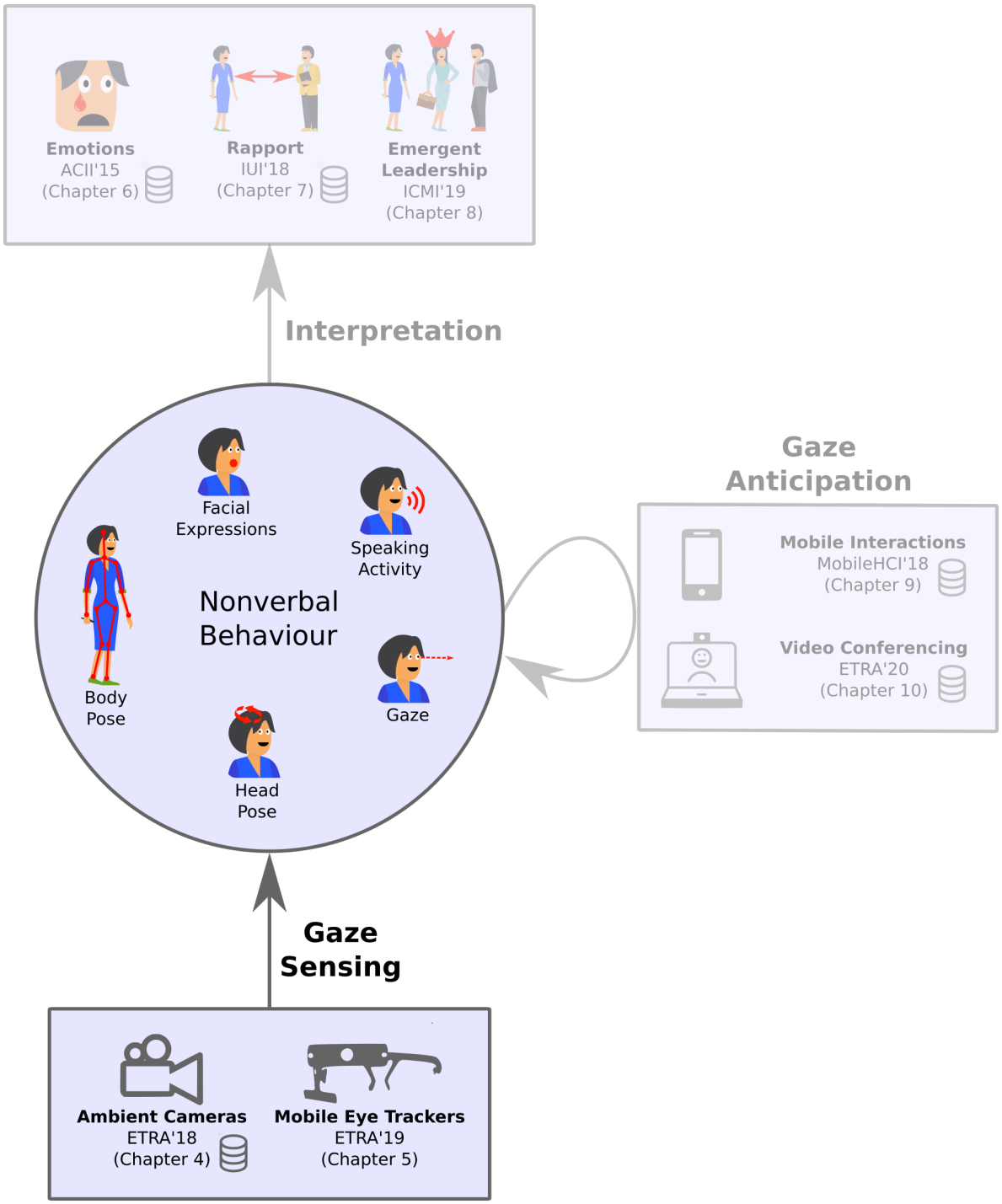
which the agent will be able to pro-actively adapt its behaviour to anticipated human gaze.

Both approaches for gaze anticipation in this thesis are designed for- and evaluated on real-world interactions outside the laboratory. Our initial work on gaze behaviour anticipation in real-world social interactions is meant to inspire further research on this novel task and lead to exciting new applications.

3.4.5 Datasets

In this thesis, we presented several novel datasets for diverse tasks in human social behaviour processing in order to evaluate our approaches but also to provide valuable resources to other researchers. We enable research on social behaviour interpretation in more realistic scenarios by introducing the first dataset of spontaneous emotional expressions in interactions of non-augmented dyads (Chapter 6), as well as the first dataset of multi-person interactions featuring rapport ratings (Chapter 7). The latter dataset is also valuable for studies on emergent leadership detection, as it was annotated with emergent leadership ratings and complements existing datasets by giving participants a different, less constrained task. We further provided eye contact annotations to this dataset (Chapter 4), allowing researchers to evaluate gaze sensing approaches in group interaction scenario. This thesis also provides publicly available dataset for the novel task of gaze behaviour anticipation in interactions. We presented the first mobile eye tracking dataset featuring multi-modal sensing including inertial measurement units and phone interaction logging in various daily-life scenarios (Chapter 9). Furthermore, we present the first dataset of dyadic video conferencing discussions collected from Youtube that are annotated for eye contact (Chapter 10). These datasets are an invitation to researchers to propose and evaluate novel methods for gaze anticipation in social interactions.

Part I


Emotions
ACII'15
(Chapter 6)
Rapport
IUI'18
(Chapter 7)
Emergent
Leadership
ICMI'19
(Chapter 8)
Interpretation
Facial
Expressions
Speaking
Activity
Nonverbal
Behaviour
Body
Pose
Gaze
Head
Pose
Gaze
Anticipation
Mobile Interactions
MobileHCI'18
(Chapter 9)
Video Conferencing
ETRA'20
(Chapter 10)
Gaze
Sensing
Ambient Cameras
ETRA'18
(Chapter 4)
Mobile Eye Trackers
ETRA'19
(Chapter 5)

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

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.

4.1 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.*, 2017), an object of interest (Smith *et al.*, 2005; Shell *et al.*, 2003; Smith *et al.*, 2013) or both (Zhang *et al.*, 2017b). 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, 2018c, 2017c), 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 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

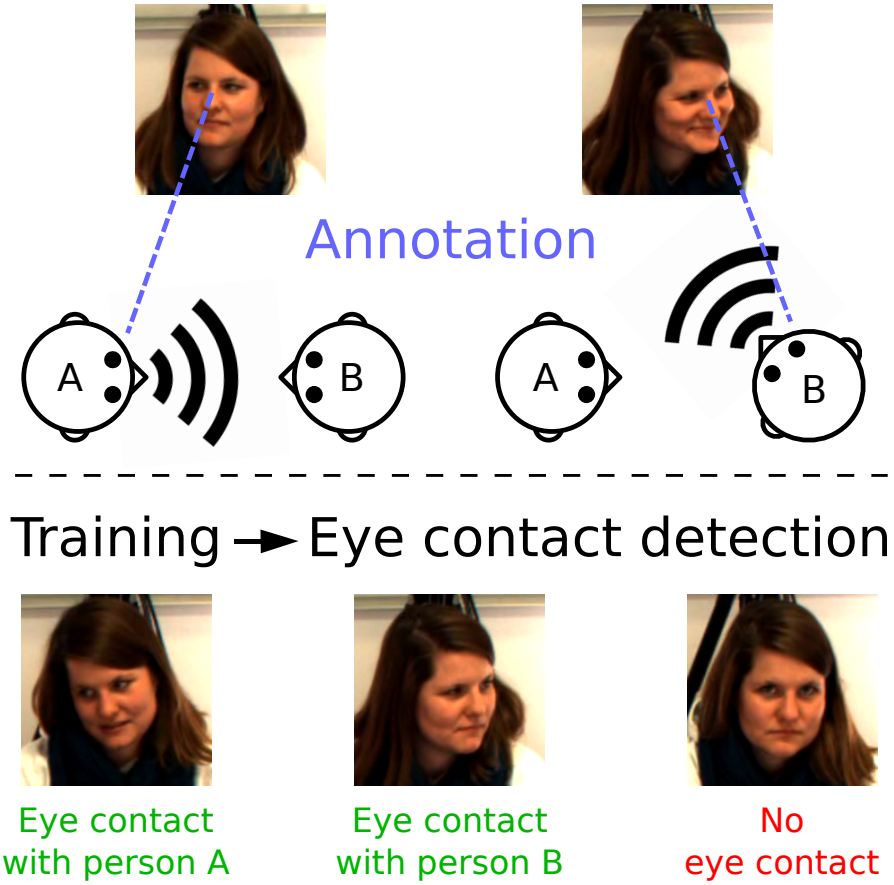


Figure 4.1: Our method exploits the correlation between gaze and speaking behaviour naturally occurring during multi-person interactions to weakly annotate images (top) that are, in turn, used to train a robust eye contact detector (bottom).

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.*, 2017b). 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.*, 2017c) 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.*, 2018a), showing that our method outperforms the state-of-the-art method (Zhang *et al.*, 2017b) 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.

4.2 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.

4.2.1 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 (Jokinen *et al.*, 2013; Ishii *et al.*, 2016). 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.

4.2.2 Gaze Estimation During Social Interactions

Gaze estimation has been of great interest for researchers in psychology (Kendon, 1967; Bavelas *et al.*, 2002) as well as affective computing (Picard, 1995; Huang *et al.*, 2016b; Andrist *et al.*, 2014a). Previous studies followed two different ways to address the challenges of gaze estimation. Most of them relied on stationary (Vertegaal *et al.*, 2001; Jokinen *et al.*, 2013) 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.*, 2014a), 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.*, 2018a). 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).

4.2.3 Eye Contact Detection

Unlike the general gaze estimation task that attempts to estimate the precise gaze direction in a continuous space (Zhang *et al.*, 2018c), 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 (Smith *et al.*, 2005; Chong *et al.*, 2017; 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).

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.*, 2017b) built on top of a learning-based gaze estimation method (Zhang *et al.*, 2017c). 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.

4.2.4 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.

4.3 Dataset

All experimental evaluations were performed on a subset of a recent dataset of three- and four-person interactions (Müller *et al.*, 2018a). 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.*, 2017b), 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.

4.3.1 Recording Setup

The dataset (Müller *et al.*, 2018a) 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 7.2, 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.

4.3.2 Gaze Annotations

Given that the dataset by (Müller *et al.*, 2018a) 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

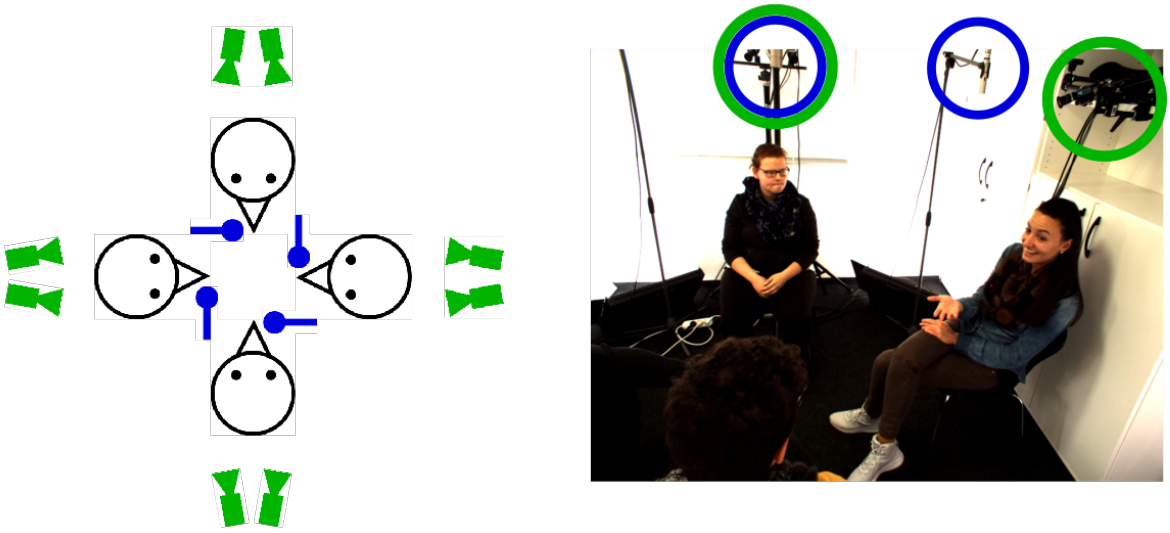


Figure 4.2: Camera setup used for the dataset recording in (Müller *et al.*, 2018a). Please note that the cameras were placed slightly above the participants to avoid occlusions.

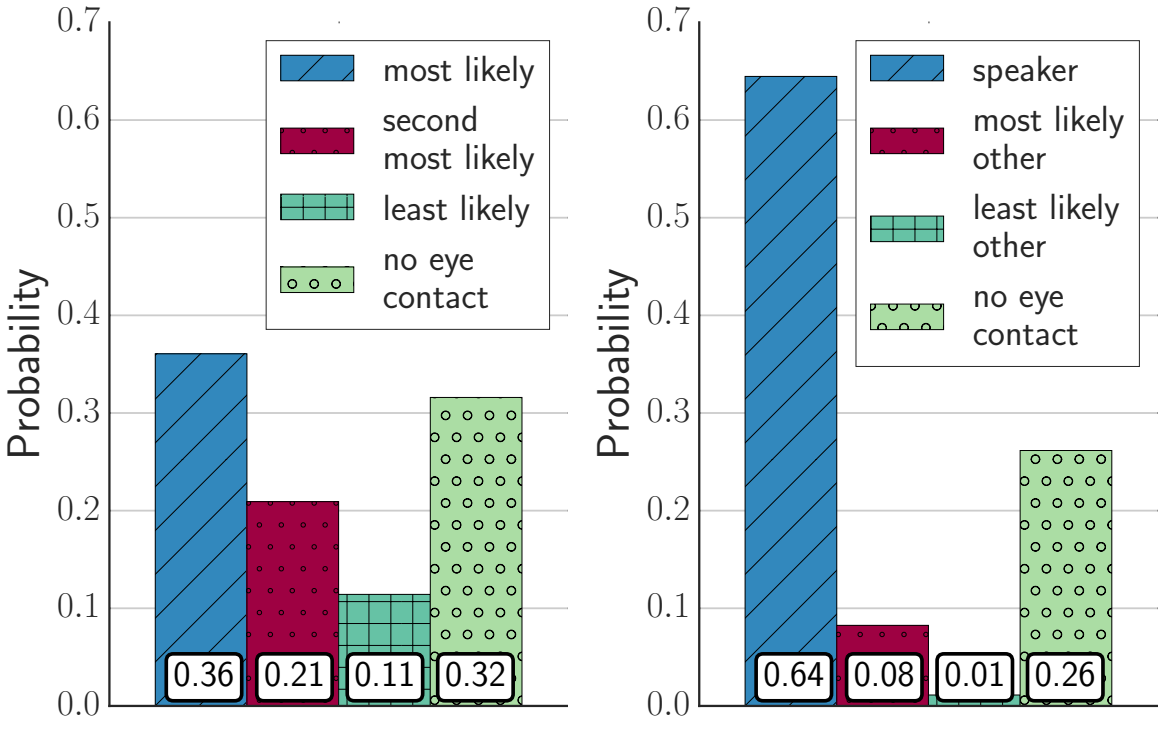


Figure 4.3: *Left*: Probability of looking to the most often, second most often, and least often looked-at person, along with looking at no face. *Right*: Probability of eye contact with the person who is currently speaking in comparison to the second and third most often looked-at person, along with looking at no face.

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.*, 2017b) 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.* (2017b), 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 4.3, 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 4.3, 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 4.3, right). This pattern lays the foundation for our method.

4.4 Method

Our method improves over the weak labelling and subsequent training of the eye contact detection method proposed in (Zhang *et al.*, 2017b). 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*.

4.4.1 Eye Contact Detection Framework

Here we briefly introduce the unsupervised eye contact pipeline in (Zhang *et al.*, 2017b). 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.*, 2018c). Subsequently, a user-independent CNN model (Zhang *et al.*, 2017c) 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

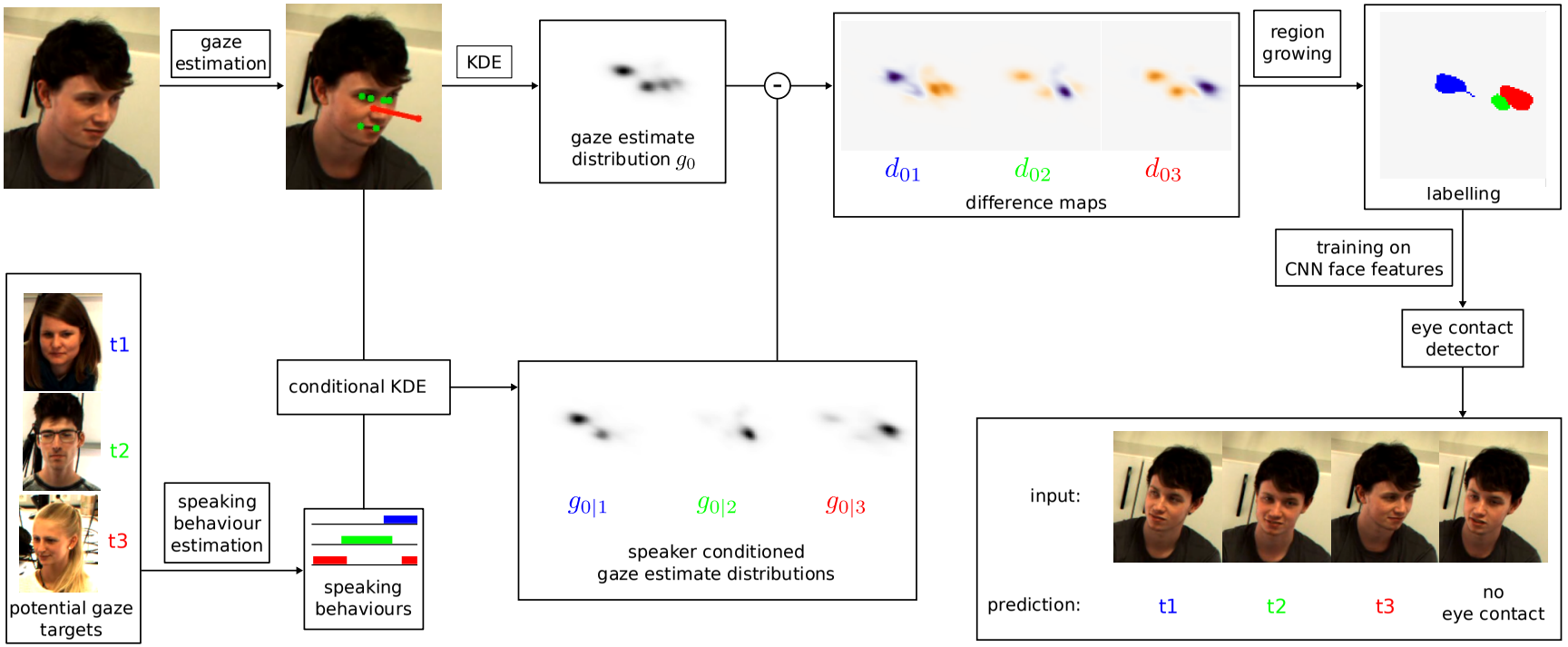


Figure 4.4: Our method takes images from multiple ambient cameras pointing at the gazer and potential gaze target persons as input. The images are the basis for coarse gaze estimates obtained using a full-face gaze estimation method (Zhang *et al.*, 2017c). Kernel density estimation (KDE) yields the 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.

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.

4.4.2 Weak Labelling Using Speaking Behaviour

In contrast to the binary classification problem considered in (Zhang *et al.*, 2017b), 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 10.4 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.*, 2017c). 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.*, 2017b). In the following, we discuss each step in detail.

4.4.2.1 *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.

4.4.2.2 *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 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.

4.4.2.3 *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:

$$\hat{p}(e_{ij}) = p(speak_j)p(ec) \tag{4.1}$$

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.*, 2017b), 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.

4.4.3 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".

4.4.4 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.*, 2017c). To make our approach more easily comparable to that of (Zhang *et al.*, 2017b), 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.

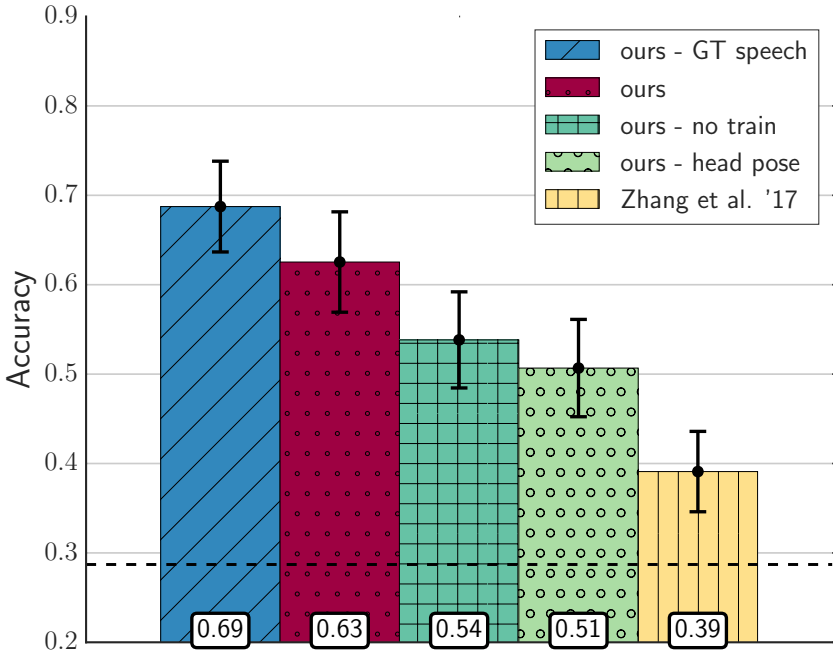


Figure 4.5: Accuracy of the different eye contact detection methods. Error bars indicate 95% confidence intervals. Random chance level is indicated with the black dashed line.

4.5 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.

4.5.1 Eye Contact Detection Performance

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

1. Unsupervised eye contact detection (Zhang *et al.*, 2017b): 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.

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 4.5 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.*, 2017b) (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 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.

4.5.2 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 4.6 shows the accuracy of our method (in red) and our upper bound (in blue). As

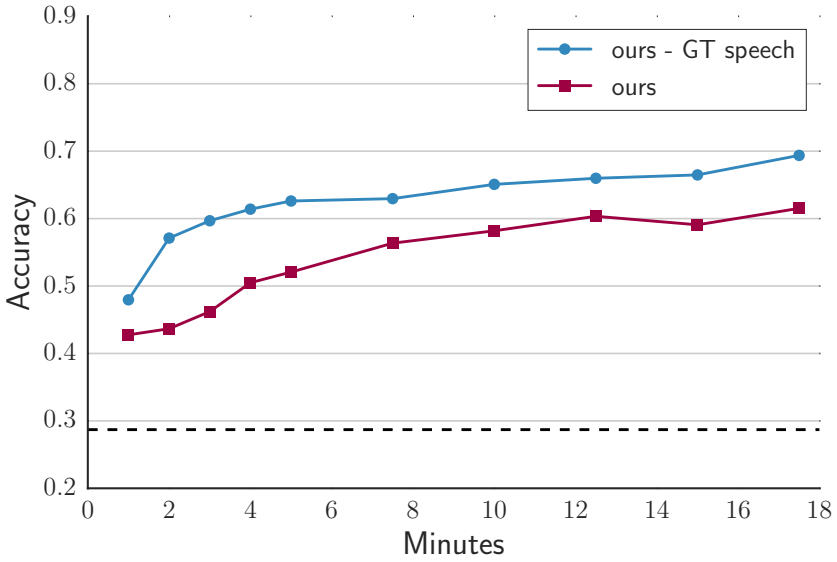


Figure 4.6: Accuracy of selected methods when only using the first x minutes of an interaction for training. The black dashed line indicates performance of a random predictor.

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.

4.5.3 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 4.7 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

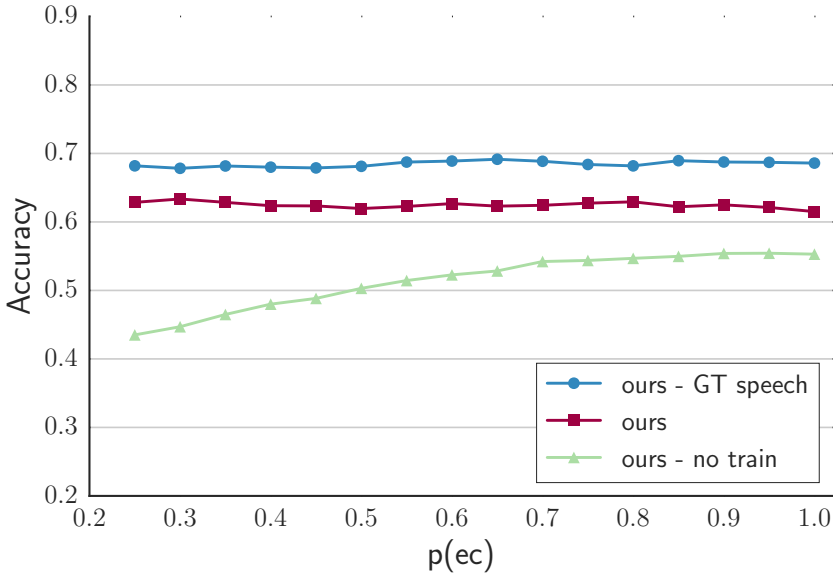


Figure 4.7: Accuracy of selected methods depending on the eye contact prior $p(ec)$.

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 4.8 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.

4.5.4 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.*, 2017b) (see Figure 4.9). 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.*, 2017c). 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.

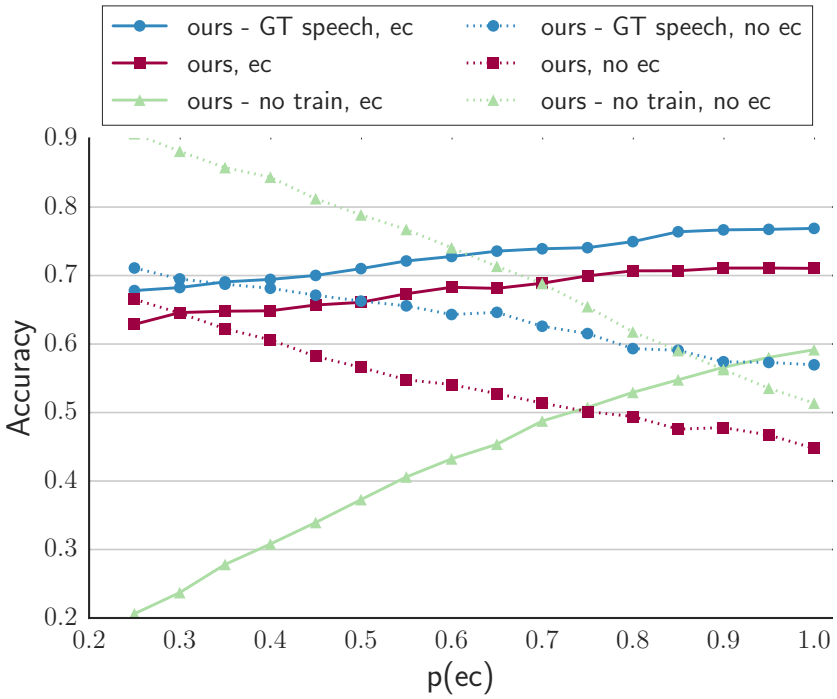


Figure 4.8: Accuracy of selected methods depending on the eye contact prior $p(ec)$ for ground truth "eye contact" (solid lines) and "no eye contact" samples (dotted lines).

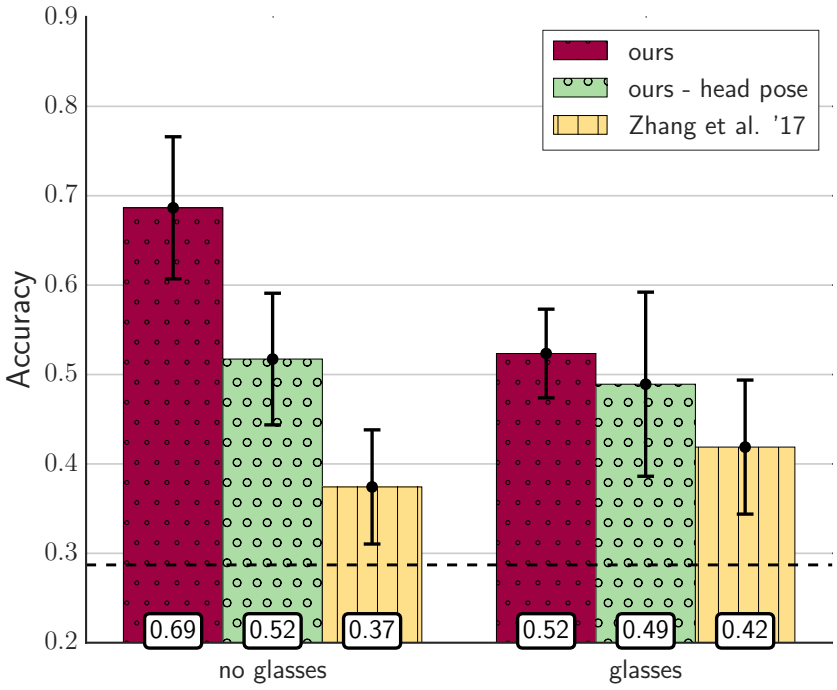


Figure 4.9: Accuracy of our method as well as the head pose proxy for participants with and without glasses. Error bars indicate 95% confidence intervals. Performance of a random predictor is indicated by the black dashed line.

4.6 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.*, 2017b): 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 4.5). 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.*, 2018a).

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.*, 2018a) and directly execute interventions, e.g. via different kinds of displays (Balaam *et al.*, 2011; Schiavo *et al.*, 2014; Damian *et al.*, 2015). As shown in our work, our method is capable of online prediction after observing the interaction for a short amount of time (see Figure 4.6). 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 4.5). 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.

4.7 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.

5 Reducing Calibration Drift in Mobile Eye Trackers by Exploiting Mobile Phone Usage

AUTOMATIC saliency-based recalibration is promising for addressing calibration drift in mobile eye trackers but existing bottom-up saliency methods neglect user's goal-directed visual attention in natural behaviour. By inspecting real-life recordings of egocentric eye tracker cameras, we reveal that users are likely to look at their phones once these appear in view. We propose two novel automatic recalibration methods that exploit mobile phone usage: The first builds saliency maps using the phone location in the egocentric view to identify likely gaze locations. The second uses the occurrence of touch events to recalibrate the eye tracker, thereby enabling privacy-preserving recalibration. Through in-depth evaluations on a recent mobile eye tracking dataset (N=17, 65 hours) we show that our approaches outperform a state-of-the-art saliency approach for automatic recalibration. As such, our approach improves mobile eye tracking and gaze-based interaction, particularly for long-term use.

5.1 Introduction

The ubiquity of smartphones and the increasing availability of mobile eye trackers enables novel interaction concepts and applications (Duchowski, 2002; Bulling and Gellersen, 2010; Pfeuffer *et al.*, 2015; Pfeuffer and Gellersen, 2016), and has lead to a growing body of research on gaze-based interaction with mobile devices (Khamis *et al.*, 2018). Unfortunately, more widespread adoption of mobile gaze-based interaction in everyday life remains challenging due to *calibration drift*. This describes the accumulating deterioration in eye tracking accuracy after an initial manual calibration was performed. For example, the eye tracking headset may slightly slip and shift on the user's head due to body movement throughout the day, rendering gaze interactions inaccurate.

Common calibration procedures are tedious and impractical to perform several times a day in everyday life. To address this problem, recent work has proposed an *automatic* recalibration approach that uses saliency maps computed from a mobile eye tracker's scene video (Sugano and Bulling, 2015b). A saliency map is a 2D probability map of where in the visual scene the user is likely to fixate on (Borji and Itti, 2013). The recalibration approach exploits the (assumed) correlation between saliency maps and gaze to continuously recalibrate the eye tracker in the background.

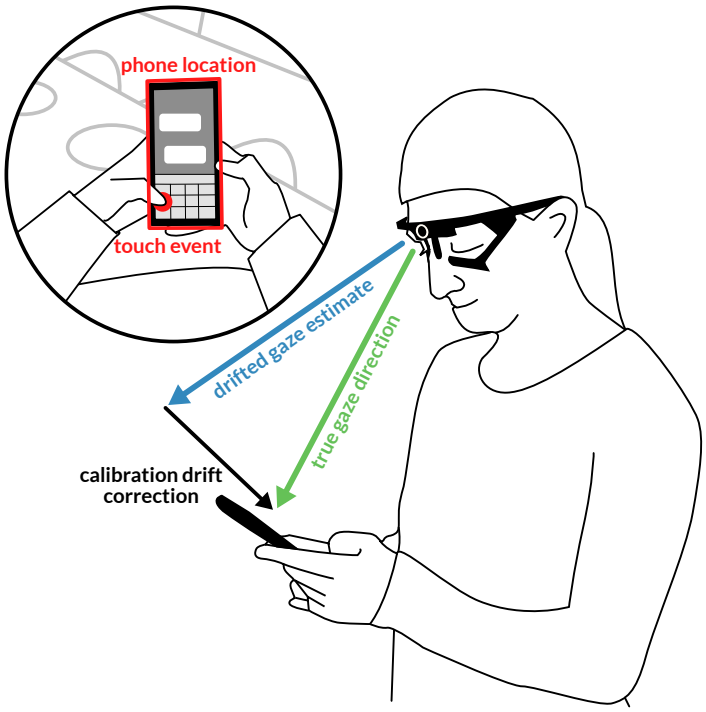


Figure 5.1: Mobile eye trackers suffer from calibration drift and inaccurate gaze estimates (blue arrow), for example caused by headset slippage. Our two novel automatic recalibration methods correct for calibration drift (black arrow) by either using the phone's location or users' touch events (red) to infer their true gaze direction (green arrow).

The performance of this approach inherently depends on the quality of the computed per-frame saliency maps. The authors in (Sugano and Bulling, 2015b) studied a free-viewing and, hence, artificial scenario in which users were walking around a building without any concrete task in mind. Natural daily-life settings, however, are dominated by task-driven attentive behaviour (e.g. grabbing something, pressing a button in an elevator, using computers or phones). In such situations, the user's task is more likely to determine attention than the visual saliency of the object (Hayhoe and Ballard, 2005). It therefore remains unclear how the original saliency-based approach performs in real-world contexts and whether it can be improved to better exploit task-driven behaviour.

This paper aims to address both questions and proposes novel improvements for a pervasive everyday mobile interaction use case. By inspecting real-life recordings of a recent mobile eye tracking dataset (Steil *et al.*, 2018b), we observed that users are likely *to attend to their phones* once these appear in the view of the egocentric camera. However, using phone presence for automatic recalibration is challenging because interaction "on the go" leads to frequent attention switches between the phone and the environment (Oulasvirta *et al.*, 2005; Steil *et al.*, 2018b). Thus, it is unlikely that users will *always* look at the phone, even if it is present in their field of view (Steil

et al., 2018b). We address this challenge by making use of phone detections in a robust state-of-the-art saliency-based recalibration approach (Sugano and Bulling, 2015b).

Moreover, we propose a second approach ("blind recalibration"), where we use the occurrence of *touch events* on the user's phone as an indicator for gazing at the assumed phone location. This approach does not require a scene camera, and may thus prove useful for privacy-sensitive applications or contexts in which recording egocentric video is not desirable (cf. (Steil *et al.*, 2018a)).

In summary, our contribution is two-fold: First, we present *two novel approaches for automatic mobile eye tracker recalibration* that use a) smartphone screen locations and b) occurrence of touch events to counter calibration drift in everyday use of mobile eye trackers. Second, we report *in-depth evaluations of these approaches* on a recent dataset collected in-the-wild (N=17, 65 hours), which show that our approaches consistently outperform the previously proposed state-of-the-art saliency-based approach.

5.2 Related Work

Our work is related to previous work on 1) phone use in everyday life and 2) automatic eye tracker calibration.

5.2.1 Phone Use in Everyday Life

We use phone interactions to recalibrate mobile eye trackers, since they come with several beneficial properties: For example, related work has shown that many phone users are *highly responsive*, attending to mobile messages 12 hours a day (84 hours a week) (Dingler and Pielot, 2015). Typing in general happens throughout the whole day, both on weekdays and weekends (Buschek *et al.*, 2018). Thus, messaging and typing already cover a large timeframe in which recalibration via phone use is possible.

People also develop *usage habits*, such as frequently checking for new messages and content updates (Oulasvirta *et al.*, 2012). Many interactions also result in repeated notifications later on (e.g. chat, email, social networks, music), bringing users back to their phones. For instance, Shirazi et al. (Sahami Shirazi *et al.*, 2014) found that 50% of interactions with incoming notifications happen within 30 seconds. This "checking behaviour" supports our approach, since self-calibration benefits from phone use spread out across time and many situations, to get up-to-date and diverse samples. Moreover, many people even interact with their phone if they have no specific task in mind, that is, if they seek stimulation in situations of *boredom* (Pielot *et al.*, 2015). Nevertheless, mobile phone use leads to frequent *switches of attention* between phone and environment (Oulasvirta *et al.*, 2005; Steil *et al.*, 2018b). Thus, exploiting phone use for recalibration needs to deal with uncertain user attention, even if the phone is in sight of the scene camera.

In summary, related work on mobile phone use in everyday life reveals unique challenges and opportunities for self-calibrating mobile eye trackers via phone use and thus motivates our research questions in this paper.

5.2.2 Automatic Eye Tracker Calibration

Despite continuing advances in eye tracking technology, e.g. by improved pupil detection algorithms (Dierkes *et al.*, 2018; Swirski and Dodgson, 2013), wider adoption of the technology is still prevented by the need for repeated manual calibration of eye trackers. Therefore, automatically calibrating (i.e. without initial calibration) and recalibrating (i.e. with initial calibration) eye trackers has been of interest to the HCI community. Initial work on automatic calibration focused on stationary settings. While (Yamazoe *et al.*, 2008) used an eyeball model, other works used mouse clicks, and more diverse associations between interaction patterns and users' visual attention (Sugano *et al.*, 2008; Huang *et al.*, 2016a; Zhang *et al.*, 2018b). Subsequently, more general self-calibration approaches exploited bottom-up saliency maps or gaze patterns obtained from other users (Alnajar *et al.*, 2013; Chen and Ji, 2015; Sugano *et al.*, 2010). A different way to self-calibrate mobile eye trackers uses corneal images (Lander *et al.*, 2017; Takemura *et al.*, 2014a). Such approaches require specialised hardware, as they rely on RGB eye cameras to extract the corneal image. Consequently they also struggle more with suboptimal lighting conditions (Takemura *et al.*, 2014a). Moreover, they cannot be used for privacy-preserving recalibration, as scene properties can be decoded from RGB eye images (Backes *et al.*, 2008; Lander *et al.*, 2017; Takemura *et al.*, 2014b). In contrast, no such approach is known for active illumination infrared eye cameras.

The closest work to ours is from Sugano et al. who were first to analyse the severe calibration drift in mobile eye trackers and proposed saliency-based recalibration to retain the quality of an initial manual calibration over a longer period of time (Sugano and Bulling, 2015b). The employed saliency maps consisted of bottom-up components along with face- and person detectors. However, they evaluation focussed on a free-viewing setting and mobile device usage was neither incorporated in the approach nor occured during the study.

5.3 Dataset

To investigate automatic recalibration in natural environments, we used a recent 20-participant mobile eye tracking dataset originally recorded to study visual attention forecasting in natural situations (Steil *et al.*, 2018b). We chose this dataset because it contains relatively long recordings of interactive behaviour with mobile phones during everyday situations, including studying in a library, working in an office, eating in a canteen, or drinking a coffee in a café. The dataset was subsequently ground-truth annotated for users' current environment (Steil *et al.*, 2018a), which we used to evaluate our methods in different daily-life situations.

5.3.1 Apparatus

For recording, participants were equipped with a state-of-the-art PUPIL mobile eye tracker (Kassner *et al.*, 2014) featuring an infrared eye (640×480 pixels) and a fisheye

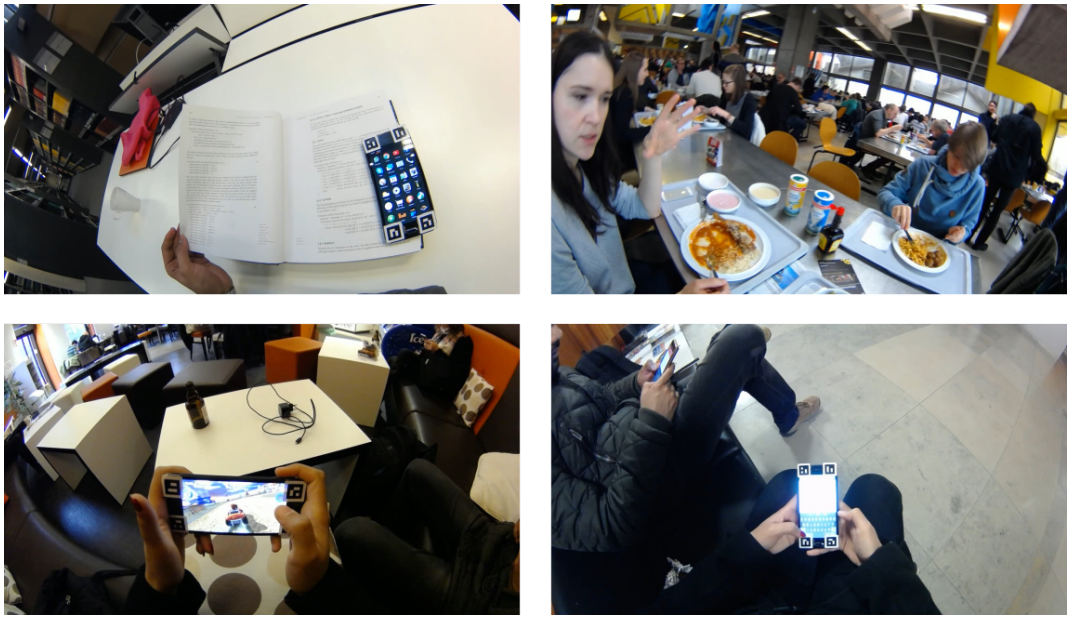


Figure 5.2: Example images from the scene camera in different situations from the dataset of Steil et al. (Steil *et al.*, 2018b).

scene camera (1280 × 720 pixels). Participants interacted with a mobile phone that was augmented with visual markers on its corners to obtain groundtruth phone location in the egocentric scene camera view. Logging software was used to monitor users' phone interactions, including all touch events. A messaging application was used as a means of communication between experimenter and participant.

5.3.2 Procedure

Each participant took part in three consecutive recording blocks, each lasting on average for 77 minutes. Before each recording block, a calibration sequence was recorded in which the participants were instructed to gaze at visual markers that were manually presented by the experimenter at at least nine different locations in order to cover the field of view of the scene camera. For 17 participants, additional calibration sequences were recorded at the end of every recording block for.We restrict our analysis to those 17 participants because the additional calibration sequences allow us to quantify the error of automatic recalibration approaches. The top of Figure 5.3 gives an overview over the structure of the dataset. During the recordings, participants were free to roam the university campus under the conditions that they did not stay at a single place for more than 30 minutes and to visit the canteen, the library and the coffee shop at least once during the recording. Apart from this, participants were not given any instructions or otherwise constrained in their behaviour. In particular, they were also allowed to put off the eye tracker during short breaks between recording blocks. Figure 5.2 shows sample images obtained using the egocentric camera.

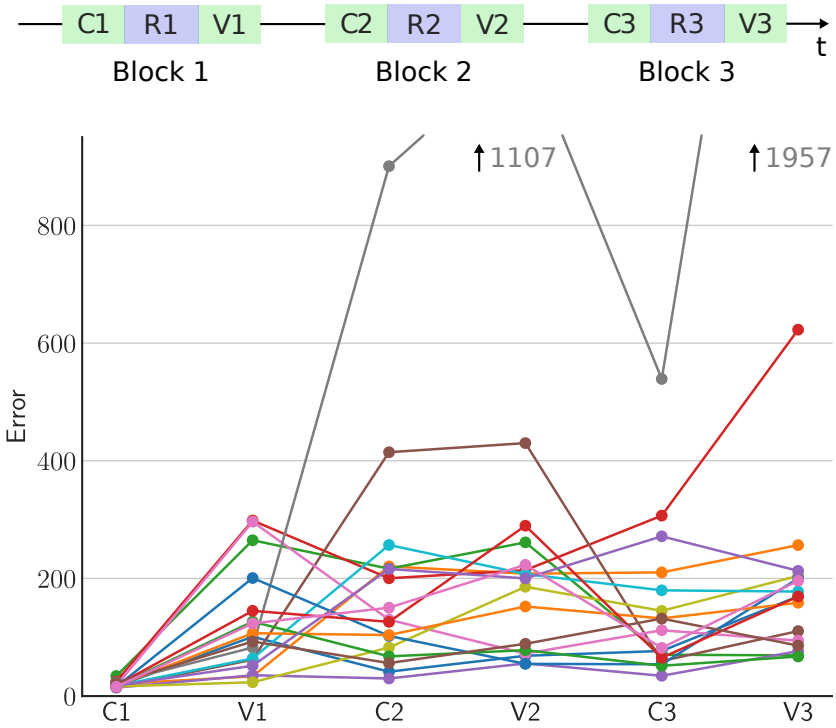


Figure 5.3: **Top:** Illustration of the dataset structure consisting of three recording blocks each comprised of calibration sequence (CX), recording (RX) and calibration sequence used for validation (VX). **Bottom:** Gaze estimation error in pixels measured on different calibration sequences when using the first calibration sequence (C1) to calibrate the eye-tracker. Lines are added to connect corresponding measurements.

5.3.3 Analysis

We used the calibration sequences recorded after each recording block to evaluate the calibration drift compared to the initial calibration recorded at the beginning. Gaze estimation is performed using a seven dimensional polynomial pupil feature based on pupil detections provided by the PUPIL software (Kassner *et al.*, 2014). We then use ridge regression to learn the mapping of pupil features to marker locations. This is in line with the approach taken in (Sugano and Bulling, 2015b), except that we perform 2D gaze estimation, as the calibration sequences on the dataset only provide 2D information. To not weight errors differently at different eccentricities of the field of view as a result of using a fisheye camera, we undistort gaze estimates and calibration markers before error measurements.

Figure 5.3 shows the gaze estimation error of the calibration obtained from the initial calibration session measured on all available calibration and validation sessions in the dataset. Each participant is represented by a line connecting the corresponding measurements. Calibration drift is present for a participant if the error at later points in time is larger than the error of the initial calibration. In line with (Sugano and Bulling, 2015b), some participants only showed a minor calibration drift while others showed

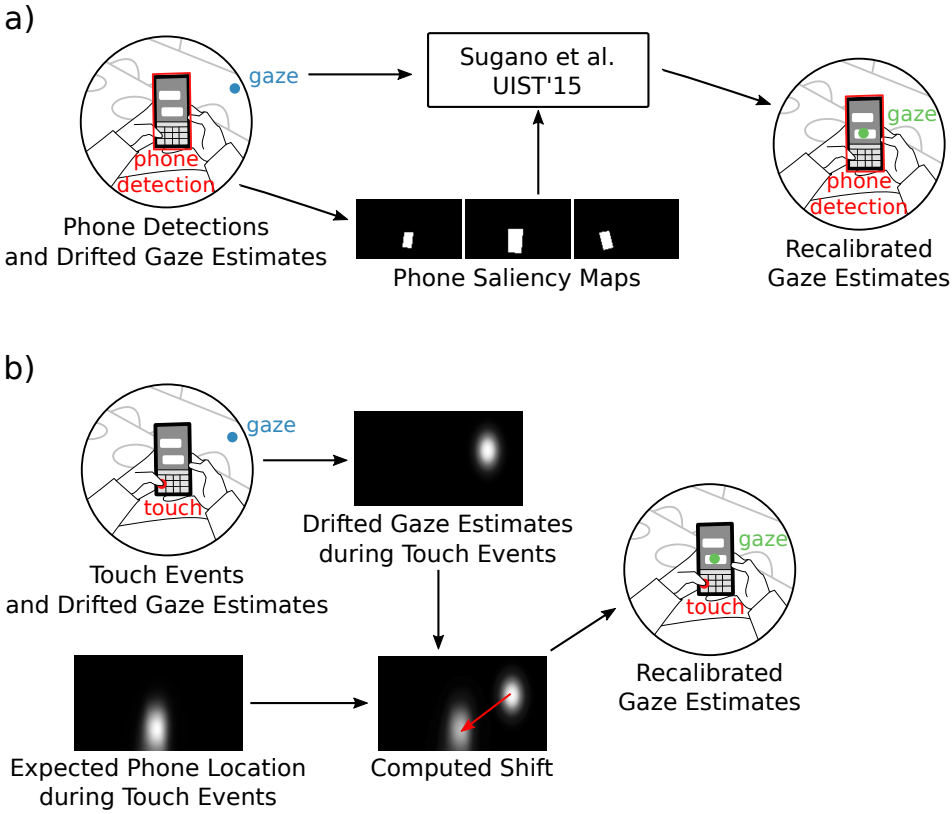


Figure 5.4: Overview over the two proposed methods. a) From phone detections we build corresponding phone saliency maps that serve as input to the method of (Sugano and Bulling, 2015b) together with an initial eye tracker calibration. b) We combine a model of where we expect the phone to be located during touch events with an aggregation of the drifted gaze estimates during touch events. From this we compute the shift we have to apply to the drifted gaze estimates in order to obtain correct gaze estimates.

a large increase in gaze estimation error over the recording blocks. One participant exhibited a particularly large increase in error, reaching a gaze estimation error of 1957 pixels during the last validation sequence. Closer inspection of this participant's eye video showed that the eye camera is severely shifted in the calibration sequence after the second and third recording blocks. We opted to keep this outlier in our analysis because such severe shifts in the relation between eye and camera are precisely the challenge in mobile eye tracking that we are trying to solve. Removing the participant from the dataset does not change the general pattern of results. The large increase in error over time that exists for many participants illustrates the need for automatic recalibration methods.

Further analysis revealed that the probability of gazing at the phone is 0.59 when the phone is detected. The presence of touch events increased this probability to 0.7. This strong relationship between phone interaction and gaze is the basis of our proposed automatic recalibration methods.

5.4 Method

We propose two different methods for automatic recalibration by exploiting phone usage: Our first approach uses *phone detections* from the scene camera. These phone detections are transferred into saliency maps, and the method of (Sugano and Bulling, 2015b) for automatic recalibration using visual saliency is applied. Our second approach recalibrates *"blindly"* without using the scene image. Here, we compensate for calibration drift by computing the shift between the (potentially drifted) gaze estimates when *touch events happen* and the expected location of the mobile phone. An overview over both methods is given in Figure 10.4. We next describe our two approaches in more detail. In all cases, initial gaze estimates are obtained as described in the previous section.

5.4.1 Approach 1: Phone Saliency Maps

To use phone detections in the scene camera, we follow the approach to automatic recalibration starting from an initial calibration as proposed by (Sugano and Bulling, 2015b) (see also (Sugano *et al.*, 2010)). We give an overview of this method and then describe our adaptation to integrate phone detections.

5.4.1.1 *Visual saliency based recalibration*

The approach of (Sugano and Bulling, 2015b) relies on the association of saliency maps extracted from the scene video with pupil positions and polynomial pupil features extracted from the eye camera. It consists of two steps, namely aggregation and robust mapping. In the aggregation step, the polynomial pupil features are clustered using the mini-batch k-means algorithm (Sculley, 2010). The clustering on pupil features also defines a clustering of the corresponding saliency maps, from which a mean saliency map is computed for every cluster. The goal of the robust mapping step is to find a mapping from the clusters of pupil features to locations in the scene video by making use of the mean saliency maps. To this end, 2D gaze predictions are obtained from the polynomial pupil features by applying the initial calibration. Subsequently, RANSAC (Fischler and Bolles, 1981) is employed to find a shift from this 2D space of initial predictions to the output space consisting of the positions of maximum values in the mean saliency maps. Applying this shift to the initial predictions removes the calibration drift. For further details we kindly refer the reader to (Sugano and Bulling, 2015b).

5.4.1.2 *Phone saliency maps*

Our approach based on phone detections relies on a saliency map in which we set the area of the detected phone in the scene video to the maximum value and everything else to zero. The area of the phone is defined as the convex polygon that has the detections of the phone corner markers as its vertices (see red polygon in Figure 10.1). In frames without phone detections, the corresponding saliency map is all zero. We call these saliency maps *phone saliency maps*. Figure 5.5 (left) shows an average phone saliency

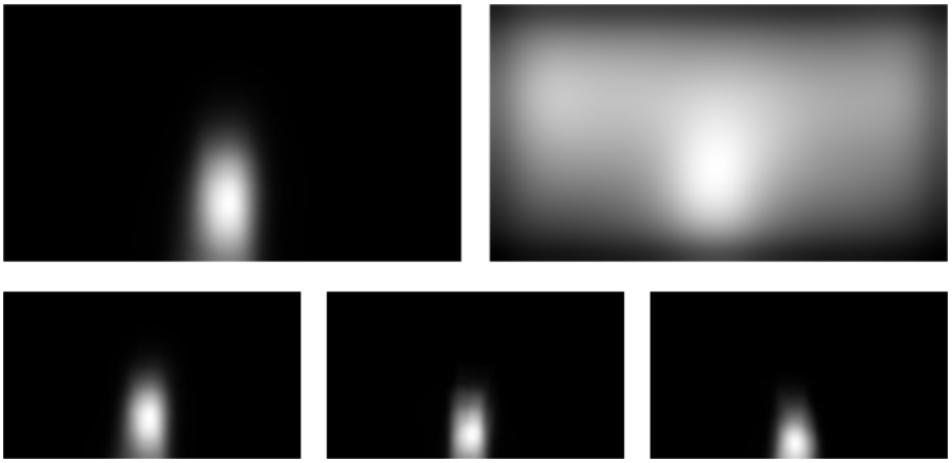


Figure 5.5: Different saliency maps averaged over all participants. Top left: Phone saliency maps for moments when touch events happen. Top right: Saliency map constructed according to (Sugano and Bulling, 2015b). Bottom, left to right: Phone saliency maps for sitting, standing and walking.

map for illustration. Phone saliency maps along with pupil detections and the initial calibration are then used as input to the approach of (Sugano and Bulling, 2015b).

For phone detection, the dataset (Steil *et al.*, 2018b) uses four visual markers attached to the phone corners, and the marker detection implemented in PUPIL (Kassner *et al.*, 2014). This simulates a robust phone detection method. However, methods for detecting screens without markers exist as well (see e.g. (Korayem *et al.*, 2016)) and could be integrated for practical deployments without markers.

5.4.2 Approach 2: Blind Recalibration

Our second proposed approach uses information about touch events taking place on the mobile phone in order to automatically recalibrate the eye tracker. This approach is motivated by privacy concerns about scene recordings using body-worn cameras (Steil *et al.*, 2018a; Koelle *et al.*, 2018), since it does not need such a scene camera for recalibration. It is based on two assumptions.

5.4.2.1 *Assumption 1: Touch events indicate attention*

We assume that when touch events take place the user is likely to look at the phone. This assumption is confirmed by our analysis on the dataset by Steil et al. (Steil *et al.*, 2018b) showing that the probability of gazing at the detected phone is 0.7 when a touch event takes place.

5.4.2.2 *Assumption 2: Common phone location in scene view*

We assume that phones are most of the time positioned at a similar area in the scene view while interacting with them via touching. The top row of Figure 5.5 supports this

assumption by showing the localised average phone saliency map when touch events take place in comparison to the average saliency map following the approach of (Sugano and Bulling, 2015b). While the maximum at the bottom middle in both cases, for the phone saliency map it is closer to the bottom. Furthermore, the bottom row of Figure 5.5 shows that the average phone saliency map during touch events only slightly changes when people are sitting, standing or walking.

5.4.2.3 *Recalibration and evaluation*

Exploiting the two assumptions, we correct calibration drift in the following way:

1. Estimating usual phone location:. We estimate the usual location of the mobile phone during touch events by averaging phone saliency maps for moments in time that are within a one second window centred on a touch event. Taking the argmax of this saliency map, we obtain the most likely location of the phone in the scene view when touch events take place (cf. Figure 5.5 top left). For our evaluation, we compute this in a leave-one-out cross-validation fashion: When testing on the data of a participant, we estimate that participant's mean saliency map on all other participants.

2. Estimating expected phone location:. For a given test recording (i.e. in practice: during use), we retrieve all the (possibly drifted) initial gaze estimates that are within a one second window centred at a touch event. By taking their median, we obtain the expected location of the phone in the space of initial gaze estimates.

3. Estimating shift for recalibration:. We can now estimate the shift (between 1. and 2.) by subtracting the expected phone location in the space of initial gaze estimates (2.) from the usual location of the phone in the scene view during touch events (1.). We recalibrate the initial gaze estimates by applying this shift.

5.5 Evaluation

We evaluated both methods for 1) short and long-term calibration, 2) performance in different environments, and 3) influence of forced phone use (i.e. chat blocks in the dataset). We give an overview of our evaluation as follows:

In all evaluations, we measure gaze estimation error on the validation sequences that were recorded at the end of each recording block. For recalibration, we always use the data recorded right before the validation sequence that is used for measuring the error.

When evaluating the influence of forced phone use or environments, we restrict the data that is used for recalibration to certain phone usage conditions or environments, respectively.

The long- and short-term calibration settings differ with respect to which initial calibration sequence is used: In the long-term case we only use the first calibration sequence for every participant to extract an initial calibration, allowing us to investigate the effect on gaze estimation accuracy over an extended period of time. In the short-term

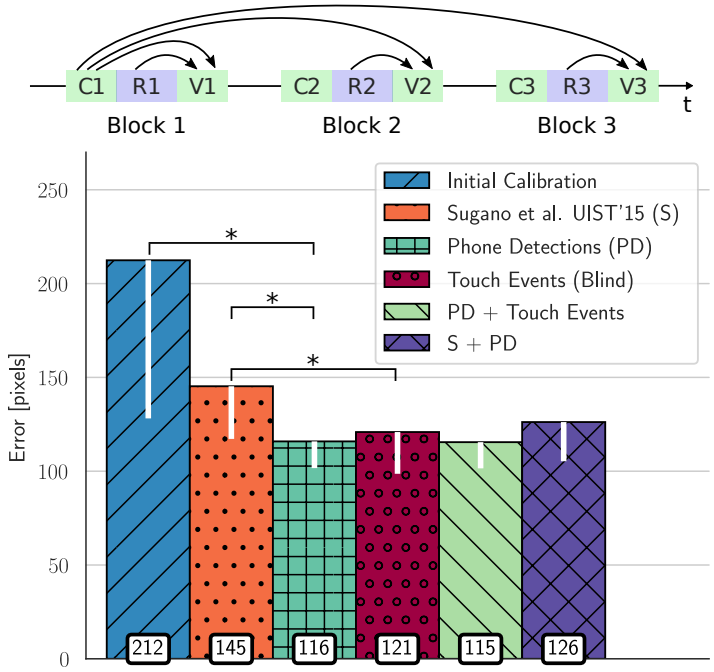


Figure 5.6: Top: Visualisation of the long-term recalibration setting. Arrows from Calibration segment C1 to validation segments VX indicate usage of the manual calibration from C1 and evaluation on VX. Arrows from the recording segments RX to validation segments VX indicate extraction of saliency maps and touch events from RX when evaluating on VX. Bottom: Our methods outperform baselines in this setting. Stars indicate statistically significant differences, white lines the lower parts of 95% confidence intervals (upper parts are symmetric).

case we always use the calibration sequence at the beginning of the recording block on which we evaluate our methods. This lets us investigate whether our methods are already useful after wearing the eye tracker for a shorter amount of time.

All evaluations use all 17 participants, with the exception of the evaluation for different environments where we make use of additional annotations which are present only for a subset of participants. The next sections report on these evaluations in detail.

5.5.1 Long-term Recalibration

We first evaluated the recalibration over an extended period of time, i.e. over the whole recording. To compare our proposed methods to the state of the art, we measured their performance after every recording block using the corresponding validation sequence. Our methods as well as the comparison methods used the calibration obtained from the calibration sequence before the first recording as a starting point. Saliency maps and touch events were always extracted from the recording block on which the error was measured. To robustly compare the different methods, we averaged the error over all

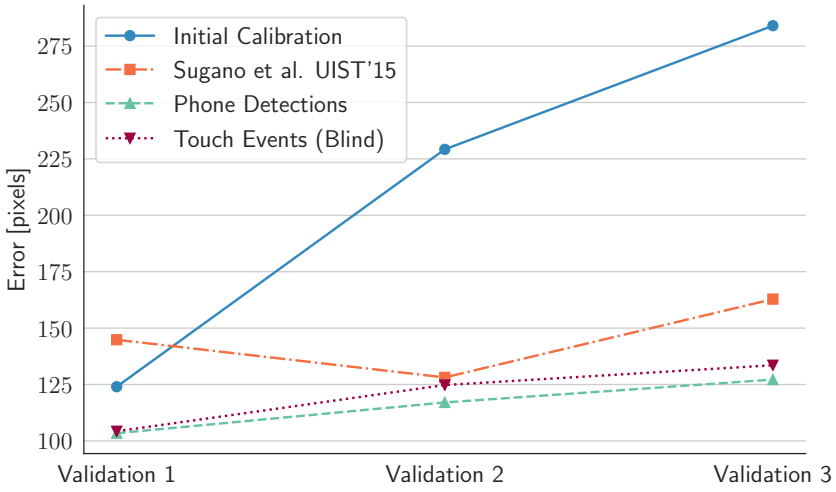


Figure 5.7: Our methods showing better performance than the baselines on every validation sequence after each of the three recording blocks, when using the manual calibration at the beginning of the first recording block as a starting point. Lines are added to connect individual measurements.

recording blocks for each subject. See Figure 5.6 for a visualisation of the evaluation scheme and the resulting performances.

As can be seen in the Figure, our method using *phone detections* achieves an error of 117 pixels, which is significantly lower than both the initial calibration at 212 pixels (t=-2.13, p=0.049, df=16, two-tailed) and the state-of-the-art method by (Sugano and Bulling, 2015b) at 145 pixels (t=-2.33, p=0.033, df=16, two-tailed). Our *blind recalibration* method achieves an error of 121 pixels, reaching statistical significance compared to the method by Sugano et al. (t=-2.45, p=0.026, df=16, two-tailed), but not quite compared to the initial calibration (t=-1.86, p=0.082, two-tailed).

We also evaluated a saliency map incorporating touch events, which was generated by doubling the magnitude of activations on the detected phone at moments in time lying within a one second window around each touch event. This approach reaches an error of 115 pixels, which is only slightly better than the plain phone saliency map. As incorporating touch events makes additional assumptions with respect to the recording setup (a phone needs to be equipped with recording software and synchronised with the eye tracker), we do not consider this approach further. Finally, we added the phone detection based saliency map to the saliency map constructed according to Sugano et al., reaching an error of 126 pixels. This combination thus did not reduce error further than our phone detection approach alone.

To quantify how stable our methods are under growing distance in time to the initial calibration, we also analysed their performances for each recording block separately (see Figure 5.7). While the error of the initial calibration increased strongly as time progressed, the error of all automatic recalibration methods was relatively stable. Our approach based on phone detections consistently achieved the lowest error, followed by our method on touch events, and by the state of the art by (Sugano and Bulling, 2015b).

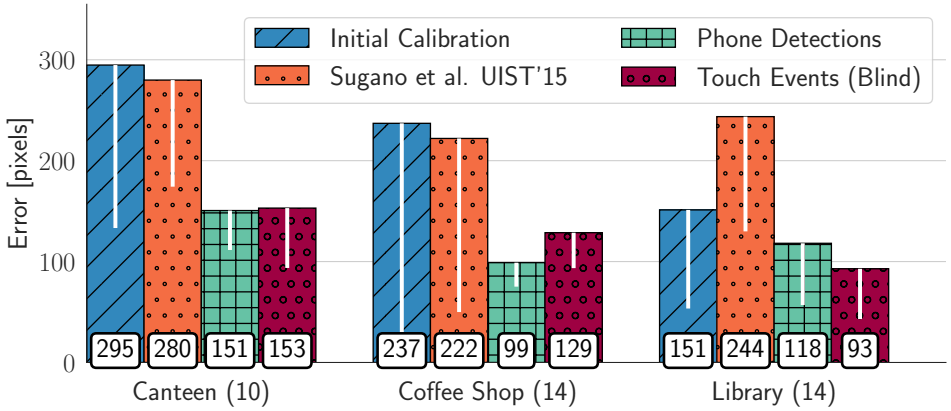


Figure 5.8: Analysis of robustness showing consistent results of our methods in different environments. The number of subjects from which the results for a specific environment are obtained is given in brackets. White lines show the lower parts of 95% confidence intervals (upper parts symmetric).

5.5.2 Performance in Different Environments

To investigate the robustness of our method with respect to different environments, we evaluated it using only data from a specific environment for recalibration. To this end, we made use of the additional annotations that were collected for 14 of the participants (Steil *et al.*, 2018a). We evaluated our recalibration methods for three environments that each participant was asked to visit at least once during the study, namely the canteen, a coffee shop and the library. These environments are interesting for evaluation, as they correspond to different tasks participants perform alongside phone interactions. Furthermore, they differ significantly with respect to the amount of other people that are present. For the canteen, on average we count 694 frames with face detections per minute, whereas it is 289 for the coffee shop and only 94 for the library. For each participant we selected one recording block in which the participant visited a specific environment for evaluation. If a participant visited the same environment in more than one recording block, we chose the recording block containing the longest visit. Additionally, to ensure that the environment "canteen" was behaviourally distinct from the other environments, we excluded four participants in this conditions who did not have a meal during their visit to the canteen.

The results of this evaluation are shown in Figure 5.8. Errors achieved in different environments cannot be compared directly, as different recording blocks are chosen for different environments. The general pattern, however, shows that our two proposed methods perform consistently better than both the initial calibration and the state-of-the-art method (Sugano and Bulling, 2015b).

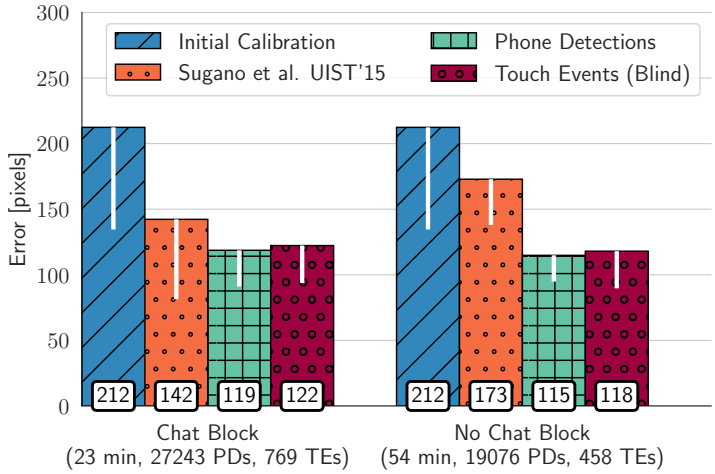


Figure 5.9: Similar patterns of results when exclusively using data from chat blocks versus non-chat block time periods. Numbers in brackets indicate average length, average number of phone detections and average number of touch events in chat blocks / outside chat blocks during a recording block. White lines show the lower parts of 95% confidence intervals (upper parts are symmetric).

5.5.3 Influence of Chat Blocks on Performance

In each recording block of the dataset, several chat blocks took place (Steil *et al.*, 2018b), in which the experimenter chatted with the participant. This implies that the participant is forced to use the phone, thereby generating phone detections and touch events. We thus investigated the influence of chat blocks on the performance of our methods: We analysed if the pattern of results stayed the same regardless of whether we restrict our saliency map generation and touch event usage to 1) the chat blocks contained in a recording, or 2) the other parts of the recording (i.e. no chat blocks).

The split of the data resulted in the following numbers of detections: On average, chat blocks took up 23 out of 77 minutes of a recording block. During the chat block portion of a recording block, there were on average 27,243 frames with phone detections and 769 touch events, while the non-chat block portion contained on average 19,076 phone detections and 458 touch events.

Figure 5.9 shows that the pattern of results is indeed the same in both conditions: Both our methods achieved a lower error than the initial calibration and the state of the art by (Sugano and Bulling, 2015b). It is important to note that direct performance comparisons between the "chat block" and "no chat block" conditions must not be drawn, since the amount of data in each of the conditions is different. The most likely explanation of the slightly worse performance of our proposed methods for chat blocks compared to the "no chat block" condition is this: Although the number of phone detections and touch events is higher during chat blocks, time spent outside of chat blocks is much higher, potentially leading to more diverse samples of phone detections and touch events.

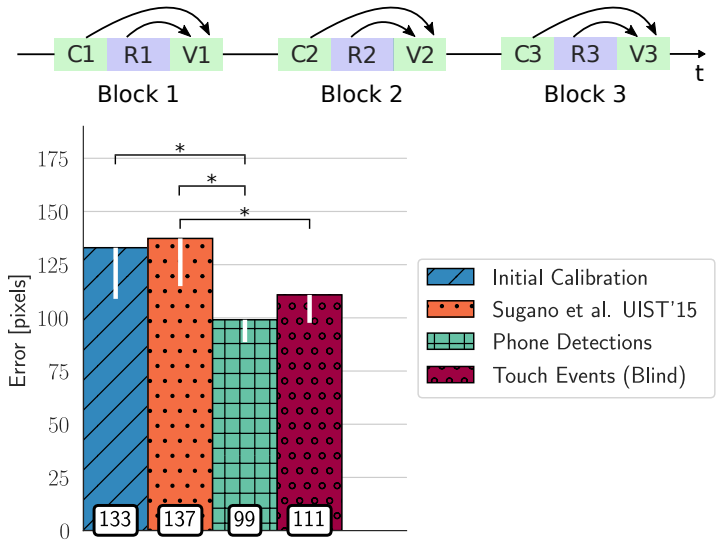


Figure 5.10: Top: Illustration of the short-term recalibration scenario (see Figure 5.6 for an explanation). Bottom: Our methods outperform baselines in this scenario. Stars indicate statistically significant differences, white lines lower parts of 95% confidence intervals (upper parts symmetric).

5.5.4 Short-term Recalibration

Finally, we evaluated whether our method was useful in rather short eye-tracking recordings by treating each recording block in the same way: We extracted an initial calibration from the calibration sequence right before the recording block started. Saliency maps and touch events for the evaluated recalibration approaches were extracted from the recording block, and the performance of methods was evaluated on the calibration sequence at the end of the recording block. Errors were averaged over all recording blocks for a given participant, yielding a more robust estimate of the performance compared to the analysis presented in Figure 5.7.

The results are shown in Figure 5.10. As can be seen from the figure, our method based on *phone detections* achieved the lowest error with 99 pixels, significantly outperforming the initial calibration at 133 pixels (t=-2.78, p=0.013, df=16, two-tailed) and the state of the art at 137 pixels error (t=-2.66, p=0.017, df=16, two-tailed). Our method based on *touch events* performed slightly worse with an error of 111 pixels. It still reached statistical significance compared to the state of the art (t=-2.22, p=0.041, df=16, two-tailed), yet not compared to the initial calibration (t=-1.35, p=0.195, df=16, two-tailed). Interestingly, the state-of-the-art saliency based method was not able to improve above the initial calibration in this evaluation.

5.6 Discussion

5.6.1 Recalibration Performance

Our approaches to automatic recalibration outperform the state of the art significantly and consistently across different evaluation scenarios. They improve eye tracking accuracy both in short- and in long-term recordings and in different situations like eating in a canteen, sitting in a library or visiting a coffee shop. Our approach based on saliency maps built from phone detections performs slightly better than our blind calibration approach based on touch events, which still significantly outperforms the state of the art.

5.6.2 Initial Manual Calibration

Our approach requires initial manual calibration (see evaluations). We also tested phone saliency maps without initial calibration but observed very inaccurate results. This is explained by the relatively narrow area near the bottom centre of the scene view in which phones occur most of the time (cf. Figure 5.5), thereby not providing diverse enough samples to estimate the full calibration parameters. Nevertheless, we have shown that these samples are still suitable to estimate and correct calibration drift. Moreover, we specifically exploited this "peak phone area" in our blind recalibration approach.

5.6.3 Dataset and Study Setting

We used the mobile eye tracking dataset provided by Steil et al. (Steil *et al.*, 2018b), which contains a rich diversity of everyday situations. One particular aspect of the study setting and dataset are the "chat blocks" in which the experimenter triggered text messaging with the participant. It is worth reflecting on whether this yields an unrealistically high degree of phone use. Considering the findings on phone use, mobile messaging, and typing in the literature (e.g. (Buschek *et al.*, 2018; Dingler and Pielot, 2015; Sahami Shirazi *et al.*, 2014)), we argue that the covered extent of chatting is not unrealistic. Moreover, we evaluated our approaches also on the parts of the dataset that involved no such study-triggered phone use and found comparable results (see Figure 5.9).

5.6.4 Applications

By significantly decreasing calibration drift, our recalibration approaches facilitate everyday use of interaction techniques that require precise gaze estimation: Examples include multi-modal mobile interaction that combines touch and gaze input, for example to redirect direct touch to a cursor at the gaze position on a table (Pfeuffer and Gellersen, 2016). Another proposed concept combines pen and gaze input in a similar way (Pfeuffer

et al., 2015). With our novel recalibration methods we take an important step towards enabling such interaction techniques in daily life.

5.6.5 Privacy

A privacy concern of mobile eye tracking is the scene camera, which might record sensitive information, in particular if it is also used to record lifelogging videos (Steil *et al.*, 2018a), and does not indicate its recording status to bystanders (Koelle *et al.*, 2018). Korayem et al. (Korayem *et al.*, 2016) used a CNN-based computer vision approach to detect displays (phone, PC, etc.) in egocentric lifelogging videos, which users perceive as sensitive content (Hoyle *et al.*, 2014, 2015). Such scenes or image regions could then be blurred or redacted. This could easily be integrated with phone-based recalibration: The combined system would detect the phone display, recalibrate the eyetracker, and redact the display area in the lifelogging video. Moreover, Steil et al. (Steil *et al.*, 2018a) used Deep Learning and both scene video and eye movement data to inform when to start/stop recording to avoid capturing sensitive content. This leads to interruptions in the scene video. Our touch-based approach could recalibrate the eye tracker during such interruptions.

In summary, if the eye tracker's scene camera recordings are stored (e.g. for lifelogging), phone detection in the scene can be exploited *both* for recalibration and privacy redaction. In contrast, if the scene recordings are not needed or momentarily interrupted, then our touch-based approach avoids the need for input from a scene camera altogether and thus helps to preserve privacy.

5.6.6 Outlook: Generalising our Approaches

While we utilised phone interactions, both our recalibration approaches could be extended to other devices, such as mobile devices like tablets, smart watches and laptops (Zhang *et al.*, 2018b), or stationary devices like public displays. For the extension of our approach based on phone saliency maps, related work on automatic detection of screens can be helpful (Korayem *et al.*, 2016).

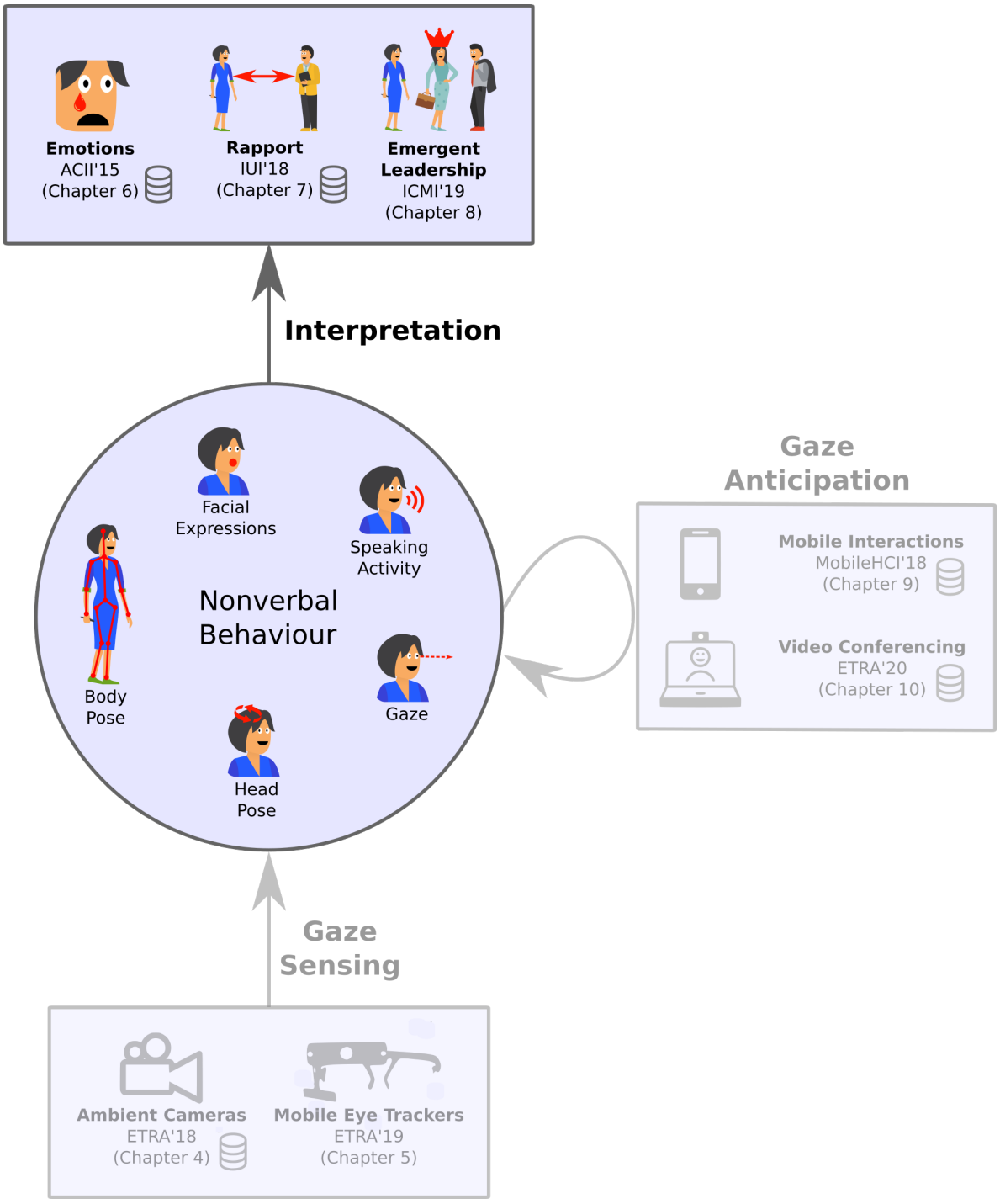
A main conceptual generalisation of our blind recalibration approach would no longer assume visual attention and interaction at the *same* location, but rather include cases with *separate* locations: For example, we might assume visual attention on a desktop monitor while keystrokes appear at the keyboard, or attention on a smart TV while using the remote control. For these cases, we need to be able to make robust assumptions about locations of these objects in the scene camera view. This might hold for some devices and contexts but not for others. For example, a laptop might commonly be located at the bottom to centre area of the camera view, while a smart TV might appear in different areas depending on where the user is sitting. These considerations present ample opportunities for future work, which could systematically collect such use cases beyond phone and touch, and investigate how to generate and exploit saliency maps for them.

Finally, our approach might be generalised beyond interaction with computing devices, such as reaching for a coffee mug, operating an elevator or a vending machine, and so on. Related, future work might exploit gaze behaviour in social situations, such as looking at faces and speakers (cf. (Müller *et al.*, 2018b; Siegfried *et al.*, 2017)), or following pointing hands or handing over objects (e.g. money at a counter).

5.7 Conclusion

In this work we presented two novel methods to recalibrate mobile eye trackers by exploiting mobile phone usage. Our first method is based on saliency maps built from phone detections obtained from the scene camera. Our second method "blindly" recalibrates the eye tracker using touch events registered on the mobile phone. We evaluated both methods against the state of the art on a recent dataset of in-the-wild mobile eye tracking recordings. Both our methods reduced calibration drift and significantly outperformed the state-of-the-art method. While our blind recalibration approach performs slightly worse than our phone detection-based one, it offers advantages in privacy-sensitive situations, as it does not rely on images obtained from a scene camera. As such, we believe our work represents an important step towards enabling gaze-based interaction techniques in daily life.

Part II


Emotions
ACII'15
(Chapter 6)
Rapport
IUI'18
(Chapter 7)
Emergent
Leadership
ICMI'19
(Chapter 8)
Interpretation
Facial
Expressions
Speaking
Activity
Nonverbal
Behaviour
Body
Pose
Gaze
Head
Pose
Gaze
Anticipation
Mobile Interactions
MobileHCI'18
(Chapter 9)
Video Conferencing
ETRA'20
(Chapter 10)
Gaze
Sensing
Ambient Cameras
ETRA'18
(Chapter 4)
Mobile Eye Trackers
ETRA'19
(Chapter 5)

6 Emotion Recognition from Embedded Bodily Expressions and Speech during Dyadic Interactions

PREVIOUS work on emotion recognition from bodily expressions focused on analysing such expressions in isolation, of individuals or in controlled settings, from a single camera view, or required intrusive motion tracking equipment. We study the problem of emotion recognition from bodily expressions and speech during dyadic (person-person) interactions in a real kitchen instrumented with ambient cameras and microphones. We specifically focus on bodily expressions that are embedded in regular interactions and background activities and recorded without human augmentation to increase naturalness of the expressions. We present a human-validated dataset that contains 224 high-resolution, multi-view video clips and audio recordings of emotionally charged interactions between eight couples of actors. The dataset is fully annotated with categorical labels for four basic emotions (anger, happiness, sadness, and surprise) and continuous labels for valence, activation, power, and anticipation provided by five annotators for each actor. We evaluate vision and audio-based emotion recognition using dense trajectories and a standard audio pipeline and provide insights into the importance of different body parts and audio features for emotion recognition.

6.1 Introduction

Emotions are an integral part of human communication and manifest themselves in vocal prosody but also in body movements, facial expressions, and gestures. Particularly body movements induced by emotional responses, colloquially referred to as body language, play a key role in non-verbal human communication that is believed to represent a substantial part of all human communication (Hogan, 2003). In contrast to facial expressions, speech, as well as physiological parameters, such as heart rate or galvanic skin response, analysis of body language and recognition of emotions from bodily expressions is less well-explored in affective computing. This is mainly due to the significant challenge of recording and annotating natural bodily expressions of emotions in everyday environments. Consequently, previous works in affective computing mainly focused on datasets recorded in artificial laboratory settings. In these settings, individual actors were either positioned directly in front of the camera (Bänziger *et al.*, 2012) or couples of actors were recorded using intrusive motion capture equipment to track their body movements (Metallinou *et al.*, 2010) (see Figure 6.2 for examples).

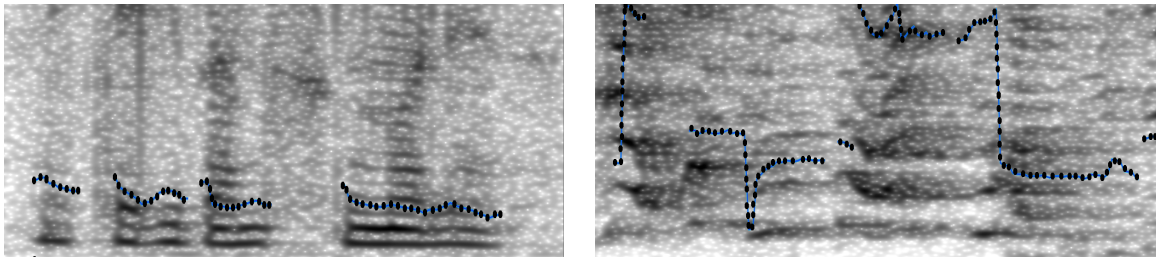
Figure 6.1: Sample bodily expressions associated with different emotions from our dataset.

In this paper we investigate multimodal emotion recognition from bodily expressions and speech recorded using unobtrusive ambient cameras and microphones in a real kitchen environment during naturalistic dyadic (person-person) interactions.

We propose an experimental setup and methodology that allows us to systematically record such bodily expressions embedded in regular interactions and background activities. To this end, we develop a set of scenarios that evolve around daily-life events and that lead to an emotionally charged conversation between two people. Each scenario is endowed with background information on the attitude of each person towards the event. We then film multiple pairs of actors role-playing and improvising each scenario in a fully functional apartment kitchen to closely resemble natural everyday living conditions (see Figure 10.1 for examples).

We took particular care to not script actors' performance, i.e. the only information we provided was a high-level background description of the scenario and emotional responses each of the actors was supposed to exhibit. In particular, we did not instruct actors how to role-play each scenario, or which bodily expressions or motions they should use to express a particular emotion. Instead, actors were free to interact with the environment and move around inside the kitchen. They were also not encumbered by

Figure 6.2: Sample scenes of emotionally charged person-person interactions from our dataset (top). Samples from GEMEP (Bänziger *et al.*, 2012) (bottom left) and CreativeIT (Metallinou *et al.*, 2010) (bottom right) datasets.



GEMEP *MPIIEmo(ours)*

Figure 6.3: Examples of audio spectrograms computed on GEMEP (Bänziger *et al.*, 2012) and our MPIIEmo dataset. Blue curves correspond to pitch trajectories extracted with the approach described by Kasi and Zahorian (2002). Note that the spectrogram on GEMEP is cleaner which enables more robust pitch extraction.

wearing motion capture equipment, which made their bodily expressions more natural. The resulting MPIIEmo dataset including all annotations will be made publicly available upon publication.

The contributions of this work are threefold. First, we present an experimental set-up and methodology to unobtrusively collect video and audio data of actors engaged in person-person interactions in an everyday environment. The set-up and methodology were specifically developed to balance between the realism of the exhibited bodily expressions as well as the ability to study emotions that are difficult to record in real-world situations. Using this methodology, we further introduce the MPIIEmo dataset that contains 224 high-quality video and audio recordings of eight couples of actors engaged in emotionally-charged, natural interactions revolving around everyday scenarios. To the best of our knowledge, this is the first dataset of full-body videos of dyads of unaugmented people in affective interactions. It features a mix of emotional expressions embedded in a regular conversation and background activities, is fully-annotated with both categorical and continuous emotion labels, and provides multiple synchronised camera views. Third, we evaluate an approach for emotion recognition from video based on dense trajectories (Wang *et al.*, 2013a) and body part detections and provide insights into the relative importance of body parts for emotion recognition. We further study emotion classification from audio (Anagnostopoulos *et al.*, 2015), highlighting difficulties of audio feature extraction on our dataset compared to the more constrained GEMEP dataset.

6.2 Related Work

Our work is related to previous work that 1) explored the close link between emotions and body movements, 2) focused on recording bodily expressions of emotions, and 3) developed computational methods for human behaviour analysis.

6.2.1 Emotions and Body Movements

Humans are skilled in expressing emotions through non-verbal signals and in interpreting signals of others but relating body movements to specific emotional expressions is challenging given the subtleness of these movements. Efforts to clarify connections between emotions and body movements have a long history in behavioural science and suggested a strong connection exists between emotions and body movements (De Meijer, 1989)(Wallbott, 1998)(Pollick *et al.*, 2001). Analysis of emotional body movements has since sparked a large body of work in social signal processing and affective computing, focusing on both encoding body movements in a similar fashion as facial expressions (Dael *et al.*, 2012)(Velloso *et al.*, 2013a) as well as inferring emotions from body movements (Kapur *et al.*, 2005)(Bernhardt and Robinson, 2007)(Bernhardt and Robinson, 2009) (see (Vinciarelli *et al.*, 2009)(Zeng *et al.*, 2009)(Kleinsmith and Bianchi-Berthouze, 2013)(Karg *et al.*, 2013) for reviews).

6.2.2 Recording Bodily Expressions of Emotions

Several previous works investigated bodily expressions of emotions of individuals, either involving directed or at least carefully executed bodily expressions while facing the camera (Bänziger *et al.*, 2012)(Gunes and Piccardi, 2007)(Glowinski *et al.*, 2011) or using body motion capture suits in controlled laboratory settings (Wang *et al.*, 2013b). Previous works that studied bodily expressions during person-person interactions were either limited to only one person showing affective behaviour (Bergmann *et al.*, 2014), to sitting people (Busso *et al.*, 2008) or also used artificial settings and required sophisticated human augmentation (Metallinou *et al.*, 2010)(Yang *et al.*, 2014b) (see Čereković (2014) for a recent survey). Previous works also often only included short snippets of isolated bodily expressions that were neither embedded in natural background activities nor interactions with the other person or the environment (Bänziger *et al.*, 2012)(Wang *et al.*, 2013b).

Our methodology is most similar to the one described by Metallinou *et al.* (2010) but aims to increase realism of the recorded data while still retaining the ability to obtain laboratory-standard recording quality and accurate ground truth annotations. Specifically, our dataset contains bodily expressions of emotions during naturalistic person-person interactions, i.e. interactions that develop around everyday events and that are therefore embedded in casual body movements as well as interactions with the other person and the environment. As described in (Bänziger *et al.*, 2012)(Metallinou *et al.*, 2010) and following guidelines from (Busso and Narayanan, 2008)(Scherer and Bänziger, 2010) we rely on recruited actors to improvise emotional expressions. Our dataset further contains two schemes for representation and annotation of emotional content, namely both categorical emotional labels (Bänziger *et al.*, 2012) and continuous affect dimensions (Metallinou *et al.*, 2011). Sample scenes from two existing datasets as well as our own are shown in Figure 10.1.

6.2.3 Human Behaviour Analysis

Computational methods to analyse human behaviour either rely on on-body sensors, such as inertial measurement units, or ambient sensors, such as video cameras. On-body sensors are widely used in human activity and gesture recognition (Bulling *et al.*, 2014). While current activity recognition systems achieve good performance for many activity recognition tasks, the majority of research focuses on recognising "which" activity is being performed at a specific point in time. More closely related to the problem investigated in this work, is qualitative activity recognition that studies means to extract qualitative information from inertial data, such as the quality or correctness of executing an activity. Such qualitative assessments are more challenging to perform automatically and have so far only been demonstrated for constrained settings, such as in sports. Specifically, previous works studied qualitative assessment of activities such as weight-lifting (Velloso *et al.*, 2013b)(Velloso *et al.*, 2013c)(Velloso *et al.*, 2011), rowing (Tessendorf *et al.*, 2011) or balance board exercises (Moller *et al.*, 2012). Recent computer vision works on human behaviour analysis mainly focused on basic recognition tasks, such as people detection (Felzenszwalb *et al.*, 2010), pose estimation (Yang and Ramanan, 2013)(Pishchulin *et al.*, 2013)(Tompson *et al.*, 2014), and recognition of fine grained details, such as appearance attributes (Zhang *et al.*, 2014), body and head orientation (Maji *et al.*, 2011), gaze direction (Zhang *et al.*, 2015), detection of facial key-points (Zhu and Ramanan, 2012), or social signals, such as holding hands or hugging (Yang *et al.*, 2012). In this work we investigate how recent advances in computer vision enable recognition of bodily expressions of emotions in video. In particular, we build on (Wang *et al.*, 2013a) that was previously used for activity recognition and (Sermanet *et al.*, 2014) for body pose estimation.

6.3 The MPIIEmo Dataset

Collecting video and audio footage of bodily expressions of emotions in everyday settings is challenging. In addition to the scarcity of such situations in daily life, legal and ethical issues pose significant challenges for the collection of real-world data. Similar to Busso and Narayanan (2008) and Scherer and Bänziger (2010) we therefore opted to rely on acted performances and recorded couples of actors interacting with each other in a naturalistic environment (an apartment kitchen).

6.3.1 Data Recording

We designed the data recording with two main objectives in mind: 1) to record video and audio footage of person-person interactions in a real kitchen setting and without on-body motion capture equipment that could affect the realism of these interactions, and 2) to record bodily expressions of emotions that are embedded in regular interactions and background activities commonly performed in the kitchen.

He is **happy** with her.

He also applied for the same job.
He is **angry**, because she
talks happily about her new job.

He is **sad** because her new job means,
she will soon move away.
They had a close relationship.
She also becomes **sad**.

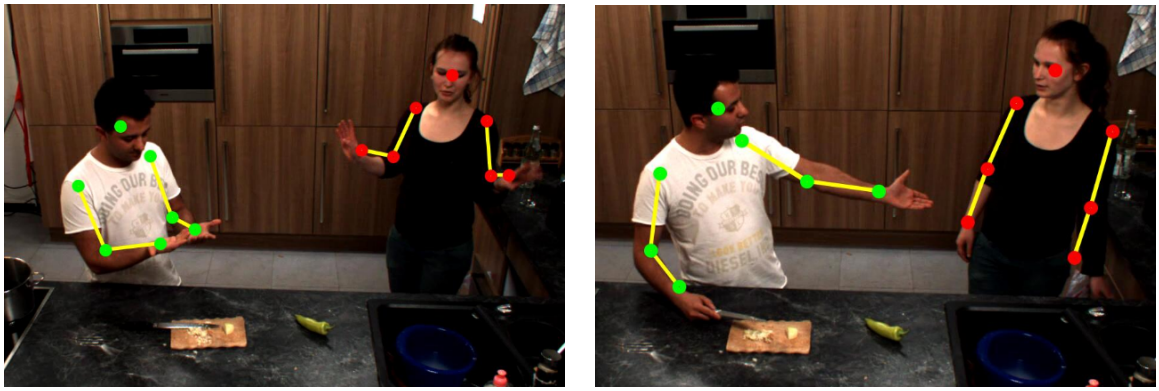
He is **surprised** because she told before,
how hard it is to get the job.
She is **proud** (occurs later in the sequence).

Figure 6.4: Sample scenario from our dataset. Each picture illustrates one subscenario. The high-level scenario description was: *She just received an offer for the job she always wanted. She enters the kitchen and tells the news* ***happily.***

6.3.1.1 *Recording setup*

We recorded video and audio footage using eight ceiling-mounted, frame-synchronized machine vision cameras recording at 29.4 fps and four microphones, covering the whole interaction space inside the kitchen (see Figure 6.6). In total, we recorded eight pairs of actors (three female only, two male only, three mixed), with each pair performing seven scenarios, each consisting of four subscenarios. This resulted in 224 video clips with a total length of 143 minutes or 252,457 frames and an average length of recorded video clips of around 38 seconds. The subscenarios were different variations of the overall scenario covering the display of different emotional responses. We designed each subscenario to correspond to a short conversations with the overall objective to record a diverse set of interactions that felt natural to the actors. According to these criteria, scenarios and subscenarios were selected from a pool of proposals by testing them in trial runs.

$t = 1$ $t = 10$



$t = 27$ $t = 40$

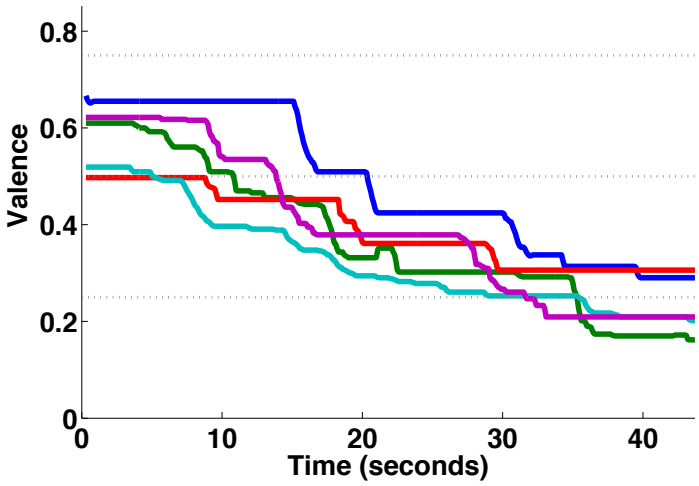


Figure 6.5: Sample frames from a sequence with annotations for valence of the female subject with pose estimates. At about 20 seconds the couple gets into an argument about throwing away the garbage, as indicated by a more negative valence rating.

Figure 6.6: Kitchen environment used for recording our dataset. The kitchen was fully functional and instrumented with ceiling-mounted cameras (red circles) and microphones (blue circles).

A sample scenario from our dataset is shown in Figure 6.4. In this scenario, one person reminds the other that it is his turn to empty the waste bin. The subscenarios then evolve around different reactions of the second person. He is either happy to be reminded, angry at the annoying reminder, angry at himself for forgetting about it again, or surprised because it's not his turn. The first person then reacts accordingly. The full list of all scenarios and subscenarios will be released with the dataset.

6.3.1.2 *Recording methodology*

Actors were recruited from local student theatre groups and selected based on their acting abilities. All actors had at least one year of theatre training and were practising improvisation as a part of it. However, most actors were much more experienced, and half of them were, among others, part of a group dedicated solely to improvisational theatre. A director from a local theatre group worked with the actors during the recording. Actors were given short descriptions (about 1-3 sentences) of the scenarios and subscenarios, including explicit statements about the emotions and feelings involved in the interaction. Four of the six basic emotions by (Ekman *et al.*, 1969) were explicitly referred to: Happiness, Anger, Sadness and Surprise. Due to the freedom of improvisation, emotional expressions not covered by those four were shown as well as mixtures of several emotions. If necessary for the actors to become more familiar with a subscenario, additional background information was provided by the experimental assistant. In case the actors had problems to access the required emotions, the director used reenactment techniques as in (Bänziger *et al.*, 2012). Otherwise the actors were free to improvise. In particular, no instructions concerning concrete verbal or non-verbal expressions or gestures were given. Actors did not wear any specific clothes, could move freely inside the kitchen, and were free to interact with the kitchen itself as well as all objects, tools etc. in it. Each subscenario was repeated until the actors and the director were satisfied with the performance.

Scale	macc	map	F 1	F 2	F 3	F 4	F 5
Happiness	91.01	88.61	32.72	26.44	21.07	17.06	11.25
Anger	94.98	91.26	24.63	20.17	17.56	15.33	11.62
Sadness	94.62	79.60	18.46	12.25	9.14	6.71	5.41
Surprise	84.66	44.86	43.35	21.40	13.53	7.09	2.41

Table 6.1: Performance evaluation of human annotators. "macc" and map" correspond to "mean accuracy" and "average precision". "F k" is the relative frequency of the emotion class when agreement of k annotators is required to mark a sample as positive.

6.3.2 Groundtruth Annotation

Two well-known models to describe emotions are the categorical and the dimensional emotion model. Categorical representations discretise the space of emotions and put labels like "Happiness" or "Surprise" to individual emotions. Recently, researchers in affective computing argued for the use of dimensional emotion models understanding affect as a real-valued vector (Metallinou and Narayanan, 2013). Our dataset provides annotations for both emotion models. We used a subset of the well-known six basic emotions (Ekman *et al.*, 1969), namely anger, happiness, sadness, and surprise. For the dimensional model, we used the four dimensions valence, activation, power and anticipation as suggested by (Fontaine *et al.*, 2007). In our setting *valence* intuitively describes whether the actor felt "good" or "bad", *activation* concerned his activeness vs. in-activeness, *power* referred to whether he felt in control of events, and *anticipation* varied between the actor being surprised and feeling he could foresee the future.

We extended the annotation tool GTrace (Cowie and Sawey, 2011) to support both emotion models. Annotations were performed by five psychology students (three female), instructed with descriptions of the different emotions as well as example video clips from a separate test recording. We then asked the annotators to label all video clips in randomized order, sequentially for both actors, and for each actor using first the dimensional and then the categorical emotion model. For the dimensional model we obtained continuous annotations across the whole sequence. For the categorical model, annotators were asked to select a subset of the six basic emotions for each actor in a clip. If an emotion was selected, its intensity was rated continuously in the same way as for the dimensional emotion model.

6.3.3 Analysis of Annotations

We analysed the quality of both dimensional and categorical ground truth annotations:

Dimensional emotion model.. We quantified the agreement between annotators by computing the median of the Pearson correlation coefficients between all pairs of annotators. The correlations were computed across all frames. We obtained a high

median correlation for valence (0.84), moderate median correlations for Power (0.67) and Activation (0.65), and a low correlation for Anticipation (0.42).

Categorical emotion model.. We aggregated annotations over several annotators to get discrete labels for all emotion categories. Prior to aggregation we normalized the intensity ratings of each annotator by subtracting the mean and dividing by the variance across all videos. When computing the mean and variance for a particular emotion we assigned a zero intensity rating to all videos where this emotion was not labelled as present. For simplicity, we afterwards discretized the intensity ratings into binary emotion category labels separately for each annotator by thresholding the normalized intensity at 1. Finally, we defined the emotion label for a frame by requiring $k = 2$ annotators to agree that the emotion is present in this frame.

Table 6.1 (left) shows the mean accuracy and mean average precision for emotion recognition achieved by our annotators. The results are obtained by using three annotators to generate ground-truth labels and comparing it with the output of the remaining two annotators, repeating the process for all combinations of annotators. To calculate mean average precision, we used the ordering of data points induced by the annotators ratings. As can be seen from the table, while Happiness, Anger and Sadness are quite accurate, Surprise has a low mean average precision. The reason for this is that Surprise (like the related Anticipation scale) is often very strictly localized in time. In consequence, person-specific delays of the annotators introduce a lot of label uncertainty. Table 6.1 (right) shows the relative frequency of the positive class on the whole dataset for different values of k. We observe that agreement across annotators varies significantly across emotions. For example there is approximately two-fold decrease between $k = 2$ and $k = 5$ for Anger (20.17 vs. 11.62) whereas we observe nearly ten-fold decrease for Surprise (21.40 vs. 2.41). We can observe a similar pattern when quantifying agreement of annotators by computing the median of their correlations. We get high correlations for Anger (0.87), Happiness (0.82) and Sadness (0.83), but a low correlation for Surprise (0.48).

6.4 Emotion Classification from Video and Audio

To establish baseline performances for emotion classification on our MPIIEmo we evaluated approaches from computer vision and speech analysis. In the visual domain, we further examine the influence of different body parts and the interlocutor. In the audio domain, we compare several features that are commonly used for emotion classification from speech with respect to their performance on MPIIEmo as well as the well-established GEMEP dataset (Bänziger *et al.*, 2012).

6.4.1 Video

We use dense trajectories, a recently introduced video descriptor which showed state-of-the-art performance for human activity recognition (Wang *et al.*, 2013a). By using pose

estimates, we can in- or exclude dense trajectories from different body parts or persons, which allows us to estimate their importance for emotion recognition in our framework.

6.4.1.1 *Pose estimation*

To estimate poses of people we train a set of body part detectors building on the convolutional neural network architecture of (Sermanet *et al.*, 2014). The detector in (Sermanet *et al.*, 2014) is trained by minimizing a multi-task loss function that combines detection accuracy and accuracy in prediction of the object bounding boxes. When applying this approach to pose estimation we substitute the bounding box prediction component with a component that predicts locations of the neighboring body part. We train detectors for the head, shoulder and wrist. Wrist and shoulder detectors are trained to also predict the location of the elbow joint. Each shoulder and wrist detection thus generates a pair of body joints corresponding to either upper arm or lower arm segments respectively.

In the first step each part detector is densely evaluated in each camera view resulting in an initial set of candidate part hypothesis. In the second step we refine the body part detections using multi-view constraints. To that end we match the body segments across camera views and generate a set of 3D body segment candidates using triangulation. In the process we also discard segments with high reconstruction error, which allows us to filter out false positive detections. In the final step we assemble 3D full-body configurations from the available pool of body segments using constraints on the relative position of head, and upper and lower arms. This process results in high-quality 3D pose estimates for both subjects in the majority of the images. The remaining failures in pose estimation correspond to cases with particularly strong occlusions or rare poses such as subjects bending under the kitchen counter.

6.4.1.2 *Identity annotation*

The pose estimates are not associated with individual actors. To add personal identities, we annotate which actor is rightmost in one of the camera views and match this annotation to one of the two estimated body configurations.

6.4.1.3 *Dense trajectories on body parts*

We compute dense trajectories from a single view (the one in Figure 6.5), and associate them with 2D person and body part bounding boxes. A trajectory is associated with a bounding box if its starting point (x, y, t) is inside the bounding box at time t. We build separate codebooks for each of the five feature channels of the dense trajectory descriptor using k-means clustering with $N = 4000$ centroids on 100,000 trajectories randomly sampled from the training set. Depending on the experimental condition, we build separate codebooks for different body parts. For training and testing, we compute histograms over a time window of 2 seconds separately for each actor.

6.4.1.4 *Classification*

We apply SVM with a RBF-χ^2 kernel k as in (Jhuang *et al.*, 2013). The L feature channels are combined by normalizing their corresponding χ^2 distances separately using the means of the χ^2 distances of the feature channels on the training set:

$$k(x, y) = exp(-\frac{1}{L}\sum_{c=1}^{L}\frac{\chi^2(x_c, y_c)}{A_c}). \tag{6.1}$$

Here $\chi^2(x, y)$ denotes the χ^2 distance between x and y, x_c the c-th feature channel of example x and A_c the mean χ^2 distance for feature channel c on the training set.

6.4.2 Audio

We compute three features commonly used for emotion recognition from speech (Anagnostopoulos *et al.*, 2015): (1) non-zero pitch values, (2) spectral centroid and spectral flatness of the timbre and (3) short time energy of the audio signal. The mean and standard deviation of these features are computed for frames of 30ms with 10ms hops, yielding an 8 dimensional feature vector for each frame. Classification is performed by using SVM with an RBF kernel and cross-validating the hyperparameters C and γ on the training set.

6.5 Experiments

6.5.1 Video

We report results on the detection of four types of emotional states. The detection is performed using a sliding window approach with a stepsize of 2 seconds. To compare detection results to human performance in Table 6.1 we generate labels from a fixed set of 3 annotators. Each window is then considered as positive for a given emotion category if at least half of the frames in that window are labelled positively by at least two annotators. A separate classifier is trained for each emotion category against other emotions and background. The regularization parameter C is selected by cross-validation in the training set. We report performance using the average precision metric as is common in human activity detection (Wang *et al.*, 2013a).

To quantify the contribution of different body parts, we compare different ways of selecting trajectories (see Table 6.2). First, we investigate differences in performance due to exclusion of trajectories associated with certain body parts (conditions *full*, *full-head*, *full-hw* in Table 6.2). We observe, that removing trajectories from the head lowers the performance for all emotions, whereas additionally removing trajectories from the wrists only results in a significant performance drop for the Happiness class. Secondly, we investigate the performances of classifiers based exclusively on trajectories associated with the head and the wrist. We find, that using trajectories from the head results in better performance and, more surprisingly it even outperforms *full* on all

Method	Happiness	Anger	Surprise	Sadness	Average
full	48.0	28.4	24.6	16.8	29.5
full-hw	41.5	26.0	23.2	15.5	26.5
full-head	46.5	26.6	23.8	15.6	28.1
wrist	44.3	25.2	21.8	16.2	26.7
head	**50.7**	**32.9**	**26.2**	**18.2**	**32.0**
head-single	46.9	27.9	20.0	15.7	27.6
posrate	21.7	18.5	13.8	10.2	16.1

Table 6.2: Mean average precision in percent for leave-one-recording-out cross-validation on our MPIIEmo dataset. "head", "wrist" and "full" denote using trajectories on the head, wrist or the full body, respectively. "full-head" denotes using all trajectories except head, and "full-hw" all trajectories except head and wrist. "head-single" denotes using trajectories from the target person only.

Dataset	Pitch	Timbre	Energy	All	MLK
GEMEP	53.9	58.9	48.7	64.1	26.0
MPIIEmo	36.5	41.1	41.0	43.2	35.0

Table 6.3: Results for emotion classification using audio features. MLK denotes the probability of the most likely class in percent points.

emotions. Finally, we pick the best performing condition to quantify the contribution of features extracted from the interlocutor. When removing those features (*head-single*), performance drops significantly.

When comparing these results with the performance of human annotators in Table 6.1, we note that human performance is strongly superior for all emotion classes.

6.5.2 Audio

We compare the performance of pitch-, timbre- and energy-based features on MPIIEmo. As a reference we also report results on the more controlled, single-actor GEMEP dataset. To align the experimental setups, we pick the 4 classes from GEMEP that are most similar to the 4 emotions on MPIIEmo (Anger, Joy, Sadness, Surprise), resulting in 39 examples. For MPIIEmo, we extract two second long training windows, with 10Hz sampling frequency, excluding all windows that had either multiple labels per actor, or no label at all (background). Note, that we construct examples without speaker separation, as first experiments using ICA indicated that this is a difficult task on MPIIEmo. As a result, the same features might appear in different classes if the two actors were given different labels in one window, making our task inherently more difficult than the

one we defined on GEMEP. To compute the test error, we use leave-one-sequence-out cross-validation on GEMEP and leave-one-couple out cross-validation on MPIIEmo. The results (Table 6.3) show, that combining all features achieves the best performance. Surprisingly, although pitch performs well on GEMEP and in prior research (Madzlan *et al.*, 2014), it is near chance on MPIIEmo. Upon closer inspection, the bad performance on MPIIEmo can be explained by the difficulties for pitch extraction arising from the more realistic recording situation with microphones being at a distance from the speakers (*cf.* Figure 6.3).

6.6 Conclusion

In this paper we proposed a new experimental setup and methodology to record bodily expressions of emotions embedded in everyday person-person conversations as well as background activities. Using this methodology, we presented the fully annotated MPIIEmo dataset that contains 224 high-resolution, multi-view video clips and audio recordings of emotionally charged interactions between eight couples of actors. We established baseline performances for emotion classification from both video and audio. We found that visual features computed from the head as well as the interlocutor were particularly important to achieve good performance, and that the more naturalistic recording setup on MPIIEmo poses challenges for audio feature extraction. To spark further research on this challenging emotion classification problem, the full dataset including all body pose estimates as well as categorical and continuous emotion annotations is publicly available.

7 Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour

Rapport, the close and harmonious relationship in which interaction partners are "in sync" with each other, was shown to result in smoother social interactions, improved collaboration, and improved interpersonal outcomes. In this work, we are first to investigate automatic prediction of low rapport during natural interactions within small groups. This task is challenging given that rapport only manifests in subtle non-verbal signals that are, in addition, subject to influences of group dynamics as well as inter-personal idiosyncrasies. We record videos of unscripted discussions of three to four people using a multi-view camera system and microphones. We analyse a rich set of non-verbal signals for rapport detection, namely facial expressions, hand motion, gaze, speaker turns, and speech prosody. Using facial features, we can detect low rapport with an average precision of 0.7 (chance level at 0.25), while incorporating prior knowledge of participants' personalities can even achieve early prediction without a drop in performance. We further provide a detailed analysis of different feature sets and the amount of information contained in different temporal segments of the interactions.

7.1 Introduction

Inter-personal conflicts are pervasive and can happen in a variety of social settings, from festivals, to family gatherings, to a bar or the classroom. Many of these conflicts are the result of low rapport between interaction or conversation partners, or more specifically, the failure of a person to establish good rapport. While a precise definition is difficult, rapport refers to the close and harmonious relationship in which interaction partners are "in sync" and can interact naturally and smoothly with each other. Failure to build rapport can lead to mutual feelings of disharmony or, in the worst case, even verbal or physical hostility. The fundamental importance of rapport for social interactions underlines the significant potential of developing intelligent user interfaces that are able to detect low rapport and to reduce or even avoid inter-personal conflicts.

While several previous works in social signal processing and affective computing investigated automatic detection of rapport during *dyadic* (person-person or person-machine) interactions from verbal and non-verbal behaviour (Cerekovic *et al.*, 2016; Hagad *et al.*, 2011; Wang and Gratch, 2009; Zhao *et al.*, 2016), few works studied the link between non-verbal behaviour and rapport in larger groups (LaFrance and

Figure 7.1: Example images of natural behaviours from the dataset.

Broadbent, 1976). No attempt has so far been made to automatically detect low rapport in *multi-person* interaction settings. This is despite the fact that much of our social life takes place in groups larger than two people, e.g. in business meetings or friend gatherings. Detecting low rapport and performing an early intervention to avoid social conflicts can therefore have a significant and practical impact.

We present the first study on detecting the failure to establish rapport in natural multi-person interactions with a small number of people. Given that no dataset exists that comprises small group interactions as well as annotations of felt rapport, we record a new dataset (see Figure 7.1 for example images). Based on this dataset, we then develop a multimodal approach to automatically detect low rapport from non-verbal behaviour. Our approach is based on state-of-the-art methods to analyse facial expression and posture, as well as speech activities and prosodic features. We further propose new features that exploit the mirroring effect by accounting for behaviour synchronisation among group members as well as cross-modal features that delineate simultaneous actions from different modalities. Our results show that while facial features perform best when the full interaction is observed, prior information about participants' personalities can boost facial features to achieve the same performance while observing only the first third of the whole interaction.

The specific contributions of our work are three-fold: 1) We collect the first dataset for small group interactions with informative audio-visual signals and rich annotations, including *felt rapport*, perceived leadership, dominance, competence and liking of the dyads in the group. 2) We propose a multimodal approach to low rapport detection that exploits both dyadic audiovisual information, such as facial action units and speech prosody, as well as group information, such as cross-modal features and mirroring effects, and potential prior knowledge of the participants' personalities. 3) We provide an in-depth performance evaluation of our method and identify key features and time segments that are most important for rapport detection in this setting.

7.2 Related Work

We first summarise prior works that aimed to predict social aspects in multi-person interactions. We then focus on rapport as a particularly important social aspect, followed by computational methods to predict rapport in dyadic interactions.

7.2.1 Automatic Analysis of Multi-Person Interactions

While a major part of research in social signal processing and affective computing focuses on the analysis of dyadic interactions (Cafaro *et al.*, 2017; Müller *et al.*, 2015; Ringeval *et al.*, 2013), a growing body of work on multi-person interactions has developed in recent years. Among the social concepts that have been studied in multi-person interactions from a computational perspective are turn-taking (Bohus and Horvitz, 2011; Laskowski, 2010; Rühlemann and Gries, 2015), laughter (McKeown *et al.*, 2015), general interest level of the group (Gatica-Perez *et al.*, 2005), and engagement of individuals inside the group (Oertel and Salvi, 2013). Cohesion is one of the more abstract concepts and describes the tendency of group members to create social bonds and stay united as a group. Cohesion is commonly understood as a global measure given by each interactant to the whole group. In contrast, rapport can be measured for each pair of people within a group and can thus provide a more detailed picture of intra-group relations. Hung and Gatica-Perez analysed audio, visual and audio-visual cues to predict cohesion levels in small groups using annotations from external observers (Hung and Gatica-Perez, 2010). More recently, Nanninga et al. specifically focused on the connection between group mimicry and task cohesion (Nanninga *et al.*, 2017). Other works focused on leadership and listener behaviour. Automatic recognition of emergent leaders is particularly relevant given that those leaders emerge from the interaction among group members (as opposed to designated leaders). The prediction of perceived leadership alongside dominance, competence and liking in groups of three to four people was also studied using information on speaking activity and speech prosody as well as activity of the body and head (Beyan *et al.*, 2017b; Sanchez-Cortes *et al.*, 2012). Other works aimed to differentiate instructed considerate from authoritarian leadership styles (Feese *et al.*, 2011) or, more recently, naturally emerging autocratic or democratic behaviour (Beyan

et al., 2017b). Classification of group members into attentive listeners, side participants, and bystanders was studied in (Oertel *et al.*, 2015).

In summary, while a number of works studied different prediction tasks in multi-person interactions, some of which are related to rapport, to the best of our knowledge we are first to predict rapport in a multi-person setting.

7.2.2 Rapport

Among the different concepts of social interactions, rapport is arguably one of the most fundamental and thus important. Failure to build rapport can result in poor social interactions, decreased collaboration, and worse interpersonal outcomes (Burns, 1984; Kelley *et al.*, 2014; Tsui and Schultz, 1985). In an early work, Tickle-Degnen and Rosenthal identified three components that are important for rapport: attention, positivity, and coordination (Tickle-Degnen and Rosenthal, 1990). The importance of these components can change over the course of a relationship as can the expression of components of rapport. For example, insults can help build rapport in later stages of a relationship (Ogan *et al.*, 2012). Izard hypothesised connections between personality traits and the ability to build rapport (Izard, 1990). For example, people with high extraversion were deemed to find it easier to build rapport given that they might more easily focus their attention on others. Furthermore, people with a tendency towards negative emotions might not be able to express the positivity component of rapport strongly enough.

Research on the link between dyadic rapport and non-verbal behaviour is extensive, so we only discuss two representative works. Harrigan, Oxman and Rosenthal analysed rapport ratings for physicians obtained from nurses (Harrigan *et al.*, 1985). They found that physicians sitting with uncrossed legs and arms in symmetrical side-by-side positions directly facing the patient received higher rapport ratings. Bernieri et al. conducted an important analysis on dyad rapport and its judgement across different situations (Bernieri *et al.*, 1996). When comparing subjective rapport ratings with ratings by external observers they found that observers had a hard time rating rapport consistently with the participants who experienced the situation. Further analyses showed that while observer judgements were mainly based on the amount of expressiveness, self-ratings were adapted to the specific situation at hand. These results indicate that observer ratings of rapport are not adapted to the specific situation. We therefore opted to use self-reported rapport ratings in this work.

7.2.3 Predicting Rapport in Dyadic Interactions

Computational approaches to rapport prediction focused on dyadic interactions, typically with the motivation to develop artificial agents that are able to build rapport with users. The first line of work investigated non-verbal cues for rapport prediction. For example, Wang and Gratch used selected facial action units (AU) to predict felt rapport in human-human and human-agent interactions (Wang and Gratch, 2009). They found that felt rapport was encoded in the absence of AUs encoding negative emotions rather

than in the presence of AUs indicating positive emotions. Hagad et al. used participants' postures and their congruences to predict rapport in dyadic interactions (Hagad *et al.*, 2011). Other works used verbal cues or mixtures of non-verbal and verbal cues for rapport prediction. A recent study by Cerekovic et al. focused on predicting self-reported and observer-rated rapport between humans and virtual agents using verbal and non-verbal cues (Cerekovic *et al.*, 2016). The results showed that self-reported rapport is rather weakly correlated with observer-judged rapport, and also harder to predict than the latter. Zhao et al. applied temporal pattern mining to extract rules for rapport management in dyads of peer-tutoring strangers and friends (Zhao *et al.*, 2014, 2016). An example of such a rule indicative of high rapport in friend dyads is the verbal violation of a social norm by one interactant while in parallel her friend is smiling. Finally, bonding is a concept related to rapport and has recently been studied in the context of dyadic human-human and human-agent interactions (Jaques *et al.*, 2016b), also depending on personality (Jaques *et al.*, 2016a).

7.3 A Dataset of Small-Group Interactions

Given the lack of suitable datasets for the development and evaluation of algorithms for rapport detection, we designed a human study to collect audio-visual non-verbal behaviour data and rapport ratings during small group interactions. Our dataset consists of 22 group discussions in German, each involving either three or four participants and each lasting about 20 minutes, resulting in a total of more than 440 minutes of audio-visual data.

7.3.1 Recording Setup

The data recording took place in a quiet office in which a larger area was cleared of existing furniture. The office was not used by anybody else during the recordings. To capture rich visual information and allow for natural bodily expressions, we used a 4DV camera system to record frame-synchronised video from eight ambient cameras. Specifically, two cameras were placed behind each participant and with a position slightly higher than the head of the participant (see the green indicators in Figure 7.2). With this configuration a near-frontal view of the face of each participant could be captured throughout the experiment, even if participants turned their head while interacting with each other. In addition, we used four Behringer B5 microphones with omnidirectional capsules for recording audio. To record high-quality audio data and avoid occlusion of the faces, we placed the microphones in front of but slightly above participants (see the blue indicators in Figure 7.2). To synchronise the audio and video streams, we clapped our hands before and after every recording session.

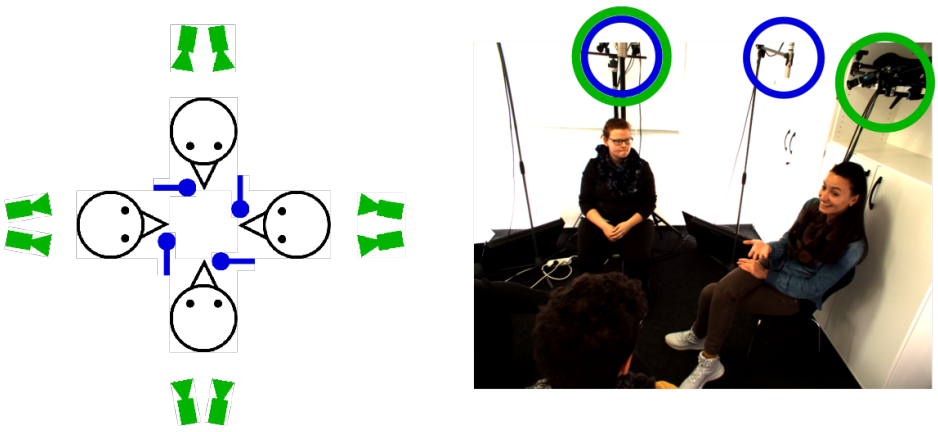


Figure 7.2: Illustration of camera and microphone positions during a recording session with four participants. Cameras are shown in green, and microphones in blue. Please note all the equipment was placed slightly above the participants to avoid occlusion for video recording.

7.3.2 Recording Procedure

We recruited 78 German-speaking participants (43 female, aged between 18 and 38 years) from a German university campus, resulting in 12 group interactions with four participants, and 10 interactions with three participants. During the group forming process, we ensured that participants in the same group did not know each other prior to the study. To prevent learning effects, every participant took part in only one interaction.

Preceding each group interaction, we told the participants that first personal encounters could result in various artifacts that we were not interested in. As a result, we would first do a pilot discussion for them to get to know each other, followed by the actual recording. We intentionally misled the participant to believe that the recording system would be turned on only *after* the pilot discussion, so that they would behave naturally. In fact, however, the recording system was running from the beginning and there was no follow-up recording. To increase engagement, we prepared a list of potential discussion topics and asked each group to choose the topic that was most controversial among group members. Afterwards, the experimenter left the room and came back about 20 minutes later to end the discussion. Participants were then asked to complete several questionnaires about the other groups members as described below. Finally, participants were debriefed, in particular about the deceit, and gave free and informed consent to their data being used and published for research purposes.

7.3.3 Data Annotation Using Questionnaires

Although in this work we were only interested in detection of low rapport, with a view to potential other future uses of our dataset, participants were asked to complete three questionnaires about different social aspects relevant for small group interactions. All

	Means	Standard Deviations
Rapport	5.41	0.46
Leadership	3.71	0.94
Dominance	4.14	0.96
Competence	5.22	0.87
Liking	5.81	0.56

Table 7.1: Means and standard deviations of the aggregated annotations obtained from seven-point Likert scales.

questionnaires were given in German to increase comprehension of the questions and, in turn, obtain more reliable scores.

- *Rapport:* Since rapport is a subjective feeling that is hard to gauge through any existing equipment, we followed previous practice using an 18-item-questionnaire (Bernieri *et al.*, 1996) to measure rapport from self reports. Responses were recorded on seven point Likert scales. Each participant rated each item for other individuals in the group, yielding two rapport scores for each dyad inside the larger group.

- *Leadership, Dominance, Competence, and Liking:* We were also interested in the correlation between rapport and other well-studied aspects in small group interactions. We thus asked participants to complete the questionnaire used in (Sanchez-Cortes *et al.*, 2012) that consists of 12 questions about four different sub-scales (leadership, dominance, competence, and liking) which we recorded using seven-point Likert scales.

- *Personality:* Finally, each participant also completed the well-established NEO-FFI questionnaire to assess personality traits, including openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism (Costa and MacCrae, 1992).

Given that we were mainly interested in the overall degree to which a participant is able to build rapport with others, we aggregated the rapport scores for a target participant by averaging those given to him by the other participants in the group. Consequently, a low rapport score indicates that a particular participant did not evoke the feeling of rapport in general for the other participants. We processed the other annotations in the same way (leadership, dominance, competence and liking).

7.3.4 Dataset Statistics

Table 7.1 summarises the means and standard deviations of the questionnaire responses over all participants. Especially for liking, competence and rapport we can observe a

AU	1	2	4	5	6	7	9	10	12
μ	0.18	0.25	0.36	0.53	0.34	0.43	0.07	0.70	0.44
σ	0.08	0.10	0.26	0.25	0.26	0.27	0.06	0.24	0.25

AU	14	15	17	20	23	25	26	45
μ	0.59	0.27	0.39	0.20	0.47	0.20	0.15	0.21
σ	0.26	0.10	0.12	0.10	0.22	0.08	0.07	0.07

Table 7.2: Statistics for average AU activations of all extracted AUs when the participant is not speaking.

tendency towards higher ratings. A more fine-grained depiction of the distribution of rapport scores is shown in Figure 7.3, which shows a tendency towards a left-skewed distribution with a peak at 5.6 and most scores between 5.0 and 6.0. The bias towards higher values in questionnaires which involve a potentially more hurtful evaluation of others (liking, competence and rapport in contrast to leadership and dominance) might be due to a general social desirability bias (Lavrakas, 2008). Given that we were particularly interested in low rapport, we grouped the data with the lower 25% percentile of rapport scores as "low rapport" and the rest as "high rapport". This results in 11 interactions without a low-rapport participant (seven of them are three-participant interactions), four interactions with a single low-rapport participant (two three-participant interactions), six interactions with two low-rapport participants (one three-participant interaction), and one interaction with three low-rapport participants (four-participant interaction).

The diversity of the dataset in terms of participants' behaviour can be illustrated, for example, by the portion of time they spoke and smiled. Figure 7.4 shows the histogram of the portion of speaking time (blue bars). While most participants spoke around 10% to 40% of the time per discussion, several participants spoke less than 10% or more than 50% of the time. Moreover, the amount of smiling is highly diverse across participants (see Figure 7.4, transparent green bars). While some participants hardly smiled at all, others smiled almost constantly. Table 7.2 shows the average activation AUs across participants. Inspecting the standard deviations, we can see that there is substantial variability in participants' average level of AU activations.

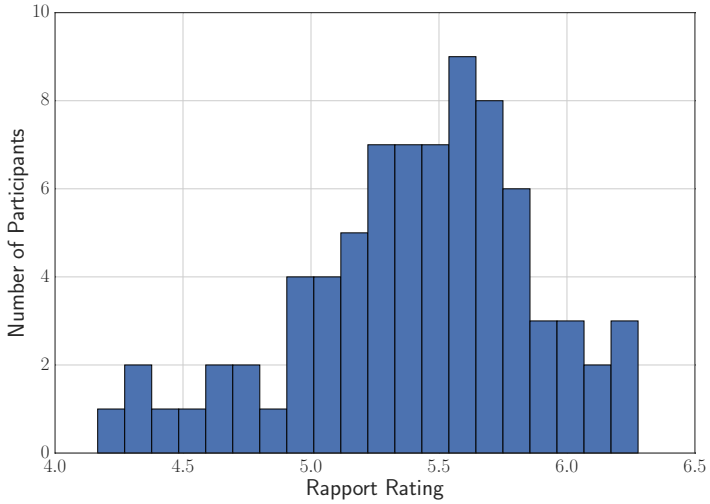


Figure 7.3: Histogram of the number of participants (y-axis) against the average received rapport ratings from other participants in an interaction (x-axis).

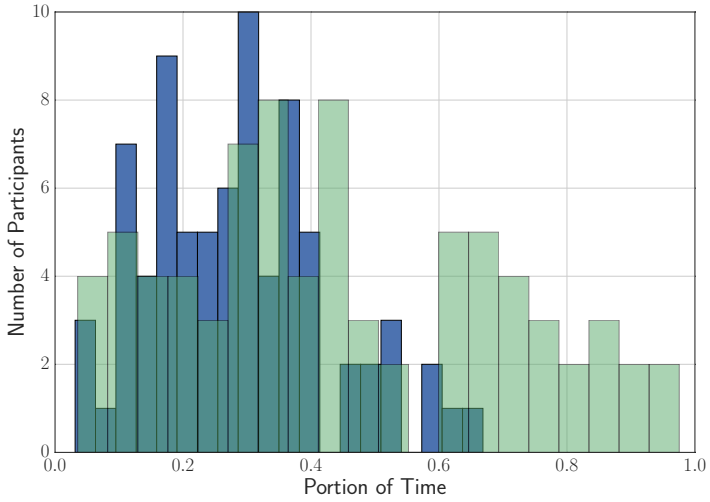


Figure 7.4: Histogram of the number of participants (y-axis) against the portion of time that participants are speaking (blue bars) and smiling (transparent green bars; detected by AU12) during the interactions.

7.4 Multimodal Method with Non-Verbal Features

Our multimodal approach to detecting low rapport relies on non-verbal features only, rather than word-related features. Specifically, it considers facial expression, hand motion, speech activity, and prosodic features. In addition, we also exploit synchronisation features and cross-modal features. The following subsections discuss each of these feature sets.

7.4.1 Non-Verbal Features

7.4.1.1 *Speech Activity Features*

Turn-taking is an important attribute in conversations, and there may be a potential link between the turn-taking behaviour in group discussion and felt rapport, for example via reflecting aspects of the coordination component of rapport (Tickle-Degnen and Rosenthal, 1990). To extract speech activity features, we annotate speaking turns from all recordings. Based on this information, we compute several features that encode the duration and frequency with which participants speak, and also different characteristics of turn-taking (see Table 9.1).

7.4.1.2 *Prosodic Speech Features*

Apart from the speech activity features, we extract a set of prosodic speech features using openSmile (Eyben *et al.*, 2013). We choose the feature set used for the IS09 emotion challenge, as it is a rather small feature set (384 features) and we assume effective features for emotion recognition might also be helpful for rapport detection. The features are extracted from individual segments when the participant speaks, and then aggregated over all segments of a speaker by taking the mean and the standard deviation, resulting in 768 features.

7.4.1.3 *Facial Features*

Facial expressions convey informative visual cues of emotions, and they are an important non-verbal channel to express one's feelings and views. Therefore, we include facial expression as one of our main features for rapport detection.

Our facial features include head orientations as well as the activation/intensity of facial action units (AUs), and additionally some higher-level facial features built on top of these basic concepts. For example, we incorporate features encoding aspects of all three components of rapport suggested by Tickle-Degnen and Rosenthal (Tickle-Degnen and Rosenthal, 1990). They include 1) the amount of positivity, 2) interpersonal synchronisation/coordination, and 3) mutual attention reflected by head orientations. An overview of the facial features is given in Table 9.1.

In practice, we used OpenFace (Baltrušaitis *et al.*, 2016). It is an automatic tool for facial expression analysis that identifies facial landmarks, head pose, and the activation/intensity level of the 17 facial AUs displayed in Table 7.2 from a video. As there

Modalities	Notation	Feature Description
Speech Activity	$TimeSpeak$	The portion of time the target participant speaks
	$TimeTurn$	The average length of speaking turns
	$RateTurn$	The number of speaking turns per minute
	$ProbTurn\|TurnTrans$	Probability of taking the turn at turn transition
Prosody	$PRSx$	Set of 768 prosodic features based on IS09 challenge feature set from (Eyben *et al.*, 2013)
Face	$PosiFace^{\mu/\sigma}$	Mean and stddev of facial positivity indicator
	$PosiFace^{\mu/\sigma}_{200s}$	Mean and stddev of facial positivity indicator during the beginning 200 seconds
	$PosiFace^{sync}$	Amount of synchronisation of facial positivity indicator with other participants
	$Facing$	How much other participants are facing the target participant
	$MutualFacing$	Amount of mutual facing with other participants
	AUx	Mean intensity of AUx
	AUx_{200s}	Mean intensity of AUx during the beginning 200 seconds
	AUx^{sync}	Amount of synchronisation of intensity of AUx with other participants
	AU^{sync}	Average amount of synchronisation of all AU intensities
	$ProbAUx$	Probability of AUx being active
	$ProbAUx_{200s}$	Probability of AUx being active during the beginning 200 seconds
	$ProbAUx^{sync}$	Amount of synchronisation of AUx activation with other participants
	$ProbAU^{sync}$	Average amount of synchronisation of all AU activations
Face and Speech Activity	$AUx_{target\|targetSpeak}$	Mean intensity of AUx of target participant when he/she is speaking
	$AUx_{target\|targetNotSpeak}$	Mean intensity of AUx of target participant when he/she is not speaking
	$AUx_{other\|targetSpeak}$	Average mean intensity of AUx of other participants when target participant is speaking
	$AUx_{target\|otherSpeak}$	Average mean intensity of AUx of target participant when another participant is speaking
	$ProbAUx_{target\|targetSpeak}$	Probability of AUx of target participant being active when he/she is speaking
	$ProbAUx_{target\|targetNotSpeak}$	Probability of AUx of target participant being active when he/she is not speaking
	$ProbAUx_{other\|targetSpeak}$	Average probability of AUx being active in other participants when target participant is speaking
	$ProbAUx_{target\|otherSpeak}$	Average probability of AUx being active in target participants when another participant is speaking
Hand Motion	$VelHand$	Average velocity of hands
	$VelHand^{sync}$	Amount of synchronisation of hand velocity with other participants
Hand Motion and Speech Activity	$VelHand_{target\|targetSpeak}$	Average hand velocity of target participant when he/she is speaking

Table 7.3: Feature notations and descriptions of different modalities.

are four cameras that cover the face of each participant from different angles, we extract the facial information from all four videos. Based on the confidence scores given by OpenFace, we selected the best view for each frame and use the facial AUs in this view for further analysis and recognition. This procedure results in high OpenFace confidence scores (>0.8 on a scale from 0 to 1) in almost all frames (97%).

Facial positivity is computed following previous practice (Chikersal *et al.*, 2017). The facial positivity indicator *PosiFace* for the target participant is set to 1, if AU12

is active, and -1 if AU15 is active in conjunction with at least one of AU1 and AU4 (Chikersal *et al.*, 2017). *PosiFace* is set to 0, when none of the above holds, or when both the positivity (AU12) and negativity AUs (AU1, AU4, AU15) are active. To reflect the intuition that the first minutes of a discussion are special, as the participants are just getting to know each other, we include additional versions of the above features that only take into account the first 200 seconds of the interaction (e.g. AUx_{200s}). Face orientation features ($Facing$ and $MutualFacing$) are constructed by thresholding of the face orientation estimated from the frontal view of the target participant. Additionally, we extract various features to describe the synchronisation of facial expressions among participants. The general approach to computing synchronisation of features between participants is detailed below.

7.4.1.4 *Synchronisation Features*

Inspired by the findings that 1) mirroring is an important phenomenon that can reflect rapport and facilitate the building of rapport (Bernieri, 1988) and that 2) synchronisation/coordination is one of three basic components of rapport (Tickle-Degnen and Rosenthal, 1990), we build features to delineate the amount of behavioural synchronisation between participants. To measure the feature synchronisation of two participants, we compute the distance between the pair of feature signals using Dynamic Time Warping (DTW) with a Sakoe-Chiba band of five seconds (Sakoe and Chiba, 1978; Chikersal *et al.*, 2017). We then compute the amount of synchronisation of a target participant i with all other participants in the interaction, by averaging the DTW distances of the target participant to others. In other words, for a feature signal F_i, the resulting average synchronisation is $\sum_{j \in N \setminus \{i\}} \text{DTW}(F_i, F_j)$, where N is the set of all participants in the interaction where the target participant i takes part, and DTW denotes the DTW function.

7.4.1.5 *Hand Motion Features*

Body posture and its coordination among people can be indicative of rapport (Bernieri, 1988). Since in our study setup all the participants were sitting, we focuse on hand motion. We use the multi-person pose estimation method OpenPose (Cao *et al.*, 2017) to extract poses from videos. OpenPose extracts the joint locations of the human body from the 2D video data. Based on the frames in which both hands are detected (on average 77%), we compute several features, such as the total amount of hand movement for each participant as well as the synchronisation of hand movements between participants (see Table 9.1 for details).

7.4.1.6 *Cross-Modal Features*

In addition to the unimodal features described above, prior research pointed out that the coordination between different modalities, such as gaze-hand coordination, can reflect human mental states (Huang *et al.*, 2016b). Moreover, cross-modal features have been applied in the context of leadership prediction (Beyan *et al.*, 2017b).

We design a number of cross-modal features, specifically to encode participants' evaluation of each other by analysing their facial expressions while others are speaking, and also to compensate for the influence on AU detection during speaking. These features include 1) AU activations and intensities while the participant is speaking or not speaking, 2) average AU activations and intensities of all other participants while the target participant is speaking, and 3) AU activations and intensities of the target participant while other participants are speaking. Apart from AUs, we combine hand motion information with speech activity into a feature that measures the amount of hand movement while the participant is speaking. This feature is intended to encode how much a participant is gesticulating during speaking.

7.4.2 Learning Low Rapport Using an Ensemble of SVMs

We train Support Vector Machines (SVM) with radial basis function kernels to classify participants' received rapport ratings into low versus medium-to-high rapport. The cost parameter C of SVM is tuned in a nested inner validation loop. We use a leave-one-interaction-out cross-validation scheme to evaluate the performance of our models. As a performance metric, we choose average precision (AP), which is common for detection problems, as it is better suited to measure the performance of models on data with class imbalances than, for example, accuracy. The necessary ranking of test examples is obtained by using probability estimates of the SVM. To marginalise out any fluctuations due to random initializations of the SVM optimisation method, we train 1,000 SVMs, from which we extract ensemble predictions by averaging.

7.5 Experimental Results

In the experimental evaluation of our proposed approach to low rapport detection, we quantify the contribution of different feature sets and individual features, as well as the amount of information that can be exploited from different temporal segments of an interaction. Finally, we show additional results concerning the relation of rapport to other previously investigated concepts in small groups.

7.5.1 Identifying Important Feature Sets

To understand the contribution of different modalities to recognition of low rapport, we evaluate our approach with different subsets of features. Figure 7.5 shows the performance comparison. The x-axis presents different feature sets. Bars with different colour represent the performances of models using different temporal segments for feature extraction, i.e. from the whole interaction (blue), and the first (yellow), middle (red), and last (purple) third of the interaction. Since we define the 25 percentile of our data with the lowest score as low rapport, the baseline method (dashed line) that ranks the test data randomly results in 0.25 AP.

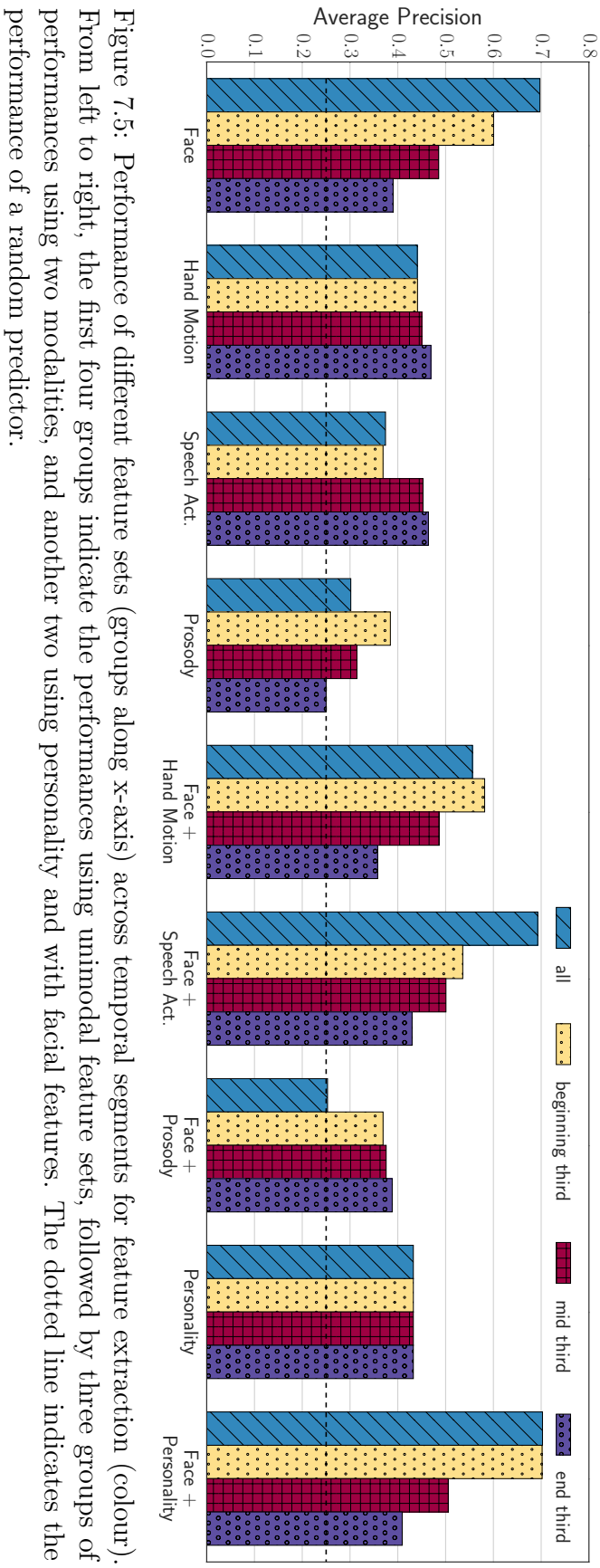


Figure 7.5: Performance of different feature sets (groups along x-axis) across temporal segments for feature extraction (colour). From left to right, the first four groups indicate the performances using unimodal feature sets, followed by three groups of performances using two modalities, and another two using personality and with facial features. The dotted line indicates the performance of a random predictor.

In this subsection, we focus on results on full interaction data only (blue bars), for which the overall highest performance is achieved by facial features (0.7 AP). They perform significantly better than the other unimodal feature sets (see the first four groups in Figure 7.5). However, hand motion, speech activity and prosodic feature sets can also outperform the baseline (0.37, 0.44, 0.30 AP, respectively), indicating that each modality carries a certain amount of useful information for low rapport detection.

Surprisingly, adding additional features to the facial features does not further improve the performance for whole-interaction data. Specifically, adding speech activity and cross-modal combinations of speech activity and facial features (Face + Speech Act.) achieves a comparable result (0.69) to Face alone (0.7). The combination of face and hand motion features (Face + Hand Motion) produces an AP of 0.56, whereas combining face and speech activity (Face + Speech Act.) or prosodic features (Face + Prosody) yields a baseline performance. All possible further combinations of feature sets fail to improve performance. This result implies that facial features play an important role for rapport detection in group interactions.

To further understand which types of facial features lead to good performance, we perform an ablation analysis (see Figure 7.6). Firstly, we split face features into four groups: 1) synchronisation features, 2) non-synchronisation features, 3) without using facial features extracted in the beginning 200s of each interaction, and 4) using only those features that were extracted in the first 200s of an interaction. Surprisingly, it turns out that facial features without synchronisation even outperform Face (comprising both sync and non-sync features), though with a marginal improvement (0.72 AP). In contrast, facial features with synchronisation only result in 0.53 AP. Thus, although facial synchronisation features carry a certain amount of information about rapport, the mirroring and behavioural coordination effects encoded in them are not indicative enough to improve over the basic facial features. Still, it could be possible that mirroring of particular member(s), or at particular points in time in the interaction (e.g. while speaking), may have a stronger indication of rapport, which needs further investigation. We also see that including the features extracted from the beginning 200s contribute to an improvement (Face vs. No 200s), though using these features alone has a low AP (0.45). This indicates that features extracted at the beginning of an interaction have a special relation to rapport, complementary to features extracted from the full interaction.

Finally, we study how well low rapport can be predicted from personality scores, a factor that can be measured without observing the actual interaction. To this end, we train a SVM on NEO-FFI scores of the target participants, which leads to an AP of 0.43. Although training on personality scores alone does not give a high rapport recognition performance, it can clearly outperform the baseline (0.25), and even speech activity (0.37) and prosodic features (0.30), and yields a comparable result with that of hand motion (0.44) from the actual interaction. This finding is interesting, as it indicates that if an intelligent user interface can gather information on personal traits, e.g. from a personal device, there is a high chance that it can make a correct prediction of rapport in a future group interaction even without access to the actual interaction signals.

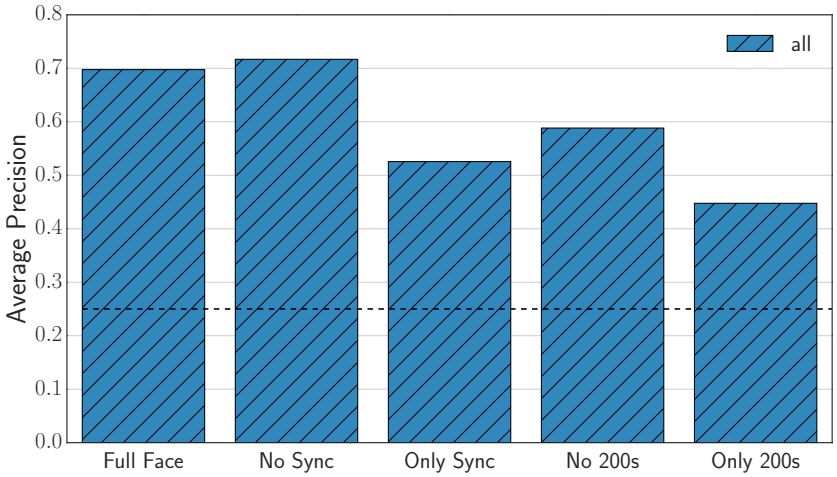


Figure 7.6: Results of ablation studies on facial feature set. From left to right: full set of facial features, without synchronisation features, only synchronisation features, without features extracted from the beginning 200s, only features extracted from the beginning 200s.

7.5.2 Prediction from Temporal Segments

In addition to understanding the contribution of different feature sets, we also evaluate the amount of information that our method is able to exploit from different temporal segments of the interactions. Specifically, we divide each interaction into three segments and train and test on each segment of the interactions only. Figure 7.5 shows the average precisions achieved in these three cases (yellow, red, and purple bars).

For our best-performing feature set on full recordings, facial features without synchronisation features, we can observe a clear trend that the amount of useful information diminishes over the time of the interaction. This indicates that rapport is encoded in facial behaviour especially at the beginning of an interaction. However, other parts of the interaction carry complementary information, as the performance of facial features for the first third (0.60 AP) is significantly lower compared to the corresponding performance for the whole interaction (0.70 AP). Moreover, it is very encouraging to see that if personality information is available, facial features extracted only from the first third of the interaction can successfully reach the best performance that can be achieved using the entire interaction. This result implies a promising application scenario, since it indicates a prior personality measurement can help to make accurate predictions with only short observations of additional behaviour. This potentially allows for effective interventions to support group interactions at an early stage.

7.5.3 Identifying Important Features

This section extends the previous evaluation to a finer granularity, by investigating the contribution of individual features on the best-performing feature set (facial features

Feature	t-score
$AU09$	2.19
$AU23$	2.04
$AU02_{before200}$	1.78
$MutualFacing$	1.77
$ProbAU25$	1.75
$AU25$	1.50
$AU14$	1.41
$PosiFace^{\sigma}$	-1.93

Table 7.4: Features from the face feature set with the highest absolute t-scores for discriminating between low and high rapport.

without synchronisation features). The presented results are obtained from whole interaction data. To identify how well individual features can discriminate between low and high rapport we compute t-scores for each feature separately. T-scores measure the linear dependency between features and the target (in our case: low vs. medium-to-high rapport). The higher the absolute value of a t-score, the more likely it is that a linear dependence exists in the population. In addition, the sign of the t-score indicates the direction of the dependency, making them straightforward to interpret. It is important to note that our trained SVMs might also use nonlinear dependencies in the data, which cannot be reflected in t-scores. A list of features with the highest absolute t-scores is given in Table 7.4.

According to these results, low rapport is especially associated with the average intensities of AU9 (nose wrinkler), AU23 (lip tightener), AU2 (outer brow raiser) during the beginning 200s, AU25 (lips part) and AU14 (dimpler), as well as the probability of AU25 being active. AU9 is often seen in disgust or anger, AU23 in sadness, and AU2 in surprise, fear, disgust or anger (Ghayoumi and Bansal, 2016). This is in line with prior work finding that low rapport is encoded in the presence of facial AUs associated with negative emotions (Wang and Gratch, 2009). AU25 on the other hand indicates speaking, meaning that a large amount of talking is indicative of low rapport. This is confirmed by the strong dependency between the amount of speaking and low rapport (t=2.9). A bit surprisingly, AU14 seems to be indicative of low rapport although this AU is often present in facial displays of happiness (Ghayoumi and Bansal, 2016). Moreover, our results show that a lot of mutual facing is indicative of low rapport. As mutual facing can be seen as a proxy for attention, this result seems to contradict the theory put forward by Tickle-Degnen and Rosenthal (Tickle-Degnen and Rosenthal, 1990) who postulated that a high degree of mutual attention is indicative of high rapport. The most likely reason for the negative connection observed in our interaction context is that mutual facing is more frequent in participants who speak a lot, resulting from the

social convention of facing the current speaker. A lot of speaking, in turn, seems to be related to low rapport. As such, this finding underlines the strong context dependency of the connection between nonverbal behaviour and rapport.

7.5.4 Understanding Correlations Among Group Attributes

As we are the first to propose the detection of low rapport in a multiparty conversation setting where all participants rate each other, it is important to investigate how this concept of rapport is related to the existing concepts that have been studied in multiparty conversations in the literature.

In particular, the attributes (leadership, dominance, competence, and liking) proposed by Sanchez-Cortes et al. (Sanchez-Cortes *et al.*, 2012) are relevant to our work and are measured via the same paradigm as ours. That is, every participant rates every other participant within the group. Moreover, especially the PLike scale suggested in their work (Sanchez-Cortes *et al.*, 2012) seems conceptually close to rapport. In contrast, it is difficult to directly compare with cohesion, as it is a group-level attribute (Chin *et al.*, 1999). To investigate the association between rapport and other group interaction attributes, we compute the aggregated score in the same way as we process rapport. Specifically, we average all ratings a participant received from other participants. We then calculate the Pearson correlation coefficient between the resulting scores.

Table 7.5 gives the correlations between different interaction attributes. It is interesting to see that rapport shows a strong correlation with competence (0.70), an obvious correlation with dominance (0.52) and liking (0.52), and a moderate correlation with leadership (0.39). The correlation analysis also reveals that although rapport is highly associated with competence, they are rather different with respect to their correlation with liking (rapport: 0.52; competence: 0.31). This implies that rapport is a complex construct associated with multiple different interaction attributes.

Finally, we computed Pearson correlation coefficients between personality scores and rapport (see Table 7.5). Although not significant in a two-tailed test, the small negative correlation of rapport with neuroticism and the small positive correlation with extraversion is in line with hypotheses on the connection between rapport and personality found in prior work (Izard, 1990). As with our previous feature analysis, it is important to keep in mind that the SVM might exploit nonlinear dependencies which are not reflected in the correlations.

In general, the correlations between rapport and different interaction attributes corroborate our hypothesis that rapport is a concept pertinent to but considerably distinct from the existing attributes proposed in previous studies (Costa and MacCrae, 1992; Sanchez-Cortes *et al.*, 2012).

	Lead	Dom	Com	Like	Rap
Lead		**0.80**	**0.41**	0.01	**0.39**
Dom			**0.50**	0.08	**0.52**
Com				**0.31**	**0.70**
Like					**0.52**
O	0.01	0.10	0.21	0.02	0.15
C	-0.09	-0.13	-0.06	-0.04	-0.13
E	0.12	0.12	-0.00	0.17	0.16
A	-0.22	-0.11	-0.07	**0.30**	0.04
N	**-0.25**	**-0.32**	-0.18	0.10	-0.21

Table 7.5: Pearson correlations coefficients between interaction attributes. The lower part of the Table shows correlations between personality scores and the rest interaction attributes. Bold coefficients indicate statistical significance at $\alpha = 0.05$, two-tailed.

7.6 Discussion

In this work we proposed a multimodal approach for detecting low rapport in small group interactions. To the best of our knowledge, we are the first to conduct such an investigation, taking into consideration individual behavioural features from separate modalities (e.g. facial expression and speech activity), cross-modal features (e.g. hand motion while speaking), as well as high-level interaction signals (e.g. behavioural mirroring). Evaluations on a novel 78-participant dataset, the first of its kind, showed that facial expressions are, in general, the most powerful signal for low rapport detection. We further demonstrated that incorporating participants' personality into our pipeline could improve performance for early prediction. This is encouraging, as recent years have seen an increase of methods to automatically predict personality traits of an individual user (Vinciarelli and Mohammadi, 2014), e.g. using mobile phones (de Montjoye *et al.*, 2013) or eye movement analysis (Hoppe *et al.*, 2015). These methods could help improve early rapport prediction without requiring additional explicit user input in the from of personality questionnaires.

The possibility to predict low rapport early and accurately enables next-generation ambient intelligent systems with the ability to support users if they fail to establish rapport with each other. Such systems could, for example, use ambient displays to encourage or amplify behaviour known to improve rapport (Balaam *et al.*, 2011). Advice for the whole group could involve proposing different interaction strategies or even socialising games to increase rapport, or encourage other people to take over or lead the discussion. Individual advice could be provided on personal screens or head-mounted displays (Damian *et al.*, 2015; Schiavo *et al.*, 2014). Beyond the small group setting, we

believe automatic detection of low rapport also has potential for applications in autism spectrum disorders, e.g. by supporting people with this disorder in properly interpreting rapport in interactions with others or even helping them to notice low rapport at all. To execute effective support strategies in these settings, it will be particularly important to detect low rapport at an early stage of the interaction. Encouragingly, our approach is able to achieve this goal when incorporating prior knowledge of personality. In addition, our results showed that facial features alone can achieve high performance given information on the entire interaction. As cameras and microphones become pervasive in personal devices, low rapport detection could become a key component in many intelligent user interfaces that aim to positively influence daily social interactions, reduce stress, avoid conflicts, and thus lead to harmonious computer-mediated interactions.

Our results also suggest that a prediction performance above chance can still be reached if certain modalities are unavailable. This implies the ability of low rapport detection to adapt to diverse interaction settings. In practice, our method therefore can suggest an alternative modality combination in case the best modality is temporarily inaccessible. Even when there is no data from the actual interaction at all, an educated guess can be made based on the prior knowledge of personality scores in order to support those who are most likely to fail in establishing rapport with others. Given all this, our method has significant potential to pave the way for rapport-aware computer-meditated communication.

Despite these promising results, there are some limitations that we plan to address in future work. Our results showed that facial expressions are the most indicative modality. However, analysing multimodal signals using a more sophisticated model, such as a neural network, might allow use of information from multiple modalities more efficiently and achieve an even higher recognition accuracy. Furthermore, the present study was conducted in a controlled laboratory environment. While this is in line with prior works on rapport during dyadic interactions and beneficial for experimental control, it will be interesting to investigate how our findings can generalise to in-the-wild situations, e.g. interactions at home, and combining the analysis of rapport with other personal or social signals that can be captured using mobile devices or on-body sensors.

7.7 Conclusion

This work proposed the first audio-visual multimodal approach to low rapport detection in small group interactions. We evaluated our method on a novel 78-participant dataset consisting of 22 three- and four- person discussions. We studied a diverse set of non-verbal behaviours, including facial expressions and orientations, hand motion, speech activities, and prosodic features as well as higher-level interaction signals, e.g. reflecting mirroring effects. Our results showed that facial features in general are most indicative to detect failure in establishing rapport in group interactions. Moreover, adding personality traits allows us to predict low rapport early on in the interaction. As such, our study advances the understanding of non-verbal behaviour and rapport establishment, pointing the way towards new intelligent user interfaces that incorporate low rapport detection to prevent disharmony in social interactions on the fly.

8 Emergent Leadership Detection Across Datasets

Automatic detection of emergent leaders in small groups from nonverbal behaviour is a growing research topic in social signal processing but existing methods were evaluated on single datasets – an unrealistic assumption for real-world applications in which systems are required to also work in settings unseen at training time. It therefore remains unclear whether current methods for emergent leadership detection generalise to similar but new settings and to which extent. To overcome this limitation, we are the first to study a cross-dataset evaluation setting for the emergent leadership detection task. We provide evaluations for within- and cross-dataset prediction using two current datasets (PAVIS and MPIIGroupInteraction), as well as an investigation on the robustness of commonly used feature channels and online prediction in the cross-dataset setting. Our evaluations show that using pose and eye contact based features, cross-dataset prediction is possible with an accuracy of 0.68, as such providing another important piece of the puzzle towards real-world emergent leadership detection.

8.1 Introduction

Emergent leaders are group members who naturally obtain a leadership position through interaction with the group, and not via a higher authority (Stein and Heller, 1979). Even without formal authority, emergent leaders are important for group performance (Druskat and Pescosolido, 2006; Kickul and Neuman, 2000), and as a result automatic identification of emergent leaders in group interactions is potentially beneficial in organisational research, in the context of assessment centres (Goodstein and Lanyon, 1999), or for robots and intelligent agents that are supposed to interact with a group naturally. Consequently, the detection of emergent leaders is a growing topic in social signal processing (Feese *et al.*, 2011; Sanchez-Cortes *et al.*, 2012; Beyan *et al.*, 2016b). These studies used nonverbal behaviour to detect emergent leaders in group interactions, which is supported by a large body of work connecting emergent leadership and nonverbal behaviour (Baird Jr, 1977; Gerpott *et al.*, 2018; Kalma, 1992).

While existent methods on emergent leadership detection in small groups showed reasonable performance, they all make the assumption that training and testing data come from the same distribution. This assumption is unrealistic for application scenarios in which a system is required to detect emergent leaders in slightly different social

Figure 8.1: Illustration of the recording setup of the MPIIGroupInteraction dataset (Müller *et al.*, 2018a). The selected view and corresponding visible participants are shown in orange.

situations for which no labelled data is available. Until now, it remains unclear whether such cross-dataset leadership detection is possible with sufficient accuracy.

Specifically, emergent leadership detection in small groups of unaugmented people has only been investigated separately on two datasets employing very similar tasks, effectively ignoring the crucial cross-dataset setting. The ELEA dataset (Sanchez-Cortes *et al.*, 2012) consists of meetings of three or four people each, in which participants are instructed to come up with a joint solution for the winter survival task. Work on ELEA investigated emergent leadership detection from recordings of the meetings, by using audio- and visual or multi-modal features (Sanchez-Cortes *et al.*, 2012, 2013), and more recently by using features obtained from a co-occurrence mining procedure (Okada *et al.*, 2019). Kindiroglu et al. investigated domain adaptation and multi-task learning for leadership- and extraversion prediction on ELEA using video blogs with personality annotations (Kindiroglu *et al.*, 2017). Their work is different to the cross-dataset setting described above, as they assumed access to leadership ground truth on ELEA.

The PAVIS dataset (Beyan *et al.*, 2016b) consists of groups of four people each either performing a winter- or a desert survival task. Research on the dataset focussed on detecting emergent leaders from nonverbal features only (Beyan *et al.*, 2016b), using multiple kernel learning (Beyan *et al.*, 2016a), or using body pose based features (Beyan *et al.*, 2017c). Further studies improved emergent leadership detection on PAVIS by using deep visual activity features (Beyan *et al.*, 2018), or by employing sequential analysis (Beyan *et al.*, 2019a). In addition, the dataset has been used to predict the leadership style of emergent leaders (Beyan *et al.*, 2017b, 2018).

Recently, the MPIIGroupInteraction dataset was recorded to study low rapport detection in small groups (Müller *et al.*, 2018a). Although emergent leadership was rated, no corresponding detection approach was proposed. This dataset is particularly interesting for emergent leadership detection, as opposed to the rather constrained tasks on ELEA and PAVIS, participants engaged in open-ended discussions.

In this paper, we move one step closer to an emergent leadership detection system that can be applied in novel social situations without additional labelling effort. We investigate emergent leadership detection across situations using two recent datasets (Beyan *et al.*, 2016b; Müller *et al.*, 2018a) both featuring small group interactions but differing in participants' tasks, language, and nationality. Our specific contributions are twofold: We are the first to study emergent leadership detection in a cross-dataset setting, thereby achieving state-of-the-art results on MPIIGroupInteraction (Müller *et al.*, 2018a). Furthermore, we conduct extensive evaluations providing insights into the usefulness of different features and the feasibility of an online prediction system.

8.2 Datasets

To study cross-dataset emergent leadership detection, we utilise the PAVIS (Beyan *et al.*, 2016b) and the MPIIGroupInteraction (Müller *et al.*, 2018a) datasets of small group interactions. We could not include ELEA because we found inconsistencies in the mapping between ground truth and videos that could not be resolved with the authors before submission.

8.2.1 PAVIS

The PAVIS dataset (Beyan *et al.*, 2016b) consists of 16 interactions of four Italian speaking unacquainted participants each. Each group performed either a winter- or a desert survival task, in which participants had to agree on a ranking of the usefulness of items in a survival situation. Each participant was recorded by a frontal-facing camera and a lapel microphone. Interactions lasted from 12 to 30 minutes, resulting in a total corpus length of 393 minutes. All recordings were divided into segments of four to six minutes and subsequently annotated for emergent leadership. In line with previous work (Beyan *et al.*, 2018), we exclude four recordings due to audio problems, resulting in 12 meetings and 48 participants. We use PAVIS as a source dataset, as the segment-based annotation yields more training data than is available on MPIIGroupInteraction (Müller *et al.*, 2018a).

8.2.2 MPIIGroupInteraction

MPIIGroupInteraction consists of 22 group interactions in German, each consisting of three- to four unacquainted participants. In contrast to the rather constrained winter- or desert survival task on the PAVIS dataset (Beyan *et al.*, 2016b), participants had an open-ended discussion. The meetings were recorded by eight frame-synchronised cameras, two of them placed behind every participants in order to cover all other participants in their field of view (see Figure 8.1). To record audio, one microphone was placed in front and slightly above participants' heads. Each group was discussing for roughly 20 minutes, resulting in more than 440 minutes of audio-visual recordings in total. After the interaction, each participant rated every other participant on a leadership scale

("PLead" as in (Sanchez-Cortes *et al.*, 2012)). We use the aggregate ratings for each participant to identify the ground truth emergent leader.

8.3 Method

To detect emergent leaders, we use Support Vector Machines and nonverbal features from gaze, body pose, face and speaking activity. We give a concise description of the method here and refer to the supplementary material for further details.

8.3.1 Nonverbal Feature Extraction

8.3.1.1 *VFOA Features*

To compute features based on the visual focus of attention (VFOA), we first perform eye contact detection, i.e. detecting at which other persons' face a target person is looking at a given moment in time. To this end, we employ the recently introduced method by Müller et al. (Müller *et al.*, 2018b), which performs unsupervised eye contact detection in small group interactions by exploiting natural conversational gaze behaviour in a weak labelling step. Based on these eye contact detections, we extract 15 VFOA features as described in (Beyan *et al.*, 2016b). While the features we compute on top of eye contact detections are the same as in (Beyan *et al.*, 2016b), in the work of Beyan et al. they are based on VFOA detections using head pose.

8.3.1.2 *Body Pose Features*

We estimate body poses of participants using OpenPose (Cao *et al.*, 2018) and follow the approach taken in (Beyan *et al.*, 2017c) for pose feature computation. This approach yields a 80-dimensional featureset consisting of statistical measures based on the angles between detected body joints.

8.3.1.3 *Facial Features*

We use OpenFace (Baltrusaitis *et al.*, 2018; Baltrušaitis *et al.*, 2015) to extract facial action units (AUs) and subsequently follow the approach described in (Müller *et al.*, 2018a) for low rapport detection. We specifically extract the means of AU activations and intensities and the mean and standard deviation of a "facial positivity indicator".

8.3.1.4 *Speaking Activity Features*

To evaluate the importance of speaking activity, we implement features used in previous work (Sanchez-Cortes *et al.*, 2013), which encode the total speaking time of a participant, the number of speaking turns of a participant, the total number of times a participant interrupts other participants, and the average duration of a participants' speaking turns.

8.3.2 Classification

In line with previous work (Beyan *et al.*, 2017c; Müller *et al.*, 2018a), we use Support Vector Machines (SVMs) with radial basis function (RBF) kernels. To obtain a single predicted leader for each interaction during test time, we obtain probability estimates using Platt scaling (Platt, 1999) and select the participant with the highest probability as the predicted emergent leader. We choose the regularisation parameter C of the SVM via cross-validation on the source dataset (PAVIS), and set the parameter γ of the rbf kernel to the default value $1/n_{feats}$.

While normalising the training data by subtracting the mean and dividing by the standard deviation computed on the whole source dataset, we normalise each test interaction in the target dataset separately. In preliminary experiments, this way of normalising data has proven to be crucial. We refer to the supplementary material for a detailed discussion.

When employing several featuresets for classification, we always use late fusion, i.e. averaging scores of classifiers applied independently on the respective featuresets. This proved to produce more reliable results than early fusion.

8.4 Experimental Results

All our evaluations are based on per-interaction accuracy of emergent leadership predictions as in (Sanchez-Cortes *et al.*, 2012, 2013). Specifically, an interaction is counted as correct, if and only if predicted and ground truth emergent leader coincide.

8.4.1 Offline Prediction

To evaluate the extent to which classifiers trained on a source dataset are able to achieve high performance on a target dataset, we train on PAVIS and test on MPIIGroupInteraction. At test time we assume to have access to a full test recording, i.e. we are predicting emergent leadership after an interaction took place ("offline" setting). In order to ensure using the same length for each of the approximately 20 minute long interactions on MPIIGroupInteraction we always use the first 19 minutes for feature extraction.

Figure 8.2 shows the obtained results for different feature sets and source- and target dataset combinations. The highest performance in the cross-dataset setting ("Source: PAVIS, Target: MPI") is achieved by a combination of VFOA and pose features with an accuracy of 0.68, slightly outperforming VFOA features only at 0.64 accuracy. Combining other featuresets (e.g. face) with VFOA and pose did not improve results, therefore we do not show these combinations in Figure 8.2. In case video recordings are not available or desired, an accuracy of 0.5 can be achieved with speaking activity features only. Both results are clearly above the random baseline of 0.29, showing the feasibility of cross-dataset prediction.

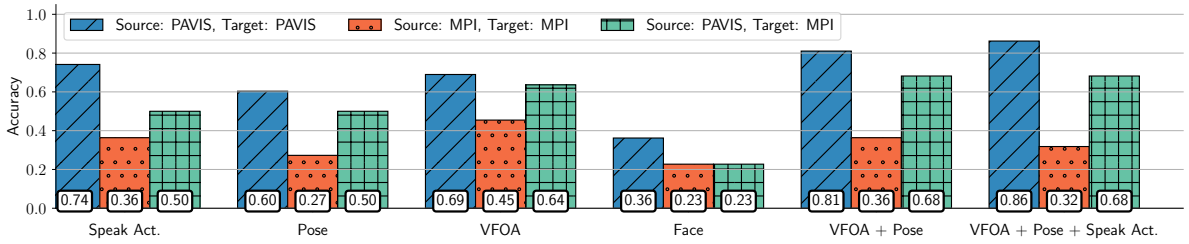


Figure 8.2: Performance of different featuresets when either training and testing on the same dataset, or training on PAVIS and testing on MPIIGroupInteraction. Random baseline for PAVIS as target is 0.25, for MPIIGroupInteraction as target 0.29.

Comparing cross-dataset to within-dataset results reveals that cross-dataset accuracies are consistently lower than within-dataset accuracies on PAVIS. More surprisingly, by training on PAVIS, we achieve higher accuracies on MPIIGroupInteraction compared to training on MPIIGroupInteraction directly. This is most likely an effect of the limited training data available on MPIIGroupInteraction. In total there are only 78 samples (one per participant), compared to 232 samples on PAVIS due to the segment based annotations.

Within datasets, we achieve the best accuracy for PAVIS with a combination of speaking activity, VFOA and pose features (0.86). The best result for the emergent leadership detection task on PAVIS, published in (Beyan *et al.*, 2017c), achieved detection scores of 0.76 for the positive and 0.93 for the negative class with a combination of pose and VFOA features. Later work by the same authors adopted a different evaluation setting, and thus can not serve as a comparison (Beyan *et al.*, 2018, 2019a). The detection scores for our predictions on PAVIS based on VFOA, pose and speaking activity features reach 0.86 for the positive and 0.95 for the negative class, exceeding the previously published results. Likely as a result of fewer training examples, within-dataset results on MPIIGroupInteraction are much lower, with a maximum accuracy of 0.45 for VFOA features.

8.4.2 Online Prediction

Some applications scenarios require information about emergent leaders already during the course of an interaction. To evaluate in this setting, we restrict the time interval from which to extract features from the test interactions. Figure 8.3 shows accuracies for classifiers that only observe data from a limited number of minutes at the beginning of the interaction. Both our best performing featureset (VFOA and pose) and speaking activity features tend to achieve higher accuracies after longer observation time. This tendency is more pronounced for the VFOA and pose featureset, which stays between 0.4 and 0.6 accuracy during the first minutes of an interaction, and clearly above 0.6 accuracy after more than 15 minutes. Thus, while prediction above chance is possible early on, longer observation is required for optimal precision.

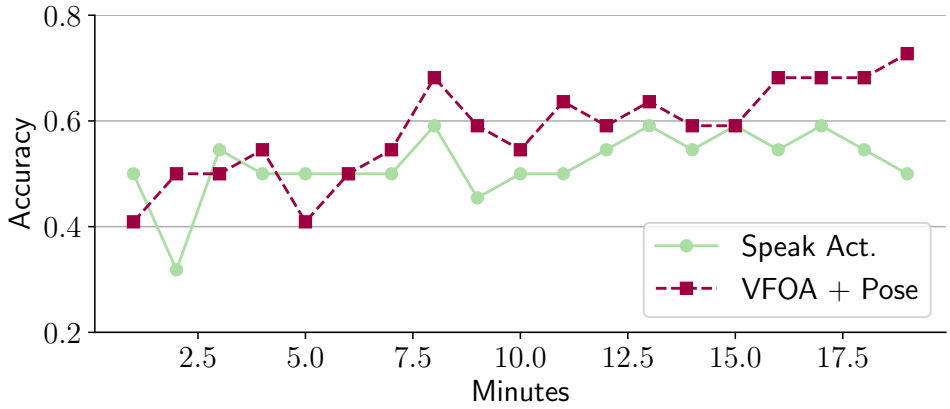


Figure 8.3: Performance of different featuresets when training on PAVIS and testing on MPIIGroupInteraction, depending on the size of the time window that is used for analysis (starting from the beginning). Random baseline is at 0.29.

8.4.3 Feature Analysis

VFOA features were the best performing individual featureset in our evaluation. To better understand which VFOA features generalise best across datasets, we quantify how well each individual feature discriminates the ground truth classes on MPIIGroupInteraction and PAVIS. For each feature, we define an unlearned classifier that simply selects the person with either the maximum or the minimum value on that feature as the emergent leader of an interaction. We decide on selection via minimum or maximum based on which strategy achieves higher accuracy. We refer to features of which we take the maximum/minimum as having positive/negative orientation respectively. This is not a valid classification approach, as we do not employ cross-validation. Instead, it is a post-hoc analysis on the connection between individual features and ground truth. See Table A.1 for the features with accuracy of at least 0.5 on both datasets (informative and good transfer) along with the features showing a difference of at least 0.2 accuracy between both datasets (weak transfer). Find the full table in the supplementary material. The features with the highest accuracies on both datasets are *totWatcher* (total time a person is watched by others), *totWatcherNoME* (totWatcher given there is no mutual eye contact (ME)) and *ratioWatcherLookSOne* (ratio between totWatcher and the time a person looks at other people). This indicates that being looked at by others is a central property of leaders on both datasets. In contrast, the low performance of *totME* on MPIIGroupInteraction in comparison to the high performance on PAVIS indicates that mutual eye contact is less robustly associated with leadership across the two datasets. The accuracy of *maxTwoWatcherNoME*, *minTwoWatcherWME* and *minTwoWatcherNoME* (the max/min time a person is looked at by two others while having/not having ME) differs strongly between the datasets while always staying below 0.5.

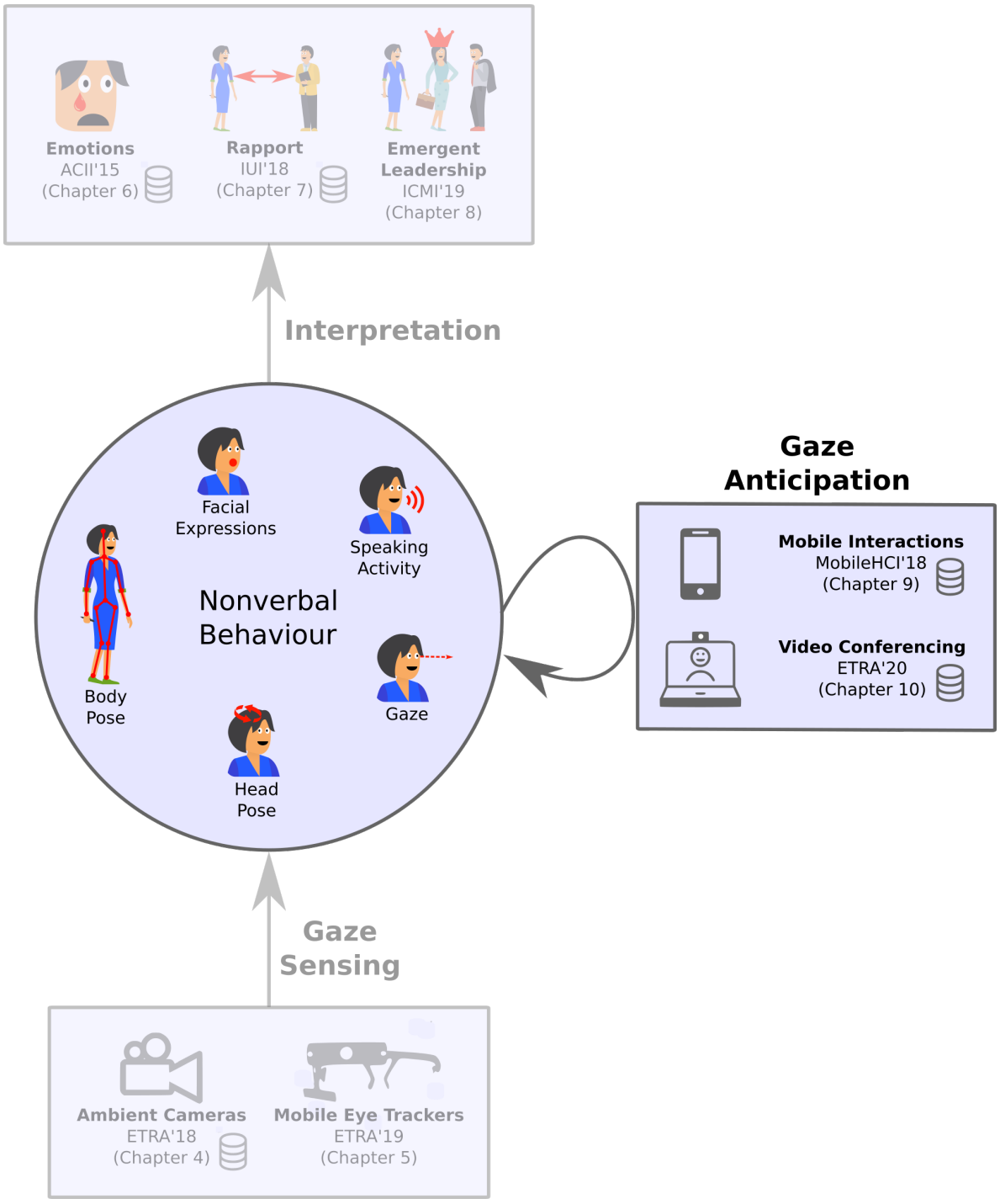
Feature	MPI		PAVIS	
	Acc.	Ori.	Acc.	Ori.
totWatcherNoME	0.59	+	0.66	+
ratioWatcherLookSOne	0.59	+	0.62	+
totWatcher	0.55	+	0.76	+
maxTwoWatcherNoME	0.45	+	0.21	+
minTwoWatcherWME	0.45	−	0.14	+
minTwoWatcherNoME	0.41	−	0.14	−
totME	0.36	+	0.60	+

Table 8.1: Accuracies for single feature based classification using selected VFOA features on PAVIS and MPIIGroupInteraction. "Ori." indicates whether the maximum or the minimum of the feature was used for prediction.

8.5 Conclusion

In this paper, we were first to investigate a cross-dataset evaluation setting for the emergent leadership detection task. We showed that it is possible to predict emergent leadership from nonverbal features on a new dataset not observed at test time, with a combination of VFOA and pose features achieving best performance. Furthermore, we analysed the feasibility of online prediction and the usefulness of single VFOA features. All in all, our initial study on cross-dataset emergent leadership prediction opens the way to investigate this important task in more realistic settings.

Part III


Emotions
ACII'15
(Chapter 6)
Rapport
IUI'18
(Chapter 7)
Emergent
Leadership
ICMI'19
(Chapter 8)
Interpretation
Facial
Expressions
Speaking
Activity
Nonverbal
Behaviour
Body
Pose
Gaze
Head
Pose
Gaze
Anticipation
Mobile Interactions
MobileHCI'18
(Chapter 9)
Video Conferencing
ETRA'20
(Chapter 10)
Gaze
Sensing
Ambient Cameras
ETRA'18
(Chapter 4)
Mobile Eye Trackers
ETRA'19
(Chapter 5)

9

Forecasting User Attention During Everyday Mobile Interactions Using Device-Integrated and Wearable Sensors

Visual attention is highly fragmented during mobile interactions, but the erratic nature of attention shifts currently limits attentive user interfaces to adapting after the fact, i.e. after shifts have already happened. We instead study *attention forecasting* – the challenging task of predicting users' gaze behaviour (overt visual attention) in the near future. We present a novel long-term dataset of everyday mobile phone interactions, continuously recorded from 20 participants engaged in common activities on a university campus over 4.5 hours each (more than 90 hours in total). We propose a proof-of-concept method that uses device-integrated sensors and body-worn cameras to encode rich information on device usage and users' visual scene. We demonstrate that our method can forecast bidirectional attention shifts and predict whether the primary attentional focus is on the handheld mobile device. We study the impact of different feature sets on performance and discuss the significant potential but also remaining challenges of forecasting user attention during mobile interactions.

9.1 Introduction

Sustained visual attention – the ability to focus on a specific piece of information for a continuous amount of time without getting distracted – has constantly diminished over the years (Rubinstein *et al.*, 2001). This trend is particularly prevalent for mobile interactions, during which user attention was shown to be highly fragmented (Oulasvirta *et al.*, 2005). Active management of user attention has consequently emerged as a key research challenge in human-computer interaction (Bulling, 2016). However, the capabilities of current mobile attentive user interfaces are still severely limited. Prior work mainly focused on estimating the point of gaze on the device screen using the integrated front-facing camera (Holland and Komogortsev, 2012; Wood and Bulling, 2014) or on using inertial sensors or application usage logs (Choy *et al.*, 2016; Exler *et al.*, 2016) to predict user engagement (Mathur *et al.*, 2016; Urh and Pejović, 2016) or boredom (Pielot *et al.*, 2015). In contrast, allocation of user attention across the device and environment has rarely been studied, and only using simulated sensors (Miettinen and Oulasvirta, 2007). Most importantly, existing attentive user interfaces are only capable to adapt *after the fact*, i.e. after an attention shift has taken place (Kern *et al.*, 2010; Mariakakis *et al.*, 2015; Gutwin *et al.*, 2017).

We envision a new generation of mobile attentive user interfaces that pro-actively adapt to imminent shifts of user attention, i.e. *before* these shifts actually occur. Pro-active adaptation promises exciting new applications. For example, future attentive user interfaces could alert users in case of a (potentially dangerous) external event that they might miss due to predicted sustained attention to the mobile device. Further, a predicted attention shift to the mobile device could trigger unlocking the device or loading the previous screen content to reduce interaction delays. Finally, pro-active adaptations could also have significant impact in interruptibility research. A future attentive user interface could show important information if user attention is predicted to continue to stay on the device or, inversely, alert users if an attention shift to the environment is predicted such that a mobile task cannot be finished in time, such as submitting a form or replying to a chat message.

The core requirement to realise such pro-active attentive user interfaces is their ability to predict users' *future* allocation of overt visual attention during interactions with a mobile device. We call this challenging new task *attention forecasting*. To facilitate algorithm development and evaluation for attention forecasting, we collected a multi-modal dataset of 20 participants freely roaming a local university campus over several hours while interacting with a mobile phone. Three annotators annotated the full dataset post-hoc with participants' current environment, indoor or outdoor location, their mode of locomotion, and whenever their attention shifted from the handheld device to the environment or back. We then developed a computational method to forecast overt visual attention during everyday mobile interactions. Our method uses device-integrated and head-worn IMU as well as computer vision algorithms for object class detection, face detection, semantic scene segmentation, and depth reconstruction. We evaluate our method on the new dataset and demonstrate its effectiveness in predicting attention

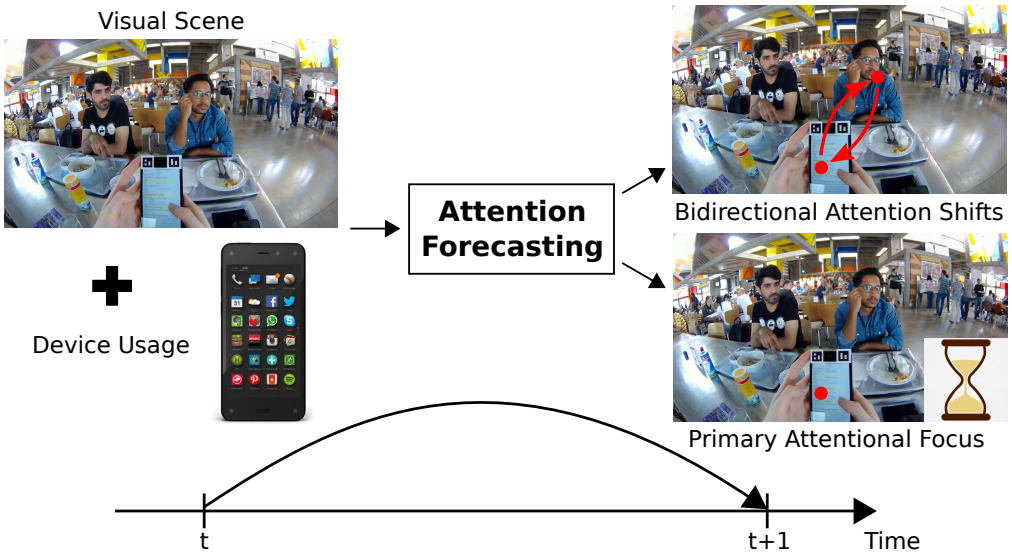


Figure 9.1: We propose a method to forecast temporal allocation of overt visual attention (gaze) during everyday interactions with a handheld mobile device. Our method uses information on users' visual scene as well as device usage to predict attention shifts between mobile device and environment and primary attentional focus on the mobile device.

shifts between the mobile device and the environment as well as whether the primary attentional focus is on the device.

The specific contributions of this chapter are three-fold. First, we propose *attention forecasting* as the challenging new task of predicting future allocation of users' overt visual attention during everyday mobile interactions. We propose a set of forecasting tasks that will facilitate pro-active adaptations to users' erratic attentive behaviour in future user interfaces. Second, we present a novel 20-participant dataset of everyday mobile phone interactions. The dataset including annotations is available at `https://www.mpii.mpg.de/MPIIMobileAttention/` (date: 12.07.2019). Third, we propose the first method to predict core characteristics of mobile attentive behaviour from device-integrated and wearable sensors. We report a detailed evaluation of our method on the new dataset, and demonstrate the feasibility of predicting attention shifts between handheld mobile device and environment and the primary attentional focus on the device.

9.2 Related Work

The work of this chapter is related to prior work on (1) user behaviour modelling and (2) gaze estimation on mobile devices as well as (3) computational modelling of egocentric attention.

9.2.1 User Behaviour Modelling on Mobile Devices

With the prevalence of sensor-rich mobile devices, modelling user behaviour, including gaze and attention, has gained significant popularity. A large body of work investigated

the use of device-integrated sensors to predict users' interruptibility (Fogarty *et al.*, 2005; Turner *et al.*, 2015; Choy *et al.*, 2016; Exler *et al.*, 2016; Turner *et al.*, 2017). In particular, Obuchi et al. detected breaks in a user's physical activities using inertial sensors on the phone to push mobile notifications during these breaks (Obuchi *et al.*, 2016). Dingler et al. used rapid serial visual presentation (RSVP) on a smartwatch in combination with eye tracking and detected when the reading flow was briefly interrupted, so that text presentation automatically paused or backtracked (Dingler *et al.*, 2016). Pielot et al. proposed a method to predict whether a participant will click on a notification and subsequently engage with the offered content (Pielot *et al.*, 2017). Others aimed to predict closely related concepts, such as user engagement (Mathur *et al.*, 2016; Urh and Pejović, 2016), boredom (Pielot *et al.*, 2015) or alertness (Abdullah *et al.*, 2016). Oulasvirta et al. investigated how different environments affected attention while users waited for a web page to load on a mobile phone (Oulasvirta *et al.*, 2005). In a follow-up work, the same authors used a Wizard-of-Oz paradigm with simulated sensors to assess the feasibility of predicting time-sharing of attention, including prediction of the number of glances, the duration of the longest glance, and the total and average durations of the glances to the mobile phone (Miettinen and Oulasvirta, 2007).

The work of this chapter is the first to propose a method to predict attentive behaviour during everyday mobile interactions from real phone-integrated and body-worn sensors. Another distinction from prior work is that our data collection constrained participants as little as possible, and specifically did not impose a scripted sequence of activities or environments.

9.2.2 Gaze Estimation on Mobile Devices

Estimating gaze on mobile devices has only recently started to receive increasing interest, driven by technical advances in gaze estimation and mobile eye tracking. In an early work, Holland and Komogortsev proposed a learning-based method for gaze estimation on an unmodified tablet computer using the integrated front-facing camera (Holland and Komogortsev, 2012). More recently, Huang et al. presented a large-scale dataset and method for gaze estimation on tablets and conducted extensive evaluations on the impact of various factors on gaze estimation performance, such as ethnic background, glasses, or posture while holding the device (Huang *et al.*, 2015). Wood and Bulling used a model-based gaze estimation approach on an off-the-shelf tablet and achieved an average gaze estimation accuracy of 6.88° at 12 frames per second (Wood and Bulling, 2014) while Vaitukaitis and Bulling combined methods from image processing, computer vision and pattern recognition to detect eye gestures using the built-in front-facing camera (Vaitukaitis and Bulling, 2012). Jiang et al. proposed a method to estimate visual attention on objects of interest in the user's environment by jointly exploiting the phone's front- and rear-facing cameras (Jiang *et al.*, 2016) while Paletta et al. investigated accurate gaze estimation on mobile phones using a computer vision method to detect the phone in an eye tracker's scene video (Paletta *et al.*, 2014). While all of these works focused on estimating gaze spatially on the device screen, we are the first to predict attention allocation temporally.

9.2.3 Computational Modelling of Egocentric Attention

While bottom-up attention modelling, i.e. solely using image features, has been extensively studied in controlled laboratory settings, egocentric settings are characterised by a mix of bottom-up and top-down influences and are therefore less well explored. Yamada et al. were among the first to predict egocentric attention using bottom-up image and egomotion information (Yamada *et al.*, 2011). Zhong et al. used a novel optical flow model to build a uniform spatio-temporal attention model for egocentric videos (Zhong *et al.*, 2016). Saliency models, which aim to predict which image regions most attract viewers' attention are an important type of computational model of visual attention (Itti and Koch, 2000). However, none of these works aimed to predict attention during mobile interactions. In addition, while we also use features extracted from egocentric video, we do not predict spatial attention distributions for the current video frame but use a short sequence of past frames (one second) to predict shifts of visual attention in the near future.

9.3 Forecasting Mobile User Attention

To be able to pro-actively adapt before users shift their attention, attentive interfaces have to predict users' future attentive behaviour. We call this new prediction task *attention forecasting*. Attention forecasting is similar in spirit to the tasks of user intention prediction as investigated, for example, in web search (Cheng *et al.*, 2010) or human-robot interaction (Ravichandar and Dani, 2017), as well as player goal or plan recognition, studied in digital games (Min *et al.*, 2016). In contrast to these lines of work, however, it specifically focuses on predicting fine-grained attentive behaviour and predictions at a moment-to-moment time scale. Attention forecasting is already highly challenging in stationary desktop interaction settings given the significant variability and strong task dependence of users' attentive behaviour. Forecasting users' attention is even more challenging during mobile interactions given the additional, as well as the large number of, potential visual attractors in the real-world environment.

In the following, we first propose a set of concrete prediction tasks within the attention forecasting paradigm and outline their potential use in future mobile attentive user interfaces. A more extensive consideration of how attention forecasting could be used in the future can be found in the discussion section. Afterwards, we propose a first proof-of-concept method that demonstrates the feasibility of predicting temporal attention allocation during everyday mobile interactions from real device-integrated and body-worn sensors.

9.3.1 Prediction Tasks

To guide future development of computational methods for attention forecasting during mobile interactions, we propose the following prediction tasks: prediction of *Attention Shifts* to the environment and to the handheld mobile device, and *Primary Attentional*

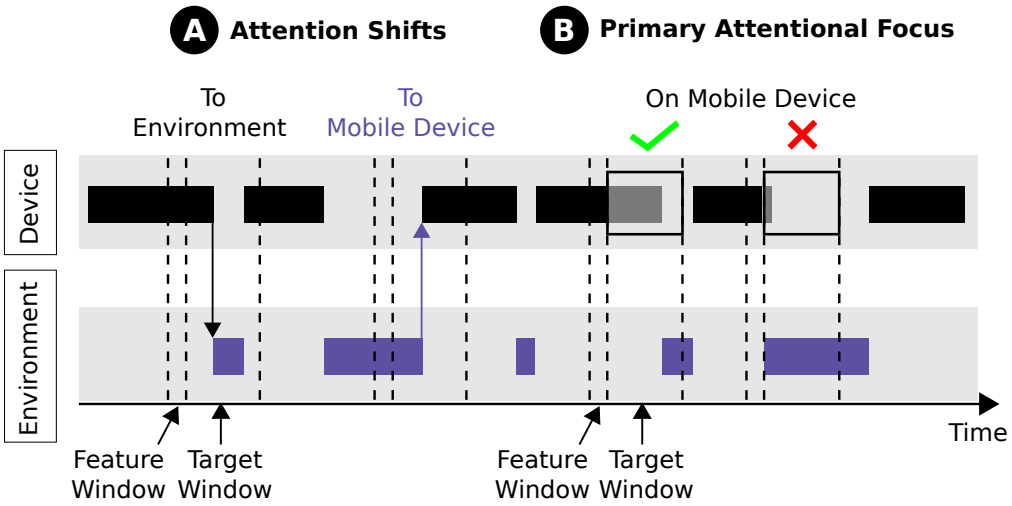


Figure 9.2: Overview of the different prediction tasks explored in this chapter: Prediction of attention shifts to the environment and (back) to the mobile device, and the primary attentional focus, i.e. whether attention is primarily on or off the device.

Focus on the device. Figure 9.2 illustrates these three prediction tasks for a sample attention allocation of a user. During the segments marked in black the user's attention is on the mobile device, while during segments marked in purple the user's attention is in the environment. In the following, we detail each of these prediction tasks.

Prediction of Attention Shifts. The first prediction task deals with attention shifts from the mobile device to the environment, and from the environment back to the device (see Figure 9.2A). Attention shifts are a key characteristic of attentive behaviour and thus an important source of information for attentive user interfaces. The task involves taking a certain time window for feature extraction, training a prediction model with this data, and using that model to predict whether an attention shift will happen during a subsequent target time window. This task assumes the user interface to already have knowledge about whether a user's attention is currently on the handheld device or not. Such knowledge can be obtained, for example, by using a method for mobile gaze estimation (Wood and Bulling, 2014). Prediction of attention shifts could be used in different ways by an attentive user interface. Attention shift prediction could be used to pro-actively support users to reorient themselves on a mobile device to smoothly get back to their previous task. Similar to Obuchi et al., who used phone data, predicted attention shifts could also be used as breakpoints for push notifications (Obuchi *et al.*, 2016). These could, for example, be shown shortly before or after an attention shift is predicted to take place. Finally, attention shift prediction could be used to automatically turn the screen on again if a shift to the handheld device is predicted to occur in the near future.

Prediction of the Primary Attentional Focus. The last task focuses on predicting whether users' attention will be primarily on the mobile device or off the device for a particular time window in the future (see Figure 9.2B). Knowledge of the primary

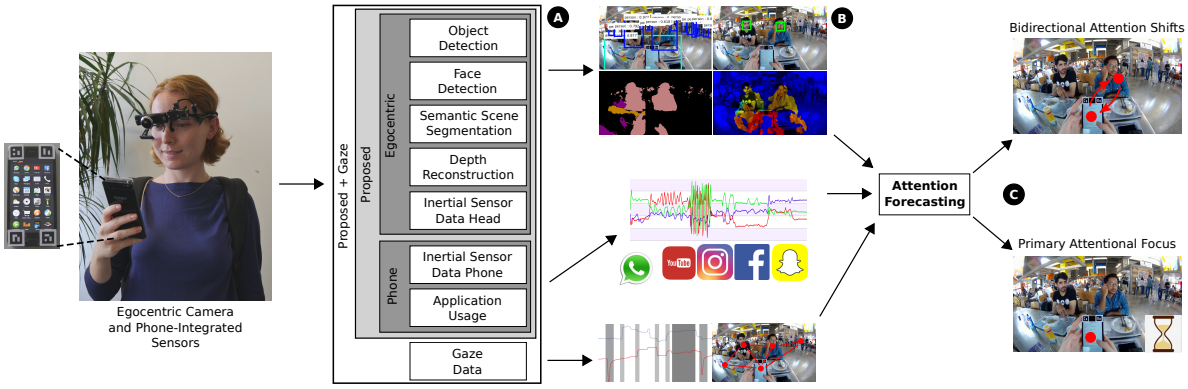


Figure 9.3: Overview of our method for attention forecasting during mobile interactions. Taking information on users' visual scene, mobile device (phone) and head inertial data, as well as on mobile app usage as input (A), our method extracts rich semantic information about the user's visual scene using state-of-the-art computer vision methods for object and face detection, semantic scene segmentation, and depth reconstruction (B). The method then extracts and temporally aggregates phone and visual features and takes eye tracking data into account to predict bidirectional attention shifts and the primary attentional focus on the phone (C).

attentional focus for an upcoming time window can be useful for different applications. For example, it could be used to highlight messages or to manage user attention in such a way that the interface needs to change content or style of presentation to keep users' attention beyond the considered time window to finish a task.

9.3.2 Proposed Method

To explore the feasibility of these prediction tasks, and to establish a baseline performance on each of them, we developed a first method for attention forecasting. Previous work demonstrated that information available on a mobile device itself, such as inertial data, GPS location, or application usage, can be used to predict engagement or interruptibility. It is therefore conceivable that such information may also be useful to predict attention shifts to the handheld mobile device. In contrast, detecting shifts to the environment requires information on the user's current environment. This suggests combining the mobile device with wearable sensors, in particular egocentric cameras worn on the user's head. Egocentric cameras represent a rich source of visual information on the user's environment as demonstrated by the rapidly growing literature on egocentric vision (Betancourt *et al.*, 2015). Combined with the fact that an ever-increasing number of egocentric cameras are used in daily life (e.g. sports cameras, cameras readily integrated in HMDs, lifelogging cameras, etc.), this makes them a not only promising but also practical sensing modality for attention forecasting.

Figure 9.3 provides an overview of our method. Inputs to our method are egocentric, mobile device (phone), and gaze data. Our method extracts information from the egocentric scene and depth videos using computer vision algorithms for object and

		Sensor	Features
Egocentric		**RGB camera**	number of detected faces and pixel counts of object classes like person, car, and monitor from the semantic segmentation, and binary occurrence indicator, numbers of detected instances of each object class from object detection, 1-hot encoded scene classes, mean, min, max, standard deviation and entropy of saliency and objectness of the scene images
		Depth camera	mean, min, max, standard deviation and entropy of the depth map from the stereo camera
		Head IMU	mean, min, max, standard deviation, norm and slope of accelerometer and gyroscope
Phone	Proposed	**Phone**	mean, min, max, standard deviation, norm and slope of accelerometer, gyroscope and orientation sensor values; 1/0 features indicating touch events, screen on/off, and activity of each of the installed applications
	Proposed + Gaze	**Gaze**	fixation positions (x, y); objectness, saliency and depth values at gaze position

Table 9.1: Overview of the different sensors and corresponding features explored in this chapter.

face detection, semantic scene segmentation labels, scene category, and reconstructed depth data as well as head motion. In addition, our method extracts features from a mobile phone, including the history of application usage and accelerometer, gyroscope, and magnetometer measurements as well as past gaze. Our method finally uses these features in a machine learning framework for attention forecasting, specifically attention shifts between the mobile phone and the environment as well as the primary attentional focus on the phone.

9.3.3 Feature Extraction

We extract features from the head-mounted egocentric RGB and depth cameras, head IMU, mobile device (phone), and past gaze data recorded using a head-mounted eye tracker (see Table 9.1 for a complete list of features used in this chapter). These features include numerical features, such as pixel counts of semantic segmentations, entropy of objectness maps, and mean depth map values, as well as binary encodings like occurrence of a touch event or whether an application on the handheld device is active. We aggregate features over a window by computing the mean, maximum, minimum, standard deviation and slope for numerical features, and the mean and the slope for binary features. Prior works on eye-based activity recognition demonstrated that gaze behaviour is characteristic for different activities (Bulling *et al.*, 2011, 2013; Steil and Bulling, 2015). It is therefore conceivable that gaze features may help to improve the performance of our method for attention forecasting. Specifically, we calculate mean,

min, max, standard deviation, norm and slope of the gaze positions (x, y) as well as objectness, saliency and depth values at that position. For evaluation purposes, and with potential future applications in mind, we group these features into four feature groups (cf. Figure 9.3 and Table 9.1): *Egocentric* (including RGB, depth, and head inertial features), *Phone* (including only phone features), *Proposed* (all features from *Egocentric* and *Phone*), as well as *Proposed + Gaze* (including fixation characteristics).

Egocentric. This feature group covers the egocentric RGB and depth camera, as well as a head inertial sensor. The depth and inertial sensors we used just for the sake of reliable feature extraction, although they can also be estimated from the egocentric camera itself (Liu *et al.*, 2015). As described above, we extract the most information from the egocentric scene video because scene information can include triggers which lead to changes of attentive behaviour. We obtain a coarse description of the scene by applying the scene recognition method of Wang et al. (Wang *et al.*, 2015a) to the video frames. This method utilises a convolutional neural network to extract scene descriptions like "office" or "library". As objects are potential targets for capturing attention, we obtain a more fine-grained description of the scene by applying the semantic scene segmentation approach of Zheng et al. (Zheng *et al.*, 2015). Semantic scene segmentation labels each pixel in a scene image as belonging to a certain object class or to background. To this end, their method combines a deep neural network with a probabilistic graphical model, trained to obtain pixel-wise segmentations of 20 different object classes including persons, monitors and cars. By encoding the occurrence of objects and also counting the number of pixels belonging to each object class, we obtain information about which objects take up the largest portion of the camera's field of view. Another important aspect of objects in a scene is the count of their instantiations. For example, gazing upon a dining hall can lead to a large number of "person" pixels, as does standing directly in front of another person. By simply counting the number of "person" pixels, these two cases cannot be distinguished. Thus, we employ the object class detection method by Ren et al. (Ren *et al.*, 2015) to obtain an estimate of the count of instances for each object class. In addition to people detection, we hypothesised that faces can help in predicting attention shifts, as they are well known to strongly draw the attention of an observer (Sato and Kawahara, 2015) and their presence is also indicative of social situations (Haxby *et al.*, 2002), constituting a highly distracting factor in the scene. To this end, we apply a face detection approach (King, 2009) and count the number of detected faces in the scene image. Moreover, we extracted depth information to obtain physical structure of the scene and mapped the depth map to the scene video via camera calibration. With the calculation of saliency and objectness maps, we collect ancillary knowledge about the scene complexity. As head poses can serve as a useful prior for gaze estimation (Valenti *et al.*, 2011), we additionally extract inertial features from the head-mounted camera.

Phone. This feature group covers inertial data, which consists of accelerometer, gyroscope and orientation information, as well as phone usage data, which consists of single app usage information, and whether touch events took place or the screen is on

or off. For that purpose we installed additional applications on the phone which were running in the background to log the movement of the phone and the user's phone usage.

9.4 Data Collection

Given the lack of a suitable dataset for algorithm development and evaluation, we conducted our own data collection. Our goal was to record natural attentive behaviour during everyday interactions with a mobile phone. The authors of (Oulasvirta *et al.*, 2005) leveraged the – at the time – long page loading times during mobile web search to analyse shifts of attention. We followed a similar approach but adapted the recording procedure in several important ways to increase the naturalness of participants' behaviour and, in turn, the realism of the prediction task. First, as page loading times have significantly decreased over the last 10 years, we instead opted to engage participants in chat sessions during which they had to perform web search tasks as in (Oulasvirta *et al.*, 2005) and then had to wait for the next chat message.

To counter side effects due to learning and anticipation, we varied the waiting time between chat messages and search tasks. Second, we did not perform a fully scripted recording, i.e. participants were not asked to follow a fixed route or perform particular activities in certain locations in the city, they were not accompanied by an experimenter, and the recording was not limited to about one hour. Instead, we observed participants passively over several hours while they interacted with the mobile phone during their normal activities on a university campus. For our study we recruited twenty participants (six females), aged between 22 and 31 years, using university mailing lists and study board postings. Participants were students with different backgrounds and subjects. All had normal or corrected-to-normal vision.

9.4.1 Apparatus

The recording system consisted of a Pupil head-mounted eye tracker (Kassner *et al.*, 2014) with an additional stereo camera, a mobile phone, and a recording laptop carried in a backpack (see Figure 9.3 left). The eye tracker featured one eye camera with a resolution of 640×480 pixels recording a video of the right eye from close proximity with 30 frames per second, and a scene camera with a resolution of 1280×720 pixels recording at 24 frames per second. The original lens of the scene camera was replaced with a fisheye lens with a 175° field of view. The eye tracker was connected to the laptop via USB. In addition, we mounted a DUO3D MLX stereo camera to the eye tracker headset. The stereo camera recorded a depth video with a resolution of 752×480 pixels at 30 frames per second as well as head movements using its integrated accelerometer and gyroscope. Intrinsic parameters of the scene camera were calibrated beforehand using the fisheye distortion model from OpenCV. The extrinsic parameters between the scene camera and the stereo camera were also calibrated. The laptop ran the recording software and stored the timestamped egocentric, stereo, and eye videos.

Given the necessity to root the phone to record touch events and application usage, similar to (Oulasvirta *et al.*, 2005) we opted to provide a mobile phone on which all necessary data collection software was pre-installed and validated to run robustly. For participants to "feel at home" on the phone, we encouraged them to install any additional software they desired and to fully customise the phone to their needs prior to the recording. Usage logs confirmed that participants indeed used a wide variety of applications, ranging from chat software, to the browser, mobile games, and maps. To robustly detect the phone in the egocentric video and thus help with the ground truth annotation, we attached visual markers to all four corners of the phone (see Figure 9.3 left). We used WhatsApp to converse with the participants and to log accurate timestamps for these conversations (Church and De Oliveira, 2013). Participants were free to save additional numbers from important contacts, but no one transferred their whole WhatsApp account to the study phone. We used the Log Everything logging software to log phone inertial data and touch events (Weber and Mayer, 2014), and the Trust Event Logger to log the current active application as well as whether the mobile phone screen was turned on or off.

9.4.2 Procedure

After arriving in the lab, participants were first informed about the purpose of the study and asked to sign a consent form. We did not reveal which parts of the recording would be analysed later so as not to influence their behaviour. Participants could then familiarise themselves with the recording system and customise the mobile phone, e.g. install their favourite apps, log in to social media platforms, etc. Afterwards, we calibrated the eye tracker using the calibration procedure implemented in the Pupil software (Kassner *et al.*, 2014). The calibration involved participants standing still and following a physical marker that was moved in front of them to cover their whole field of view.

To obtain some data from similar places on the university campus, we asked participants to visit three places at least once (a canteen, a library, and a café) and to not stay in any self-chosen place for more than 30 minutes. Participants were further asked to stop the recording after about one and a half hours so we could change the laptop's battery pack and recalibrate the eye tracker. Otherwise, participants were free to roam the campus, meet people, eat, or work as they normally would during a day at the university. We encouraged them to log in to Facebook, check emails, play games, and use all pre-installed applications on the phone or install new ones. Participants were also encouraged to use their own laptop, desktop computer, or music player if desired.

As illustrated in Figure 9.4, 12 chat blocks (CB) were distributed randomly over the whole recording. Each block consisted of a conversation via WhatsApp during which the experimental assistant asked the participant six random questions (Q1–Q6) out of a pool of 72 questions. Some questions could be answered with a quick online search, such as "How many states are members of the European Union?" or "How long is the Golden Gate Bridge?". Similar to Oulasvirta et al. (Oulasvirta *et al.*, 2005) we also asked simple demographic questions like "What is the colour of your eyes?" or

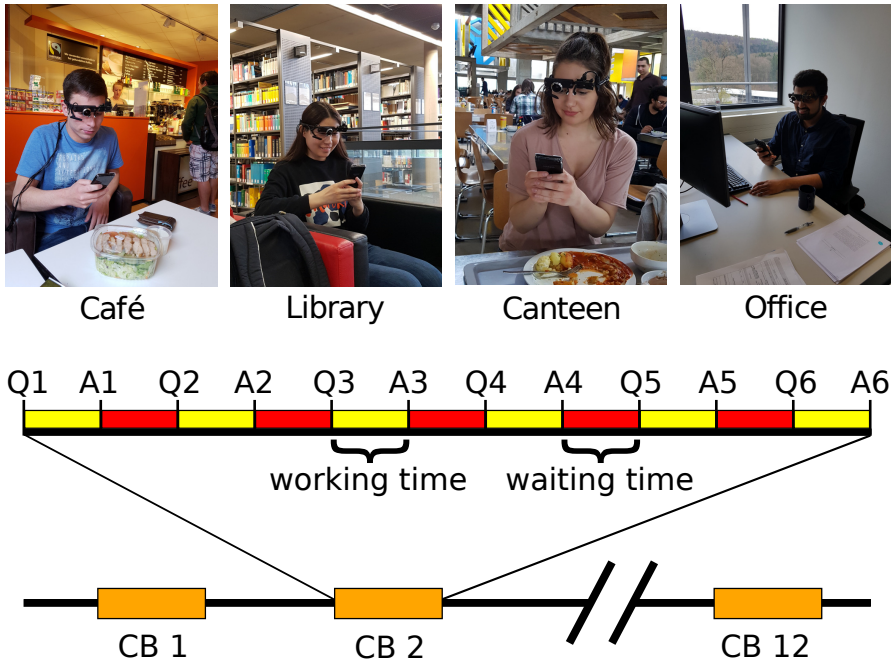


Figure 9.4: Participants were engaged in 12 chat blocks (CB) in different environments that were randomly distributed over their recording, which lasted in total about 4.5 hours. In each block, participants had to answer six questions, some of which required a short online search (Q1–Q6, working time), followed by waiting for the next question (waiting time).

"What is your profession?" that could be answered without an online search. After each answer (A1–A6), participants had to wait for the next question. This waiting time was varied randomly between 10, 15, 20, 30, and 45 seconds by the experimental assistant. This was to avoid learning effects and to create a similar situation as in (Oulasvirta *et al.*, 2005). This question-answering procedure was repeated until the sixth answer had been received, thus splitting each chat block into six working time segments (yellow) and five waiting time segments (red) (cf. Figure 9.4). At the end of the recording, participants returned to the lab and completed a questionnaire about demographics and their mobile phone usage behaviour. In total, we recorded 1,440 working and 1,200 waiting segments over all participants. Statistics about our dataset are listed in Table 9.2.

9.4.3 Data Preprocessing

Fixations were detected from the raw gaze data using a dispersion-based algorithm with a duration threshold of 150 ms and an angular threshold of 1° (Kassner *et al.*, 2014). The 3D position of the mobile phone in the scene camera was estimated using visual markers (see Figure 9.3 left). The position of the mobile phone surface was logged if at least two markers were visible in the scene camera. However, we only used the mobile phone detection as an aid for the ground truth annotation.

	mean	std	total
Working segments per question (sec)			
Working time	40.29	11.27	–:–
Time on mobile device	29.96	7.31	–:–
Waiting segments per question (sec)			
Waiting time	25.28	7.45	–:–
Time on mobile device	11.02	4.26	–:–
Attention shifts (quantity)			
Shifts to environment	248.85	107.22	4,957
Shifts to mobile device	259.90	106.88	5,178
Fixation time on/off screen (hh:mm)			
On	00:46	00:12	15:24
Off	00:13	00:05	04:36
Environments (hh:mm)			
Café	00:11	00:06	03:55
Corridor	00:12	00:12	04:08
Library	00:11	00:07	03:51
Canteen	00:08	00:06	02:50
Office	00:23	00:12	07:37
Street	00:04	00:06	01:20
Indoor/Outdoor (hh:mm)			
Indoor	01:06	00:17	22:08
Outdoor	00:06	00:08	01:56
Modes of locomotion (hh:mm)			
Sit	01:02	00:14	20:49
Stand	00:05	00:05	01:44
Walk	00:04	00:04	01:31

Table 9.2: Statistics of the ground truth annotated chat block sequences with mean, standard deviation (std) and total time.

9.4.4 Data Annotation

Classifier training requires precise annotations of when an attention shift takes place and how long an attention span lasts. Findlay and Gilchrist showed that in real-world settings, covert attention rarely deviates from the gaze location (Findlay and Gilchrist, 2003). Thus, we leveraged gaze as a reliable indicator of the user's current attentional focus. Annotations were performed using videos extracted from the monocular egocentric video for the working/waiting time segments overlaid with gaze data provided by the eye tracker. Three annotators were asked to annotate each chat block with information on participants' current environment (office, corridor, library, street, canteen, café), whether they were indoors or outdoors, their mode of locomotion (sitting, standing or walking), as well as when their attention shifted from the mobile device to the environment or back.

9.5 Experiments

We conducted several experiments to evaluate the performance of our method for the different prediction tasks described before: attention shifts between the handheld mobile device and the environment and primary attentional focus on the device. We evaluated our method for different time segments, i.e. while answering questions (*working*) and while *waiting* for the next question, as well as for the aforementioned four different feature groups. For all experiments, we extracted features from a one-second window (feature window) and aimed to predict for a subsequent target window. The choice of the one-second feature window was informed by preliminary experiments in which it showed superior performance compared to longer time windows. For the target window size we investigated one, five, and ten seconds, reflecting that different applications might benefit from different time horizons when forecasting user attention. Performance was calculated using the weighted F1 score. The F_1 score $= 2 * \frac{precision*recall}{precision+recall}$ is the harmonic mean of precision $\frac{TP}{TP+FP}$ and recall $\frac{TP}{TP+FN}$, where TP, FP, and FN represent frame-based true positive, false positive, and false negative counts, respectively.

We trained a random forest using the different features using a leave-one-person-out evaluation scheme, i.e., the data of n-1 participants was used for training, and of the last participant, for testing. This procedure was repeated for all participants and the resulting F1 scores averaged over all iterations. All hyperparameters (number of features, maximum depth and minimum samples at leaf nodes) were optimised via cross-validation on the training set. We used a random subset of samples with a 50/50 distribution of positive and negative samples to avoid class imbalance.

9.5.1 Performance for Different Prediction Tasks

Figure 9.5 summarises the performance of our proposed method for different target window sizes and the different prediction tasks. As can be seen from the figure, the performance for predicting shifts to the environment decreases with increasing target window size, while for attention shifts to the mobile device an increase can be observed. A possible interpretation for this is that these shifts are often caused by distractors in the environment which result in a immediate reaction by the user. When trying to predict shifts to the environment over a longer time interval in the future, such environmental distractors might not yet be present in the feature window. To pro-actively pause interactions on a currently used device, a one-second target window for the prediction of shifts to the environment is sufficient, and it is not meaningful to choose a larger target window because the corresponding features do not contain the features necessary for a correct prediction.

On the other hand, a shift of attention back to the mobile device often lasts longer than just one second, as it might involve turning the head and picking up the mobile device, resulting in higher performance for longer target time intervals. For the reduction of interaction delay when the attention shifts back to the device, a larger target window is needed anyway to restart the system or to load the previous screen content. Moreover, predicted shifts to the mobile device can be used to avoid potential dangerous situations

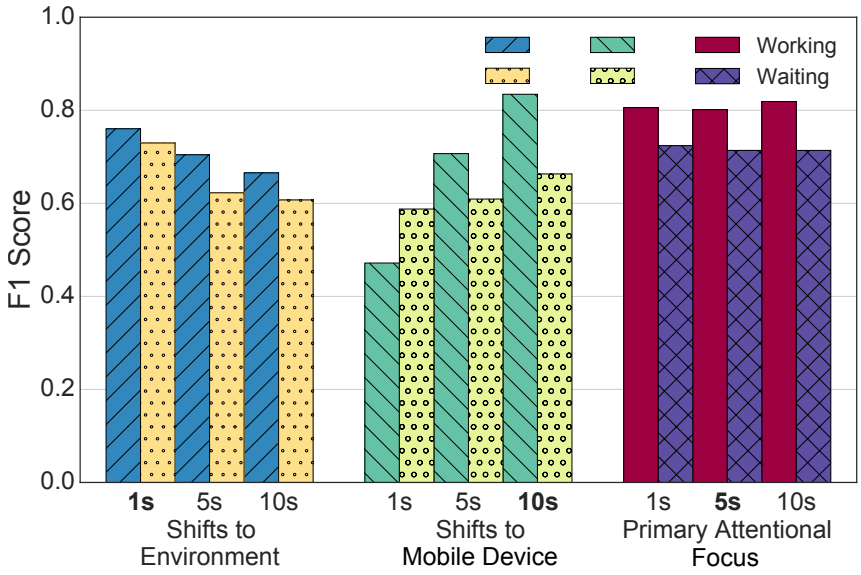


Figure 9.5: Performance analysis for shifts to environment, shifts to mobile device, and primary attentional focus for different target sizes (1s, 5s, 10s).

when the user shifts his/her attention to the device, e.g. when driving a car, an alert could warn the user to keep their attention on the street. In such situations, predicting a shift to the device sufficiently early to still be able to intervene is required. We therefore chose a target window size of ten seconds for shifts to the mobile device.

The primary attentional focus prediction is robust across target window size. Thus, longer target windows can be used to show notifications, or break long attention span prediction during dangerous situations. We opted for a five-second target window for predicting the primary attentional focus.

9.5.2 Prediction of Attention Shifts

We first compared the performance of different feature sets for both attention shift prediction tasks. Figure 9.6 shows the prediction performance of our method depending on feature sets used for both *working* and *waiting* time segments. As can be seen from the figure, performance for predicting shifts to the environment is above chance level (F1 score 0.5) for all feature sets. This shows the effectiveness of our method for this challenging task. However, we can see differences in the prediction performance between the working and waiting time segments and feature sets. As expected, the *Egocentric* sensor modality (F1 0.80) performs competitively against the *Proposed* feature combination (F1 0.76) during working but also during waiting time segments. During working segments performance is generally higher than during waiting segments except for the phone feature combination. A possible explanation for this is that during working time, the task defines a certain phone interaction pattern (e.g. app usage, phone movement) with minor variability, whereas during waiting time the phone interaction can be chosen more freely (e.g. surfing the internet, using Facebook, playing games, chatting, etc.) and can induce different tendencies to switch one's attention to the

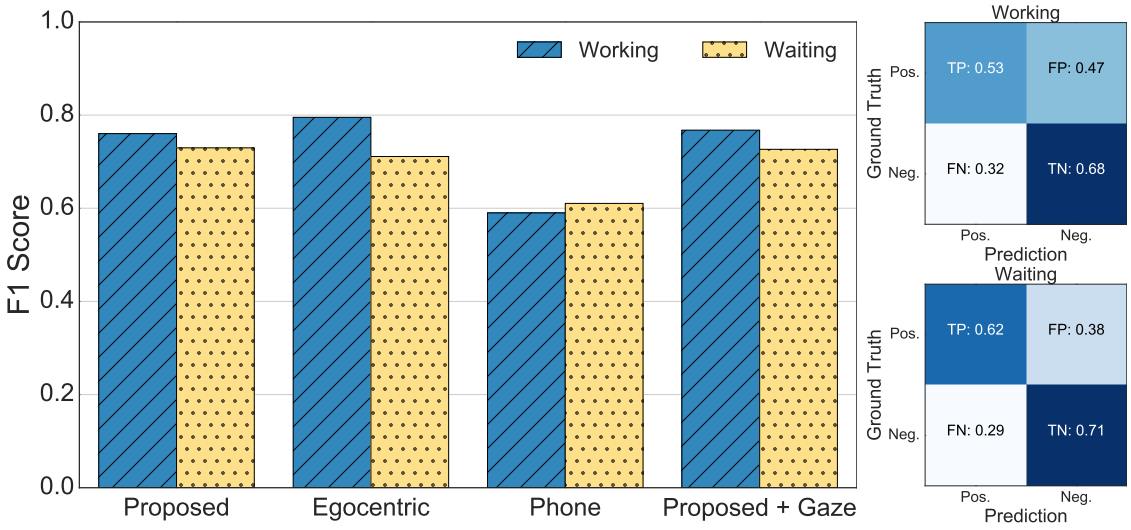


Figure 9.6: Performance for predicting *shifts to the environment* during working and waiting time segments for the different feature sets for a one-second target window, and confusion matrices for our proposed feature set.

environment. A detailed feature analysis showed that especially during working time, detected faces from the scene camera are a helpful feature for the prediction of attention shifts to the environment. The egocentric features, which are part of our proposed feature set, are the dominant ones for this task because shifts to the environment are mainly driven by attractors in our field of view. However, having access to the smartphone state can also help the classifier. The confusion matrices for predicting shifts to the environment show that the classifier achieves a good performance mainly on the negative training examples (i.e. no shift happening).

To further analyse the performance of our method for different environments, we evaluated our feature set in six environments each (see Figure 9.7) during working and waiting time segments for the one-second target window. For the corridor and library environments our proposed feature set even exceeds an F1 score of 0.70, while the performance over all environments during working is higher than during waiting segments except for office environments. For the street environment, it is below 0.6 for working, and during waiting time segments even below 0.4, where participants are mainly focusing on the street and do not check their mobile devices as often as in the other environments.

For shifts to the mobile device the results are different from those for predicting shifts to the environment (see Figure 9.8). With our proposed feature set we reach F1 scores of 0.66 during waiting and F1 scores of 0.83 during working time segments for the ten-second target window, respectively. The competitive performance of phone features for the attention shift forecasting is caused by participants' natural device usage behaviour, which is characterised by picking up and moving the device or turning on its screen. Participants often held their phones in their hands out of the view of the camera, so there was a movement of the device followed by the shift to the device and a touch sequence to unlock the phone. A detailed feature analysis confirmed that both

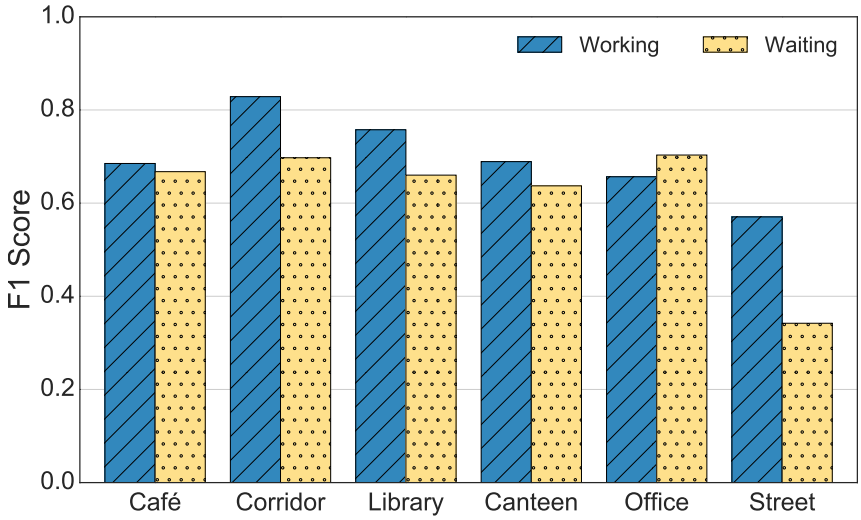


Figure 9.7: Performance for predicting shifts to the environment for different real-world environments of our proposed feature set during working and waiting time segments.

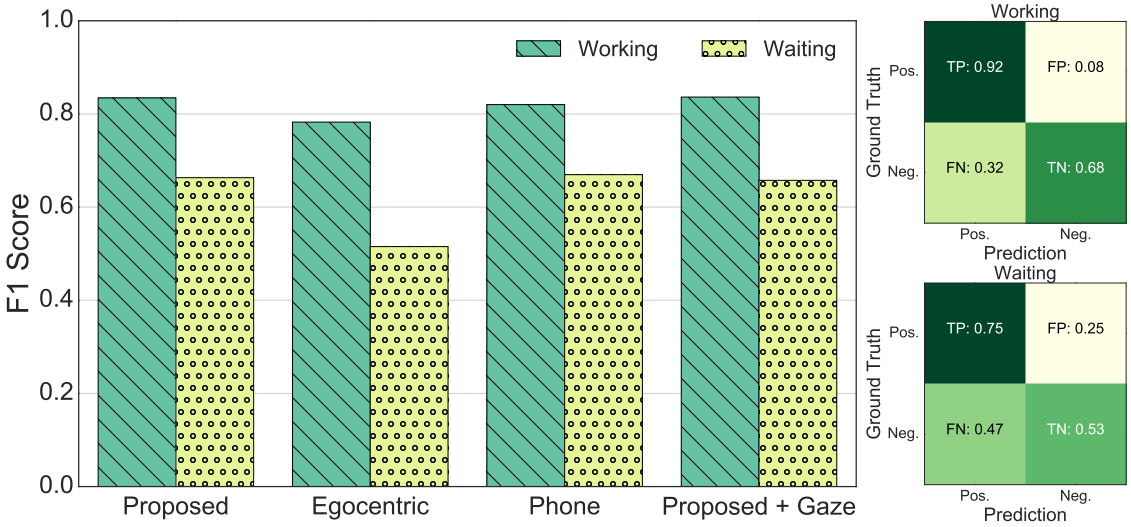


Figure 9.8: Performance for predicting *shifts to the mobile device* during working and waiting time segments for the different feature sets for a ten-second target window, and confusion matrices for our proposed feature set.

actions were registered by the phone sensors and logging apps with F1 scores higher than 0.8 (phone IMU and application usage). Features from the egocentric camera only resulted in chance-level performance, which indicates that the visual environment of the participant does not play a role in determining whether the attention will go back to the screen. This is in line with our reasoning given above, indicating that poorly observable top-down factors influence shifts to the phone, as compared to better observable properties of the visual environment that might capture attention in a way that is more influenced by bottom-up processes. In contrast to the prediction of shifts to the environment, the most errors occur for the negative examples, as indicated by the confusion matrices.

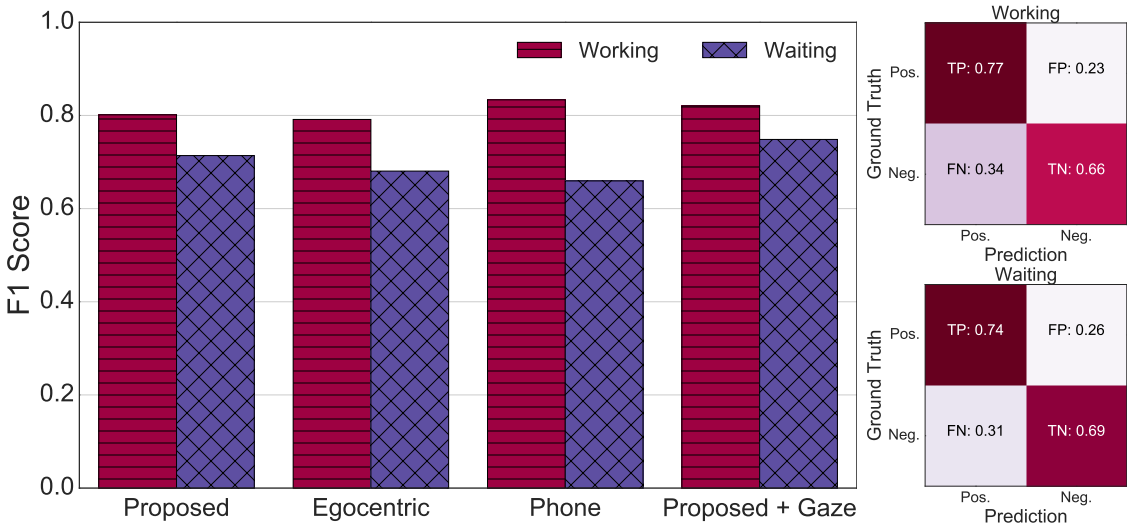


Figure 9.9: Performance for *primary attentional focus* on mobile device during working and waiting time segments for the different feature sets for a five-second target window, and confusion matrices for our proposed feature set.

9.5.3 Prediction of the Primary Attentional Focus

Finally, we analysed the performance of our method for predicting the primary attentional focus on the mobile device. As can be seen from Figure 9.9, for this prediction task, our method reaches an F1 score of more than 0.7 for both working and waiting time segments. It can also be seen that combining features is helpful in all cases. A detailed feature analysis shows that head IMU, depth, and face features from the egocentric feature subsets, as well as the phone IMU, and app usage features, contribute especially to the good performance of our method. Phone features show performance competitive to our proposed features during working but a lower performance during waiting time segments. From a detailed feature analysis it can be seen that users' app usage patterns on the mobile device contributed especially to the performance. The proposed feature combination can even be improved when taking gaze information into account, reaching an F1 performance larger than 0.8 during working and 0.75 during waiting time segments. Thus, for this kind of prediction task, a full eye tracking system is a meaningful setup. The increasing availability of mobile eye tracking as well as gaze estimation using the cameras readily integrated into laptop, tablets, and public displays (Wood and Bulling, 2014; Zhang *et al.*, 2015; Sugano *et al.*, 2016; Zhang *et al.*, 2018c) makes gaze another interesting source of information on users' future attentive behaviour. The corresponding confusion matrices show that our approach performs clearly above chance on all ground truth classes.

9.6 Discussion

The experiments demonstrated that our method can predict several key aspects of attentive behaviour during everyday mobile interactions using a combination of egocentric

and device-integrated sensors. Specifically, we showed that we can predict shifts between the handheld mobile device and environment, as well as the primary attentional focus, above chance level. These results are promising for future mobile attentive user interfaces, particularly given the large variability in natural user behaviour and the large number of possible visual attractors in users' environments, and thus the difficulty of these prediction tasks.

Importance of Different Features. For predicting shifts to the environment, egocentric features contributed most to the performance (see Figure 9.6). A detailed feature analysis showed that face features especially, but also head IMU, semantic scene and depth features, contributed positively. In contrast, phone features showed the best performance for predicting attention shifts back to the mobile device (see Figure 9.8). The chance-level performance for the egocentric features suggested that shifts to the mobile device were less influenced by the environment, especially during waiting time segments. This was to be expected given that such shifts are typically triggered by events on the mobile device, such as an incoming chat message or notification.

Our method performed robustly for predicting attention shifts in different environments, with performance peaking for working and waiting time segments in the corridor (see Figure 9.7). Results for predicting the primary attentional focus (a binary classification task) suggested that information readily available on the handheld device is most informative for predicting on-device focus, and that performance could be improved further by contextualising attentive behaviour using information on the visual scene (see Figure 9.9). A particularly interesting direction for future work is attention span prediction, i.e. the regression task of predicting the actual duration of attention on the mobile device and in the environment. Preliminary experiments on our dataset (not shown here) suggested that this task is currently too challenging – at least with the sensors and features used in this chapter. It will be interesting to study this task in more detail in the future and to see which sensors and features will help to increase performance on this task above chance level.

Potential Applications. Automatic forecasting of user attention opens up a range of exciting new applications that could have paradigm-changing impacts on our everyday interactions with mobile devices. Predicted attention shifts to a mobile device could, for example, be used to reduce interaction delays. The device could turn back on pro-actively and load the previous screen content for a smooth transition, or help users to reorient themselves on the device screen. However, attentive user interfaces are also faced with situations where predicted attention shifts to a mobile device should be prevented. Especially within face-to-face conversations in the real world, user interfaces could help us to keep our focus by giving an alert to avoid unkind behaviour when there is a predicted shift to one's own mobile phone. While driving, crossing a road, or walking down a busy street, it is also desirable for mobile device users to avoid attention shifts to the mobile device, to prevent potentially hazardous situations. Attention shift prediction, for example combined with a detection of dangerous situations using an

body-worn egocentric camera, could suppress on-device alerts or notifications to avoid such attention shifts.

For attention shifts to the environment, attention forecasting could be used to proactively support the users and automatically pause a video even before the attention drifts away, so that the user does not miss a second. Similar to face-to-face conversations, predicted shifts to the environment could be prevented by attentive user interfaces during Skype meetings, so as to keep eye contact. Alternatively, if a user really wants to finish a task, the attentive user interface could help the user to keep their attention on the device by changing the content or style of content presentation.

If the primary attentional focus is predicted to be on the mobile device, previously missed messages or notifications could be shown to the user. Moreover, the user interface could suggest the next task to be performed by the user. Similar to avoiding attention shifts in dangerous situations, future user interfaces could break longer attentional focus spans when potential threats are detected via a scene camera. The aforementioned prediction of attention span would further extend application opportunities by allowing for temporally more fine-grained and targeted adaptations.

Limitations and Future Work. Despite these promising results, the work presented in this chapter also has several limitations. First, while we only considered visual triggers, attention shifts to the environment can also be triggered by auditory stimuli. An interesting direction for future work is to analyse both visual and auditory information for predicting mobile attention allocation. Second, we only considered prediction of temporal attention characteristics, namely timing of attention shifts and primary attentional focus. Future mobile attentive user interfaces could also predict "where" user attention will shift (Zhang *et al.*, 2017a). Third, while all our predictions were clearly above chance level, performance has to further increase to make attention forecasting practically useful. To improve performance, additional sensors for heart rate, galvanic skin response (GSR) or brain activity could be used. Given the rapid development in sensor technology, some of the wearables used may no longer be needed in the future, or they may be replaced by more sophisticated ones, providing even better features for attention forecasting. Also, the method itself could be improved, for example, by using spatio-temporal CNN features extracted from each frame (Tran *et al.*, 2015) that demonstrated superior performance in a variety of computer vision tasks. Particularly interesting are features extracted from intermediate layers, as for example used for vision-based (Ma *et al.*, 2016; Huang *et al.*, 2018) or wearable sensor-based (Ordóñez and Roggen, 2016) activity recognition. Fourth, the current hardware setup is rather bulky (head-mounted mobile eye tracker, multiple cameras, mobile phone, laptop backpack), which might have influenced participants' attentive behaviour. Therefore, investigating in-the-wild studies with participants' awareness about the recording will be an interesting future project (Risko and Kingstone, 2011; Nasiopoulos *et al.*, 2015). Fully integrating the required cameras is an important direction for future work, but likely to be feasible given recent advances in fully embedded head-mounted eye tracking (Tonsen *et al.*, 2017).

9.7 Conclusion

In this chapter we explored *attention forecasting* – the task of predicting future allocation of users' overt visual attention during interactions with a handheld mobile device. We proposed three prediction tasks with direct relevance for future mobile attentive user interfaces, as well as a first computational method to predict key characteristics of attentive behaviour from device-integrated and wearable sensors. We evaluated our method on a novel 20-participant dataset and demonstrated its effectiveness in predicting attention shifts between the mobile device and the environment, as well as the primary attentional focus on the mobile device. Our results demonstrate not only the feasibility but also the significant challenge of attention forecasting, and point towards a new class of user interfaces that pro-actively support, guide or even optimise for users' ever-changing attentive behaviour.

10 Anticipating Averted Gaze in Dyadic Interactions

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.

10.1 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).

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 (Zhang *et al.*, 2017b; Müller *et al.*, 2018b). 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

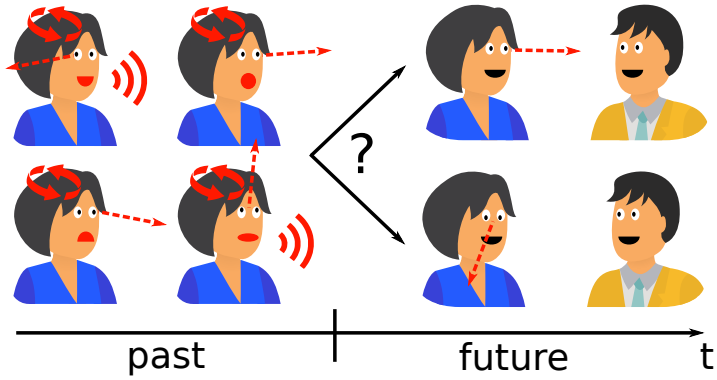


Figure 10.1: We study the challenging task of averted gaze anticipation in conversations: Given past observation of a person's gaze, head pose, facial expressions and speaking behaviour, we predict averted gaze in the near future.

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.*, 2017a) and mobile device interactions (Steil *et al.*, 2018b), 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 10.1 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 time interval in the future. Our method exploits the dependence of subsequent states of eye contact on previous states of eye contact, gaze, head pose, and facial expressions as well as the well-known link between speaking status and eye contact (Kendon, 1967; Ho *et al.*, 2015).

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 will be made publicly available upon acceptance of this work and can become a valuable resource for research on eye contact detection and anticipation.

10.2 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.

10.2.1 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 recent research has shown that while speakers likely gaze at their interlocutor at the end of speaking turns, listeners begin speaking with averted gaze (Ho *et al.*, 2015). Romaniuk (2009) suggested that interviewers avoid mutual gaze with their interviewee "during the production of interviewee laughter", which is an intended regulatory feature to dissuade or evade a response to the perceived inappropriate or distracting laughter. The importance of averted gaze in particular is further underlined by a study showing that averted gaze is connected to cognitive load (Glenberg *et al.*, 1998). In this study, participants were given questions that induced different amounts cognitive load. 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 (Ekman, 1992; Adams Jr and Kleck, 2003; Zuckerman *et al.*, 1981). A key finding is that coordinated gaze behavior 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 (Santarcangelo and Dyer, 1988; Leroy *et al.*, 2009; Streeck, 1993; Argyle and Cook, 1976; Maglio *et al.*, 2000; Jokinen *et al.*, 2010). 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.* (2018b) combined gaze and speech to improve eye contact detection in group interactions. Finally, recent work demonstrated that combining information on the visual focus of attention of people with other features such as facial expressions or body pose can be used to detect leadership (Müller and Bulling, 2019; Beyan *et al.*, 2017c) or rapport (Müller *et al.*, 2018a) in group interactions.

10.2.2 Gaze Estimation and Eye Contact Detection

Analysing social gaze in conversations either requires specialised mobile eye tracking equipment (Tonsen *et al.*, 2017; Kassner *et al.*, 2014) 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 (Yamazoe *et al.*, 2008; Valenti *et al.*, 2011; Wood *et al.*, 2015), appearance based methods directly regress the gaze from the image input (Zhang *et al.*, 2019b; Lu *et al.*, 2012). 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

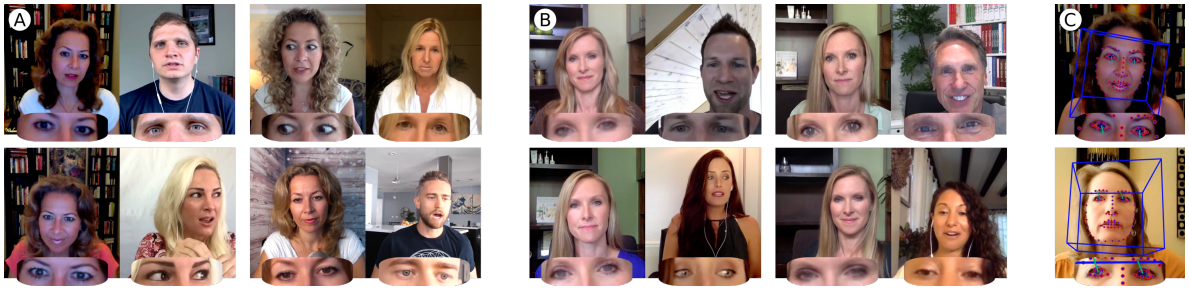


Figure 10.2: Example images from the dataset with enlarged eye regions for better visibility. A: Images from the Youtube channel “Wisdom From North”, B: Images from the Youtube channel "The Spa Dr." For each image, the host is shown on the left and the guest on the right. C: Examples of head pose estimates, keypoint detections and gaze estimates obtained from OpenFace.

not. While Smith et al. (Smith *et al.*, 2013) detected eye contact with the camera an image was taken from, Zhang et al. (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 on the sample task of detecting eye contact in group interactions (Müller *et al.*, 2018b).

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.

10.2.3 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 that indicate at which locations people are most likely to look at when viewing an image (Itti *et al.*, 1998; Harel *et al.*, 2007; Kümmerer *et al.*, 2016) or user interface (Xu *et al.*, 2016). In contrast to saliency prediction, *scanpath prediction* introduces a temporal component by attempting to predict sequences of plausible fixations for a given image (Liu *et al.*, 2013; Assens Reina *et al.*, 2017). Both tasks, however, assume a fixed input image as stimulus and assume a free-viewing task, thereby disregarding the effects of context and top-down influences on gaze behaviour. A study by Borji et al. introduced such top-down effects by modelling gaze behaviour during driving in a computer game (Borji *et al.*, 2013). More recently, there has also been interest in predicting gaze on egocentric videos – a task that requires the system to take into account bottom-up as well as top-down factors and integrate them across time (Zhang *et al.*, 2018e; Huang *et al.*, 2018; Li *et al.*, 2018).

Only few previous works explored the even more challenging task of anticipating (or forecasting) gaze behaviour in the future. Zhang et al. predicted future gaze in egocentric videos by generating future video frames and predicting temporal saliency on

these (Zhang *et al.*, 2017a). 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.*, 2018b). There, the authors focused on forecasting attention during everyday mobile interactions by combining visual scene information obtained using a head-mounted camera with information on app usage and device-integrated mobile phone sensors. Using this approach, 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.

10.3 Dataset

To the best of our knowledge, there currently doesn't exist any dataset of natural dyadic interactions with fine-grained eye contact annotations. To study gaze in conversational behaviour, 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 using a video conferencing tool 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, many recorded 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, finally 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 10.2 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 10.2 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.

10.3.1 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".

10.3.2 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 10.2 C for an illustration of OpenFace output).

10.3.2.1 *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.

10.3.2.2 *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

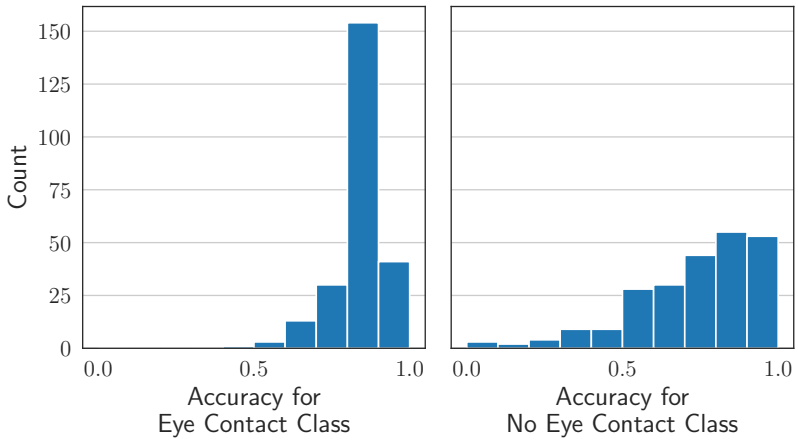


Figure 10.3: Left: Histogram of accuracies of our semi-automatic eye contact detection approach obtained on ground truth eye contact samples. Right: Corresponding accuracies obtained on ground truth no eye contact samples.

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".

10.3.2.3 *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 10.3 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.

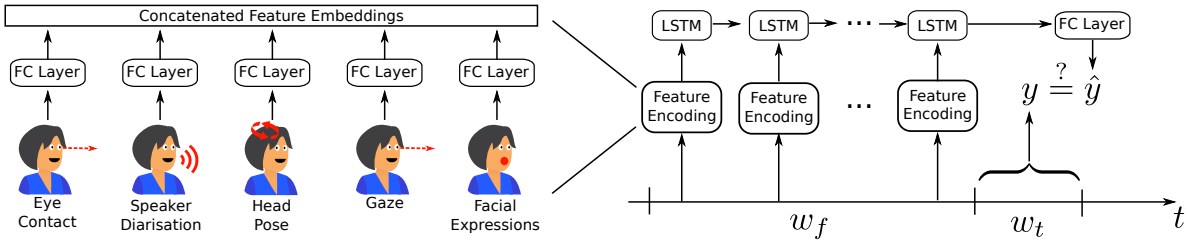


Figure 10.4: Overview of our eye contact anticipation method. Left: In the feature encoding network, each feature modality is fed through a fully connected layer (FC Layer) separately and the resulting representations are concatenated. Right: features are extracted on a feature window w_f and fed through 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.

10.4 Method

Figure 10.4 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.

10.4.1 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"), include eye contact, raw gaze, head pose and facial expressions as well as the speaking status. We do not extract features from the interactant, as they did not lead to improvements in preliminary experiments.

10.4.1.1 *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 10.3.2 (one-hot encoding, 2 dimensions).

10.4.1.2 *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.*, 2018b). 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.

10.4.2 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.

10.5 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.

10.5.1 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.

10.5.2 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 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.

10.5.3 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 10.5 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.

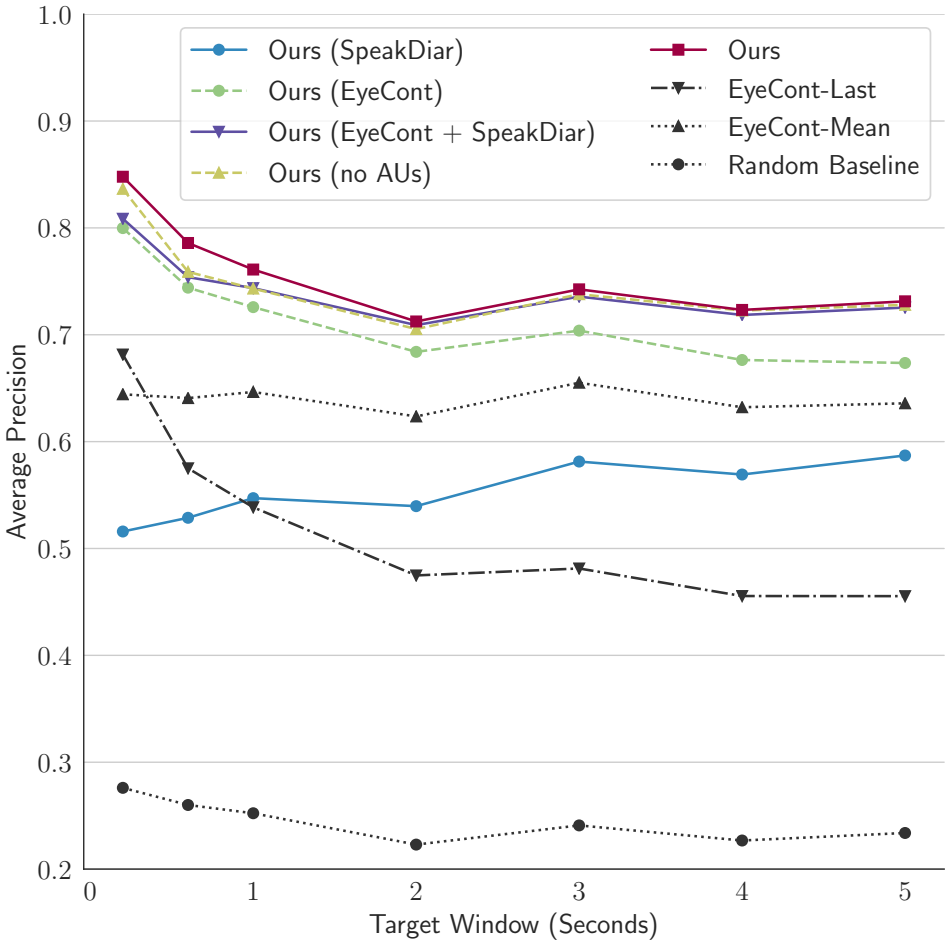


Figure 10.5: Average precision achieved in the person-specific evaluation for different feature channel ablations of our method and baselines across different target window sizes.

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 (*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.

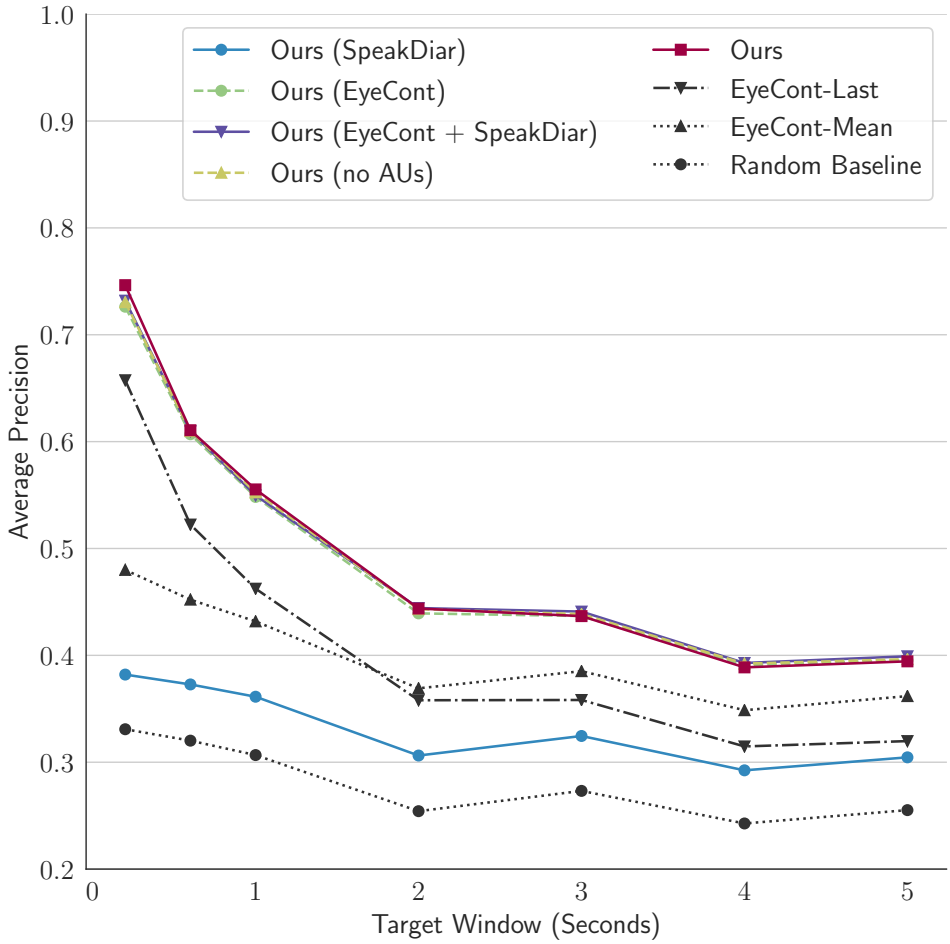


Figure 10.6: Average precision achieved in the person-independent evaluation for different feature channel ablations of our method and baselines across different target window sizes.

10.5.4 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 10.6. 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

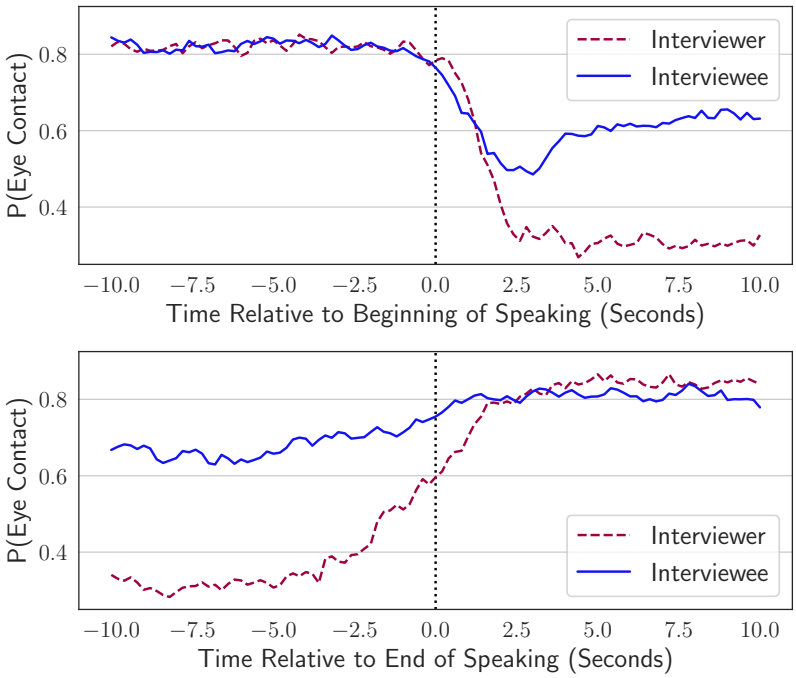


Figure 10.7: Temporal evolution of the probability of interviewer or interviewee having eye contact at the start (top), or end (bottom) of speaking turns.

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).

10.5.5 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 10.7). 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 10.7) 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.*, 2018b). 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.

10.6 Discussion

10.6.1 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.

10.6.2 On Potential Applications

Being able to automatically anticipate averted gaze during interactions opens up multiple possibilities for exciting new applications. For example, 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's gaze will be averted in the near future, potentially resulting in the user overlooking what 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 in systems supporting human-human interactions. Current research on real-time feedback in social interactions is limited to intervening after a specific target behaviour has been observed (Damian *et al.*, 2015; Schiavo *et al.*, 2014). In contrast, with the ability to anticipate averted gaze before it actually occurs, feedback systems could intervene earlier, not allowing the undesired behaviour to occur in the first place. Feedback could be given in different ways. One possibility is explicit feedback, 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.

10.6.3 On Possible Improvements and Extensions

While our work represents an important step towards automated methods to anticipate averted gaze, several possibilities for future improvements and extensions remain. First of all, there is room for improvement in ground truth quality. The semi-automatic eye contact annotations we used for training achieved an average accuracy of between 80% and 90% but performance of averted gaze anticipation could probably still be improved by providing better eye contact detections as input. Furthermore, a highly accurate, fully automatic eye contact detection approach would make the eye contact labelling step obsolete 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, both in terms of performance in challenging everyday settings and generalisability across users (Zhang *et al.*, 2017b; Müller *et al.*, 2018b), these methods still need to improve further to provide close to gold-standard predictions. Further performance improvements might be gained by utilising additional input features. For example, the link between the difficulty of a question and gaze aversion (Glenberg *et al.*, 1998) could be exploited by a sophisticated 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.*, 2017d) and robots can make use of gaze aversion mechanisms to make a more thoughtful impression and effectively manage the conversational floor (Andrist *et al.*, 2014b). 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.*, 2018b). 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.

10.7 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.

Emergent Leadership Detection Across Datasets

A

A.1 Nonverbal Features

A.1.1 VFOA Features

The first step in VFOA feature computation is to detect eye contact between the participants. To this end, we employ the recently introduced method by Müller et al. (Müller *et al.*, 2018b), which performs unsupervised eye contact detection in small group interactions. This method exploits that people usually look at the current speaker to obtain a weak labelling which is used to train a person-specific eye contact detector. The output of this method are frame-wise predictions indicating with which other person the target person has eye contact, or whether the target person has no eye contact at all. For optimal results, we input ground-truth annotated speaker segmentations to the method on MPIIGroupInteraction. On PAVIS we resort to speaking activity detection via thresholding facial action units (cf. (Müller *et al.*, 2018b)), as we found the speaker segmentations provided with the dataset to not be perfectly synchronised with the video. Evaluating on the eye contact annotations provided by the authors of (Müller *et al.*, 2018b), we obtain an accuracy of 0.7 on MPIIGroupInteraction. To eliminate jitter, we apply a median filter of five frames to the predictions.

Based on these eye contact detections, we compute the 15 VFOA features described in (Beyan *et al.*, 2016b). As the original implementation is not available from the authors, we implement the following features ourselves using the description in (Beyan *et al.*, 2016b): **totWatcher:** total time a person is watched by others, **totME:** total time a person has mutual eye contact (MEC) with others, **totWatcherNoME:** total time a person is being watched by others without having MEC, **totNoLook:** total time a person is not looking at any other person, **lookSomeOne:** total time a person looks at other people, **totInitiatorME:** proportion of MECs of a person that are initiated by her, **stdInitiatorME:** the standard deviation of lengths of MECs that are initiated by the person, **totInterCurrME:** average time between intiation of a MEC and the start of the MEC, **stdInterCurrME:** standard deviation of totInterCurrME, **totWatchNoME:** total time a person is looking at others without MEC, **maxTwoWatcherWME:** maximum time a person is looked at by two others, **minTwoWatcherWME:** minimum time a person is looked at by two others, **maxTwoWatcherNoME:** maximum time a person is looked at by two others without having MEC with them, **minTwoWatch-**

erNoME: minimum time a person is looked at by two others without having MEC with them, **ratioWatcherLookSOne:** ratio between totWatcher and lookSomeOne.

A.1.2 Body Pose Features

In line with (Beyan *et al.*, 2017c), body pose features are computed from frames with significant activity. We detect such frames by a two-step thresholding approach on the difference images of subsequent greyscale frames: In the first step a pixel is classified as moving if its value exceeds the threshold $T_1 = 30$ in the difference image. The second step is classifying a frame as having significant activity if the number of moving pixels in it exceeds a threshold T_2. We set T_2 such that we obtain the same proportion of frames with significant activity as described in (Beyan *et al.*, 2017c) (roughly 8.1%). For MPIIGroupInteraction we set T_2 for each interaction separately to not leak information between interactions at test time.

On frames with significant activity, we compute the 80-dimensional featureset described in (Beyan *et al.*, 2017c), consisting of statistical measures extracted from the angles between vectors that are defined by 2D joint positions. We use code provided to us by the authors of (Beyan *et al.*, 2017c).

A.1.3 Speaking Activity Features

We implement the four speaking activity features from (Sanchez-Cortes *et al.*, 2013), namely the total speaking time of a participant (SPL), the number of speaking turns of a participant (SPT), the total number of times a participant interrupts other participants (SPI), and the average duration of a participants' speaking turns (ASP). We normalise SPL, SPT and SPI with the length of the time interval from which we extract the feature. On both PAVIS as well as MPIIGroupInteraction, we extract speaking activity features from ground truth speaker segmentations.

A.2 Classification

A.2.1 Data Normalisation

The standard way to normalise both train and test data is via mean and standard deviation computed on the training data (Friedman *et al.*, 2001). This prevents information leakage from the test set at training time (e.g. when normalising train and test data jointly), and also leakage from "future" test samples at test time (when normalising the whole test set at once). However, in our case training and testing data distributions differ and our data is structured by interactions made up of three to four individual participants. As a consequence, while normalising the training data as usual, we normalise each test interaction separately (i.e. independently from the training data as well as other test interactions). In this way, no information "from the future" is leaked while testing and we comply to the fact of different training and testing distributions. In

Feature	MPI		PAVIS	
	Acc.	Ori.	Acc.	Ori.
totWatcherNoME	0.59	+	0.66	+
ratioWatcherLookSOne	0.59	+	0.62	+
totWatcher	0.55	+	0.76	+
totWatchNoME	0.55	−	0.43	−
totInitiatorME	0.45	−	0.40	−
lookSomeOne	0.45	−	0.34	−
stdInitiatorME	0.45	+	0.34	+
totNoLook	0.45	+	0.34	+
stdInterCurrME	0.45	−	0.41	−
maxTwoWatcherNoME	0.45	+	0.21	+
minTwoWatcherWME	0.45	−	0.14	+
maxTwoWatcherWME	0.41	+	0.36	+
minTwoWatcherNoME	0.41	−	0.14	−
totInterCurrME	0.41	−	0.43	−
totME	0.36	+	0.60	+

Table A.1: Accuracies for single feature based classification using VFOA features on PAVIS and MPIIGroupInteraction. "Ori." indicates whether the maximum or the minimum of the feature was used for prediction.

preliminary experiments, we found this way of normalising to be crucial. The standard way of normalising described above resulted in much worse performance.

A.2.2 Alternative Classification Methods

Apart from SVMs, we also evaluated several dedicated domain adaptation methods including Transfer Component Analysis (Pan *et al.*, 2011), Correlation Alignment (Sun *et al.*, 2016), Random Walk Adaptation (van Laarhoven and Marchiori, 2017) as well as transductive methods like label propagation (Zhu and Ghahramani, 2002). Neither of these methods could consistently improve over the plain SVM approach in our experiments.

A.3 Feature Analysis

Table A.1 shows the results of single feature based classification for all 15 VFOA features.

List of Figures

List of Tables

Bibliography

S. Abdullah, E. L. Murnane, M. Matthews, M. Kay, J. A. Kientz, G. Gay, and T. Choudhury (2016). Cognitive Rhythms: Unobtrusive and Continuous Sensing of Alertness Using a Mobile Phone, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing 2016*. Cited on page 126.

A. Abele (1986). Functions of gaze in social interaction: Communication and monitoring, *Journal of Nonverbal Behavior*, vol. 10(2), pp. 83–101. Cited on pages 26 and 145.

R. B. Adams Jr and R. E. Kleck (2003). Perceived gaze direction and the processing of facial displays of emotion, *Psychological science*, vol. 14(6), pp. 644–647. Cited on pages 145 and 147.

S. M. Aglioti, P. Cesari, M. Romani, and C. Urgesi (2008). Action anticipation and motor resonance in elite basketball players, *Nature neuroscience*, vol. 11(9), p. 1109. Cited on page 15.

M. Al-Rubaie and J. M. Chang (2019). Privacy-preserving machine learning: Threats and solutions, *IEEE Security & Privacy*, vol. 17(2), pp. 49–58. Cited on page 31.

X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe (2016). Salsa: A novel dataset for multimodal group behavior analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38(8), pp. 1707–1720. Cited on page 55.

X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe (2015). Analyzing free-standing conversational groups: A multimodal approach, in *Proc. of the ACM International Conference on Multimedia 2015*. Cited on page 10.

F. Alnajar, T. Gevers, R. Valenti, and S. Ghebreab (2013). Calibration-Free Gaze Estimation Using Human Gaze Patterns, in *Proc. of the IEEE International Conference on Computer Vision 2013*. Cited on pages 7 and 60.

N. Ambady and R. Rosenthal (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis., *Psychological bulletin*, vol. 111(2), p. 256. Cited on page 1.

D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, *et al.* (2016). Deep speech 2: End-to-end speech recognition in english and mandarin, in *Proc. of the International Conference on Machine Learning 2016*. Cited on page 32.

C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, *Artificial Intelligence Review*, vol. 43(2), pp. 155–177. Cited on pages 80 and 89.

S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu (2014a). Conversational Gaze Aversion for Humanlike Robots, in *Proc. of the 2014 ACM/IEEE International Conference on Human-robot Interaction 2014*. Cited on page 42.

S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu (2014b). Conversational gaze aversion for humanlike robots, in *Proc. of the ACM/IEEE International Conference on Human-robot Interaction 2014*. Cited on page 161.

M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander (1999). OPTICS: ordering points to identify the clustering structure, in *ACM Sigmod record 1999*. Cited on page 45.

E. Arabadzhiyska, O. T. Tursun, K. Myszkowski, H.-P. Seidel, and P. Didyk (2017). Saccade landing position prediction for gaze-contingent rendering, *ACM Transactions on Graphics (TOG)*, vol. 36(4), pp. 1–12. Cited on page 16.

O. Aran and D. Gatica-Perez (2013). Cross-domain personality prediction: from video blogs to small group meetings, in *Proc. of the ACM on International Conference on Multimodal Interaction 2013*. Cited on page 10.

M. Argyle and M. Cook (1976). Gaze and mutual gaze. Cited on page 147.

M. Assens Reina, X. Giro-i Nieto, K. McGuinness, and N. 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 2017*. Cited on page 148.

U. Avci and O. Aran (2016). Predicting the performance in decision-making tasks: From individual cues to group interaction, *IEEE Transactions on Multimedia*, vol. 18(4), pp. 643–658. Cited on page 10.

S. O. Ba and J.-M. Odobez (2010). Multiperson visual focus of attention from head pose and meeting contextual cues, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33(1), pp. 101–116. Cited on page 8.

M. Backes, M. Dürmuth, and D. Unruh (2008). Compromising Reflections -or- How to Read LCD Monitors around the Corner, in *Proc. of the IEEE Symposium on Security and Privacy 2008*. Cited on pages 7 and 60.

C. Bai, M. Bolonkin, S. Kumar, J. Leskovec, J. Burgoon, N. Dunbar, and V. Subrahmanian (2019). Predicting dominance in multi-person videos, in *Proc. of the International Joint Conference on Artificial Intelligence 2019*. Cited on page 10.

R. Bailey, A. McNamara, N. Sudarsanam, and C. Grimm (2009). Subtle Gaze Direction, *ACM Transactions on Graphics*, vol. 28(4), pp. 100:1–100:14. Cited on page 160.

J. E. Baird Jr (1977). Some nonverbal elements of leadership emergence, *Southern Speech Communication Journal*, vol. 42(4), pp. 352–361. Cited on pages 13 and 113.

M. Balaam, G. Fitzpatrick, J. Good, and E. Harris (2011). Enhancing Interactional Synchrony with an Ambient Display, in *Proc. of the ACM Conference on Human Factors in Computing Systems 2011*. Cited on pages 55 and 111.

T. Baltrušaitis, M. Mahmoud, and P. Robinson (2015). Cross-dataset learning and person-specific normalisation for automatic action unit detection, in *Proc. of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition 2015*. Cited on page 116.

T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency (2018). Openface 2.0: Facial behavior analysis toolkit, in *Proc. of the IEEE International Conference on Automatic Face & Gesture Recognition 2018*. Cited on pages 2, 8, 27, 116, 150, and 152.

T. Baltrušaitis, P. Robinson, and L.-P. Morency (2016). OpenFace: an open source facial behavior analysis toolkit, in *Proc. of the IEEE Winter Conference on Applications of Computer Vision 2016*. Cited on pages 45, 49, and 102.

T. Bänziger, M. Mortillaro, and K. R. Scherer (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception., *Emotion*, vol. 12(5), p. 1161. Cited on pages 3, 77, 79, 80, 81, 85, 87, and 169.

J. B. Bavelas, L. Coates, and T. Johnson (2002). Listener responses as a collaborative process: The role of gaze, *Journal of Communication*, vol. 52(3), pp. 566–580. Cited on pages 41, 42, and 147.

K. Bergmann, R. Böck, and P. Jaecks (2014). Emogest: investigating the impact of emotions on spontaneous co-speech gestures, *Multimodal Corpora: Combining applied and basic research targets*. Cited on page 81.

D. Bernhardt and P. Robinson (2007). Detecting affect from non-stylised body motions, in *Proc. of the International Conference on Affective Computing and Intelligent Interaction 2007*. Cited on pages 11 and 81.

D. Bernhardt and P. Robinson (2009). Detecting emotions from connected action sequences, in *Visual Informatics: Bridging Research and Practice 2009*. Cited on pages 11 and 81.

F. J. Bernieri (1988). Coordinated movement and rapport in teacher-student interactions, *Journal of Nonverbal Behavior*, vol. 12(2), pp. 120–138. Cited on page 104.

F. J. Bernieri, J. S. Gillis, J. M. Davis, and J. E. Grahe (1996). Dyad rapport and the accuracy of its judgment across situations: A lens model analysis., *Journal of Personality and Social Psychology*, vol. 71(1), pp. 110–129. Cited on pages 29, 96, and 99.

A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg (2015). The Evolution of First Person Vision Methods: A Survey, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25(5), pp. 744–760. Cited on page 129.

C. Beyan, F. Capozzi, C. Becchio, and V. 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 2016*. Cited on pages 2, 14, 42, and 114.

C. Beyan, F. Capozzi, C. Becchio, and V. 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 2017*. Cited on page 42.

C. Beyan, F. Capozzi, C. Becchio, and V. Murino (2017b). Prediction of the Leadership Style of an Emergent Leader Using Audio and Visual Nonverbal Features, *IEEE Transactions on Multimedia*, vol. 20(2), pp. 441–456. Cited on pages 8, 9, 10, 14, 39, 42, 50, 55, 95, 104, and 114.

C. Beyan, N. Carissimi, F. Capozzi, S. Vascon, M. Bustreo, A. Pierro, C. Becchio, and V. Murino (2016b). Detecting emergent leader in a meeting environment using nonverbal visual features only, in *Proc. of the ACM International Conference on Multimodal Interaction 2016*. Cited on pages 13, 24, 25, 28, 29, 43, 113, 114, 115, 116, and 163.

C. Beyan, V.-M. Katsageorgiou, and V. Murino (2017c). Moving as a Leader: Detecting Emergent Leadership in Small Groups using Body Pose, in *Proc. of the ACM International Conference on Multimedia 2017*. Cited on pages 3, 14, 21, 24, 29, 114, 116, 117, 118, 147, and 164.

C. Beyan, V.-M. Katsageorgiou, and V. Murino (2019a). A Sequential Data Analysis Approach to Detect Emergent Leaders in Small Groups, *IEEE Transactions on Multimedia*. Cited on pages 14, 114, and 118.

C. Beyan, M. Shahid, and V. Murino (2018). Investigation of Small Group Social Interactions Using Deep Visual Activity-Based Nonverbal Features, in *Proc. of the ACM Multimedia Conference 2018*. Cited on pages 14, 114, 115, and 118.

C. Beyan, A. Zunino, M. Shahid, and V. Murino (2019b). Personality Traits Classification Using Deep Visual Activity-based Nonverbal Features of Key-Dynamic Images, *IEEE Transactions on Affective Computing*. Cited on pages 10 and 31.

U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, and D. Manocha (2019). Step: Spatial temporal graph convolutional networks for emotion perception from gaits, *arXiv preprint arXiv:1910.12906*. Cited on page 11.

A. Bhattacharyya, M. Fritz, and B. Schiele (2018). Long-term on-board prediction of people in traffic scenes under uncertainty, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2018*. Cited on page 15.

P. Bilinski and F. Bremond (2016). Human violence recognition and detection in surveillance videos, in *Proc. of the IEEE International Conference on Advanced Video and Signal Based Surveillance 2016*. Cited on page 10.

M. Black, A. Katsamanis, C.-C. Lee, A. C. Lammert, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan (2010). Automatic classification of married couples' behavior using audio features, in *Proc. of the IEEE Winter Conference on Applications of Computer Vision 2010*. Cited on page 11.

D. Bohus and E. Horvitz (2011). Multiparty Turn Taking in Situated Dialog: Study, Lessons, and Directions, in *Proc. of the Annual SIGdial Meeting on Discourse and Dialogue 2011*. Cited on page 95.

A. Borji and L. Itti (2013). State-of-the-Art in Visual Attention Modeling, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35(1), pp. 185–207. Cited on page 57.

A. Borji, D. N. Sihite, and L. Itti (2013). What/where to look next? Modeling top-down visual attention in complex interactive environments, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44(5), pp. 523–538. Cited on page 148.

E. Bozkurt, H. Khaki, S. Keçeci, B. B. Türker, Y. Yemez, and E. Erzin (2017). The JESTKOD database: an affective multimodal database of dyadic interactions, *Language Resources and Evaluation*, vol. 51(3), pp. 857–872. Cited on page 12.

A. Bulling (2016). Pervasive Attentive User Interfaces, *IEEE Computer*, vol. 49(1), pp. 94–98. Cited on page 124.

A. Bulling, U. Blanke, and B. Schiele (2014). A Tutorial on Human Activity Recognition Using Body-worn Inertial Sensors, *ACM Computing Surveys*, vol. 46(3), pp. 33:1–33:33. Cited on page 82.

A. Bulling and H. Gellersen (2010). Toward Mobile Eye-Based Human-Computer Interaction, *IEEE Pervasive Computing*, vol. 9(4), pp. 8–12. Cited on page 57.

A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster (2011). Eye Movement Analysis for Activity Recognition Using Electrooculography, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 33(4), pp. 741–753. Cited on page 130.

A. Bulling, C. Weichel, and H. Gellersen (2013). EyeContext: Recognition of High-Level Contextual Cues from Human Visual Behaviour, in *Proc. of the ACM Conference on Human Factors in Computing Systems 2013*. Cited on page 130.

Bundespsychotherapeutenkammer (2018). *Ein Jahr nach der Reform der Psychotherapeuten-Richtlinie: Wartezeiten 2018*, https://www.bptk.de/wp-content/uploads/2019/01/20180411_bptk_studie_wartezeiten_2018.pdf. Cited on page 1.

J. K. Burgoon, L. K. Guerrero, and K. Floyd (2016). *Nonverbal communication*, Routledge. Cited on pages 1 and 11.

M. Burns (1984). Rapport and relationships: The basis of child care, *Journal of Child Care*, vol. 2(2), pp. 47–57. Cited on pages 3, 12, 23, and 96.

D. Buschek, B. Bisinger, and F. Alt (2018). ResearchIME: A Mobile Keyboard Application for Studying Free Typing Behaviour in the Wild, in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems 2018*. Cited on pages 25, 59, and 72.

C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan (2008). IEMOCAP: Interactive emotional dyadic motion capture database, *Language resources and evaluation*, vol. 42(4), pp. 335–359. Cited on pages 11 and 81.

C. Busso and S. Narayanan (2008). Recording audio-visual emotional databases from actors: a closer look, in *2nd International Workshop on Emotion, ICLRE'08 2008*. Cited on pages 81 and 82.

J. Butepage, M. J. Black, D. Kragic, and H. Kjellstrom (2017). Deep representation learning for human motion prediction and classification, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2017*. Cited on page 15.

A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André, and M. Valstar (2017). The NoXi Database: Multimodal Recordings of Mediated Novice-expert Interactions, in *Proc. of the ACM International Conference on Multimodal Interaction 2017*. Cited on page 95.

C. Caldeira, Y. Chen, L. Chan, V. Pham, Y. Chen, and K. Zheng (2017). Mobile apps for mood tracking: an analysis of features and user reviews, in *AMIA Annual Symposium Proceedings 2017*. Cited on page 32.

M. Campbell, A. J. Hoane Jr, and F.-h. Hsu (2002). Deep blue, *Artificial intelligence*, vol. 134(1-2), pp. 57–83. Cited on page 1.

Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields, in *arXiv preprint arXiv:1812.08008 2018*. Cited on pages 2 and 116.

Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh (2017). Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2017*. Cited on page 104.

F. Capozzi, C. Beyan, A. Pierro, A. Koul, V. Murino, S. Livi, A. P. Bayliss, J. Ristic, and C. Becchio (2019). Tracking the Leader: Gaze Behavior in Group Interactions, *iScience*, vol. 16, p. 242. Cited on page 145.

O. Celiktutan and H. Gunes (2017). Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability, *IEEE Transactions on Affective Computing*, vol. 8(1), pp. 29–42. Cited on page 10.

A. Čereković (2014). An insight into multimodal databases for social signal processing: acquisition, efforts, and directions, *Artificial Intelligence Review*, vol. 42(4), pp. 663–692. Cited on page 81.

A. Cerekovic, O. Aran, and D. Gatica-Perez (2016). Rapport with Virtual Agents: What Do Human Social Cues and Personality Explain?, *IEEE Transactions on Affective Computing*, vol. 8(3), pp. 382–395. Cited on pages 12, 29, 93, and 97.

Y.-W. Chao, J. Yang, B. Price, S. Cohen, and J. Deng (2017). Forecasting human dynamics from static images, in *Proc. of the IEEE conference on computer vision and pattern recognition 2017*. Cited on page 15.

J. Chen and Q. Ji (2015). A Probabilistic Approach to Online Eye Gaze Tracking Without Explicit Personal Calibration, *IEEE Transactions on Image Processing*, vol. 24(3), pp. 1076–1086. Cited on pages 7 and 60.

Z. Cheng, B. Gao, and T.-Y. Liu (2010). Actively Predicting Diverse Search Intent from User Browsing Behaviors, in *Proc. of the World Wide Web Conference 2010*. Cited on page 127.

P. Chikersal, M. Tomprou, Y. J. Kim, A. W. Woolley, and L. Dabbish (2017). Deep Structures of Collaboration: Physiological Correlates of Collective Intelligence and Group Satisfaction, in *Proc. of the International Conference on Computer-Supported Collaborative Work 2017*. Cited on pages 103 and 104.

W. W. Chin, W. D. Salisbury, A. W. Pearson, and M. J. Stollak (1999). Perceived Cohesion in Small Groups: Adapting and Testing the Perceived Cohesion Scale in a Small-Group Setting, *Small group research*, vol. 30(6), pp. 751–766. Cited on page 110.

H.-k. Chiu, E. Adeli, B. Wang, D.-A. Huang, and J. C. Niebles (2019). Action-agnostic human pose forecasting, in *Proc. of the IEEE Winter Conference on Applications of Computer Vision 2019*. Cited on pages 2 and 15.

E. Chong, K. Chanda, Z. Ye, A. Southerland, N. Ruiz, R. M. Jones, A. Rozga, and J. M. Rehg (2017). 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*, vol. 1(3), pp. 43:1–43:20. Cited on pages 39 and 42.

E. Chong, Y. Wang, N. Ruiz, and J. M. Rehg (2020). Detecting Attended Visual Targets in Video, *arXiv preprint arXiv:2003.02501*. Cited on page 8.

T. Choudhury and A. Pentland (2003). Sensing and modeling human networks using the sociometer, in *Proc. of the IEEE International Symposium on Wearable Computers 2003*. Cited on page 10.

M. Choy, D. Kim, J.-G. Lee, H. Kim, and H. Motoda (2016). Looking Back on the Current Day: Interruptibility Prediction Using Daily Behavioral Features, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing 2016*. Cited on pages 124 and 126.

K. Church and R. De Oliveira (2013). What's up with WhatsApp? Comparing Mobile Instant Messaging Behaviors with Traditional SMS, in *Proc. of the ACM International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI) 2013*. Cited on page 133.

D. Ciresan, U. Meier, and J. Schmidhuber (2012). Multi-column Deep Neural Networks for Image Classification, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2012*. Cited on page 28.

I. Cohen, A. Garg, T. S. Huang, *et al.* (2000). Emotion recognition from facial expressions using multilevel HMM, in *Neural information processing systems 2000*. Cited on page 11.

P. T. Costa and R. R. MacCrae (1992). *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI): Professional manual*. Cited on pages 24, 99, and 110.

R. Cowie and M. Sawey (2011). GTrace-General trace program from Queen's, Belfast. Cited on page 86.

N. Dael, M. Mortillaro, and K. R. Scherer (2012). The body action and posture coding system (BAP): Development and reliability, *Journal of Nonverbal Behavior*, vol. 36(2), pp. 97–121. Cited on page 81.

D. Damen, H. Doughty, G. Maria Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.* (2018). Scaling egocentric vision: The epic-kitchens dataset, in *Proc. of the European Conference on Computer Vision 2018*. Cited on page 15.

I. Damian, C. S. S. Tan, T. Baur, J. Schöning, K. Luyten, and E. 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 2015*. Cited on pages 55, 111, and 160.

M. De Meijer (1989). The contribution of general features of body movement to the attribution of emotions, *Journal of Nonverbal behavior*, vol. 13(4), pp. 247–268. Cited on pages 11 and 81.

Y.-A. de Montjoye, J. Quoidbach, F. Robic, and A. Pentland (2013). Predicting Personality Using Novel Mobile Phone-Based Metrics, in *Proc. of the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction 2013*. Cited on pages 30 and 111.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). Imagenet: A large-scale hierarchical image database, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2009*. Cited on page 28.

K. Dierkes, M. Kassner, and A. Bulling (2018). A novel approach to single camera, glint-free 3D eye model fitting including corneal refraction, in *Proc. of the ACM Symposium on Eye Tracking Research and Applications 2018*. Cited on page 60.

T. Dingler and M. Pielot (2015). I'll Be There for You: Quantifying Attentiveness Towards Mobile Messaging, in *Proc. of the International Conference on Human-Computer Interaction with Mobile Devices and Services 2015*. Cited on pages 25, 59, and 72.

T. Dingler, R. Rzayev, V. Schwind, and N. Henze (2016). RSVP on the Go: Implicit Reading Support on Smart Watches Through Eye Tracking, in *Proc. of the ACM International Symposium on Wearable Computers 2016*. Cited on page 126.

V. U. Druskat and A. T. Pescosolido (2006). The impact of emergent leader's emotionally competent behavior on team trust, communication, engagement, and effectiveness, *Research on Emotions in Organizations*, vol. 2, pp. 25–55. Cited on pages 3, 13, 24, and 113.

N. F. Duarte, M. Raković, J. Tasevski, M. I. Coco, A. Billard, and J. Santos-Victor (2018). Action anticipation: Reading the intentions of humans and robots, *IEEE Robotics and Automation Letters*, vol. 3(4), pp. 4132–4139. Cited on page 15.

A. T. Duchowski (2002). A breadth-first survey of eye-tracking applications, *Behavior Research Methods, Instruments, & Computers*, vol. 34(4), pp. 455–470. Cited on page 57.

P. Ekman (1992). Facial expressions of emotion: an old controversy and new findings, *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 335(1273), pp. 63–69. Cited on page 147.

P. Ekman, E. R. Sorenson, and W. V. Friesen (1969). Pan-cultural elements in facial displays of emotion, *Science*, vol. 164(3875), pp. 86–88. Cited on pages 85 and 86.

A. Exler, M. Braith, A. Schankin, and M. Beigl (2016). Preliminary Investigations about Interruptibility of Smartphone Users at Specific Place Types, in *Adj. Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing 2016*. Cited on pages 124 and 126.

F. Eyben, F. Weninger, F. Gross, and B. Schuller (2013). Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor, in *Proc. of the ACM International Conference on Multimedia 2013*. Cited on pages 102 and 103.

Y. A. Farha and J. Gall (2019). Uncertainty-Aware Anticipation of Activities, *arXiv preprint arXiv:1908.09540*. Cited on page 15.

T. Farroni, G. Csibra, F. Simion, and M. H. Johnson (2002). Eye contact detection in humans from birth, *Proc. of the National Academy of Sciences*, vol. 99(14), pp. 9602–9605. Cited on page 39.

S. N. Fatima and E. Erzin (2017). Cross-Subject Continuous Emotion Recognition Using Speech and Body Motion in Dyadic Interactions., in *Proc. of the Annual Conference of the International Speech Communication Association 2017*. Cited on page 12.

S. Feese, A. Muaremi, B. Arnrich, G. Troster, B. Meyer, and K. Jonas (2011). Discriminating Individually Considerate and Authoritarian Leaders by Speech Activity Cues, in *Proc. of the IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing 2011*. Cited on pages 10, 95, and 113.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan (2010). Object Detection with Discriminatively Trained Part-Based Models, *IEEE TPAMI*. Cited on page 82.

W. Feng, A. Kannan, G. Gkioxari, and C. L. Zitnick (2017). Learn2smile: Learning non-verbal interaction through observation, in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems 2017*. Cited on page 149.

J. M. Findlay and I. D. Gilchrist (2003). *Active Vision: The Psychology of Looking and Seeing*, no. 37, Oxford University Press. Cited on page 135.

M. A. Fischler and R. C. Bolles (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, *Communications of the ACM*, vol. 24(6), pp. 381–395. Cited on page 64.

J. Fogarty, S. E. Hudson, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. C. Lee, and J. Yang (2005). Predicting Human Interruptibility with Sensors, *ACM Transactions on Computer-Human Interaction*, vol. 12(1), pp. 119–146. Cited on page 126.

J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth (2007). The world of emotions is not two-dimensional, *Psychological science*, vol. 18(12), pp. 1050–1057. Cited on page 86.

N. Fourati, C. Pelachaud, and P. Darmon (2019). Contribution of temporal and multi-level body cues to emotion classification, in *Proc. of the International Conference on Affective Computing and Intelligent Interaction 2019*. Cited on page 11.

J. Friedman, T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, vol. 1, Springer Series in Statistics, New York. Cited on page 164.

E. Y. Fu, M. X. Huang, H. V. Leong, and G. Ngai (2018). Cross-species learning: A low-cost approach to learning human fight from animal fight, in *Proc. of the ACM International Conference on Multimedia 2018*. Cited on page 10.

A. Furnari, S. Battiato, and G. Maria Farinella (2018). Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation, in *Proc. of the European Conference on Computer Vision 2018*. Cited on page 15.

A. Furnari and G. M. Farinella (2019). What Would You Expect? Anticipating Egocentric Actions with Rolling-Unrolling LSTMs and Modality Attention, in *Proc. of the IEEE International Conference on Computer Vision 2019*. Cited on page 15.

D. Gatica-Perez, L. McCowan, D. Zhang, and S. Bengio (2005). Detecting Group Interest-Level in Meetings, in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing 2005*. Cited on pages 8, 9, 10, 21, 40, 42, 50, 55, and 95.

F. H. Gerpott, N. Lehmann-Willenbrock, J. D. Silvis, and M. Van Vugt (2018). In the eye of the beholder? An eye-tracking experiment on emergent leadership in team interactions, *The Leadership Quarterly*, vol. 29(4), pp. 523–532. Cited on pages 13 and 113.

M. Ghayoumi and A. K. Bansal (2016). Unifying Geometric Features and Facial Action Units for Improved Performance of Facial Expression Analysis, *arXiv preprint arXiv:1606.00822*. Cited on page 109.

T. Giannakopoulos (2015). pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis, *PloS one*, vol. 10(12). Cited on page 153.

A. M. Glenberg, J. L. Schroeder, and D. A. Robertson (1998). Averting the gaze disengages the environment and facilitates remembering, *Memory & cognition*, vol. 26(4), pp. 651–658. Cited on pages 26, 145, 147, 159, and 161.

D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer (2011). Toward a minimal representation of affective gestures, *IEEE TAC*, vol. 2(2), pp. 106–118. Cited on page 81.

M. S. Gobel, H. S. Kim, and D. C. Richardson (2015). The dual function of social gaze, *Cognition*, vol. 136, pp. 359–364. Cited on page 4.

L. D. Goodstein and R. I. Lanyon (1999). Applications of Personality Assessment to the Workplace: A Review, *Journal of Business and Psychology*, vol. 13(3), pp. 291–322. Cited on page 113.

S. Gorga and K. Otsuka (2010). Conversation scene analysis based on dynamic bayesian network and image-based gaze detection, in *International Conference on Multimodal*

Interfaces and the Workshop on Machine Learning for Multimodal Interaction 2010. Cited on page 8.

P. A. Granhag and L. A. Strömwall (2002). Repeated interrogations: Verbal and non-verbal cues to deception, *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 16(3), pp. 243–257. Cited on page 31.

H. Griffith, S. Biswas, and O. Komogortsev (2018). Towards Reduced Latency in Saccade Landing Position Prediction Using Velocity Profile Methods, in *Proc. of the Future Technologies Conference 2018*. Cited on page 16.

J. Guan, L. Yin, J. Sun, S. Qi, X. Wang, and Q. Liao (2020). Enhanced Gaze Following via Object Detection and Human Pose Estimation, in *International Conference on Multimedia Modeling 2020*. Cited on page 8.

J. Guan, Y. Yuan, K. M. Kitani, and N. Rhinehart (2019). Generative hybrid representations for activity forecasting with no-regret learning, *arXiv preprint arXiv:1904.06250*. Cited on page 15.

H. Gunes and M. Piccardi (2007). Bi-modal emotion recognition from expressive face and body gestures, *Journal of Network and Computer Applications*, vol. 30(4), pp. 1334 – 1345. Cited on page 81.

C. Gutwin, S. Bateman, G. Arora, and A. Coveney (2017). Looking Away and Catching Up: Dealing with Brief Attentional Disconnection in Synchronous Groupware, in *Proc. of the ACM Conference on Computer Supported Cooperative Work and Social Computing 2017*. Cited on page 124.

J. L. Hagad, R. Legaspi, M. Numao, and M. Suarez (2011). Predicting Levels of Rapport in Dyadic Interactions through Automatic Detection of Posture and Posture Congruence, in *Proc. of the IEEE International Conference on Social Computing 2011*. Cited on pages 10, 12, 93, and 97.

A. F. d. 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*, vol. 371(1686), p. 20150080. Cited on page 147.

D. W. Hansen and Q. Ji (2009). In the eye of the beholder: A survey of models for eyes and gaze, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32(3), pp. 478–500. Cited on page 147.

J. Harel, C. Koch, and P. Perona (2007). Graph-based visual saliency, in *Advances in Neural Information Processing Systems 2007*. Cited on page 148.

J. A. Harrigan, T. E. Oxman, and R. Rosenthal (1985). Rapport expressed through nonverbal behavior, *Journal of Nonverbal Behavior*, vol. 9(2), pp. 95–110. Cited on pages 12 and 96.

J. V. Haxby, E. A. Hoffman, and M. I. Gobbini (2002). Human Neural Systems for Face Recognition and Social Communication, *Biological Psychiatry*, vol. 51(1), pp. 59–67. Cited on page 131.

M. Hayhoe and D. Ballard (2005). Eye movements in natural behavior, *Trends in cognitive sciences*, vol. 9(4), pp. 188–194. Cited on page 58.

J. Hemminahaus and S. Kopp (2017). Towards adaptive social behavior generation for assistive robots using reinforcement learning, in *Proc. of the ACM/IEEE International Conference on Human-Robot Interaction 2017*. Cited on page 30.

L. Hirvenkari, J. Ruusuvuori, V.-M. Saarinen, M. Kivioja, A. Peräkylä, and R. Hari (2013). Influence of turn-taking in a two-person conversation on the gaze of a viewer, *PloS one*, vol. 8(8), p. e71569. Cited on page 41.

S. Ho, T. Foulsham, and A. Kingstone (2015). Speaking and listening with the eyes: gaze signaling during dyadic interactions, *PloS one*, vol. 10(8), p. e0136905. Cited on pages 41, 42, 146, and 147.

S. Hochreiter and J. Schmidhuber (1997). Long Short-Term Memory, *Neural Computation*, vol. 9(8), pp. 1735–1780. Cited on pages 146 and 152.

K. Hogan (2003). *Can't Get Through: Eight Barriers to Communication*, Pelican Publishing. Cited on page 77.

C. Holland and O. Komogortsev (2012). Eye Tracking on Unmodified Common Tablets: Challenges and Solutions, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications 2012*. Cited on pages 124 and 126.

S. Hoppe, T. Loetscher, S. Morey, and A. Bulling (2015). Recognition of Curiosity Using Eye Movement Analysis, in *Adj. Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing 2015*. Cited on page 111.

S. Hoppe, T. Loetscher, S. Morey, and A. Bulling (2018). Eye Movements During Everyday Behavior Predict Personality Traits, *Frontiers in Human Neuroscience*, vol. 12. Cited on page 31.

R. Hoyle, R. Templeman, D. Anthony, D. Crandall, and A. Kapadia (2015). Sensitive Lifelogs: A Privacy Analysis of Photos from Wearable Cameras, in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems 2015*. Cited on page 73.

R. Hoyle, R. Templeman, S. Armes, D. Anthony, D. Crandall, and A. Kapadia (2014). Privacy Behaviors of Lifeloggers Using Wearable Cameras, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing 2014*. Cited on page 73.

C.-M. Huang and B. Mutlu (2016). Anticipatory robot control for efficient human-robot collaboration, in *Proc. of the ACM/IEEE International Conference on Human-Robot Interaction 2016*. Cited on page 15.

M. X. Huang, T. C. Kwok, G. Ngai, S. C. Chan, and H. V. Leong (2016a). Building a Personalized, Auto-Calibrating Eye Tracker from User Interactions, in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems 2016*. Cited on pages 7 and 60.

M. X. Huang, J. Li, G. Ngai, and H. V. Leong (2016b). StressClick: Sensing Stress from Gaze-Click Patterns, in *Proc. of the ACM Conference on Multimedia 2016*. Cited on pages 42 and 104.

Q. Huang, A. Veeraraghavan, and A. Sabharwal (2015). TabletGaze: Unconstrained Appearance-Based Gaze Estimation in Mobile Tablets, *arXiv preprint arXiv:1508.01244*. Cited on page 126.

Y. Huang, M. Cai, Z. Li, and Y. Sato (2018). Predicting Gaze in Egocentric Video by Learning Task-dependent Attention Transition, in *Proc. of the European Conference on Computer Vision 2018*. Cited on pages 142 and 148.

H. Hung and D. Gatica-Perez (2010). Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior, *IEEE Transactions on Multimedia*, vol. 12(6), pp. 563–575. Cited on pages 2, 10, and 95.

E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele (2017). Arttrack: Articulated multi-person tracking in the wild, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2017*. Cited on page 2.

R. Ishii, K. Otsuka, S. Kumano, and J. Yamato (2016). Prediction of Who Will Be the Next Speaker and When Using Gaze Behavior in Multiparty Meetings, *ACM Transactions on Interactive Intelligent Systems*, vol. 6(1), pp. 4:1–4:31. Cited on page 41.

L. Itti and C. Koch (2000). A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention, *Vision Research*, vol. 40(10), pp. 1489–1506. Cited on page 127.

L. Itti, C. Koch, and E. Niebur (1998). A model of saliency-based visual attention for rapid scene analysis, *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11), pp. 1254–1259. Cited on page 148.

C. E. Izard (1990). Personality, Emotion Expressions, and Rapport, *Psychological Inquiry*, vol. 1(4), pp. 315–317. Cited on pages 96 and 110.

A. M. Jansen, E. Giebels, T. J. van Rompay, and M. Junger (2018). The influence of the presentation of camera surveillance on cheating and pro-social behavior, *Frontiers in psychology*, vol. 9, p. 1937. Cited on page 31.

N. Jaques, Y. L. Kim, and R. Picard (2016a). Personality, Attitudes, and Bonding in Conversations, in *Proc. of the International Conference on Intelligent Virtual Agents 2016*. Cited on page 97.

N. Jaques, D. McDuff, Y. L. Kim, and R. Picard (2016b). Understanding and Predicting Bonding in Conversations Using Thin Slices of Facial Expressions and Body Language, in *Proc. of the International Conference on Intelligent Virtual Agents 2016*. Cited on page 97.

H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black (2013). Towards understanding action recognition, in *Proc. of the IEEE International Conference on Computer Vision 2013*. Cited on page 89.

Z. Jiang, J. Han, C. Qian, W. Xi, K. Zhao, H. Ding, S. Tang, J. Zhao, and P. Yang (2016). VADS: Visual Attention Detection with a Smartphone, in *Proc. of the IEEE International Conference on Computer Communications 2016*. Cited on page 126.

K. Jokinen, H. Furukawa, M. Nishida, and S. Yamamoto (2013). Gaze and Turn-taking Behavior in Casual Conversational Interactions, *ACM Transactions on Interactive Intelligent Systems*, vol. 3(2), pp. 12:1–12:30. Cited on pages 41 and 42.

K. Jokinen, K. Harada, M. Nishida, and S. Yamamoto (2010). Turn-alignment using eye-gaze and speech in conversational interaction, in *Proc. of the Annual Conference of the International Speech Communication Association 2010*. Cited on page 147.

R. Kachouie, S. Sedighadeli, R. Khosla, and M.-T. Chu (2014). Socially assistive robots in elderly care: a mixed-method systematic literature review, *International Journal of Human-Computer Interaction*, vol. 30(5), pp. 369–393. Cited on page 32.

A. Kalma (1992). Gazing in triads: A powerful signal in floor apportionment, *British Journal of Social Psychology*, vol. 31(1), pp. 21–39. Cited on pages 13 and 113.

A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. F. Driessen (2005). Gesture-based affective computing on motion capture data, in *Proc. of the International Conference on Affective Computing and Intelligent Interaction 2005*. Cited on pages 11 and 81.

M. Karg, A.-a. Samadani, R. Gorbet, K. Kuhnlenz, J. Hoey, and D. Kulic (2013). Body Movements for Affective Expression: A Survey of Automatic Recognition and Generation, *IEEE Trans. Affect. Comp.*, (99), p. 1. Cited on pages 11 and 81.

K. Kasi and S. A. Zahorian (2002). Yet another algorithm for pitch tracking, in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing 2002*. Cited on pages 80 and 169.

P. Kasprowski, K. Harężlak, and M. Stasch (2014). Guidelines for the eye tracker calibration using points of regard, in *Information Technologies in Biomedicine, Volume 4 2014*, pp. 225–236, Springer. Cited on page 6.

M. Kassner, W. Patera, and A. 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 2014*. Cited on pages 2, 6, 60, 62, 65, 132, 133, 134, and 147.

S. Kawase (2014). Assignment of leadership role changes performers' gaze during piano duo performances, *Ecological Psychology*, vol. 26(3), pp. 198–215. Cited on page 2.

J. Kellerman, J. Lewis, and J. D. Laird (1989). Looking and loving: The effects of mutual gaze on feelings of romantic love, *Journal of research in personality*, vol. 23(2), pp. 145–161. Cited on pages 2, 31, and 145.

J. M. Kelley, G. Kraft-Todd, L. Schapira, J. Kossowsky, and H. Riess (2014). The Influence of the Patient-Clinician Relationship on Healthcare Outcomes: A Systematic Review and Meta-Analysis of Randomized Controlled Trials, *PloS one*, vol. 9(4), p. e94207. Cited on pages 3, 12, 23, and 96.

A. Kendon (1967). Some functions of gaze-direction in social interaction, *Acta psychologica*, vol. 26, pp. 22–63. Cited on pages 2, 26, 41, 42, 145, 146, and 147.

P. Kenny, T. D. Parsons, J. Gratch, A. Leuski, and A. A. Rizzo (2007). Virtual patients for clinical therapist skills training, in *International Workshop on Intelligent Virtual Agents 2007*. Cited on page 31.

D. Kern, P. Marshall, and A. Schmidt (2010). Gazemarks: Gaze-Based Visual Placeholders to Ease Attention Switching, in *Proc. of the ACM Conference on Human Factors in Computing Systems 2010*. Cited on page 124.

M. Khamis, F. Alt, and A. Bulling (2018). The Past, Present, and Future of Gaze-enabled Handheld Mobile Devices: Survey and Lessons Learned, in *Proc. of the International Conference on Human-Computer Interaction with Mobile Devices and Services 2018*. Cited on page 57.

J. Kickul and G. Neuman (2000). Emergent Leadership Behaviors: The Function of Personality and Cognitive Ability in Determining Teamwork Performance and KSAs, *Journal of Business and Psychology*, vol. 15(1), pp. 27–51. Cited on pages 3, 13, 24, and 113.

J. Kim and E. André (2008). Emotion recognition based on physiological changes in music listening, *IEEE transactions on pattern analysis and machine intelligence*, vol. 30(12), pp. 2067–2083. Cited on pages 11 and 29.

A. A. Kindiroglu, L. Akarun, and O. Aran (2017). Multi-domain and multi-task prediction of extraversion and leadership from meeting videos, *EURASIP Journal on Image and Video Processing*, vol. 2017(1), p. 77. Cited on pages 10 and 114.

D. E. King (2009). Dlib-ml: A Machine Learning Toolkit, *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758. Cited on page 131.

D. P. Kingma and J. Ba (2014). Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*. Cited on page 153.

P. Klasnja, S. Consolvo, T. Choudhury, R. Beckwith, and J. Hightower (2009). Exploring privacy concerns about personal sensing, in *International Conference on Pervasive Computing 2009*. Cited on page 31.

C. L. Kleinke (1986). Gaze and eye contact: a research review., *Psychological bulletin*, vol. 100(1), p. 78. Cited on pages 1 and 39.

A. Kleinsmith and N. Bianchi-Berthouze (2013). Affective body expression perception and recognition: A survey, *IEEE TAC*, vol. 4(1), pp. 15–33. Cited on page 81.

M. L. Knapp, J. A. Hall, and T. G. Horgan (2013). *Nonverbal communication in human interaction*, Cengage Learning. Cited on page 1.

M. Koelle, K. Wolf, and S. Boll (2018). Beyond LED Status Lights - Design Requirements of Privacy Notices for Body-worn Cameras, in *Proc. of the International Conference on Tangible, Embedded, and Embodied Interaction 2018*. Cited on pages 65 and 73.

M. Korayem, R. Templeman, D. Chen, D. Crandall, and A. Kapadia (2016). Enhancing Lifelogging Privacy by Detecting Screens, in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems 2016*. Cited on pages 65 and 73.

A. Krizhevsky, I. Sutskever, and G. E. Hinton (2012). Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems 2012*. Cited on page 28.

U. Kubasova, G. Murray, and M. Braley (2019). Analyzing Verbal and Nonverbal Features for Predicting Group Performance, *arXiv preprint arXiv:1907.01369*. Cited on page 10.

M. Kümmerer, T. S. Wallis, and M. Bethge (2016). DeepGaze II: Reading fixations from deep features trained on object recognition, *arXiv preprint arXiv:1610.01563*. Cited on page 148.

M. LaFrance and M. Broadbent (1976). Group Rapport: Posture Sharing as a Nonverbal Indicator, *Group & Organization Studies*, vol. 1(3), pp. 328–333. Cited on page 93.

C. Lander, M. Löchtefeld, and A. Krüger (2017). hEYEbrid: A Hybrid Approach for Mobile Calibration-free Gaze Estimation, *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1(4), pp. 149:1–149:29. Cited on pages 7 and 60.

R. Larson and M. Csikszentmihalyi (2014). The experience sampling method, in *Flow and the foundations of positive psychology 2014*, pp. 21–34, Springer. Cited on page 29.

K. Laskowski (2010). Modeling Norms of Turn-Taking in Multi-Party Conversation, in *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics 2010*. Cited on page 95.

P. J. Lavrakas (2008). *Encyclopedia of survey research methods*. Cited on page 100.

C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan (2011). Emotion recognition using a hierarchical binary decision tree approach, *Speech Communication*, vol. 53(9-10), pp. 1162–1171. Cited on page 11.

M. Leroy, E. Mathiot, and A. Morgenstern (2009). *Pointing gestures, vocalizations and gaze: two case studies*. Cited on page 147.

Y. Li, M. Liu, and J. 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 2018*. Cited on page 148.

F. Liu, C. Shen, and G. Lin (2015). Deep Convolutional Neural Fields for Depth Estimation from a Single Image, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2015*. Cited on page 131.

H. Liu, D. Xu, Q. Huang, W. Li, M. Xu, and S. Lin (2013). Semantically-Based Human Scanpath Estimation with HMMs, in *Proc. of the IEEE International Conference on Computer Vision 2013*. Cited on page 148.

M. Liu, S. Tang, Y. Li, and J. Rehg (2019). Forecasting Human Object Interaction: Joint Prediction of Motor Attention and Egocentric Activity, *arXiv preprint arXiv:1911.10967*. Cited on page 15.

Y. Liu, O. Sourina, and M. K. Nguyen (2011). Real-time EEG-based emotion recognition and its applications, in *Transactions on computational science XII 2011*, pp. 256–277, Springer. Cited on page 29.

F. Lu, Y. Sugano, T. Okabe, and Y. Sato (2012). Head pose-free appearance-based gaze sensing via eye image synthesis, in *Proc. of the International Conference on Pattern Recognition 2012*. Cited on page 147.

M. Ma, H. Fan, and K. M. Kitani (2016). Going Deeper into First-Person Activity Recognition, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2016*. Cited on page 142.

Y. Ma, X. Zhu, S. Zhang, R. Yang, W. Wang, and D. Manocha (2019). Trafficpredict: Trajectory prediction for heterogeneous traffic-agents, in *Proc. of the AAAI Conference on Artificial Intelligence 2019*. Cited on page 15.

N. A. Madzlan, J. G. Han, F. Bonin, and N. Campbell (2014). Automatic recognition of attitudes in video blogs - prosodic and visual feature analysis, in *Proc. of the Annual Conference of the International Speech Communication Association 2014*. Cited on page 91.

P. P. Maglio, T. Matlock, C. S. Campbell, S. Zhai, and B. A. Smith (2000). Gaze and Speech in Attentive User Interfaces, in *Proc. of the International Conference on Multimodal Interfaces 2000*. Cited on page 147.

S. Maji, L. Bourdev, and J. Malik (2011). Action Recognition from a Distributed Representation of Pose and Appearance, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2011*. Cited on page 82.

H. M. Mandalia and M. D. D. Salvucci (2005). Using support vector machines for lane-change detection, in *Proc. of the Human Factors and Ergonomics Society Annual Meeting 2005*. Cited on page 15.

A. Mariakakis, M. Goel, M. T. I. Aumi, S. N. Patel, and J. O. Wobbrock (2015). SwitchBack: Using Focus and Saccade Tracking to Guide Users' Attention for Mobile Task Resumption, in *Proc. of the ACM Conference on Human Factors in Computing Systems 2015*. Cited on page 124.

M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor (2017). ResnetCrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification, in *Proc. of the IEEE International Conference on Advanced Video and Signal Based Surveillance 2017*. Cited on page 10.

A. Mathur, N. D. Lane, and F. Kawsar (2016). Engagement-Aware Computing: Modelling User Engagement from Mobile Contexts, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing 2016*. Cited on pages 124 and 126.

G. McKeown, W. Curran, J. Wagner, F. Lingenfelser, and E. André (2015). The Belfast storytelling database: A spontaneous social interaction database with laughter focused annotation, in *Proc. of the International Conference on Affective Computing and Intelligent Interaction 2015*. Cited on page 95.

A. Mehrabian and S. R. Ferris (1967). Inference of attitudes from nonverbal communication in two channels., *Journal of consulting psychology*, vol. 31(3), p. 248. Cited on page 1.

M. Mehu and K. R. Scherer (2012). A psycho-ethological approach to social signal processing, *Cognitive processing*, vol. 13(2), pp. 397–414. Cited on page 5.

A. Metallinou, A. Katsamanis, and S. Narayanan (2013). Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information, *Image and Vision Computing*, vol. 31(2), pp. 137–152. Cited on pages 11, 22, and 31.

A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan (2011). Tracking changes in continuous emotion states using body language and prosodic cues, in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing 2011*. Cited on page 81.

A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan (2010). The USC CreativeIT database: a multimodal database of theatrical improvisation, *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality 18 May 2010*, p. 55. Cited on pages 3, 11, 77, 79, 81, and 169.

A. Metallinou and S. Narayanan (2013). Annotation and processing of continuous emotional attributes: Challenges and opportunities, in *Proc. of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition 2013*. Cited on page 86.

A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan (2012). Context-sensitive learning for enhanced audiovisual emotion classification, *IEEE Transactions on Affective Computing*, vol. 3(2), pp. 184–198. Cited on page 2.

M. Miettinen and A. Oulasvirta (2007). Predicting Time-Sharing in Mobile Interaction, *User Modeling and User-Adapted Interaction*, vol. 17(5), pp. 475–510. Cited on pages 124 and 126.

W. Min, B. W. Mott, J. P. Rowe, B. Liu, and J. C. Lester (2016). Player Goal Recognition in Open-World Digital Games with Long Short-Term Memory Networks, in *Proc. of the International Joint Conference on Artificial Intelligence 2016*. Cited on page 127.

S. Mohammadi, H. Kiani, A. Perina, and V. Murino (2015). Violence detection in crowded scenes using substantial derivative, in *Proc. of the IEEE International Conference on Advanced Video and Signal Based Surveillance 2015*. Cited on page 10.

S. Mohammadi, A. Perina, H. Kiani, and V. Murino (2016). Angry crowds: Detecting violent events in videos, in *Proc. of the European Conference on Computer Vision 2016*. Cited on page 10.

A. Moller, L. Roalter, S. Diewald, J. Scherr, M. Kranz, N. Hammerla, P. Olivier, and T. Plotz (2012). Gymskill: A personal trainer for physical exercises, in *Proc. Percom 2012*. Cited on page 82.

A. Morales, F. M. Costela, R. Tolosana, and R. L. Woods (2018). Saccade Landing Point Prediction: A Novel Approach based on Recurrent Neural Networks, in *Proc. of the International Conference on Machine Learning Technologies 2018*. Cited on page 16.

P. Müller and A. Bulling (2019). Emergent Leadership Detection Across Datasets, in *Proc. of the International Conference on Multimodal Interaction 2019*. Cited on pages 19, 145, and 147.

P. Müller, D. Buschek, M. X. Huang, and A. Bulling (2019). Reducing Calibration Drift in Mobile Eye Trackers by Exploiting Mobile Phone Usage, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications 2019*. Cited on page 19.

P. Müller, M. X. Huang, and A. Bulling (2018a). Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behavior, in *Proc. of the ACM International Conference on Intelligent User Interfaces 2018*. Cited on pages 9, 19, 25, 29, 41, 42, 43, 44, 55, 114, 115, 116, 117, 147, 167, and 170.

P. Müller, M. X. Huang, X. Zhang, and A. Bulling (2018b). 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 2018*. Cited on pages 19, 74, 116, 145, 147, 148, 153, 159, 161, and 163.

P. Müller, E. Sood, and A. Bulling (2020). Anticipating Averted Gaze in Dyadic Interactions, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications 2020*. Cited on page 19.

P. M. Müller, S. Amin, P. Verma, M. Andriluka, and A. 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 2015*. Cited on pages 19, 55, and 95.

M. Nanninga, Y. Zhang, N. Lehmann-Willenbrock, Z. Szlávik, and H. Hung (2017). Estimating Verbal Expressions of Task and Social Cohesion in Meetings by Quantifying Paralinguistic Mimicry, in *Proc. of the ACM International Conference on Multimedia Interaction 2017*. Cited on pages 10 and 95.

C. G. Narber, W. Lawson, and J. G. Trafton (2015). Anticipation of Touch Gestures to Improve Robot Reaction Time, in *2015 AAAI Fall Symposium Series 2015*. Cited on page 15.

E. Nasiopoulos, E. F. Risko, T. Foulsham, and A. Kingstone (2015). Wearable Computing: Will It Make People Prosocial?, *British Journal of Psychology*, vol. 106(2), pp. 209–216. Cited on page 142.

M. Obuchi, W. Sasaki, T. Okoshi, J. Nakazawa, and H. Tokuda (2016). Investigating Interruptibility at Activity Breakpoints Using Smartphone Activity Recognition API, in *Adj. Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing 2016*. Cited on pages 126 and 128.

OECD (2018). *Private Equity Investment in Artificial Intelligence*, `www.oecd.org/going-digital/ai/private-equity-investment-in-artificial-intelligence.pdf`. Cited on page 32.

C. Oertel, K. A. Funes Mora, J. Gustafson, and J.-M. Odobez (2015). Deciphering the Silent Participant: On the Use of Audio-Visual Cues for the Classification of Listener Categories in Group Discussions, in *Proc. of the ACM on International Conference on Multimodal Interaction 2015*. Cited on page 96.

C. Oertel and G. 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 2013. Cited on pages 10, 39, 43, and 95.

A. Ogan, S. L. Finkelstein, E. Walker, R. Carlson, and J. Cassell (2012). Rudeness and Rapport: Insults and Learning Gains in Peer Tutoring., in *Proc. of the International Conference on Intelligent Tutoring Systems 2012*. Cited on pages 30 and 96.

Y. Ohshima and A. Nakazawa (2019). Eye Contact Detection from Third Person Video, in *Asian Conference on Pattern Recognition 2019*. Cited on pages 8 and 9.

S. Okada, L. S. Nguyen, O. Aran, and D. Gatica-Perez (2019). Modeling Dyadic and Group Impressions with Intermodal and Interperson Features, *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15(1s), p. 13. Cited on pages 10 and 114.

F. J. Ordóñez and D. Roggen (2016). Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition, *Sensors*, vol. 16(1), p. 115. Cited on page 142.

K. Otsuka, K. Kasuga, and M. Köhler (2018). Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks, in *Proc. of the ACM International Conference on Multimodal Interaction 2018*. Cited on pages 8 and 21.

K. Otsuka, H. Sawada, and J. Yamato (2007). Automatic inference of cross-modal nonverbal interactions in multiparty conversations: " who responds to whom, when, and how?" from gaze, head gestures, and utterances, in *Proc. of the International Conference on Multimodal Interfaces 2007*. Cited on page 8.

K. Otsuka, Y. Takemae, and J. Yamato (2005). A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances, in *Proc. of the International Conference on Multimodal Interfaces 2005*. Cited on page 8.

K. Otsuka and J. Yamato (2008). Fast and robust face tracking for analyzing multiparty face-to-face meetings, in *International Workshop on Machine Learning for Multimodal Interaction 2008*. Cited on page 8.

A. Oulasvirta, T. Rattenbury, L. Ma, and E. Raita (2012). Habits Make Smartphone Use More Pervasive, *Personal and Ubiquitous Computing*, vol. 16(1), pp. 105–114. Cited on page 59.

A. Oulasvirta, S. Tamminen, V. Roto, and J. Kuorelahti (2005). Interaction in 4-Second Bursts: The Fragmented Nature of Attentional Resources in Mobile HCI, in *Proc. of the ACM Conference on Human Factors in Computing Systems 2005*. Cited on pages 58, 59, 124, 126, 132, 133, and 134.

L. Paletta, H. Neuschmied, M. Schwarz, G. Lodron, M. Pszeida, S. Ladstätter, and P. Luley (2014). Smartphone Eye Tracking Toolbox: Accurate Gaze Recovery on

Mobile Displays, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications 2014*. Cited on page 126.

S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang (2011). Domain Adaptation via Transfer Component Analysis, *IEEE Transactions on Neural Networks*, vol. 22(2), pp. 199–210. Cited on page 165.

J. W. Penney (2016). Chilling effects: Online surveillance and Wikipedia use, *Berkeley Tech. LJ*, vol. 31, p. 117. Cited on page 31.

A. Pentland (2010). *Honest signals: how they shape our world*, MIT press. Cited on page 1.

K. Pfeuffer, J. Alexander, M. K. Chong, Y. Zhang, and H. Gellersen (2015). Gaze-Shifting: Direct-Indirect Input with Pen and Touch Modulated by Gaze, in *Proc. of the ACM Symposium on User Interface Software and Technology 2015*. Cited on pages 57 and 72.

K. Pfeuffer and H. Gellersen (2016). Gaze and Touch Interaction on Tablets, in *Proc. of the ACM Symposium on User Interface Software and Technology 2016*. Cited on pages 57 and 72.

R. W. Picard (1995). Affective computing. Cited on page 42.

P. E. Pidcoe and P. A. Wetzel (2006). Oculomotor tracking strategy in normal subjects with and without simulated scotoma, *Investigative ophthalmology & visual science*, vol. 47(1), pp. 169–178. Cited on page 16.

M. Pielot, B. Cardoso, K. Katevas, J. Serrà, A. Matic, and N. Oliver (2017). Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users, *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1(3), p. 91. Cited on page 126.

M. Pielot, T. Dingler, J. S. Pedro, and N. Oliver (2015). When Attention Is Not Scarce – Detecting Boredom from Mobile Phone Usage, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing 2015*. Cited on pages 59, 124, and 126.

L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele (2013). Strong Appearance and Expressive Spatial Models for Human Pose Estimation, in *Proc. of the IEEE International Conference on Computer Vision 2013*. Cited on page 82.

J. Platt (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, *Advances in Large Margin Classifiers*, vol. 10(3), pp. 61–74. Cited on page 117.

F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford (2001). Perceiving affect from arm movement, *Cognition*, vol. 82(2), pp. B51–B61. Cited on pages 11 and 81.

T. Randhavane, A. Bera, K. Kapsaskis, U. Bhattacharya, K. Gray, and D. Manocha (2019). Identifying emotions from walking using affective and deep features, *arXiv preprint arXiv:1906.11884*. Cited on page 11.

H. C. Ravichandar and A. P. Dani (2017). Human Intention Inference Using Expectation-Maximization Algorithm with Online Model Learning, *IEEE Transactions on Automation Science and Engineering*, vol. 14(2), pp. 855–868. Cited on page 127.

A. Recasens, A. Khosla, C. Vondrick, and A. Torralba (2015). Where are they looking?, in *Advances in Neural Information Processing Systems 2015*. Cited on pages 8, 9, and 42.

A. Recasens, C. Vondrick, A. Khosla, and A. Torralba (2017). Following gaze in video, in *Proc. of the IEEE International Conference on Computer Vision 2017*. Cited on page 9.

S. Ren, K. He, R. Girshick, and J. Sun (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, in *Proc. of the Advances in Neural Information Processing Systems (NIPS) 2015*. Cited on page 131.

F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions, in *Proc. of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition 2013*. Cited on page 95.

E. F. Risko and A. Kingstone (2011). Eyes wide shut: implied social presence, eye tracking and attention, *Attention, Perception, & Psychophysics*, vol. 73(2), pp. 291–296. Cited on pages 31, 42, and 142.

H. Ritschel, A. Seiderer, K. Janowski, S. Wagner, and E. André (2019). Adaptive linguistic style for an assistive robotic health companion based on explicit human feedback, in *Proc. of the ACM International Conference on Pervasive Technologies Related to Assistive Environments 2019*. Cited on page 30.

T. Romaniuk (2009). The 'Clinton Cackle': Hillary Rodham Clinton's Laughter in News Interviews, *Crossroads of Language, Interaction, and Culture*, vol. 7, pp. 17–49. Cited on page 147.

F. Rossano (2013). Gaze in Conversation, *The handbook of conversation analysis*, p. 308. Cited on pages 28, 147, and 159.

J. S. Rubinstein, D. E. Meyer, and J. E. Evans (2001). Executive Control of Cognitive Processes in Task Switching, *Journal of Experimental Psychology: Human Perception and Performance*, vol. 27(4), p. 763. Cited on page 124.

C. Rühlemann and S. Gries (2015). Turn order and turn distribution in multi-party storytelling, *Journal of Pragmatics*, vol. 87, pp. 171–191. Cited on page 95.

A. Sahami Shirazi, N. Henze, T. Dingler, M. Pielot, D. Weber, and A. Schmidt (2014). Large-scale Assessment of Mobile Notifications, in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems 2014*. Cited on pages 59 and 72.

H. Sakoe and S. Chiba (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26(1), pp. 43–49. Cited on page 104.

T. A. Salthouse and C. L. Ellis (1980). Determinants of Eye-Fixation Duration, *The American journal of psychology*, pp. 207–234. Cited on page 154.

D. Sanchez-Cortes, O. Aran, D. B. Jayagopi, M. S. Mast, and D. Gatica-Perez (2013). Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition, *Journal on Multimodal User Interfaces*, vol. 7(1-2), pp. 39–53. Cited on pages 13, 114, 116, 117, and 164.

D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez (2012). A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups, *IEEE Transactions on Multimedia*, vol. 14(3), pp. 816–832. Cited on pages 3, 10, 13, 24, 28, 29, 95, 99, 110, 113, 114, 116, and 117.

S. Santarcangelo and K. Dyer (1988). Prosodic aspects of motherese: Effects on gaze and responsiveness in developmentally disabled children, *Journal of Experimental Child Psychology*, vol. 46(3), pp. 406–418. Cited on page 147.

T. Santini, D. C. Niehorster, and E. Kasneci (2019). Get a grip: slippage-robust and glint-free gaze estimation for real-time pervasive head-mounted eye tracking, in *Proc. of the ACM Symposium on Eye Tracking Research & Applications 2019*. Cited on page 7.

G. Saponaro, G. Salvi, and A. Bernardino (2013). Robot anticipation of human intentions through continuous gesture recognition, in *Proc. of the International Conference on Collaboration Technologies and Systems 2013*. Cited on page 15.

S. Sato and J. I. Kawahara (2015). Attentional Capture by Completely Task-Irrelevant Faces, *Psychological Research*, vol. 79(4), pp. 523–533. Cited on page 131.

H. Sattar, S. Müller, M. Fritz, and A. Bulling (2015). Prediction of Search Targets from Fixations in Open-World Settings, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2015*. Cited on page 6.

K. R. Scherer and T. Bänziger (2010). On the use of actor portrayals in research on emotional expression, in *Blueprint for affective computing: A sourcebook 2010*. Cited on pages 81 and 82.

G. Schiavo, A. Cappelletti, E. Mencarini, O. Stock, and M. Zancanaro (2014). Overt or Subtle? Supporting Group Conversations with Automatically Targeted Directives, in *Proc. of the ACM International Conference on Intelligent User Interfaces 2014*. Cited on pages 32, 55, 111, and 160.

B. Schuller, G. Rigoll, and M. Lang (2003). Hidden Markov model-based speech emotion recognition, in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing 2003*. Cited on page 11.

M. S. Schwartz and F. Andrasik (2017). *Biofeedback: A practitioner's guide*, Guilford Publications. Cited on page 160.

C. Schwedes and D. Wentura (2012). The revealing glance: Eye gaze behavior to concealed information, *Memory & cognition*, vol. 40(4), pp. 642–651. Cited on page 6.

P. Schydlo, M. Rakovic, L. Jamone, and J. Santos-Victor (2018). Anticipation in human-robot cooperation: A recurrent neural network approach for multiple action sequences prediction, in *proc. of the ieee international conference on robotics and automation 2018*. Cited on page 15.

D. W. Scott (2015). *Multivariate density estimation: theory, practice, and visualization*, John Wiley & Sons. Cited on page 48.

D. Sculley (2010). Web-scale k-means clustering, in *Proc. of the International Conference on World Wide Web 2010*. Cited on page 64.

T. Selker, A. Lockerd, and J. Martinez (2001). Eye-R, a glasses-mounted eye motion detection interface, in *Proc. of the ACM Extended Abstracts on Human Factors in Computing Systems 2001*. Cited on page 42.

A. Senju and M. H. Johnson (2009). The eye contact effect: mechanisms and development, *Trends in cognitive sciences*, vol. 13(3), pp. 127–134. Cited on pages 4 and 146.

P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun (2014). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks, in *Proc. of the International Conference on Learning Representations 2014*. Cited on pages 23, 82, and 88.

J. S. Shell, R. Vertegaal, D. Cheng, A. W. Skaburskis, C. Sohn, A. J. Stewart, O. Aoudeh, and C. Dickie (2004). ECSGlasses and EyePliances: using attention to open sociable windows of interaction, in *Proc. of the ACM Symposium on Eye Tracking Research & Applications 2004*. Cited on page 42.

J. S. Shell, R. Vertegaal, and A. 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 2003*. Cited on pages 39 and 42.

Y. Shen, B. Ni, Z. Li, and N. Zhuang (2018). Egocentric activity prediction via event modulated attention, in *Proc. of the European Conference on Computer Vision 2018*. Cited on pages 2 and 15.

R. Siegfried, Y. Yu, and J.-M. Odobez (2017). Towards the Use of Social Interaction Conventions as Prior for Gaze Model Adaptation, in *Proc. of ACM International Conference on Multimodal Interaction 2017*. Cited on pages 40, 42, 47, and 74.

D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.* (2017). Mastering the game of go without human knowledge, *Nature*, vol. 550(7676), pp. 354–359. Cited on pages 1 and 32.

G. Skantze, A. Hjalmarsson, and C. Oertel (2014). Turn-taking, feedback and joint attention in situated human–robot interaction, *Speech Communication*, vol. 65, pp. 50–66. Cited on page 4.

B. A. Smith, Q. Yin, S. K. Feiner, and S. 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 2013*. Cited on pages 39, 40, 42, and 148.

J. D. Smith, R. Vertegaal, and C. Sohn (2005). ViewPointer: lightweight calibration-free eye tracking for ubiquitous handsfree deixis, in *Proc. of the ACM Symposium on User Interface Software and Technology 2005*. Cited on pages 39 and 42.

H. Soo Park and J. Shi (2015). Social Saliency Prediction, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2015*. Cited on page 8.

J. Steil and A. Bulling (2015). Discovery of Everyday Human Activities from Long-Term Visual Behaviour Using Topic Models, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing 2015*. Cited on page 130.

J. Steil, I. Hagestedt, M. X. Huang, and A. Bulling (2019a). Privacy-Aware Eye Tracking Using Differential Privacy, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications 2019*. Cited on pages 2 and 31.

J. Steil, M. Koelle, W. Heuten, S. Boll, and A. Bulling (2018a). PrivacEye: Privacy-Preserving First-Person Vision Using Image Features and Eye Movement Analysis, Technical report. Cited on pages 59, 60, 65, 69, and 73.

J. Steil, M. Koelle, W. Heuten, S. Boll, and A. Bulling (2019b). PrivacEye: Privacy-Preserving Head-Mounted Eye Tracking Using Egocentric Scene Image and Eye Movement Features, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications 2019*. Cited on pages 22 and 31.

J. Steil, P. Müller, Y. Sugano, and A. Bulling (2018b). 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 2018*. Cited on pages 19, 58, 59, 60, 61, 65, 70, 72, 146, 149, 161, and 168.

R. T. Stein and T. Heller (1979). An empirical analysis of the correlations between leadership status and participation rates reported in the literature., *Journal of Personality and Social Psychology*, vol. 37(11), pp. 1993–2002. Cited on pages 3, 13, and 113.

B. Stephens-Fripp, F. Naghdy, D. Stirling, and G. Naghdy (2017). Automatic affect perception based on body gait and posture: A survey, *International Journal of Social Robotics*, vol. 9(5), pp. 617–641. Cited on page 11.

R. Stiefelhagen (2002). Tracking Focus of Attention in Meetings, in *Proc. of the IEEE International Conference on Multimodal Interfaces 2002*. Cited on pages 8 and 42.

J. Streeck (1993). Gesture as communication I: Its coordination with gaze and speech, *Communications Monographs*, vol. 60(4), pp. 275–299. Cited on page 147.

Y. Sugano and A. Bulling (2015a). Self-Calibrating Head-Mounted Eye Trackers Using Egocentric Visual Saliency, in *Proc. of the ACM Symposium on User Interface Software and Technology 2015*. Cited on pages 2, 6, 7, and 22.

Y. Sugano and A. Bulling (2015b). Self-Calibrating Head-Mounted Eye Trackers Using Egocentric Visual Saliency, in *Proc. of the ACM Symposium on User Interface Software and Technology 2015*. Cited on pages 57, 58, 59, 60, 62, 63, 64, 65, 66, 68, 69, 70, and 168.

Y. Sugano, Y. Matsushita, and Y. Sato (2010). Calibration-free gaze sensing using saliency maps, in *Proc. of the IEEE Conference on Computer Wision and Pattern Recognition 2010*. Cited on pages 7, 60, and 64.

Y. Sugano, Y. Matsushita, Y. Sato, and H. Koike (2008). An Incremental Learning Method for Unconstrained Gaze Estimation, in *Proc. of the European Conference on Computer Vision 2008*. Cited on pages 7 and 60.

Y. Sugano, X. Zhang, and A. Bulling (2016). AggreGaze: Collective Estimation of Audience Attention on Public Displays, in *Proc. of the ACM Symposium on User Interface Software and Technology 2016*. Cited on page 140.

B. Sun, J. Feng, and K. Saenko (2016). Return of frustratingly easy domain adaptation., in *Proc. of the AAAI Conference on Artificial Intelligence 2016*. Cited on page 165.

L. Swirski and N. Dodgson (2013). A fully-automatic, temporal approach to single camera, glint-free 3d eye model fitting, in *Proc. of the European Conference on Eye Movements 2013*. Cited on page 60.

C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi (2017). Inception-v4, inception-resnet and the impact of residual connections on learning, in *Proc. of the AAAI Conference on Artificial Intelligence 2017*. Cited on page 32.

T. Takegami, T. Gotoh, and G. Ohyama (2002). An algorithm for an eye tracking system with self-calibration, *Systems and computers in Japan*, vol. 33(10), pp. 10–20. Cited on page 7.

K. Takemura, S. Kimura, and S. Suda (2014a). Estimating point-of-regard using corneal surface image, in *Proc. of the ACM Symposium on Eye Tracking Research and Applications 2014*. Cited on pages 7 and 60.

K. Takemura, T. Yamakawa, J. Takamatsu, and T. Ogasawara (2014b). Estimation of a focused object using a corneal surface image for eye-based interaction, *Journal of Eye Movement Research*, vol. 7(3), pp. 1–9. Cited on page 60.

B. Tessendorf, F. Gravenhorst, B. Arnrich, and G. Tröster (2011). An imu-based sensor network to continuously monitor rowing technique on the water, in *Proc. of the IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing 2011*. Cited on page 82.

J. Theeuwes, A. F. Kramer, S. Hahn, and D. E. Irwin (1998). Our eyes do not always go where we want them to go: Capture of the eyes by new objects, *Psychological Science*, vol. 9(5), pp. 379–385. Cited on page 6.

L. Tickle-Degnen and R. Rosenthal (1990). The Nature of Rapport and Its Nonverbal Correlates, *Psychological Inquiry*, vol. 1(4), pp. 285–293. Cited on pages 10, 12, 96, 102, 104, and 109.

J. Tompson, A. Jain, Y. LeCun, and C. Bregler (2014). Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2014*. Cited on page 82.

M. Tonsen, J. Steil, Y. Sugano, and A. 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*, vol. 1(3), pp. 106:1–106:21. Cited on pages 142 and 147.

D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri (2015). Learning Spatiotemporal Features with 3D Convolutional Networks, in *Proc. of the IEEE International Conference on Computer Vision 2015*. Cited on page 142.

P. Tsui and G. L. Schultz (1985). Failure of rapport: Why psychotherapeutic engagement fails in the treatment of Asian clients., *American Journal of Orthopsychiatry*, vol. 55(4), pp. 561–569. Cited on pages 3, 12, 23, and 96.

L. D. Turner, S. M. Allen, and R. M. Whitaker (2015). Interruptibility Prediction for Ubiquitous Systems: Conventions and New Directions from a Growing Field, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing 2015*. Cited on page 126.

L. D. Turner, S. M. Allen, and R. M. Whitaker (2017). Reachable but Not Receptive: Enhancing Smartphone Interruptibility Prediction by Modelling the Extent of User Engagement with Notifications, *Pervasive and Mobile Computing*, vol. 40, pp. 480–494. Cited on page 126.

G. Urh and V. Pejović (2016). TaskyApp: Inferring Task Engagement via Smartphone Sensing, in *Adj. Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing 2016*. Cited on pages 124 and 126.

V. Vaitukaitis and A. Bulling (2012). Eye Gesture Recognition on Portable Devices, in *Proc. of the International Workshop on Pervasive Eye Tracking and Mobile Gaze-Based Interaction 2012*. Cited on page 126.

R. Valenti, N. Sebe, and T. Gevers (2011). Combining Head Pose and Eye Location Information for Gaze Estimation, *IEEE Transactions on Image Processing*, vol. 21(2), pp. 802–815. Cited on pages 131 and 147.

P. Vamplew and A. Adams (1995). Recognition and anticipation of hand motions using a recurrent neural network. Cited on page 15.

M. Van Bommel, J.-W. Van Prooijen, H. Elffers, and P. A. van Lange (2014). Intervene to be seen: The power of a camera in attenuating the bystander effect, *Social Psychological and Personality Science*, vol. 5(4), pp. 459–466. Cited on page 31.

T. van Laarhoven and E. Marchiori (2017). Unsupervised Domain Adaptation with Random Walks on Target Labelings, *arXiv preprint arXiv:1706.05335*. Cited on page 165.

S. Van Vuuren and L. R. Cherney (2014). A virtual therapist for speech and language therapy, in *Proc. of the International Conference on Intelligent Virtual Agents 2014*. Cited on page 32.

A. Veenstra and H. Hung (2011). Do they like me? Using video cues to predict desires during speed-dates, in *Proc. of the IEEE International Conference on Computer Vision Workshops 2011*. Cited on page 11.

E. Velloso, A. Bulling, and H. Gellersen (2011). Towards Qualitative Assessment of Weight Lifting Exercises Using Body-Worn Sensors, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing 2011*. Cited on page 82.

E. Velloso, A. Bulling, and H. Gellersen (2013a). AutoBAP: Automatic Coding of Body Action and Posture Units from Wearable Sensors, in *Proc. of the International Conference on Affective Computing and Intelligent Interaction 2013*. Cited on page 81.

E. Velloso, A. Bulling, and H. Gellersen (2013b). MotionMA: Motion Modelling and Analysis by Demonstration, in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems 2013*. Cited on page 82.

E. Velloso, A. Bulling, H. Gellersen, W. Ugulino, and H. Fuks (2013c). Qualitative Activity Recognition of Weight Lifting Exercises, in *Proc. of the Augmented Human International Conference 2013*. Cited on page 82.

R. Vertegaal, R. Slagter, G. van der Veer, and A. Nijholt (2001). Eye Gaze Patterns in Conversations: There is More to Conversational Agents Than Meets the Eyes, in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems 2001*. Cited on pages 40, 41, and 42.

A. Vinciarelli and G. Mohammadi (2014). A Survey of Personality Computing, *IEEE Transactions on Affective Computing*, vol. 5(3), pp. 273–291. Cited on page 111.

A. Vinciarelli, M. Pantic, and H. Bourlard (2009). Social signal processing: Survey of an emerging domain, *Image and Vision Computing*, vol. 27(12), pp. 1743–1759. Cited on pages 5, 33, and 81.

T. Vogt, E. André, and J. Wagner (2008). Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation, in *Affect and emotion in human-computer interaction 2008*, pp. 75–91, Springer. Cited on page 11.

H. Vrzakova, R. Bednarik, Y. I. Nakano, and F. Nihei (2016). Speakers' Head and Gaze Dynamics Weakly Correlate in Group Conversation, in *Proc. of the ACM Symposium on Eye Tracking Research & Applications 2016*. Cited on pages 8, 40, 42, 51, and 55.

B. N. Waber, D. Olguin Olguin, T. Kim, A. Mohan, K. Ara, and A. Pentland (2007). Organizational engineering using sociometric badges, *Available at SSRN 1073342*. Cited on page 10.

J. Walker, K. Marino, A. Gupta, and M. Hebert (2017). The pose knows: Video forecasting by generating pose futures, in *Proc. of the IEEE International Conference on Computer Vision 2017*. Cited on page 15.

H. G. Wallbott (1998). Bodily expression of emotion, *European journal of social psychology*, vol. 28(6), pp. 879–896. Cited on pages 11 and 81.

H. Wang, A. Kläser, C. Schmid, and C.-L. Liu (2013a). Dense trajectories and motion boundary descriptors for action recognition, *International Journal of Computer Vision*, vol. 103(1), pp. 60–79. Cited on pages 23, 80, 82, 87, and 89.

L. Wang, S. Guo, W. Huang, and Y. Qiao (2015a). Places205-VGGNet Models for Scene Recognition, *arXiv preprint arXiv:1508.01667*. Cited on page 131.

N. Wang and J. Gratch (2009). Rapport and Facial Expression, in *Proc. of the International Conference on Affective Computing and Intelligent Interaction Workshops 2009*. Cited on pages 2, 10, 12, 93, 96, and 109.

S. Wang, R. L. Woods, F. M. Costela, and G. Luo (2017). Dynamic gaze-position prediction of saccadic eye movements using a Taylor series, *Journal of vision*, vol. 17(14), pp. 3–3. Cited on page 16.

W. Wang, G. Athanasopoulos, G. Patsis, V. Enescu, and H. Sahli (2014). Real-time emotion recognition from natural bodily expressions in child-robot interaction, in *Proc. of the European Conference on Computer Vision 2014*. Cited on page 11.

W. Wang, V. Enescu, and H. Sahli (2013b). Towards real-time continuous emotion recognition from body movements, in *Human Behavior Understanding 2013*, pp. 235–245. Cited on page 81.

W. Wang, V. Enescu, and H. Sahli (2015b). Adaptive real-time emotion recognition from body movements, *ACM Transactions on Interactive Intelligent Systems*, vol. 5(4), pp. 1–21. Cited on pages 11 and 22.

D. Weber and S. Mayer (2014). *LogEverything*, `https://github.com/hcilab-org/LogEverything/`. Cited on page 133.

K. Weber, H. Ritschel, I. Aslan, F. Lingenfelser, and E. André (2018). How to shape the humor of a robot-social behavior adaptation based on reinforcement learning, in *Proc. of the ACM International Conference on Multimodal Interaction 2018*. Cited on page 30.

M. Wiggers (1982). Judgments of facial expressions of emotion predicted from facial behavior, *Journal of Nonverbal Behavior*, vol. 7(2), pp. 101–116. Cited on page 10.

H. Woo, Y. Ji, H. Kono, Y. Tamura, Y. Kuroda, T. Sugano, Y. Yamamoto, A. Yamashita, and H. Asama (2017). Lane-change detection based on vehicle-trajectory prediction, *IEEE Robotics and Automation Letters*, vol. 2(2), pp. 1109–1116. Cited on page 15.

E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling (2015). Rendering of eyes for eye-shape registration and gaze estimation, in *Proc. of the IEEE International Conference on Computer Vision 2015*. Cited on page 147.

E. Wood and A. Bulling (2014). EyeTab: Model-Based Gaze Estimation on Unmodified Tablet Computers, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications 2014*. Cited on pages 124, 126, 128, and 140.

D. Wyatt, T. Choudhury, and J. Bilmes (2007a). Conversation detection and speaker segmentation in privacy-sensitive situated speech data, in *Proc. of the Annual Conference of the International Speech Communication Association 2007*. Cited on page 31.

D. Wyatt, T. Choudhury, and H. Kautz (2007b). Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort, in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2007*. Cited on page 31.

P. Xu, Y. Sugano, and A. 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 2016*. Cited on page 148.

Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao (2018). Gaze prediction in dynamic 360 immersive videos, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2018*. Cited on pages 16 and 26.

K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki (2011). Attention Prediction in Egocentric Video Using Motion and Visual Saliency, in *Proc. of the Pacific-Rim Symposium on Image and Video Technology 2011*. Cited on page 127.

H. Yamazoe, A. Utsumi, T. Yonezawa, and S. 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 2008*. Cited on pages 7, 60, and 147.

Y. Yang, S. Baker, A. Kannan, and D. Ramanan (2012). Recognizing proxemics in personal photos, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2012*. Cited on page 82.

Y. Yang and D. Ramanan (2013). Articulated Human Detection with Flexible Mixtures of Parts, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35. Cited on page 82.

Z. Yang, A. Metallinou, E. Erzin, and S. Narayanan (2014a). Analysis of interaction attitudes using data-driven hand gesture phrases, in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing 2014*. Cited on page 11.

Z. Yang, A. Metallinou, and S. Narayanan (2014b). Analysis and Predictive Modeling of Body Language Behavior in Dyadic Interactions From Multimodal Interlocutor Cues, *IEEE Transactions on Multimedia*, vol. 16(6), pp. 1766–1778. Cited on page 81.

Z. Ye, Y. Li, Y. Liu, C. Bridges, A. Rozga, and J. 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 2015*. Cited on page 42.

Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31(1), pp. 39–58. Cited on page 81.

L. Zhang, M. Morgan, I. Bhattacharya, M. Foley, J. Braasch, C. Riedl, B. Foucault Welles, and R. J. Radke (2019a). Improved Visual Focus of Attention Estimation and Prosodic Features for Analyzing Group Interactions, in *Proc. of the ACM International Conference on Multimodal Interaction 2019*. Cited on pages 8 and 21.

M. Zhang, K. T. Ma, J. H. Lim, Q. Zhao, and J. Feng (2018a). Anticipating where people will look using adversarial networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41(8), pp. 1783–1796. Cited on pages 16 and 26.

M. Zhang, K. Teck Ma, J. Hwee Lim, Q. Zhao, and J. Feng (2017a). Deep Future Gaze: Gaze Anticipation on Egocentric Videos Using Adversarial Networks, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2017*. Cited on pages 2, 16, 26, 142, 146, and 149.

N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev (2014). PANDA: Pose Aligned Networks for Deep Attribute Modeling, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2014*. Cited on page 82.

X. Zhang, M. X. Huang, Y. Sugano, and A. Bulling (2018b). Training Person-Specific Gaze Estimators from User Interactions with Multiple Devices, in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems 2018*. Cited on pages 7, 60, and 73.

X. Zhang, Y. Sugano, and A. Bulling (2017b). Everyday Eye Contact Detection Using Unsupervised Gaze Target Discovery, in *Proc. of the ACM Symposium on User Interface Software and Technology 2017*. Cited on pages 2, 3, 21, 39, 40, 41, 42, 43, 45, 47, 49, 50, 51, 53, 55, 145, 148, and 161.

X. Zhang, Y. Sugano, M. Fritz, and A. Bulling (2015). Appearance-Based Gaze Estimation in the Wild, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2015*. Cited on pages 39, 82, and 140.

X. Zhang, Y. Sugano, M. Fritz, and A. Bulling (2017c). 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 2017*. Cited on pages 21, 39, 41, 42, 45, 46, 47, 49, 53, and 167.

X. Zhang, Y. Sugano, M. Fritz, and A. Bulling (2018c). MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Cited on pages 2, 39, 42, 45, and 140.

X. Zhang, Y. Sugano, M. Fritz, and A. Bulling (2019b). MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41(1), pp. 162–175. Cited on page 147.

Y. Zhang, J. Beskow, and H. Kjellström (2017d). Look but Don't Stare: Mutual Gaze Interaction in Social Robots, in *Proc. of the International Conference on Social Robotics 2017*. Cited on page 161.

Y. Zhang, J. Olenick, C.-H. Chang, S. W. Kozlowski, and H. Hung (2018d). The I in team: Mining personal social interaction routine with topic models from long-term team data, in *Proc. of the ACM International Conference on Intelligent User Interfaces 2018*. Cited on page 10.

Z. Zhang, D. J. Crandall, C. Yu, and S. Bambach (2018e). 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 2018*. Cited on page 148.

R. Zhao, A. Papangelis, and J. Cassell (2014). Towards a Dyadic Computational Model of Rapport Management for Human-Virtual Agent Interaction, in *Proc. of the International Conference on Intelligent Virtual Agents 2014*. Cited on pages 12 and 97.

R. Zhao, T. Sinha, A. W. Black, and J. Cassell (2016). Socially-Aware Virtual Agents: Automatically Assessing Dyadic Rapport from Temporal Patterns of Behavior, in

Proc. of the International Conference on Intelligent Virtual Agents 2016. Cited on pages 12, 93, and 97.

S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr (2015). Conditional Random Fields as Recurrent Neural Networks, in *Proc. of the IEEE International Conference on Computer Vision 2015*. Cited on page 131.

S.-h. Zhong, Y. Liu, T.-Y. Ng, and Y. Liu (2016). Perception-Oriented Video Saliency Detection via Spatio-Temporal Attention Analysis, *Neurocomputing*, vol. 207, pp. 178–188. Cited on page 127.

X. Zhu and Z. Ghahramani (2002). Learning from labeled and unlabeled data with label propagation, Technical report. Cited on page 165.

X. Zhu and D. Ramanan (2012). Face Detection, Pose Estimation, and Landmark Localization in the Wild, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition 2012*. Cited on page 82.

M. Zuckerman, B. M. DePaulo, and R. Rosenthal (1981). Verbal and Nonverbal Communication of Deception, in *Advances in experimental social psychology 1981*, vol. 14, pp. 1–59, Elsevier. Cited on page 147.