Mobile Eye Tracking for Everyone

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

by

Julian Steil, M.Sc.

Saarbrücken
2019

Day of Colloquium	8th of November, 2019
Dean of the Faculty	Prof. Dr. Sebastian Hack

Examination Committee

Chair	Prof. Dr. Vera Demberg
Reviewer, Advisor	Prof. Dr. Andreas Bulling
Reviewer	Prof. Dr. Antonio Krüger
Reviewer	Prof. Dr. Enkelejda Kasneci
Academic Assistant	Dr. Florian Daiber

To My Father

Abstract

EYE tracking and gaze-based human-computer interfaces have become a practical modality in desktop settings, since remote eye tracking is efficient and affordable. However, remote eye tracking remains constrained to indoor, laboratory-like conditions, in which lighting and user position need to be controlled. Mobile eye tracking has the potential to overcome these limitations and to allow people to move around freely and to use eye tracking on a daily basis during their everyday routine. However, mobile eye tracking currently faces two fundamental challenges that prevent it from being practically usable and that, consequently, have to be addressed before mobile eye tracking can truly be used by everyone: Mobile eye tracking needs to be advanced and made fully functional in unconstrained environments, and it needs to be made socially acceptable.

Numerous sensing and analysis methods were initially developed for remote eye tracking and have been successfully applied for decades. Unfortunately, these methods are limited in terms of functionality and correctness, or even unsuitable for application in mobile eye tracking. Therefore, the majority of fundamental definitions, eye tracking methods, and gaze estimation approaches cannot be borrowed from remote eye tracking without adaptation. For example, the definitions of specific eye movements, like classical fixations, need to be extended to mobile settings where natural user and head motion are omnipresent. Corresponding analytical methods need to be adjusted or completely reimplemented based on novel approaches encoding the human gaze behaviour.

Apart from these technical challenges, an entirely new, and yet under-explored, topic required for the breakthrough of mobile eye tracking as everyday technology is the overcoming of social obstacles. A first crucial key issue to defuse social objections is the building of acceptance towards mobile eye tracking. Hence, it is essential to replace the bulky appearance of current head-mounted eye trackers with an unobtrusive, appealing, and trendy design. The second high-priority theme of increasing importance for everyone is privacy and its protection, given that research and industry have not focused on or taken care of this problem at all. To establish true confidence, future devices have to find a fine balance between protecting users' and bystanders' privacy and attracting and convincing users of their necessity, utility, and potential with useful and beneficial features. The solution of technical challenges and social obstacles is the prerequisite for the development of a variety of novel and exciting applications in order to establish mobile eye tracking as a new paradigm, which ease our everyday life.

This thesis addresses core technical challenges of mobile eye tracking that currently prevent it from being widely adopted. Specifically, this thesis proves that 3D data used for the calibration of mobile eye trackers improves gaze estimation and significantly reduces the parallax error. Further, it presents the first effective fixation detection method for head-mounted devices that is robust against the prevalence of user and gaze target motion.

In order to achieve social acceptability, this thesis proposes an innovative and unobtrusive design for future mobile eye tracking devices and builds the first prototype with fully frame-embedded eye cameras combined with a calibration-free deep-trained appearance-based gaze estimation approach. To protect users' and bystanders' privacy in the presence of head-mounted eye trackers, this thesis presents another first-of-its-kind prototype. It is able to identify privacy-sensitive situations to automatically enable and disable the eye tracker's first-person camera by means of a mechanical shutter, leveraging the combination of deep scene and eye movement features.

Nevertheless, solving technical challenges and social obstacles alone is not sufficient to make mobile eye tracking attractive for the masses. The key to success is the development of convincingly useful, innovative, and essential applications. To extend the protection of users' privacy on the software side as well, this thesis presents the first privacy-aware VR gaze interface using differential privacy. This method adds noise to recorded eye tracking data so that privacy-sensitive information like a user's gender or identity is protected without impeding the utility of the data itself. In addition, the first large-scale online survey is conducted to understand users' concerns with eye tracking.

To develop and evaluate novel applications, this thesis presents the first publicly available long-term eye tracking datasets. They are used to show the unsupervised detection of users' activities from eye movements alone using novel and efficient video-based encoding approaches as well as to propose the first proof-of-concept method to forecast users' attentive behaviour during everyday mobile interactions from phone-integrated and body-worn sensors. This opens up possibilities for the development of a variety of novel and exciting applications.

With more advanced features, accompanied by technological progress and sensor miniaturisation, eye tracking is increasingly integrated into conventional glasses as well as virtual and augmented reality (VR/AR) head-mounted displays, becoming an integral component of mobile interfaces. This thesis paves the way for the development of socially acceptable, privacy-aware, but highly functional mobile eye tracking devices and novel applications, so that mobile eye tracking can develop its full potential to become an everyday technology for everyone.

Zusammenfassung

Seitdem externe, am Bildschirm befestigte Eyetracker effizient und erschwinglich sind, haben sich Eye-Tracking- und blickbasierte Mensch-Maschine Schnittstellen zu einer verlässlichen Eingabemodalität für Desktopanwendungen entwickelt. Allerdings können solche Eyetracker nur innerhalb von Gebäuden, unter nahezu laborähnlichen Bedingungen, in denen die Körperhaltung der Anwender und die Lichtverhältnisse einer ständigen Kontrolle unterliegen, eingesetzt werden. Mobiles Eye-Tracking hat das Potenzial diese Einschränkungen aufzuheben, wobei es den Nutzern ermöglicht, sich frei zu bewegen und Eye-Tracking im Alltag zu verwenden. Dem mobilen Eye-Tracking stehen derzeit jedoch zwei grundlegende Herausforderungen gegenüber, die seine praktische Verwendung verhindern: Mobiles Eye-Tracking muss weiterentwickelt werden, um in jeder Umgebung voll funktionsfähig zu sein und um gesellschaftlich akzeptiert zu werden. Damit mobiles Eye-Tracking von jedem genutzt werden kann, müssen diese Herausforderungen jedoch zuvor bewältigt werden.

Zahlreiche Aufnahme- und Analysemethoden wurden ursprünglich für das externe Eye-Tracking entwickelt und werden seit Jahrzehnten erfolgreich angewandt. Diese Methoden sind jedoch in ihrer Funktionalität und Genauigkeit beschränkt bzw. für die Anwendung im mobilen Eye-Tracking ungeeignet. Die Mehrheit der grundlegenden Definitionen, Eye-Tracking-Methoden und Blickbestimmungsansätze kann daher nicht ohne Anpassung vom externen Eye-Tracking übernommen werden. Beispielsweise müssen die Definitionen einzelner Augenbewegungen, wie die der klassischen Fixation, erweitert werden, da das mobile Eye-Tracking mit natürlichen Bewegungen des Nutzers und seines Kopfes einhergeht. Entsprechende Analysemethoden müssen angepasst oder von Grund auf neu implementiert werden, basierend auf neuen Ansätzen, die das menschliche Blickverhalten kodieren.

Darüber hinaus sind die gesellschaftlichen Vorbehalte ein völlig neues und noch wenig erforschtes Gebiet, deren Überwindung jedoch für den Durchbruch des mobilen Eye-Trackings als Alltagstechnologie unbedingt erforderlich ist. Ein erster wichtiger Schritt, um gesellschaftliche Bedenken auszuräumen, ist der Aufbau von Akzeptanz gegenüber dem mobilen Eye-Tracking. Daher ist es unerlässlich, das wuchtige Design der derzeit genutzten, am Kopf getragenen Eyetracker durch ein unauffälliges, ansprechendes und modernes Design zu ersetzen. Der zweite wichtige Schritt ist der Schutz der Privatsphäre, der für die Gesellschaft von zunehmender Bedeutung ist, da sich weder die Forschung noch die Industrie in der Vergangenheit hinreichend oder überhaupt nicht mit diesem Thema befasst haben. Um wirkliches Vertrauen schaffen zu können, müssen zukünftige Eyetracker eine ausgewogene Balance zwischen dem Schutz der Privatsphäre der Träger und umstehender Personen, sowie nützlicher und vorteilhafter Anwendungen finden, um

die Menschen zu begeistern und sie von ihrer Notwendigkeit, ihrem Nutzen und ihrem Potenzial zu überzeugen. Die Lösung dieser technischen und gesellschaftlichen Probleme ist die Grundvoraussetzung für die Entwicklung zahlreicher neuer und spannender Anwendungen, um mobiles Eye-Tracking als neues Paradigma zu etablieren, das unseren Alltag erleichtert.

Diese Dissertation bewältigt zentralen technischen Herausforderungen des mobilen Eye-Trackings, die derzeit seine breite Anwendung verhindern. Diese Arbeit zeigt insbesondere, dass 3D-Daten, die für die Kalibrierung von mobilen Eyetrackern verwendet werden, die Blickbestimmung verbessern und den Parallaxenfehler signifikant reduzieren. Darüber hinaus präsentiert diese Dissertation das erste effektive Fixationserkennungsverfahren für am Kopf getragene Eyetracker, das stabil gegenüber jeglichen Bewegungen der Nutzer und deren anvisierten Objekten in der Umgebung ist.

Um gesellschaftliche Akzeptanz zu erreichen, stellt diese Arbeit ein innovatives und unauffälliges Design für zukünftige mobile Eyetracker vor, sowie den ersten Prototyp mit vollständig in die Brillenfassung integrierten Augenkameras. Zur Blickbestimmung kommt dabei ein auf dem Erscheinungsbild der Augen basierendes, kalibrationsfreies und tief trainiertes Verfahren zum Einsatz. Um die Privatsphäre der Nutzer und umstehender Personen bei der Verwendung von mobilen Eyetrackern zu schützen, wird in dieser Dissertation ein weiterer Prototyp vorgestellt. Dabei handelt es sich um einen Eyetracker, der private Situationen erkennen kann, um die auf die Umgebung gerichtete Kamera des Eyetrackers mittels einer mechanischen Klappe automatisch zu aktivieren oder zu deaktivieren. Zu diesem Zweck werden Informationen aus der Umgebung und den Augenbewegungseigenschaften miteinander kombiniert.

Dennoch reicht es nicht aus, sich auf die Bewältigung technischer Herausforderungen und den Abbau gesellschaftlicher Vorbehalte zu beschränken, um das mobile Eye-Tracking für die breite Masse attraktiv zu machen. Der Schlüssel zum Erfolg ist die Entwicklung nützlicher, innovativer und grundlegender Anwendungen, um die Menschen zu überzeugen. Um die Privatsphäre der Nutzer auch auf der Softwareseite zu schützen, wird in dieser Arbeit die erste datenschutzfreundliche VR-Oberfläche mit integrierter Eye-Tracking-Funktion vorgestellt, die Differential Privacy verwendet. Bei dieser Methode werden die aufgezeichneten Eye-Tracking-Daten durch die Hinzugabe von Rauschwerten so verändert, dass datenschutzrelevante Informationen wie das Geschlecht oder die Identität eines Nutzers nicht mehr aus den Daten bestimmt werden können und damit geschützt bleiben, ohne die Nutzung der Daten für eine bestimmte Anwendung selbst zu beeinträchtigen. Darüber hinaus wurde die erste große Online-Umfrage durchgeführt, um die Bedenken der Nutzer hinsichtlich des Eye-Trackings zu verstehen.

Um zukünftige Anwendungen zu entwickeln und zu evaluieren, präsentiert diese Dissertation die ersten öffentlich zugänglichen Langzeitdatensätze, die mit einem mobilen Eyetracker aufgenommen wurden. Sie ermöglichen die unüberwachte Erkennung von Aktivitäten durch die bloße Analyse der Augenbewegungen der Nutzer. Dabei kommen neue und effiziente videobasierte Kodierungsverfahren zum Einsatz. Außerdem werden die Datensätze dazu verwendet, die Wirksamkeit einer Methode zur Vorhersage des Blickverhaltens von Nutzern in Bezug auf ihre Aufmerksamkeit während alltäglicher Interaktionen mit mobilen Geräten nachzuweisen, wobei in die Geräte integrierte und am Körper getragene Sensoren verwendet werden. Dies eröffnet die Möglichkeit zur Entwicklung einer Vielzahl neuer und spannender Anwendungen.

Mit weiterentwickelten Funktionen, die mit dem technischen Fortschritt und der Miniaturisierung von Sensoren einhergehen, wird Eye-Tracking zunehmend in gewöhnliche Brillen sowie in am Kopf getragene Virtual und Augmented Reality (VR/AR) Geräte integriert und somit zu einem festen Bestandteil mobiler Nutzerschnittstellen. Diese Dissertation ebnet den Weg für die Entwicklung gesellschaftlich akzeptierter, datenschutzfreundlicher, aber hochfunktionaler mobiler Eyetracker und deren zukünftiger Anwendungen, damit das mobile Eye-Tracking sein volles Potenzial entfalten und eine alltagstaugliche Technologie für jedermann werden kann.

Acknowledgements

First and foremost I would like to thank my supervisor Prof. Dr. Andreas Bulling for his full support from my master's degree to the PhD thesis. He opened my mind to the vast, diverse, and interesting research field of mobile eye tracking. This work would not have been possible without his perfect guidance and keen sense of the appropriate selection of upcoming topics in our research community. Likewise, I thank Assoc. Prof. Dr. Yusuke Sugano and Dr. Michael Xuelin Huang for their mentoring, and countless fruitful discussions. Their experience in research and motivation helped me to improve the quality of my work.

I would also like to thank the Japan Science and Technology Agency (JST) for funding my PhD position (CREST Grant No.: JPMJCR14E1).

I am truly grateful to Prof. Dr. Antonio Krüger and Prof. Dr. Enkelejda Kasneci for agreeing to serve as reviewers of this thesis as well as to Prof. Dr. Vera Demberg and Dr. Florian Daiber for being part of the thesis committee.

I had the pleasure to work in an inspiring and creative environment at the Max Planck Institute for Informatics. For this, I want to thank all my colleagues for the great time and pleasant working atmosphere. I would especially like to thank Prof. Dr. Bernt Schiele for his leadership of the department and his good advice. I also thank our department secretary Connie Balzert for her extensive assistance. Special thanks go to my former officemates Sabrina Hoppe and Philipp Müller for their valuable support and inspiring conversations.

Also I would like to express my sincere gratitude to all my collaborators and especially to Marion Koelle, Inken Hagestedt, and Marc Tonsen for our very successful collaborations.

I thank Margaret De Lap who has proofread not only this dissertation but also my publications in the last few years.

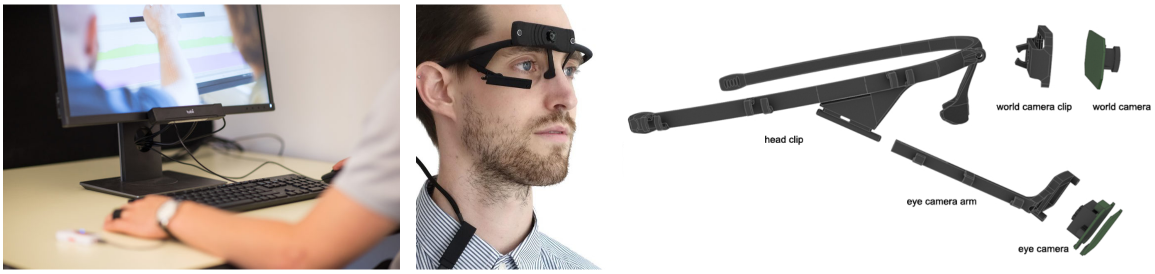
Last but not least, I would like to thank my family and friends for their honest understanding and never-ending support, giving me the possibility to work intensively on this thesis over the last four and a half years.

I want to dedicate this work to the memory of my father Peter Steil. He always had the dream that my brother and I would have every opportunity to become what we want to be. I think we are on the right track!

Table of Contents

1 Introduction

This chapter provides an introduction to state-of-the-art eye tracking devices, sensing and analysis methods, and identifies their immanent technical challenges. Further, it presents currently available eye tracking applications, pointing out their limitations, and introduces already existing methods to overcome social concerns with eye tracking. Finally, it presents the aims as well as an outline of the thesis, together with a list of publications that resulted from this work.



(a)

(b)

Figure 1.1: Camera-based eye tracking can be differentiated into two methods: (a) remote eye tracking[1] and (b) mobile eye tracking[2,3].

1.1 Eye Tracking Devices

There are three major methods to track the human eye: 1) electro-oculography (EOG), 2) the scleral search coil method, and 3) camera-based methods. EOG uses electrodes placed around the eye to measure the skin potential, which changes when the user moves the eyes in another direction (Kaufman *et al.*, 1993). The scleral search coil method exposes a user to an alternating magnetic field, in which the eye position may be accurately recorded from the voltage generated in a coil of wire embedded in a scleral contact lens worn by the user (Robinson, 1963). In the following this thesis will restrict itself to camera-based methods predominantly used in state-of-the-art eye tracking devices.

Camera-based methods are generally divided into remote (static) and mobile (head-mounted) eye tracking methods. This work will point out the fundamental differences between the already well established remote eye tracking and the more recent mobile eye tracking sensing modality, as it focuses on the latter. Both methods have their raison d'être, as they are not mutually exclusive and might complement each other in their different fields of application. However, the development of mobile eye tracking first arose from the technical process of miniaturisation and the already gained knowledge of remote eye tracking.

Mobile and remote eye tracking differ in several important aspects: Remote eye trackers do not require any attachments to the human body, but a fixed mount below or an integration into a computer or laptop monitor to track a user's eye movements (see Figure 1.1a). They consist of a user-facing camera and at least one additional infrared (IR) light source. In contrast, the components of state-of-the-art mobile eye trackers are one or two eye cameras capturing close-up images of a user's eye(s), actively illuminating the eye(s) with IR LEDs, and a front-facing scene (world) camera mounted on a glasses frame (see Figure 1.1b). The cameras are either connected to a portable laptop or to a powerful mobile phone (cf. Pupil Mobile[4]).

[1] `https://imotions.com/blog/eye-tracking-work/`, date: 12.07.2019

[2] `http://idstats.co/2018/09/13/eye-tracking/`, date: 12.07.2019

[3] `https://docs.pupil-labs.com/`, date: 12.07.2019

[4] `https://pupil-labs.com/blog/2017-08/introducing-pupil-mobile/`, date: 12.07.2019

1.1.1 Remote Eye Tracking

Remote eye tracking has a long history in psychology for the detection and analysis of human processing of visual information (Duchowski, 2002; Mele and Federici, 2012). In particular, remote eye tracking was used to study users' scene perception, visual search strategies, and gaze behaviour during reading tasks (Rayner and Pollatsek, 1992; Rayner, 1998; Sattar *et al.*, 2015). In recent years, there has been an increase in its use for purposes other than research. For example, remote eye tracking can be used to improve the functionality of e-learning systems, increasing the learners' motivation and attention by stimulating their interests (Al-Khalifa and George, 2010). In terms of marketing, it is important to understand users' gaze behaviour towards advertisements (Rayner *et al.*, 2001) and how to constantly adapt web content to increase the users' attention toward e.g. online shops (Alt *et al.*, 2012). Further, it is used for driving assistance to estimate drivers' fatigue (Eriksson and Papanikolopoulos, 2001) or cognitive workload (Palinko *et al.*, 2010). In addition to mouse or keyboard interactions, remote eye tracking enabled a novel hands-free interaction modality (Zhai *et al.*, 1999; Jacob and Karn, 2003). From users' eye gaze patterns, their interest can be sensed, which makes it possible to converse with an interactive system managing computer information output accordingly, e.g. to plan a city trip (Qvarfordt and Zhai, 2005). Besides user interest, eye movements even contain information to determine users' personality traits (Hoppe *et al.*, 2018). Another factor for the success of remote eye tracking nowadays is that it has become available and affordable for everyone while being unobtrusive, as it can be mounted below or integrated into stationary displays. For everyday life, applications like gaming remote eye tracking devices operating at more than 60 Hz can be bought for less than $200 (e.g. Tobii 4C[5]), whereas remote eye trackers for research purposes operating at about 1000 Hz cost more than $10,000 (e.g. EyeLink 1000 Plus[6]). Even the mid-range devices are able to achieve an accuracy error (difference between true and estimated gaze point) of around 1° (Holmqvist *et al.*, 2011; Nyström *et al.*, 2013; Blignaut *et al.*, 2014; Ooms *et al.*, 2015). However, remote eye tracking comes with severe limitations. It constrains head movements due to a narrow field of view, assumes that users always keep their heads still at a fixed distance to the display, and requires stable lighting conditions (San Agustin *et al.*, 2010a; Stellmach and Dachselt, 2013). Although remote eye trackers can record at a much higher frame rate than mobile devices, tracking a user's eyes remains restricted to distinct locations and towards predefined surfaces.

1.1.2 Mobile Eye Tracking

In comparison to remote eye tracking, mobile eye tracking enables the tracking of a user's eyes regardless of head pose or motion in natural unconstrained environments and allows the user to freely move around. However, this new freedom comes with novel technical challenges and social obstacles.

[5] `https://gaming.tobii.com/product/tobii-eye-tracker-4c/`, date: 12.07.2019

[6] `https://www.sr-research.com/products/eyelink-1000-plus/`, date: 12.07.2019

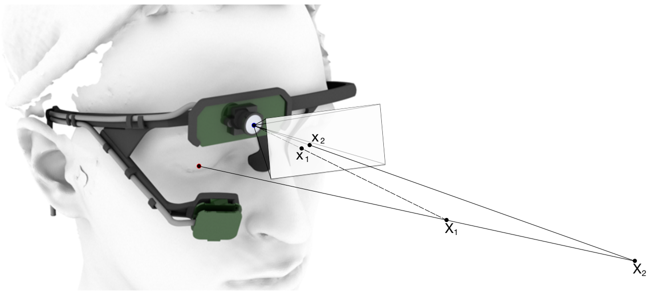


Figure 1.2: The parallax error[7] is caused by the assumption that the eyeball centre position is exactly the same as the origin of the scene camera coordinate system, resulting in a misalignment of the computed gaze point and true gaze depending on the distance to the object the user is looking at, whereas the eye tracking was calibrated for one specific distance.

Previously developed methods for sensing and analysis need to be rethought, as they cannot be borrowed from remote eye tracking and applied without adaptations. These methods were designed for constrained laboratory-like, indoor conditions, in which lighting and user position need to be controlled.

As opposed to remote eye tracking where only one camera is focusing on the user, mobile eye tracking needs one egocentric front-facing camera as well as at least one eye camera recording a user's eye movements from close proximity. Hence, mobile eye tracking is more invasive than the remote alternative, as a user has to wear the tracker on their head, which causes novel privacy concerns. Especially the scene camera, which releases the user from a stationary desktop, can pose a significant threat to user and bystander privacy.

To record a user's pupil position, most of the state-of-the-art devices additionally use an IR LED in combination with the eye camera(s). A mapping function calculated in a calibration procedure is then used to determine the corresponding gaze point in the camera coordinate system given by the image recorded by the scene camera.

Although the accuracy of mobile eye tracking devices also reaches around 1° (MacInnes *et al.*, 2018) using scene cameras at 25-60 frames per second (fps) and eye cameras at 30-200 fps, they have to deal with the problem of the so-called parallax error (see Figure 1.2). This error is caused by a misalignment of the calculated gaze point and the true gaze depending on the distance to the object the user is looking at, whereas the eye tracker was calibrated to a fixed distance. As mobile eye tracking is not restricted to display interactions, where it is easy to identify an object of interest on a screen given the calculated gaze ray on a small distance, 3D gaze vectors are used to find the intersecting objects in real-world scenarios. Most of the prior work assumes that the 3D gaze vectors are originating from the origin of the scene camera coordinate system to obtain 3D gaze vectors (Munn and Pelz, 2008; Pfeiffer and Renner, 2014; Takemura *et al.*, 2014). In this case, estimated 2D gaze positions can be simply back-projected as 3D vectors from the camera coordinate system into the scene. This

[7] adapted from `https://www.shapeways.com/product/LQJJK2CHQ/pupil-mobile-eye-tracking-headset`, date: 12.07.2019

is equivalent to assuming that the eyeball centre position is exactly the same as the origin of the scene camera coordinate system. However, in practice there is always an offset between the scene camera origin and the eyeball position. This offset induces the parallax error. Previous works tried to compensate for the parallax error by using an error function (Mardanbegi and Hansen, 2012) or predictive error models (Barz *et al.*, 2016). A straightforward way to eliminate the parallax error is to use a binocular eye tracking headset and stereo geometric calculations (Duchowski, 2007). However, the majority of head-mounted eye trackers still rely on monocular settings, which ease the data synchronisation and analysis, and decrease the price of mobile eye trackers.

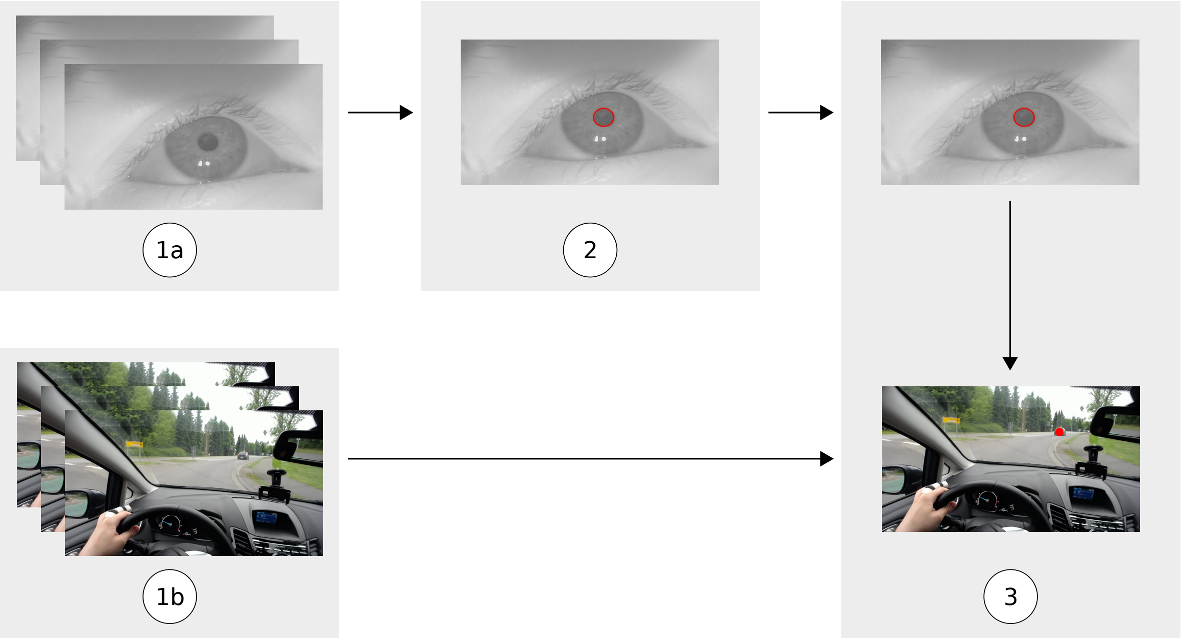


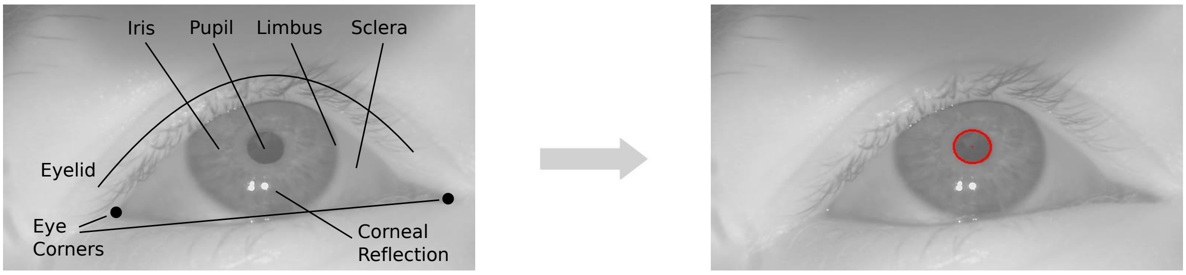
Figure 1.3: To estimate a user's gaze from eye tracking devices, three steps are essential: (1a, 1b) Recording of video frames from an eye and a scene camera; (2) Detection of a user's pupil in the eye image; (3) Mapping a user's pupil position to a gaze position in the scene camera coordinate system.

1.2 Sensing

Sensing information from a mobile eye tracking device consists of three essential and consecutive steps as displayed in Figure 1.3. In the first step, video frames from an eye and a scene camera need to be recorded (see Figure 1.3 (1a) and (1b)). In the second step, a user's pupil and pupil centre are detected in the eye camera image (see Figure 1.3 (2)). In the last step, a user's pupil position is mapped to a gaze position in the scene (see Figure 1.3 (3)). For this, a mapping function or a model is used. The core procedure to collect the necessary data is the so-called calibration, where a user's pupil or eyeball positions, and marked positions in a user's field of view in the scene camera coordinate system that the user is instructed to look at, are collected. Finally, a gaze estimation method is applied to correlate the data, generating a mapping function. State-of-the-art pupil detection (see Figure 1.4) and gaze estimation (see Figure 1.5) methods including the ones used in this thesis are explained in the following two sections.

1.2.1 Pupil Detection

Detecting and tracking a user's eye movement is an essential prerequisite towards gaze estimation. The visual data gained from camera-based eye tracking enables a variety of detection methods. This section presents the three most common techniques (Hansen and Ji, 2009) and focuses on the state-of-the-art approaches relying on IR images.



(a) Eye frame

(b) Eye frame with pupil extraction

Figure 1.4: The extraction of a user's pupil is an essential prerequisite towards gaze estimation.

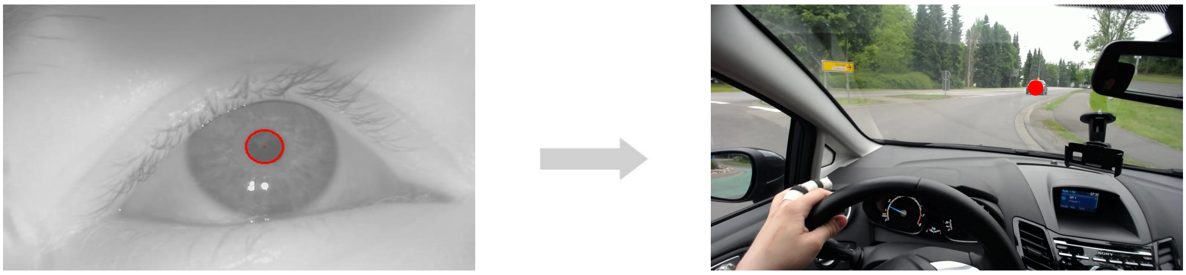
Shape-Based. Considering the human eye, iris and pupil contours as well as eyelids can be well described by their shape. Therefore, the shape-based technique exploits a predefined geometric model of the human eye to find corresponding matches in provided images, resulting in the extraction of the eye or pupil area. The geometric models can be subdivided by their complexity, allowing deformations and transformations by scale or rotation. Simple models assume a fixed elliptical shape depending on the viewing angle, which can be described by five shape parameters. To fit the ellipse model, previous works used edge detection techniques to extract the pupil boundaries and threshold the image intensities to estimate the centre of the pupil ellipse (Kim and Ramakrishna, 1999; Peréz *et al.*, 2003), Hough transformation to extract the iris or the pupil (Nixon, 1985; Young *et al.*, 1995), image gradients (Kothari and Mitchell, 1996), or the RANSAC method (Hansen and Pece, 2005). More complex approaches allow eye shape deformation taking the eye corners (Lam and Yan, 1996; Zhang, 1996) (see Figure 1.4a) or the modelling of the eyelids into account (Wood and Bulling, 2014; Fuhl *et al.*, 2016a). A prominent example is given by Yullie et al., who modelled the iris and eyelids with eleven parameters (Yuille *et al.*, 1992). However, the more parameters used to model the human eye, the higher the computational complexity.

Feature-Based. In comparison to the shape-based approach, the feature-based approach initially tries to detect characteristic features of the eye in facial images to locate the eye shape. Common features are the limbus (border of the cornea and the sclera), the pupil, and corneal reflections (see Figure 1.4a). To detect edges and lines, as well as their orientation, length and scale, Herpers et al. applied different computer vision methods in combination with a prior eye shape model (Herpers *et al.*, 1996). Other works described an eyeball model using six landmarks, e.g. eye corner points (Feng and Yuen, 2001), or used specially trained neural networks (Reinders *et al.*, 1996), convolution (D'Orazio *et al.*, 2004), or linear and nonlinear filtering (Sirohey and Rosenfeld, 2001). Besides discriminating lines, the pupil is a highly reliable feature for detecting the human eye, as the pupil may be darker than its surrounding iris and sclera (see Figure 1.4b). Among the first, Stiefelhagen et al. and Yang et al. (Stiefelhagen *et al.*, 1997a,b; Yang *et al.*, 1998) proposed an iterative threshold algorithm to locate a user's pupil shape. While it is possible to use pupil-only tracking, with the use of additional IR light sources the corneal reflections, the so-called glint (a bright light spot visible on a user's iris) can be detected. This feature offers an additional reference point to compensate for small head movements (Duchowski, 2007; Hammoud, 2008; Hansen and Ji, 2009).

Appearance-Based. The appearance-based or holistic method directly detects a user's eye by its photometric appearance as characterised by filter responses or colour distribution of a given image. The appearance-based approaches can be either applied in the spatial (Hallinan, 1991) or in a transformed domain (Huang and Wechsler, 1999) using image template-based methods or statistical techniques to analyse the intensity distribution across the entire image. To detect different eye representations of different subjects, lighting conditions, and face orientations, a large amount of data is necessary to train corresponding models and classifiers. Huang et al. and Zhu et al. used support vector machines (SVMs) (Huang *et al.*, 1998; Zhu *et al.*, 2002a); Samaria and Young employed stochastic modelling (hidden Markov models (HMMS)) (Samaria and Young, 1994) to detect a user's eyes from face images. Recent works of Zhang et al. and Karafka et al. (Zhang *et al.*, 2015; Krafka *et al.*, 2016; Zhang *et al.*, 2017b) used deep convolutional neural networks (CNNs) to train models with images of unconstrained settings recording a user's face from a laptop or tablet camera. A recent work of Fuhl et al. trained a CNN for automatic pupil detection from eye images in real-world scenarios (Fuhl *et al.*, 2016b).

Hybrid Models. Besides the three most common techniques, hybrid models aim at combining the advantages of the different detection approaches to overcome their shortcomings and advance the performance of the detection of a user's eye and pupil (Xie *et al.*, 1994; Hansen *et al.*, 2002; Zhu *et al.*, 2002b).

Pupil Detection on IR Images. In comparison to the methods above, which were mainly developed for remote eye tracking devices where a user's eyes first need to be detected from face images, mobile eye tracking devices record a user's eye region from close proximity. Thus, mobile eye tracking devices directly focus on the precise detection of a user's pupil shape and centre. Especially driven from the feature-based approach, trying to detect the pupil often fails in the presence of other dark regions using RGB or common grayscale images relying on existing light of the environment (*passive* approach); mobile eye trackers are therefore equipped with an additional IR light source (*active* approach), e.g. an IR LED close to the eye camera, to achieve a higher contrast between the pupil and the surrounding iris and sclera. The IR light (780-900nm) is invisible for the human eye and thus does not disturb the user. From the high-contrast images thereby gained, multiple approaches were invented to properly fit an ellipse around the pupil shape. Recently, Fuhl et al. (Fuhl *et al.*, 2016d) compared the six most popular pupil detection algorithms applicable to mobile eye tracking data: 1) Starburst (Li *et al.*, 2005), 2) Świrski (Świrski *et al.*, 2012), 3) SET (Javadi *et al.*, 2015), 4) ExCuSe (Fuhl *et al.*, 2015), 5) ELSE (Fuhl *et al.*, 2016c), and 6) Pupil Labs (Kassner *et al.*, 2014). The ELSE algorithm was able to outperform the other approaches with respect to detection rate on different real-world datasets including changing lighting conditions, occlusions, and reflections.



(a) Eye frame with detected pupil

(b) Scene frame with gaze point

Figure 1.5: For gaze point estimation, a user's pupil position is (a) extracted and (b) mapped in the scene camera coordinate system.

1.2.2 Gaze Estimation and Calibration

The primary task of eye tracking is estimating a user's gaze. The human gaze can be either defined as the so-called point of regard (POR) or as the gaze direction. Thus, two different methods need to be discriminated: 1) interpolation-/regression-based and 2) geometric-/model-based. Both approaches have the aim to find a suitable mapping from the pupil position to the point in the scene the user is currently looking at. They are the most commonly used methods in commercially available eye tracking devices nowadays. The data used for the mapping is collected during the calibration process.

Regression-/Interpolation-Based. The regression- or interpolation-based approaches take infrared eye images as input and calculate a parametric mapping function from the pupil positions to the PORs (called gaze positions in the following) from a sufficient number of data pairs. Given the final mapping function for each recorded pupil position, the corresponding gaze position can be interpolated (see Figure 1.5b). Devices which rely on the interpolation-based approach need a more extensive calibration in terms of calibration points (e.g. a 9-point calibration is used by Pupil Labs[8]) than those using the model-based approach (e.g. one-point calibration (Villanueva and Cabeza, 2008) used in Tobii Pro Glasses 2[9]). Besides polynomial mapping functions (Cerrolaza *et al.*, 2012), there are also approaches applying neuronal networks (Baluja and Pomerleau, 1994; Ji and Yang, 2002) or homography transformation (Yu and Eizenman, 2004).

Geometric-/Model-Based. The second gaze estimation method models the human eye as a sphere, relying on prior knowledge of personal data (e.g. eyeball size, cornea radius) to calculate the gaze as a 3D direction vector (see Figure 1.6). For this, the cornea centre is estimated so that the optical and visual axis (line of sight) can be calculated. The intersection of one of the lines with an object of interest in the field of view of a user provides the gaze position, while the line of sight is defined as the true gaze direction. 3D gaze estimation has been widely studied in remote settings as it

[8] `https://docs.pupil-labs.com/`, date: 12.07.2019

[9] `https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/`, date: 12.07.2019

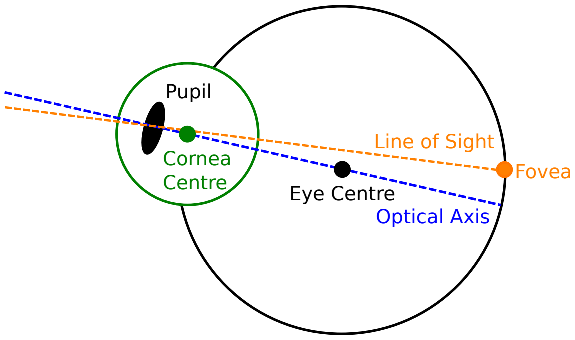


Figure 1.6: Geometric model of the human eyeball including the visual (line of sight) and optical axes.

requires special hardware, such as multiple IR light sources or stereo cameras (Beymer and Flickner, 2003; Nagamatsu *et al.*, 2010). Given these hardware constraints, it remains unclear whether model-based gaze estimation can be done properly using a head-mounted eye tracker. Recently, Świrski and Dodgson proposed a method to recover 3D eyeball poses from a monocular eye camera (Świrski and Dodgson, 2013). However, for the evaluation of their method they used noiseless synthetic data, which makes it difficult to judge whether their approach also works under real-world conditions.

Calibration Alternatives. Besides the model- and interpolation-based gaze estimation approaches, there are also other calibration alternatives which try to sample pupil and gaze data in a more natural, unobtrusive, and less tedious way, applying auto-calibration or even calibration-free approaches.

Instead of following and gazing at a target point for a certain amount of time, Pfeuffer et al. exploited a moving target and recorded users' smooth pursuit eye movements (Pfeuffer *et al.*, 2013) which has the advantage of determining whether a user is really following the target. Khamis et al. applied a similar approach but showed a moving text on a display (Khamis *et al.*, 2016). Other approaches used a gaming scenario and gaze as input modality (Flatla *et al.*, 2011), or relied on interaction events such as mouse clicks or key presses (Sugano *et al.*, 2008), assuming that the user is looking at specific interface elements (Huang *et al.*, 2014), as a proxy for a user's gaze position on a computer screen in order to sample calibration data.

A different approach leverages visual saliency maps (Koch and Ullman, 1987) to auto-calibrate or recalibrate the eye tracking devices (Chen and Ji, 2011; Sugano and Bulling, 2015). This approach assumes that a user is looking at the most salient object in the scene. A recent work of Müller et al. exploits a user's mobile phone location or touch events during device interactions to recalibrate a mobile eye tracker to reduce the so-called calibration drift, a gaze estimation error which can be caused by headset slippage (Müller *et al.*, 2019).

Purely data-driven approaches use a large amount of training data to train person-independent gaze estimators without the need for a person-specific calibration (Funes Mora and Odobez, 2013; Schneider *et al.*, 2014; Sugano *et al.*, 2014).

Another calibration-free approach without the need of training data uses corneal imaging. To estimate a user's gaze on near-by surfaces and objects Lander et al. record one eye with a single high-resolution RGB camera from close proximity and apply natural feature tracking (Lander *et al.*, 2017a,b) or in combination with an additional IR camera a homography matrix (Lander *et al.*, 2018a). These two approaches work well for small distances achieving gaze estimation errors of 4.03° and 2.19° respectively, which are sufficient accurate for the analysis of a user's attentive behaviour (Lander *et al.*, 2018b) and activities (Lander and Krüger, 2018). Their lightweight mobile system consists of a RaspberryPi or a mobile phone to record the video data. However, using an off-the-shelf RGB webcam results in a rather bulky appearance of the eye camera occluding parts of users' field of view. As the corneal image reflection only covers one-third of users' field of view (Lander *et al.*, 2017a) scene information from far periphery are lost. Thus, these approaches are restricted to a specific distance to the objects a user is looking at. Further, corneal imaging suffers from serious privacy concerns as the current recording systems do not communicate the recording status so that bystanders have no opportunity to recognise a potential recording and to take action themselves to protect their privacy compared to a clearly observable scene camera.

In contrast to the corneal imaging approaches of Lander et al. which require pre-knowledge about considered objects, AR markers on objects, or access to shown content of displays to successfully track and map features from the corneal image reflection to the considered surface, this thesis exploits multi-modal recording systems recording a user's eyes and scene separately. The minimal effort of an initial calibration enables mobile eye tracking in fully unconstrained environments and allows the recording of high-resolution scene image data, highly accurate gaze estimation (Kassner *et al.*, 2014) as well as the application of more sophisticated methods for scene analysis and scene feature extraction.

1.3 Analysis

The analysis of a user's gaze behaviour is the connecting element between sensing eye tracking data and the development of real-world applications, and the most important task towards the goal of mobile eye tracking for everyone. It consists of the detection, encoding, and visualisation of distinct eye movements necessary to achieve a deep understanding of a user's gaze behaviour.

1.3.1 Eye Movement Detection

A user's gaze behaviour can be differentiated into three basic eye movements: 1) blinks (closing and opening the eyelid), 2) fixations (static states of the eye), and 3) saccadic movements (gaze shifts between fixations). Besides the three main eye movements, the human gaze behaviour can be further described by smooth pursuit movements, when tracing a moving target, and vestibulo-ocular reflex (VOR) movements, which correct the eye position during head and body motion to keep the object of interest in focus.

The detection of blinks has importance as an input modality for human-computer interfaces (Jacob and Karn, 2003), especially for people with disabilities (Grauman *et al.*, 2001, 2003), but also in assistance systems to identify a person's state of vigilance, fatigue, or drowsiness (Suzuki *et al.*, 2006; Schleicher *et al.*, 2008; Yang *et al.*, 2012; McIntire *et al.*, 2014), e.g. when driving a car (Braunagel *et al.*, 2015), or for the detection of mental disorders, such as schizophrenia (Li *et al.*, 2002). A common approach is to detect a blink when the pupil cannot be detected in a number of consecutive eye camera frames (Kim *et al.*, 2011). Other methods learn from eye blink patterns (Le *et al.*, 2013), consider the motion vector of eyelid movement (Fogelton and Benesova, 2016), rely on the average illumination intensity of eye images (Moriyama *et al.*, 2002), or threshold the difference between two consecutive frames to identify the pupil and eyelid (Jiang *et al.*, 2013). A previous work of Appel et al. exploits the fact that during a blink the dark pupil gets partly occluded so that the frame brightness increases, reaching its maximum at the blink apex (Appel *et al.*, 2016).

Most of the existing methods developed for remote eye tracking rely on gaze point analysis to identify fixations and saccades. These methods can be categorised into velocity-based, dispersion-based, and data-driven approaches.

The velocity-based methods are the most widely used (Andersson *et al.*, 2017), leveraging a velocity threshold to separate fixations from saccades (Salvucci and Goldberg, 2000), where eye movements with a velocity below the threshold are classified as fixations, and above it, as saccades. Since smooth pursuit movements are too slow to be identified as saccades and too fast to be classified as fixations, an additional threshold is used to differentiate smooth pursuit from saccades (Ferrera, 2000; Komogortsev and Karpov, 2013).

Dispersion-based algorithms assume that gaze estimates belonging to a fixation should be located in a cluster (Salvucci and Goldberg, 2000; Blignaut, 2009; Holmqvist *et al.*, 2011). For this, these algorithms measured the degree of gaze estimates' scattering to identify fixations.

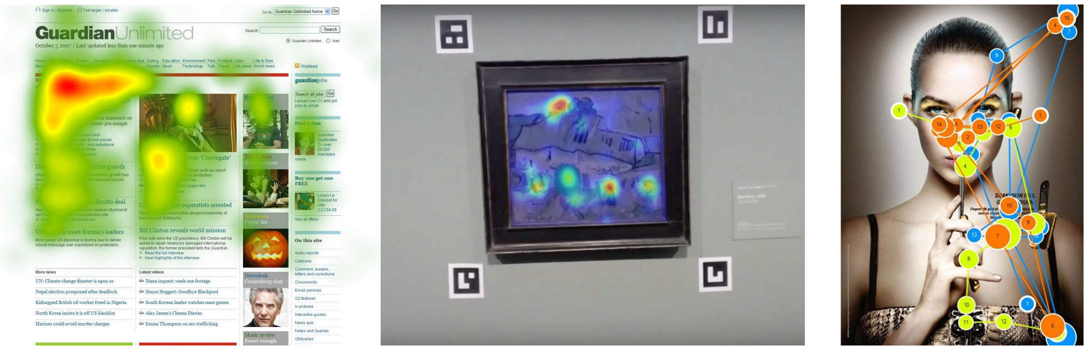
Both velocity- and dispersion-based approaches suffer from the necessity of handcrafted thresholds. Thus, a number of recent works have applied data-driven approaches to improve eye movement detection, including smooth pursuits (Vidal *et al.*, 2012a) and fixations. For fixation detection, prior works have proposed the use of projection clustering (Urruty *et al.*, 2007), principle component analysis (Kasneci *et al.*, 2015), eigenvector analyses (Berg *et al.*, 2009), Bayesian decision theory (Santini *et al.*, 2016), or detailed geometric properties of signal components (Vidal *et al.*, 2012a). Only a few previous works have addressed the challenging task of discriminating between multiple eye movement types at once using a random forest classifier (Zemblys *et al.*, 2017) or end-to-end trained neural networks (Hoppe and Bulling, 2016; Zemblys *et al.*, 2018).

1.3.2 Eye Movement Encoding and Visualisation

Throughout the history of eye tracking research, several key variables have emerged as meaningful characteristics of ocular behaviours, including fixations, saccades, pupil dilation, and a user's scan paths (Rayner, 1998). Fixations in particular represent the instances in which information acquisition and processing are able to occur (Rayner, 1998). From recorded eye movement data, researchers are able to determine whether a user is properly reading or only scanning a text for a particular word or phrase, or whether a user is looking at an appropriate object, or searching for a specific item on a web page (Crowe and Narayanan, 2000). Therefore, a user's gaze behaviour needs to be encoded using statistical methods or visualise the corresponding attentive behaviour (Blascheck *et al.*, 2014).

Remote eye trackers were mainly used to study a user's eye movement behaviour towards selected display content. An appropriate tool to visualise a user's gaze allocation is heatmaps (Špakov and Miniotas, 2007) around predefined areas of interest (AOIs) showing images or text (see Figure 1.7a). They can be used to display the attentive behaviour of a single user, or multiple users aggregating their gaze (Sugano *et al.*, 2016). Darker colours indicate low user attention, whereas brighter colours indicate high attention. Heatmaps were especially used to study a user's gaze behaviour on web pages (Cowen *et al.*, 2002), to identify objects of interest (Velásquez, 2013), or to tailor web content (Alt *et al.*, 2012).

The aggregation of users' visual attention in daily life environments is significantly more challenging using a mobile eye tracking setup because of a missing reference point to align users' attentive behaviour. Therefore, either augmented reality (AR) markers (see Figure 1.7b) are used to define a surface in the field of view of a user (Kandemir and Kaski, 2012) or researchers pre-record an environment to create a 3D environment representation. Such a representation is then used to map a user's gaze towards the scene (Paletta *et al.*, 2013a,b,c; Schrammel *et al.*, 2014). A further approach uses visual simultaneous localisation and mapping (SLAM) (Engel *et al.*, 2014) to estimate 3D PORs in a real environment (Takemura *et al.*, 2014).



(a) (b) (c)

Figure 1.7: Heatmap encoding of human gaze behaviour on (a) a web site on a stationary display[10] and (b) an artwork in a mobile setup[11] as well as (c) gaze plot encoding of multiple users[12].

Other state-of-the-art metrics, which arose from reading activity research (Poole and Ball, 2005; Ehmke and Wilson, 2007; Albert and Tullis, 2013), are 1) fixation count (FC) – the number of times a user fixates on a scene object, 2) total fixation duration (TFD) – the sum of the duration of all fixations a user has laid on a scene object, and 3) average fixation duration (AFD) – the average duration of a fixation on a scene object. These three metrics generally represent a user's relative engagement with an object of interest (Poole and Ball, 2005; Albert and Tullis, 2013). As Chen and Wang have stated: "More fixations on an object suggest that it is more noticeable and important. A longer duration may indicate that the fixated object is more engaging in some way" (Chen and Wang, 2016). Thus, the most popular feature to infer user interest is the measurement of a user's fixation duration (Castagnos *et al.*, 2009, 2010; Castagnos and Pu, 2010; Chen and Pu, 2010) or attention time (Xu *et al.*, 2008) towards predefined AOIs.

Besides attentional means, scan-paths and saccadic movements are another group of features that describe a user's gaze behaviour. They can be represented as so-called n-grams (Reani *et al.*, 2018), which encode each eye movement event as a character forming words of length n (Bulling *et al.*, 2011b). Nakano and Ishii (Nakano and Ishii, 2010), for example, used 3-gram patterns to estimate whether a user is engaged in a conversation.

In comparison to heatmaps visualising a user's attention intensity and distribution, so-called gaze plots combine fixations and saccade movements. They display a user's scan path as a temporal sequence and express the duration of included fixations as circles (see Figure 1.7c). The longer the fixation, the larger the circle.

Another statistical means is a user's pupil dilation. It is typically used as a measure to gauge a user's interest or arousal in the content they are looking at (Granka *et al.*, 2004) or their cognitive activity, such as high cognitive load (Marshall, 2002; Palinko *et al.*, 2010).

[10] `https://blog.ezoic.com/ux-metrics-changing-view-visitors/heatmap-eye/`, date: 12.07.2019

[11] `https://twitter.com/pupil_labs/`, date: 12.07.2019

[12] `http://eyetracking.com.ua/eng/visualization/8.html`, date: 12.07.2019

Given the data gained from mobile eye tracking devices, another approach to visualise a user's gaze behaviour tries to incorporate a spatio-temporal component visualising the gaze behaviour of a user in horizontal timelines as a sequence of gaze point images, similar to thumbnails of images (Kurzhals *et al.*, 2016b). This AOI-free approach enables the analysis and comparison of the viewing behaviour of multiple users over time and can also be used for interactive semi-automatic labelling of mobile eye tracking data, applying clustering on extracted scene information around a user's gaze position (Kurzhals *et al.*, 2017).

1.4 Social Concerns

Mobile and remote eye tracking is increasingly used in both commercial and academic practice to analyse the utility of websites, mobile devices, software, and games[13] (Jacob and Karn, 2003; Poole and Ball, 2006; Ehmke and Wilson, 2007; Bergstrom and Schall, 2014). However, in the future especially, mobile eye tracking devices will be not only used as analysis tools but in real life, such as during everyday social interactions. A study of Ali et al. (Ali *et al.*, 2016) identified privacy and comfort as main concerns when using wearable cameras as integrated into mobile eye trackers. In the context of modern human-computer interaction, a broad discussion arose as to how new emerging technology can be made socially acceptable (Koelle *et al.*, 2018a). In order to reach a similar status as mobile phones or smart watches, mobile eye trackers need to achieve general social acceptance and find a proper balance between functionality and privacy protection.

1.4.1 Acceptability

As early as 1994, Nielsen (Nielsen, 1994) identified social acceptability as key to system acceptability. The social acceptability of emerging technology has been investigated in terms of accessibility by particular user groups (Profita *et al.*, 2016; Shinohara, 2017) and in various contexts (Koelle *et al.*, 2015), such as in medical use cases (DeBlasio and Walker, 2009; Ziefle and Röcker, 2010).

According to Montero et al. (Montero *et al.*, 2010) there are two dimensions of social acceptability: 1) the user's social acceptance and 2) the spectator's social acceptance.

The first dimension deals with the question of how comfortable a user's interaction is with the head-mounted displays (HMDs) and how awkward the task feels in the current environment. The second dimension deals with the question of how "normal" the HMD appears to bystanders, or whether it stands out.

The findings of Montero et al. (Montero *et al.*, 2010) still hold today, as current mobile trackers are still rather uncomfortable to wear, especially during long-term recordings. A main reason for this is the wide use of heavyweight high-quality image sensors, which lead to a bulky appearance of head-mounted eye tracking devices, and cause discomfort or even pain. With their size, they often even occlude a user's field of view, decreasing a user's confidence and assurance during device interaction. Required wires to connect mobile eye trackers with a recording computer, such as a laptop carried in a user's backpack to record and process the data and as a power supply, further limit a user's mobility. Social acceptability is also an issue; it comprises the perceptions of people faced with the new technology, determining whether it looks "cool" or "weird" (Goffman, 2006). Finally, the obtrusive design leads to low social acceptance in daily life and unnatural behaviour of both the wearer and bystanders (Risko and Kingstone, 2011; Nasiopoulos *et al.*, 2015), fundamentally limiting the practical application of mobile eye tracking as a tool in behavioural and social sciences.

[13] `https://www.tobiipro.com/fields-of-use/user-experience-interaction/`, date: 12.07.2019

1.4.2 Privacy

Despite concerns about unnatural user behaviour, eyewear devices, such as HMDs or AR glasses, have recently emerged as a new research platform (Bulling and Kunze, 2016) to analyse users' attention allocation (Eriksen and Yeh, 1985; Sugano *et al.*, 2016), for computational user modelling (Fischer, 2001; Itti and Koch, 2001), or for hands-free interaction (Hansen *et al.*, 2003; Vertegaal *et al.*, 2003). But despite this potential and advances in eye tracking technology (Bulling and Gellersen, 2010; Tonsen *et al.*, 2017) there has been no breakthrough for customer usage. Eye tracking technology has so far been limited to special users such as disabled people, or niche applications in research (Liu *et al.*, 2018). However, this is about to change with eye tracking being an integral part of upcoming head-mounted augmented and virtual reality (VR) displays, where eye tracking leads to an improved VR experience by enabling natural interaction, accuracy in correcting the parallax error (Luo *et al.*, 2005; Jones *et al.*, 2008), or savings in computational power by exploiting foveated rendering (Patney *et al.*, 2016; Hsu *et al.*, 2017). The wider availability of eye tracking – hitting the mass market with potentially millions of users worldwide – comes with new security and privacy risks. However, privacy threats caused by the information gained from the eye camera and scene camera need to be differentiated.

Eye Camera Privacy Threats. Using an HMD with integrated eye tracking technology, a user is continuously monitored by an eye camera. Thus, previous works applied eye movements as promising biometrics for privacy applications and user authentication (Kasprowski and Ober, 2003), using a point stimulus (Kasprowski, 2004; Bednarik *et al.*, 2005; Kasprowski and Ober, 2005) or images (Maeder and Fookes, 2003) that users were instructed to follow or to look at. Recent works proposed mathematical models (Komogortsev *et al.*, 2010; Komogortsev and Holland, 2013), eye movement patterns (Eberz *et al.*, 2016), or even a continuous authentication for VR headset users (Zhang *et al.*, 2018b). However, leveraging eye movement behaviour for privacy applications and user authentication has the downside of posing a potential threat to users' privacy.

In contrast to prior work, this thesis is the first to practically explore recorded eye movements as potential threat to users' privacy given the rich information content available in human eye movements (Bulling *et al.*, 2011a). The rapidly increasing capabilities of interactive systems to sense, analyse, and exploit this information in everyday life (Hansen *et al.*, 2003; Vertegaal *et al.*, 2003; Stellmach and Dachselt, 2012) further increase the threat for users that privacy-sensitive information can be inferred. For example, previous works show that a user's interest in a scene (Hess and Polt, 1960) and a user's cognitive load (Matthews *et al.*, 1991) are related to the measured pupil size. Considering a user's health status, mental disorders such as Alzheimer's (Hutton *et al.*, 1984), Parkinson's (Kuechenmeister *et al.*, 1977), or schizophrenia (Holzman *et al.*, 1974) can be detected even in an early stage by the analysis of a user's eye movement behaviour. Similar approaches also demonstrated that a user's activities

(Bulling *et al.*, 2013; Steil and Bulling, 2015), cognitive states (Bulling and Zander, 2014; Faber *et al.*, 2017), or personality traits can be recognised (Hoppe *et al.*, 2018).

In addition, several researchers have shown that gender and age can be inferred from eye movements, e.g. by analysing the spatial distribution of gaze on images like faces (Cantoni *et al.*, 2015; Sammaknejad *et al.*, 2017).

This wide range of future applications points out and confirms the enormous potential of eye movement analysis, but also poses significant privacy risks for potential users, resulting in a rejection of mobile eye tracking devices if there is no acceptable trade-off between utility and privacy protection.

Scene Camera Privacy Threats. Inversely, instead of inferring private attributes of a user via an eye camera, HMDs equipped with a scene camera can detract from others' perceptions of a user's trustworthiness and acceptability. Scene cameras enable the recording of personal information, such as login credentials, banking information, or text messages; they can also infringe on the privacy of bystanders (Perez *et al.*, 2017), and the latter privacy risks will be intensified by the unobtrusive integration of eye tracking in ordinary glasses frames (Tonsen *et al.*, 2017), which can be unsettling (Denning *et al.*, 2014). However, the privacy concerns and attitudes of users and bystanders towards HMDs with integrated cameras were found to be affected by context, situation, usage intentions (Koelle *et al.*, 2015), and user group (Profita *et al.*, 2016). In order to defend bystanders' privacy, previous works have suggested conveying their privacy preferences to nearby capture devices via wireless connections (Krombholz *et al.*, 2015) or a piece of cloth using *Privacy Fabric* (Krombholz *et al.*, 2017). Other works tried to prevent unauthorised recordings by compromising the recorded imagery, e.g., using infrared light signals (Harvey, 2010; Yamada *et al.*, 2013), disturbing face recognition (Harvey, 2012), masking objects or faces (Raval *et al.*, 2014; Shu *et al.*, 2016), "blacklisting" sensitive spaces (Templeman *et al.*, 2014) and screen content (Korayem *et al.*, 2016), restricting the visibility of content on two-dimensional surfaces by a marker framework (Raval *et al.*, 2014), or training a neuronal network to identify security risks, such as ATMs, keyboards, and credit cards (Erickson *et al.*, 2014).

While all of these methods improved privacy, they either only did so post-hoc, i.e. after images had already been captured, or required bystanders to take action, which might be impractical due to the costs and effort involved (Denning *et al.*, 2014). Another negative effect on social acceptability which remains unaddressed is the bystanders' assumption that HMDs are always recording (Koelle *et al.*, 2015).

1.5 Eye Tracking Applications

Eye tracking applications for remote eye trackers have a long history in usability, user experience, and psychology research. Various eye tracking metrics were developed to correlate users' attentive behaviour with usability problems of computer interfaces (Goldberg and Kotval, 1999; Jacob and Karn, 2003) and to understand users' interaction behaviour with mobile devices, large screens (Dybdal *et al.*, 2012) and user interface elements on websites (Ehmke and Wilson, 2007; Bergstrom and Schall, 2014). In psychology research, eye tracking was used to study users' attentive behaviour to predict visual search targets (Zelinsky *et al.*, 2013; Sattar *et al.*, 2015, 2017b) and intents (Bednarik *et al.*, 2012), and to recognise cognitive processes (Salvucci and Anderson, 1998; Steichen *et al.*, 2013) or mental disorders (Holzman *et al.*, 1974; Kuechenmeister *et al.*, 1977; Hutton *et al.*, 1984). Especially for human-computer interaction, users' gaze served as an input modality for attentive user interfaces (Zhai *et al.*, 1999), such as public displays (Vidal *et al.*, 2013; Zhang *et al.*, 2014), interactive dialogue systems (Qvarfordt and Zhai, 2005), or as multi-modal input combined with users' touch events (Pfeuffer *et al.*, 2014).

This thesis focuses on applications for mobile eye tracking devices. Specifically, Chapter 7 presents a privacy-preserving application playing the role of a trailblazer for recent research focusing on this topic (Chuang *et al.*, 2019). Further, as mobile eye tracking can be differentiated into diagnostic and interactive applications (Duchowski, 2002), Chapters 8 and 9 present an application relying on an eye camera alone, and another application which combines information gained from the eye and the scene, respectively.

1.5.1 Eye-Only-Based Applications

Similar to remote eye tracking, the development of applications for mobile eye tracking devices started using information gained from eye movement behaviour alone.

With the EOG-based eye movement feature encoding of Bulling et al., it became possible to recognise human activities, such as reading in transit (Bulling *et al.*, 2012); office activities (Bulling *et al.*, 2011b) or cognitive processes, such as visual memory recall processes (Bulling and Roggen, 2011); and contextual cues, like social interaction, concentrated cognitive work, physical activity, and spatial information from long-term visual behaviour (Bulling *et al.*, 2013).

In addition to reading detection, Ishimaru et al. exploited eye blink frequency and head motion patterns to recognise whether a user is talking, watching TV, or solving mathematical problems (Ishimaru *et al.*, 2014) using a video-based method. Mobile eye tracking devices are also able to automatically detect task transition and non-transition states and to estimate increasing levels of perceptual and cognitive load by measuring pupillary response, blinks and saccadic movements during display interaction (Chen *et al.*, 2013).

1.5.2 Joint Eye- and Scene-Based Applications

The combination of eye and scene information gained from mobile eye trackers provides the necessary data to analyse users' gaze-based visual and attentive behaviour in natural environments (Land and Tatler, 2009).

As reading is the most investigated activity in remote eye tracking research (Rayner, 2012) novel applications were developed for mobile eye tracking. Using video-based approaches, Kunze et al. developed gaze-based applications which may be used in a reading assistant to detect the documents a user is reading (Kunze *et al.*, 2013c), to automatically estimate how many words a user reads per day (Kunze *et al.*, 2013b), or to infer a user's language experience (Kunze *et al.*, 2013a).

Ogaki et al. coupled eye motion and ego-motion features from a first-person scene camera to recognise user activities (Ogaki *et al.*, 2012). Others only rely on ego-motion cues to detect users' engagement in the environment (Su and Grauman, 2016) or users' actions in sports (Kitani *et al.*, 2011). A recent work of Ma et al. combined ego-motion, gaze interaction, and scene content information for deeply trained action and activity recognition from an egocentric scene perspective (Ma *et al.*, 2016).

In terms of psychology research, users' attentive behaviour has been used to infer the relevance of real-world objects (Kandemir and Kaski, 2012) or to automatically discriminate among objects according to their interest (Adiba *et al.*, 2016). Hoppe et al. analysed users' gaze interaction behaviour to automatically recognise different levels of curiosity (Hoppe *et al.*, 2015) as well as users' personality traits (Hoppe *et al.*, 2018).

Users' gaze derived from head-mounted eye trackers further demonstrated fast, accurate, and natural interaction with both ambient displays (Stellmach and Dachselt, 2013; Lander *et al.*, 2015; Turner *et al.*, 2014), including target selection via looking (Stellmach and Dachselt, 2012) and eye typing on a virtual keyboard (Majaranta and Räihä, 2002), and body-worn ones, such as smartwatches (Akkil *et al.*, 2015; Esteves *et al.*, 2015).

1.6 Aims of the Thesis

The aim of this thesis is to lay the foundation for the breakthrough of mobile eye tracking, for everyone. To reach this aim and to start exploiting the full potential of mobile eye tracking numerous teething problems have to be solved first. Technical challenges and social obstacles are addressed to make mobile eye tracking fully functional in unconstrained settings, socially acceptable, and privacy-aware. Based on these solutions, this thesis has developed novel, useful, and exciting applications to convince future users of the value of mobile eye tracking.

1.6.1 Technical Challenges

The paradigm change from remote to mobile eye tracking offers great potential to apply eye tracking even in unconstrained environments, but at the cost of a number of technical burdens. Specifically, this thesis investigates innovative solutions for sensing and calibration to improve the gaze estimation accuracy of mobile eye trackers as well as eye movement analysis in mobile settings, given the lack of suitably robust fixation detection in the presence of user and gaze target motion.

Sensing. To record eye movement and egocentric scene video data in this thesis, Pupil Labs eye trackers, as well as their IR interpolation-based pupil detection software (Kassner *et al.*, 2014), were used. In contrast to commonly used mapping functions relying on 2D pupil and 2D gaze positions (cf. Section 1.2.2), this work expected there to be considerable advantages to the idea of using 3D information to improve the gaze estimation accuracy. Therefore, in Chapter 3 this thesis investigates whether using 3D data for the calibration of a mobile eye tracker can improve the gaze estimation performance. Contrary to existing approaches, this thesis formulates 3D gaze estimation as a direct mapping task from 2D pupil positions in the eye camera image to 3D gaze directions in the scene camera coordinate system. This requires the collection of 2D pupil positions as well as 3D target points for the calibration procedure, with the objective to reduce the gaze estimation error in comparison to common 2D-to-2D calibration methods.

Analysis. The use of mobile eye tracking enables new freedom of movement, but it forces research to deal with the appearance of novel eye movements which rarely appear in a remote setup (cf. Section 1.3.1). Definitions of eye movements, like classical fixations, need to be extended to mobile settings, where maintaining gaze on a particular real-world target consequently involves a complex combination of fixations, smooth pursuit, and VOR movements so that robust fixation detection is profoundly challenging. The interaction between head and eye movements (Doshi and Trivedi, 2012) as well as vision (Gegenfurtner, 2016) to understand the natural human attentive behaviour shifts VOR movements into focus. These eye movements are omnipresent in mobile settings and play the dominant role, keeping objects of interest in the centre of the user's field of view (Laurutis and Robinson, 1986; Fetter, 2007; Daye *et al.*, 2015).

In the absence of a suitable method which is able to deal with natural head and body motion, one aim of this thesis was the development of robust fixation detection in mobile settings. In order to meet this objective, in Chapter 4 this work investigates the joint analysis of scene and gaze information, as prior work has already pointed out that the analysis of a user's eye movements from the eye camera alone is not enough (Kinsman *et al.*, 2012). The proposed method relies on the assumption that, independent of user and gaze target motion, a target's appearance remains about the same during a fixation. Specifically, the method extracts image information from small regions around the current gaze position and analyses the appearance similarity of consecutive gaze patches across video frames to detect fixations in mobile settings.

Apart from the gaze-based analysis of human attentive fixation behaviour, another goal of this thesis focuses on the analysis of users' eye movements recorded from the eye camera alone. With this goal in mind, in Chapter 8 this thesis investigates whether the EOG-based encoding approaches of Bulling et al. (Bulling *et al.*, 2011b) can be adapted towards an application to video-based eye tracking data. This essential step enables the development of novel encoding approaches to express complex eye movement sequences as a bag-of-words representation applicable to discover users' activities without scene information.

However, for the development of useful and exciting applications, this thesis complements the investigated pupil-based eye movement encoding with visual features from the scene, either as CNN features (see Chapter 6) or context information such as from semantic scene segmentation, object detection, and depth reconstruction (see Chapter 9).

1.6.2 Social Obstacles

In addition to technical challenges, future eyewear devices need to overcome two crucial social obstacles which are inevitable in enabling eye tracking for everyone: 1) social acceptance and 2) privacy protection for both wearers and bystanders.

Acceptability. To achieve social acceptability of eye-tracking-enabled HMD this thesis aims to move on from the current bulky to a novel unobtrusive design, improving a user's comfort (Montero *et al.*, 2010) and trust in the novel technology, but becoming invisible for bystanders (cf. Section 1.4.1). Thus, Chapter 5 presents the design and implementation of *InvisibleEye*, a novel lightweight sensing and design approach for mobile eye tracking that uses millimetre-size RGB cameras that can be fully embedded into normal glasses frames. To compensate for the cameras' low image resolution of only a few pixels, a calibration-free approach is investigated, which uses multiple cameras in parallel and learning-based gaze estimation to regress a user's gaze position in the scene camera coordinate system, instead of using classical approaches for gaze estimation and calibration (cf. Section 1.2.2).

Privacy. Besides social acceptability, privacy is a core argument for future customers accepting or rejecting new technology (cf. Section 1.4.2). A careful investigation of previous research on privacy-related eye tracking identified three major limitations this thesis aims to address.

First, there is a lack of even basic understanding of users' privacy concerns with eye tracking in general and eye movement analysis in particular. Therefore, a goal of this thesis was to conduct a large-scale online survey to provide the first comprehensive account of with whom, for which services, and to what extent users are willing to share their gaze data (see Appendix C).

Second, eye tracking methods currently fail to preserve the privacy of user attributes, such as their gender or identity inferrable from their eye movement behaviour. Thus, another aim of this work was to make the first crucial step towards a new generation of eye tracking systems that respect and actively protect private information that can be inferred from a user's eyes. The approach taken in this work is used to protect a user's privacy in eye tracking based on the paradigm of differential privacy (DP). DP is a well-studied framework in the privacy research community (Dwork *et al.*, 2014). It adds a specific amount of noise to aggregated eye movement data, which decreases the chance of inferring privacy-sensitive information to a minimum while, at the same time, still allowing the use of the data for desired applications, such as activity recognition (Steil and Bulling, 2015) (see Chapter 7).

Third, there is still no solution for how to increase users' trust in HMDs equipped with scene cameras which are able to record users' sensitive data, such as pin entry or cash withdrawal at the ATM, and bystanders, e.g. during social interactions. However, preserving users' and bystanders' privacy is essential for the social acceptability of mobile eye tracking devices. Koelle et al. identified in their studies the necessity to inform bystanders whether a device is working or not, using e.g. a self-made or 3D-printed solution like the "Glass Privacy Cover"[14] (Koelle *et al.*, 2015). Thus, this work presents and investigates *PrivacEye*, the first prototype system and method that combines the analysis of egocentric scene image features with eye movement analysis to detect privacy-sensitive everyday situations and automatically enables and disables the eye tracker's first-person camera using a non-spoofable mechanical shutter (see Chapter 6). If a privacy-sensitive situation is detected, the scene camera is occluded. To open the shutter without visual input, the proposed method detects changes in users' eye movements alone to gauge changes in the "privacy level" of the current situation. The developed prototype aimed to protect both user and bystander privacy with no action required, while increasing the social acceptability by communicating the current status of the egocentric scene camera using a physical shutter (Koelle *et al.*, 2018b).

1.6.3 Novel Applications

Recent technology developments showed that even a well-designed product, like Google Glass, fails if it cannot answer the question: "What problem does it solve or why would

[14] `https://www.thingiverse.com/thing:96237`, date: 12.07.2019

I need it?”[15] Hence, while technical and social challenges can be overcome by more sophisticated sensing and analysis approaches and improved hardware design, the key to encourage use of mobile eye tracking devices is the development of innovative and essential applications (cf. Section 1.5). Therefore, a major aim of this thesis was the development of three highly useful and promising future applications 1) to protect users' privacy, 2) to recognise their activities, and 3) to forecast their attentive behaviour.

Privacy-Aware Eye Tracking. As already introduced in Section 1.6.2, the first of these applications is a novel privacy-preserving mechanism based on differential privacy. Chapter 7 evaluates the effectiveness of this application in a realistic use case to prove the preservation of users' privacy without impeding data utility. It endeavours to protect users' private attributes inferrable from eye movements recorded by the eye camera alone (cf. Section 1.4.2). This highly useful application contributes to a sense of security, resulting in increased trust and reduced social reservations. However, this application is essential and needs to become state-of-the-art in all eye tracking devices, fostering the success of functional applications offering a greater degree of privacy and security.

As HMD devices are either equipped with only an eye camera or with an eye and scene camera, this thesis had the goal to develop two additional future applications which have an inherent potential to attract and convince users of the effectiveness of mobile eye tracking, focusing on long-term usage.

Activity Recognition. The second application proposes an unsupervised method for eye-based discovery of everyday activities combining the bag-of-words visual behaviour representation, introduced in Section 1.6.1, with a latent Dirichilet allocation (LDA) topic model, taking no information from a scene camera into account. In Chapter 8 the proposed application is evaluated against state-of-the-art supervised methods investigating a variety of novel eye movement encoding approaches. Beyond activity recognition, this application enables long-term lifelogging (Ishiguro *et al.*, 2010; Bulling *et al.*, 2013), diary functions using video summarisation (Salvucci and Anderson, 2001; Lee *et al.*, 2012) and video captioning (Sah *et al.*, 2017), routine or event detection (Zhong *et al.*, 2004; Itti and Baldi, 2005; Kolski *et al.*, 2007), or it can serve as a memory aid (Hodges *et al.*, 2006; Piasek *et al.*, 2014) enhancing users' cognitive performance (Silva *et al.*, 2013).

Attention Forecasting. With the third application, this thesis aims to address a problem of our ever-accelerating world, where the ability of people 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). Everyone wants to self-quantify and optimise themselves in all areas of life, but the increase in external stimuli leads to highly fragmented attention, which is particularly prevalent during mobile interactions (Oulasvirta *et al.*, 2005). Thus, the active management of user attention will be a key task of mobile eye tracking devices (Bulling, 2016). In comparison to prior work which developed attentive user interfaces only capable of

[15] `https://www.forbes.com/sites/ianaltman/2015/04/28/why-google-glass-failed-and-why-apple-watch-could-too/`, date: 12.07.2019

adapting after the fact, i.e. after users have already shifted their attention towards a novel stimulus (Kern *et al.*, 2010; Mariakakis *et al.*, 2015; Gutwin *et al.*, 2017), in Chapter 9 this thesis proposes a method to predict attentive behaviour during everyday mobile interactions. Specifically, the work investigates the prediction abilities of the method for bidirectional attention shifts between a mobile device and the environment, respectively, as well as users' primary attentional focus for the near future from real phone-integrated and body-worn sensors combining eye and scene information. This new generation of mobile attentive user interfaces pro-actively adapts to imminent shifts of user attention, i.e. before these shifts actually occur, and enables a variety of exciting new applications. To reduce the interaction delay with mobile devices, predicted attention shifts to mobile devices could trigger unlocking the device and loading of previous screen content, whereas predicted attention shifts to the environment can be used to alert users to keep their attention on the current task. With the knowledge that users' attention will stay on the device for a specific amount of time, attentive user interfaces could display important information or alert the user in the case of potentially dangerous external events that they might miss during their interaction.

1.7 Outline of the Thesis

This thesis encompasses eight scientific publications that address the aims described in the last section. Figure 1.8 visualises the aims of the thesis and the chapters that cover them. Table 1.1 links each of these chapters to the corresponding publication.

Chapter 2 provides a summary of contributions, an outlook on upcoming future work and concludes with an emphasis on the significance of the thesis. Part I of this thesis, consisting of Chapter 3 and 4, covers the technical challenges of mobile eye tracking introducing a novel calibration approach and a new method to analyse users' fixational behaviour in mobile settings. Chapters 5 and 6 form Part II of this work. It contributes to the overcoming of social obstacles, presenting a novel unobtrusive design approach for head-mounted eye trackers as well as a privacy-preserving method which automatically detects privacy-sensitive situations. Chapters 7, 8, and 9 comprise Part III of the thesis and present innovative and essential novel applications to protect users' privacy, to recognise their activities, and to forecast their temporal attentive behaviour.

Finally, the appendix details the findings introduced in Chapters 3, 6, and 7. Appendix A provides an extended evaluation of the 3D calibration approach described in Chapter 3. Appendix B contains study descriptions and error case analyses of the method to detect privacy-sensitive situations in everyday life explained in Chapter 6. Appendix C collects the full results and detailed numbers from statistical tests of a large-scale online survey on privacy aspects of eye tracking.

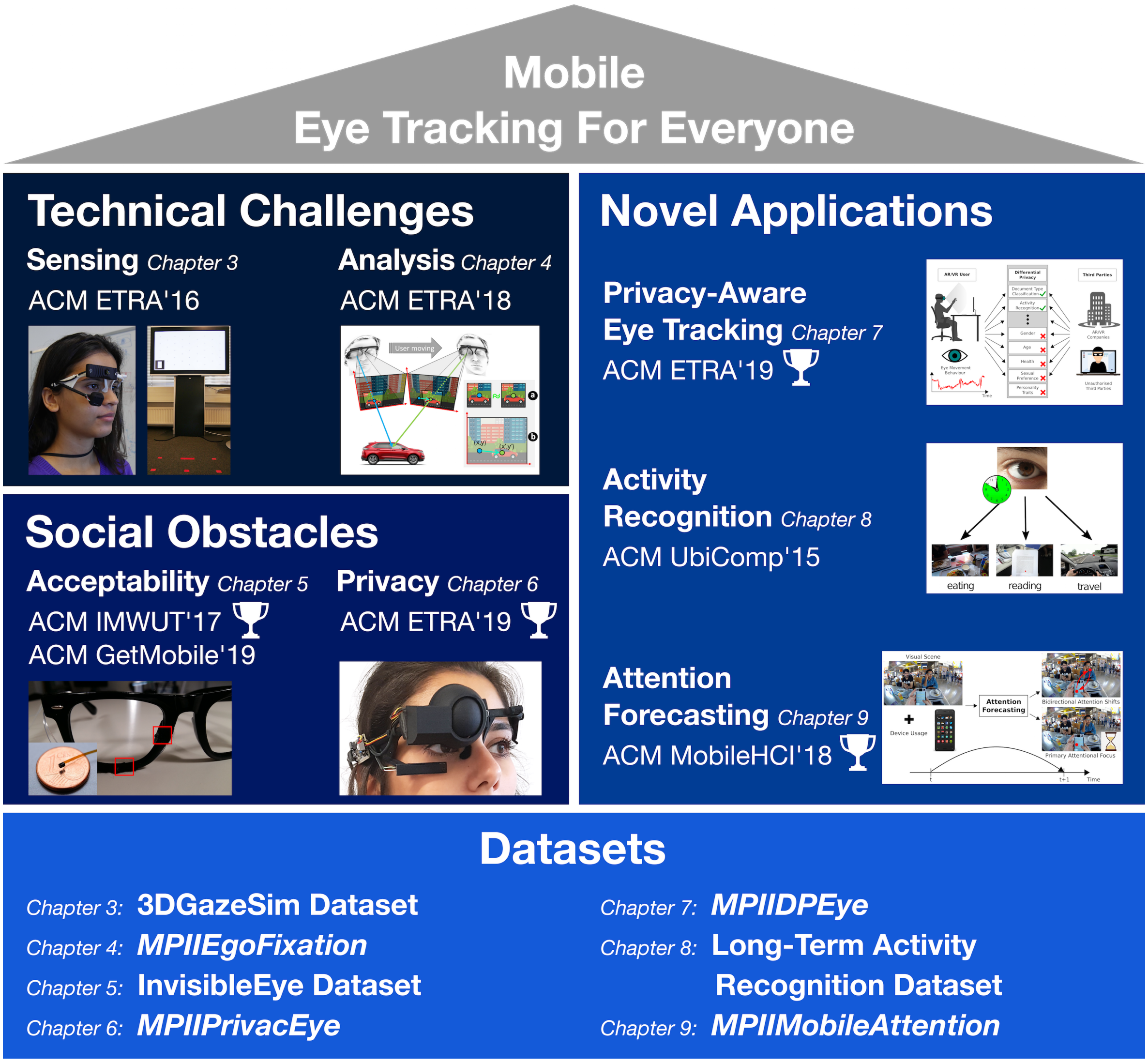


Figure 1.8: Outline of the chapters included in this thesis according to the aims presented in Section 1.6 (see Table 1.1 for the corresponding publications).

Chapter	Publication
3	*3D Gaze Estimation from 2D Pupil Positions on Monocular Head-Mounted Eye Trackers* Mohsen Mansouryar, Julian Steil, Yusuke Sugano, and Andreas Bulling; In Proc. of the ACM International Symposium on Eye Tracking Research and Applications (**ETRA**), 2016. (Mansouryar *et al.*, 2016)
4	*Fixation Detection for Head-Mounted Eye Tracking Based on Visual Similarity of Gaze Targets* Julian Steil, Michael Xuelin Huang, and Andreas Bulling; In Proc. of the ACM International Symposium on Eye Tracking Research and Applications (**ETRA**), 2018. (Steil *et al.*, 2018a)
5	*InvisibleEye: Mobile Eye Tracking Using Multiple Low-Resolution Cameras and Learning-Based Gaze Estimation* 🏆 **Distinguished Paper Award** Marc Tonsen, Julian Steil, Yusuke Sugano, and Andreas Bulling; Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (**IMWUT**), Vol. 1, No. 3, 2017. (Tonsen *et al.*, 2017)
	InvisibleEye: Fully Embedded Mobile Eye Tracking Using Appearance-Based Gaze Estimation Julian Steil, Marc Tonsen, Yusuke Sugano, and Andreas Bulling; ACM GetMobile: Mobile Computing and Communications (**GetMobile**), Vol. 23, No. 2, 2019. (Steil *et al.*, 2019c)
6	*PrivacEye: Privacy-Preserving Head-Mounted Eye Tracking Using Egocentric Scene Image and Eye Movement Features* 🏆 **Best Video/Demo Award** Julian Steil, Marion Koelle, Wilko Heuten, Susanne Boll, and Andreas Bulling; In Proc. of the ACM International Symposium on Eye Tracking Research and Applications (**ETRA**), 2019. (Steil *et al.*, 2019b)
7	*Privacy-Aware Eye Tracking Using Differential Privacy* 🏆 **Best Paper Award** Julian Steil, Inken Hagestedt, Michael Xuelin Huang, and Andreas Bulling; In Proc. of the ACM International Symposium on Eye Tracking Research and Applications (**ETRA**), 2019. (Steil *et al.*, 2019a)
8	*Discovery of Everyday Human Activities from Long-Term Visual Behaviour Using Topic Models* Julian Steil and Andreas Bulling; In Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (**UbiComp**), 2015. (Steil and Bulling, 2015)
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)

Table 1.1: Publications and corresponding chapters included in this thesis.

2 Thesis Summary

This chapter summarises the contributions of this thesis consisting of solutions for fundamental *technical challenges*, the overcoming of current *social obstacles*, and the development of exciting, *novel applications*, laying foundations for mobile eye tracking for everyone. In the following, this thesis discusses each challenge, summarises how they were addressed, and points out the specific contributions which were made. Additionally, the corresponding self-recorded datasets which are essential to test and evaluate the contributions are shortly explained. This chapter concludes with an outlook on upcoming future work and an emphasis on the significance of this thesis enabling further advances in mobile eye tracking.

2.1 Summary of Contributions

This section presents the most important results and contributions this thesis made to enable mobile eye tracking for everyone. The summary is structured according to the aims of this work introduced in Section 1.6 and illustrated in Figure 1.8. Detailed descriptions and discussions can be found in the particular publication chapters referenced in Table 1.1:

- **Technical Challenges** – In order to advance mobile eye tracking and to make it fully functional in unconstrained environments, there are two questions this thesis answered first:
  1) Can 3D calibration data improve the gaze estimation of a mobile eye tracker? (see Section 2.1.1 and Chapter 3)
  2) How can fixations be robustly detected in mobile settings? (see Section 2.1.2 and Chapter 4)
- **Social Obstacles** – The second goal of this thesis was to overcome social obstacles hindering the everyday usability and interaction with mobile eye tracking devices as well as to diminish privacy concerns of users and bystanders. To achieve social acceptability, this work developed a first prototype which improves on the bulky and rather functional design of state-of-the-art mobile eye trackers using millimetre-size RGB cameras that can be fully embedded into normal glasses frames (see Section 2.1.3 and Chapter 5). In order to increase the users' trust in the privacy-protection abilities of mobile eye trackers, this thesis built another prototype equipped with a mechanical shutter which automatically closes in everyday privacy-sensitive situations (see Section 2.1.4 and Chapter 6).
- **Novel Applications** – After solving technical challenges and overcoming social obstacles, this work developed exciting, novel applications which can be applied to the data recorded with mobile eye tracking devices.

  Complementing the hardware solution in Chapter 6, this thesis presented the first method for privacy-aware eye tracking using the differential privacy paradigm, minimising the chance to infer privacy-sensitive information from eye movement data while maintaining data utility (see Section 2.1.5 and Chapter 7). Further, this work particularly focused on the eye-movement-based recognition of everyday activities, leveraging novel encoding methods for blinks, fixations, and saccades (see Section 2.1.6 and Chapter 8). Finally, a variety of future applications were discussed, enabled by the core function of forecasting temporal allocation of users' overt visual attention from phone-integrated and head-mounted sensors (see Section 2.1.7 and Chapter 9).

2.1.1 3D Data for Calibration

Challenge. To enable gaze analysis on real-world objects and scenes, 3D gaze estimation is a key component. However, prior works suffer from parallax error caused by the offset between the scene camera origin and eyeball position (Mardanbegi and Hansen, 2012; Duchowski *et al.*, 2014). The approximated 3D gaze is obtained by projections from estimated 2D gaze positions in the scene camera image to a corresponding 3D scene reconstruction (Munn and Pelz, 2008; Pfeiffer and Renner, 2014; Takemura *et al.*, 2014), which leads to inaccurate results. Solutions for this problem have been widely studied in remote eye tracking with the aid of special hardware, such as multiple IR light sources and stereo cameras (Beymer and Flickner, 2003; Nagamatsu *et al.*, 2010). The mobile eye tracking use case has only been evaluated with synthetic eye images (Świrski and Dodgson, 2013). Thus, it is still unclear whether 3D gaze estimation can be done properly in real-world environments and whether calibration at a single depth is sufficient or not when using lightweight head-mounted eye tracking devices.

Contributions. Instead of opting for a 3D-to-3D mapping which links 3D eyeball poses to a 3D position in the scene, this thesis proposed a **method which directly maps 2D pupil positions in the eye camera to 3D gaze target points in scene camera coordinate space**. This hybrid 2D-to-3D mapping approach is evaluated for both simulated and real data against 3D-to-3D and error-prone state-of-the-art 2D-to-2D gaze estimation approaches (Mansouryar *et al.*, 2016). As displayed in Chapter 3, the **2D-to-3D mapping approach significantly reduced the parallax error** and outperformed the 2D-to-2D mapping in the simulation environment. Combining calibration data from five different depths in the real-world setting, the **2D-to-3D mapping approach achieved an angular error of less than 1.3° and outperformed the 2D-to-2D as well as the 3D-to-3D approach**. In Appendix A a more detailed analysis and discussion with corresponding performance plots can be found. With the collected dataset described below, this thesis provides a solid basis for future research on 3D gaze estimation with lightweight head-mounted devices.

Dataset. To evaluate the three introduced mapping approaches, two studies to record data from 1) a simulation environment and 2) 14 participants in a real-world environment were conducted. In the simulated environment, calibration and test data was recorded from a 5×5 calibration and 4×4 test point grid in five calibration depths between one and two meters. The simulation environment relied on a basic model of the human eye consisting of a pair of spheres (Lefohn *et al.*, 2003) and pin-hole camera models for the eye and scene camera. The real-world data was recorded with a Pupil head-mounted eye tracker (Kassner *et al.*, 2014). Similar to the simulation environment, 14 participants looked first at a calibration and then on a test point grid displayed on a public display at corresponding distances. The visual stimuli in the calibration and test grid were designed as AR markers so that their 3D positions could be obtained. The open-source simulation environment and the resulting 3DGazeSim dataset are available at `https://www.mpii.de/3DGazeSim/` (date: 12.07.2019).

2.1.2 Fixation Detection in Mobile Settings

Challenge. Accurate fixation detection in mobile settings is becoming increasingly important (Kurzhals *et al.*, 2017) with the spread of lightweight and affordable head-mounted eye trackers (Kassner *et al.*, 2014; Tonsen *et al.*, 2017). Although sufficient to detect fixations in stationary settings, but not in mobile settings, state-of-the-art dispersion-based, velocity-based (Holmqvist *et al.*, 2011), and data-driven approaches (Urruty *et al.*, 2007) rely solely on gaze data without taking any scene information into account. The omnipresent head and scene dynamics add an additional challenge to detecting fixations in mobile settings, as there is no fixed frame of reference, so that gaze estimates feign to shift within the scene camera coordinate system during a fixation. Consequently, maintaining gaze on a particular real-world target involves a complex combination of fixations, smooth pursuit, and VOR movements.

Contributions. To address this challenge, in Chapter 4 this thesis presented **the first robust and effective fixation detection method for head-mounted eye trackers** which combines gaze data and egocentric scene content information. The method exploits the fact that independent of user or gaze target motion, target appearance remains about the same during a fixation (Steil *et al.*, 2018a). To detect fixations, image information from small regions around the current gaze position are extracted and the appearance similarity of these gaze patches is analysed across consecutive video frames using a deep convolutional image patch similarity network (Zagoruyko and Komodakis, 2015). The **method outperformed widely used velocity- and dispersion-based algorithms**, particularly with respect to the total number of correctly detected fixations as well as insertion and merge event errors. These **results highlight the significant potential of joint analysis of scene information and gaze data** for fixation detection.

Given the advance of mobile eye tracking and the emerging attention to mobile computing, the method contribution presented in this thesis opens up numerous opportunities for applications and user experience studies as well as gaze behaviour and augmented reality research.

Dataset. To evaluate the proposed fixation detection method, a subset of a recent mobile eye tracking dataset (Sugano and Bulling, 2015) was annotated in a fine-grained way, on the individual video frame level. The annotated subset contains five videos, each lasting five minutes, resulting in over 2,300 annotated fixations and more than 40,000 frames. The chosen dataset is particularly suitable because the participants were always in motion, leading to a large amount of head motion and scene dynamics, which is both challenging and interesting for the fixation detection task.

The final *MPIIEgoFixation* annotation dataset is publicly available at `https://www.mpi-inf.mpg.de/MPIIEgoFixation/` (date: 12.07.2019).

2.1.3 Unobtrusive Mobile Eye Tracking

Challenge. Despite increasing importance and functionality of mobile eye tracking devices, they still suffer from the fundamental problem of a lack of social acceptability, mainly caused by their obtrusive and bulky appearance. This leads to unnatural behaviour of both the wearers and bystanders (Risko and Kingstone, 2011; Nasiopoulos *et al.*, 2015). Currently available eye trackers rely preferably on heavyweight high-quality image sensors, which often occlude a user's field of view. Although these components guarantee high-resolution recordings and real-time data processing with reliable gaze estimation, they cause discomfort or even pain, especially during long-term recordings, while the external perception of the device's design decreases a user's confidence and assurance during device interaction.

Contributions. To address these challenges, in Chapter 5 this thesis presented ***InvisibleEye*, a novel design approach for mobile eye trackers**. The use of millimetre-size RGB cameras reduces the image sensor size significantly, enabling full integration of eye tracking into normal glasses frames without occluding a user's field of view. To compensate for the low resolution of the frame-embedded cameras this thesis developed a **calibration-free learning-based gaze estimation approach** which directly regresses a user's gaze direction from eye images using multiple cameras in parallel.

InvisibleEye was evaluated on **three large-scale, increasingly realistic, and thus challenging datasets**. In the first experiment, the design space for fully embedded mobile eye tracking was investigated using synthetic eye image data, opting for the minimum required number and positions of eye cameras. Based on the findings of the first experiment, the **first hardware prototype** was built, attaching four Awaiba NanEye cameras to a pair of safety glasses, and **evaluated on real images collected in a controlled laboratory environment**. Finally, a **second binocular hardware prototype** was constructed, featuring three Pupil eye camera pairs mounted in a 3D-printed frame **recording each eye respectively in a mobile setting**. In this most challenging setup, *InvisibleEye* achieved a top person-specific **gaze estimation accuracy of 1.79° and 2.04° using three camera pairs at 5×5-pixel and 3×3-pixel resolution**, respectively (Tonsen *et al.*, 2017; Steil *et al.*, 2019c).

With the novel design approach, this thesis underlines the significant potential for finally realising the vision of invisible mobile eye tracking.

Dataset. For the first experiment, highly-realistic and perfectly annotated eye region images were generated using the computer graphics eye region model of UnityEyes (Wood *et al.*, 2016b). A set of 1,600 different eyeball poses for varying eye regions, camera angles, and lighting conditions was recorded.

The first prototype was used to record a first-of-its-kind 17-participant (12 male, 5 female) dataset with 280,000 close-up eye images that were captured from multiple views. Each image was annotated with a corresponding ground-truth gaze direction. Participants were instructed to look at a series of gaze targets shown on a computer screen.

Using the second prototype, another dataset of 240,000 eye images was recorded with four participants (4 male) in a mobile setting. Participants were asked to position themselves at an arbitrary distance of up to 3 meters in front of a calibration marker attached to a wall while performing a series of head movements gazing at the marker.

The dataset recorded with the first prototype is available at `http://www.mpi-inf.mpg.de/invisibleeye/` (date: 12.07.2019).

2.1.4 Detection of Privacy-Sensitive Situations

Challenge. First-person cameras increasingly integrated into head-mounted eyewear devices can pose a significant threat to user and bystander privacy. Especially threatening is the "always-on" characteristic of these devices, which leads to social frictions on the part of bystanders and potential capturing of unintended and potentially sensitive imagery by the wearer. Pure manual control of the scene camera increases users' workload and causes stress. There is the permanent danger that users may forget to turn off the scene camera, so that privacy-sensitive situations or even bystanders are recorded. Another significant drawback of current eyewear devices is the lack of a clear indication of whether the scene camera is recording or not. This additionally increases the privacy threat (Koelle *et al.*, 2015, 2018a) and uncomfortable feeling for bystanders (Bohn *et al.*, 2005; Denning *et al.*, 2014; Ens *et al.*, 2015).

Contributions. To address these challenges, this thesis developed ***PrivacEye*****, a prototype system and method that is able to automatically de-activate and re-activate a front-facing scene camera using a physical, non-spoofable shutter** (Steil *et al.*, 2019b). The key idea and core novelty of the method is the **combination of information gained from both eye and scene camera to detect privacy-sensitive everyday situations** to close the camera shutter without the need of users or bystanders taking action. As no scene information is available if the camera shutter is closed, users' eye movement behaviour alone triggers the reopening of the shutter. With a quantitative evaluation of *PrivacEye* in Chapter 6, the **best performance of 73% accuracy could be achieved in a person-specific realistic sequential analysis over a 17-participant long-term dataset**. Appendix B details results of an in-depth error case analysis. To provide **insights on perceived social acceptability, trustworthiness, and desirability, user feedback was collected from 12 semi-structured interviews**. With *PrivacEye* this thesis presented the first privacy-preserving head-mounted eye tracking method that opens up a new and promising direction for future work contributing to users' and bystanders' sense of security, trust, and social acceptability.

Dataset. As none of the eye movement datasets published in the recent years focused on privacy-related attributes, this thesis made use of a previously recorded dataset, called *MPIIMobileAttention* (Steil *et al.*, 2018b), presented in Chapter 9.4. This dataset covers a rich set of representative real-world situations, including sensitive environments and tasks of 20 students during a regular day at university. In order to gain knowledge about participants' privacy sensitivity, they were re-invited to fully annotate their recorded data themselves with continuous annotations of location, activity, scene content, and their subjective privacy sensitivity level, rated on a 7-point Likert scale ranging from 1 (privacy-sensitive) to 7 (non-sensitive). 17 out of the 20 participants finished the annotation of their own recordings, resulting in about 70 hours of annotated video data. The corresponding *MPIIPrivacEye* dataset is available at `https://www.mpi-inf.mpg.de/MPIIPrivacEye/` (date: 12.07.2019).

2.1.5 Protection of Private Information Inferable from Eye Tracking Data

Challenge. With the increasing integration of eye tracking into VR and AR headsets, the protection of private user information inferable from recorded eye tracking data has emerged as an urgent and important topic. The equally rapidly improving capabilities of these devices to sense, analyse, and exploit (Hansen *et al.*, 2003; Vertegaal *et al.*, 2003; Stellmach and Dachselt, 2012) rich information contained in human eye movements (Bulling *et al.*, 2011a), such as personal preferences, goals, or intentions, further foster the rise of eye tracking technology. To pave the way for privacy-aware eye tracking, two major limitations need to be overcome. First, it is essential to understand users' privacy concerns about eye tracking in general. Second, based on this gained knowledge, novel eye tracking methods need to be developed which preserve users' private information while maintaining the general functionality of already available applications.

Contributions. To address both limitations, **the first large-scale online survey on privacy aspects of eye tracking and eye movement analysis** was conducted in this thesis. The survey results, condensed in Chapter 7, provided the first comprehensive account of with whom, for which services, and to what extent users are willing to share their eye tracking data (Steil *et al.*, 2019a).

Informed by the survey, this thesis presented the **first privacy-aware method using differential privacy**. The proposed approach adds noise to the recorded eye tracking data, which minimises the chances to infer privacy-sensitive information while, at the same time, still preserving the utility of the data. To prove the effectiveness of this method in a realistic and authentic future application, this thesis opted for a reading task in VR. The results reveal that **differential privacy decreases an attacker's probability of inferring a user's gender or re-identifying a user to chance level while at the same time the data utility**, evaluated as the ability to detect the document classes the participants read, **is maintained**.

With the survey and differential privacy approach for eye tracking data, this thesis laid important foundations for future research on privacy-aware gaze interfaces that respect and actively protect private information that can be inferred from our eyes.

Survey and Dataset. In the survey, 124 people from 29 different countries participated, answering more than 100 questions. The majority of participants were young and educated people with a technical background most likely to experience AR or VR technology. All questions, detailed numbers, and plots can be found in Appendix C.

To evaluate the privacy-preserving eye tracking method, a novel dataset was recorded, called *MPIIDPEye*. 20 participants (10 male, 10 female) read three documents in VR wearing an Oculus DK2 headset equipped with Pupil eye tracking add-ons (Kassner *et al.*, 2014). The recording of each of these documents was about a 10-minute read, depending on a user's reading skill (∼10 hours in total). Ground truth information was given by the participant's gender, the document type, and IDs assigned to each participant.

The complete dataset is available at `https://www.mpi-inf.mpg.de/MPIIDPEye/` (date: 12.07.2019).

2.1.6 Efficient Encoding of Eye Movements for Activity Recognition

Challenge. Human vision is a valuable and rich source of information with significant potential for activity recognition and computational behaviour analysis. However, previous works focused on supervised methods and the recognition of predefined activity classes based on short-term eye movement recordings. Therefore, the following challenging questions need to be answered:

1) How much information about daily routines is contained in long-term human gaze behaviour?
2) How can this information be extracted, encoded, and modelled efficiently?
3) Is it possible to discover human activities from eye tracking data in an unsupervised manner?

Contributions. To answer the first question the **first video-based 10-participant long-term eye tracking dataset with more than 80 hours of egocentric and eye video data** (Steil and Bulling, 2015) was recorded and annotated in this thesis. The second question was addressed by **stepping from EOG-based to video-based eye movement analysis**, adapting eye movement encoding methods initially introduced by Bulling et al. (Bulling *et al.*, 2012). In addition, as covered in Chapter 8, this thesis developed a **novel approach for combining the encodings of blinks, fixations, and saccades, enabling a bag-of-words representation of users' gaze behaviour**. The third question is solved by a **novel method which combines bag-of-words eye movement behaviour representations with a latent Dirichlet allocation (LDA) topic model that is able to discover everyday human activities**. In comparison to prior work, the proposed **method is fully unsupervised**, i.e. it does not require manual annotation of gaze behaviour **and is able to deal with an arbitrary number of activity classes**. It not only extracts information from saccade sequences, but learns a more holistic model of gaze behaviour from saccades, fixations, and blinks. The performance evaluation of the proposed method shows its ability to **discover everyday activities with a performance competitive with that of previously published supervised methods**, achieving a maximum performance of an F1 score over 90% for watching media and a top average performance of 74.75% for reading. These results reveal the significant information contained in users' gaze behaviour and open up a new venue for future eye tracking applications.

Dataset. To evaluate whether the proposed fully unsupervised method is competitive against supervised approaches, a new long-term gaze dataset was collected. It contains natural gaze behaviour of 10 participants, recording each participant for about eight hours, respectively. The data was collected with a state-of-the-art head-mounted Pupil eye tracker continuously worn by the participants for a full day of their normal life. The dataset was annotated for evaluation purposes with eight (non-mutually exclusive) sample activity classes: 1) outdoor (7.8 hours), 2) social interaction (14.3 hours), 3) focused work (31.3 hours), 4) travel (8.3 hours), 5) reading (39.5 hours), 6) computer work (28.7 hours), 7) watching media (18.3 hours), 8) eating (7 hours), and periods with no specific activity (11.4 hours).

The continuously fully annotated dataset is available at `https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/human-activity-recognition/discovery-of-everyday-human-activities-from-long-term-visual-behaviour-using-topic-models/` (date: 12.07.2019).

2.1.7 Forecasting Human Attentive Behaviour During Mobile Interactions

Challenge. In particular during mobile interactions, users' visual attention is highly fragmented (Oulasvirta *et al.*, 2005). Nevertheless, pro-active management of users' challenging erratic nature of attention shifts promises a new generation of exciting, novel applications. The necessary precondition for these applications is the forecasting of users' temporal attention, which is very difficult to address even in stationary settings given the strong task dependence and inherent variability of users' gaze behaviour. Current mobile attentive user interfaces are not able to cope with the large number of real-world visual attractors and thus they are not able to adapt to imminent shifts before these shifts actually occur (Kern *et al.*, 2010; Mariakakis *et al.*, 2015; Gutwin *et al.*, 2017).

Contributions. To address this lack of application, this thesis proposed the **first proof-of-concept method that predicts users' temporal attentive behaviour during everyday mobile interactions** from real phone-integrated and body-worn sensors. Specificity, **three forecasting tasks** were presented that will facilitate pro-active adaptations to users' erratic attentive behaviour in future user interfaces: **1) attention shifts to the environment, 2) attention shifts to the mobile device, and 3) the primary attentional focus**.

The proposed method uses device-integrated and head-worn IMUs (Inertial Measurement Units) as well as computer vision algorithms for object detection, face detection, semantic scene segmentation, and depth reconstruction. In Chapter 9 the forecasting abilities for all three tasks were evaluated for different feature sets gained from the egocentric scene, the mobile phone, and users' gaze. The **method achieved robust performances with F1 scores of 0.8** for prediction shifts to the environment and back to the phone. Similar results could be achieved for the prediction of the primary attentional focus, where a combination of all available sensor sets reached an improved performance (Steil *et al.*, 2018b).

This thesis envisions a new generation of attentive user interfaces that exploit predicted attention shifts to the environment to show push notifications shortly before a predicted shift will happen or even to alert a user to keep their attention on finishing a task. Attention shifts to the mobile device could be used to unlock the device and loading the previous screen content to reduce interaction delays or to reorient the user on the mobile device, as well as to prevent attention shifts during face-to-face conversations. The primary attentional focus prediction tells whether the attention will be mainly on the mobile device or not, and can be used to alert the user in the case of potentially dangerous external events the user could miss while staring at the mobile device, or simply to show important information.

Dataset. To develop and evaluate the forecasting abilities for the three proposed tasks, the 90-hour multi-modal *MPIIMobileAttention* dataset was collected. 20 participants were recorded during everyday activities freely roaming on a local university campus for about 4.5 hours each while interacting with a mobile phone. During the recording, the participants wore a Pupil head-mounted eye tracker (Kassner *et al.*, 2014) with an additional stereo camera and carried a mobile phone in their hands as well as a recording laptop in a backpack. In contrast to prior work, this dataset collection did not impose a scripted sequence of activities or environments, to constrain participants as little as possible. During three 1.5-hour recording sessions, participants were engaged in chat sessions using WhatsApp, during which they had to perform web search tasks. In a chat session which consisted of six questions, users' attention was measured while finishing a given task in the so-called working time, as well as whether they can keep their attention on the mobile phone while waiting for the next question in the so-called waiting time. In total, 1,440 working and 1,200 waiting segments were recorded and fully annotated in terms of a participant's environment, mode of locomotion, and attention shift direction (from the environment to the mobile device or vice versa).

The complete *MPIIMobileAttention* dataset is available at `https://www.mpii.mpg.de/MPIIMobileAttention/` (date: 12.07.2019).

2.2 Future Work

This work identified current technical challenges and social obstacles that hinder mobile eye tracking from achieving increased functionality, usability, and social acceptability. It pointed out the urgent necessity for the development and implementation of novel and attractive future applications to convince users of the effectiveness of head-mounted devices with integrated eye tracking technology. Specifically, this work is a central starting point for new and promising future research which will push mobile eye tracking technology to the next level. However, not every problem could be solved in this work. Therefore, future work should address the following challenges:

- **Sensing** – The success of head-mounted eye tracking devices as consumer products strongly depends on the four following major factors which need to be solved by future work: 1) miniaturisation, 2) sensor combinations, 3) usability, and 4) real-time processing.

  The miniaturisation of visual sensors like the scene and eye cameras has a direct impact on users' social acceptance of eye tracking and the development of novel applications.

  This thesis already gave a foretaste of how sensor fusion will look in the future. The integration of IMUs into mobile eye trackers will become standard, so that they can be applied in combination with eye and scene cameras. The integration of improved sensors and novel ones, like GPS (Global Positioning System) and GSR (Galvanic Skin Response), will follow, opening a broad range of novel and exciting applications, and performance improvements to already available applications.

  Considering future calibration and gaze estimation approaches, given the irresistible demand for easing usability of eye tracking devices, and with larger datasets either recorded in real-world environments by real participants or in terms of data augmentation by artificially created eye tracking data, marker-based calibration methods and mapping functions as presented in Chapter 1 will become increasingly rare. They will be gradually replaced by calibration-free gaze estimation, which is deeply trained on a huge amount of eye tracking data.

  Especially the fourth factor, the availability of real-time processing, will open doors for the development of novel eye tracking applications. In the majority of works, recorded data was analysed post-hoc. The lack of powerful microprocessors was the bottleneck for real-time operations and CNN models for feature extraction necessary for a variety of applications. However, with the direct implementation in hardware (cf. Qualcom's Snapdragon 845) this problem will be solved. With real-time access, powerful deep-learning models, e.g. for object detection or semantic scene segmentation, can be applied to extract new sets of features, pushing the application performance and accuracy to undreamt-of levels.

- **Analysis** – This thesis focused on the detection of fixations in mobile settings in particular and showed that their classical definition does not hold in the presence of natural scene and body motion, so that it needs to be rethought and extended.

But this is only the beginning of future work focusing on the joint analysis of eye movements and visual scene content towards a deeper understanding of human gaze behaviour in unconstrained settings. In this context, VOR movements will play a central role in research and gaze-based interaction with real-world objects as a core component of fixations in mobile settings, covering user and gaze target motion to keep the objects of interest in focus. Further, this thesis was restricted to the currently available state-of-the-art eye tracking technology. However, technical development will enable higher frame rate recordings of mobile eye cameras, so that even smaller eye movements and expressions like microsaccades and tremors could be detected. Especially in psychology, information about these eye movements will trigger novel research in human gaze behaviour analysis, allowing research in unconstrained settings.

In addition, future eye movement analysis research will not continue relying on hand-crafted thresholds, but will be strongly influenced by deep-trained data-driven approaches, enabling a joint detection of different eye movements (Hoppe and Bulling, 2016; Zemblys *et al.*, 2018).

- **Social Acceptability** – For social acceptability of mobile eye tracking the miniaturisation of visual sensors with wireless recording opportunities in an unobtrusive, lightweight design will be crucial. With *InvisibleEye* this thesis paved the way for a novel unobtrusive design which was picked up and released by a first company[16]. However, wired connections to a laptop or mobile phone are still necessary. With the 5G network expansion larger data packages or even continuously recorded data could be transmitted and saved on servers. This will provide additional freedom in terms of locomotion and improvement of device usability. Future head-mounted devices will be oriented to this design approach and further extend the functionality, most likely using AR components.

- **Privacy** – The protection of users' and bystanders' privacy will be a dominating topic in the future. This year, the symposium on eye tracking research and applications (ETRA) organised the first panel discussion, which focused on the joint reflection of eye tracking and privacy and its importance for future work. If research does not intervene now, eye tracking will amount to an uncontrollable risk if users' gaze behaviour can be recorded at nearly unlimited scale as integral parts of VR or AR devices or even normal glasses frames. Legal regulation will provide guidelines for data hygiene, such as the GDPR (General Data Protection Regulation). However, for end-users, more trustworthy solutions than legislative texts need to be established on the hardware and software side.

  This thesis contributed the first proof-of-concept to detect privacy-sensitive situations. These findings serve as a basis but need to be improved, enabling generalised models which only rely on minimal user adaptations, given the challenge that privacy is highly person-specific.

[16] `https://pupil-labs.com/blog/2019-01/pupil-invisible-beta-launch/`, date: 12.07.2019

The application of differential privacy to eye movement data and the provided proof of existence of a proper trade-off between the protection of users' privacy and the preservation of data utility, opens up a completely new research area. This work is limited to the investigated task of reading in VR. However, there remains a huge variety of tasks, such as activity recognition or lifelogging, where it is necessary to show the effectiveness of privacy-aware methods to foster the emergence of novel concepts and approaches. To achieve these aims, new collaborations with the privacy and security community need to be built up, bringing together a powerful set of privacy-preserving methods and a large number of novel datasets and use cases, respectively.

- **Novel Applications** – The lack of a sufficient number of useful and exciting applications still prevents head-mounted eye tracking devices from enticing users. However, eye tracking has gained momentum to reach its breakthrough, becoming pervasive potentially in VR or AR first. In VR headsets, eye tracking technology is already used to save computational power by exploiting foveated rendering (Patney *et al.*, 2016; Hsu *et al.*, 2017) and would lead to increased functionality if integrated into AR. With real-time processing of eye movements and scene content information, direct user feedback and improved interaction will become possible.

  This thesis presented and discussed application opportunities based on forecasting users' attentive behaviour. With their implementation, a huge variety of novel and useful applications would enter the market, attracting and convincing users of the necessity of head-mounted eye tracking devices. The approach taken can even be extended, combining temporal attention forecasting with the spatial prediction of users' gaze in real-world environments. Knowledge of which object a user is currently looking at can serve for human memory enhancement, e.g. detecting faces, and reporting information about scene content to the user (Ishiguro *et al.*, 2010).

  To convince future consumers of eye tracking functionality, smart glasses need to provide the same functionality as smartwatches, especially focusing on activity recognition. For that purpose, future applications will be able to identify a user's current activity from a calculated feature pattern, which even takes into account sensors other than eye and scene cameras, like IMUs, resulting in much higher accuracy and the detectability of much more complex or composite activities.

  Before mobile eye tracking hits the mass market, its applications will grow in other sectors first. The automotive industry is a well-known use case where eye tracking is integrated to facilitate better safety functions that improve user experience[17]. In a medical context, eye tracking devices are already able to detect mental disorders (Holzman *et al.*, 1974; Kuechenmeister *et al.*, 1977; Hutton *et al.*, 1984). But they could also be exploited as a medical tool for continuous stress level monitoring. Further, they will also gain specific importance in Industry 4.0 in terms of remote maintenance work, productivity improvement, and documentation. One task of imminent importance for advertising companies will be the detection

[17] `http://smarteye.se/applied-solutions/`, date: 12.07.2019

of interest from users' attentive behaviour. This is currently limited to remote eye tracking devices and gaze-based computer screen interaction providing customised services to recommend images, documents, videos, or e-commerce products (Xu *et al.*, 2008; Cheng *et al.*, 2010a; Faro *et al.*, 2010; Jung *et al.*, 2013; Velásquez, 2013). With future mobile eye tracking devices, users' interest could be to detected even in real-world shopping environments like malls or supermarkets.

In conclusion, there will be a tremendous amount of eye-tracking-based functionality available in the future. However, the implementation of these new applications is constrained by the corresponding and necessary hardware development.

- **Datasets** – Mobile eye tracking can only be successful if much longer-term datasets are recorded to enable the development of novel methods or models with increased reliability and performance gained from high-quality training data. Therefore, the rule of thumb for eye tracking datasets must be the following: The more data the better, the longer the data recordings the better, and the more fine-grained annotations to the data the better!

  To satisfy the hunger of data-driven eye movement analysis approaches or eye tracking applications, currently available eye tracking datasets could even be too small. As generalisability of these trained models will be a key component of upcoming applications, future research needs to focus on long-term data recordings collected from a broad age, ethnic, and cultural spectrum of participants. With the rise of eye tracking in VR and AR in the near future, much larger datasets recorded at scale will become available.

  However, depending on the task, they may even need to be annotated frame-by-frame causing an enormous workload. Thus, future research will need to put more effort into the development of semi-automatic pre-annotation tools, relieving annotators who can then focus on the post-editing of provided annotations.

  Finally, there is the urgent need to make recorded datasets publicly available, either in anonymised form, or with differential privacy noise, as proposed in this thesis. Available eye tracking data is the basis for novel ideas, methods, and applications.

2.3 Significance of the Thesis

Mobile eye tracking is emerging as the key component for AR and VR headsets. However, to make mobile eye tracking ready for the masses, research needs to address technical challenges, overcome social obstacles, and develop novel applications to convince consumers of its necessity. This thesis identified the core problems of mobile eye tracking and provided efficient and effective solutions for a broad variety of research topics enabling mobile eye tracking for everyone.

Based on the summary presented in Section 2.1, the following conclusions can be drawn:

- **Sensing** – This thesis bridged the gap between state-of-the-art 2D-to-2D and currently still error-prone 3D-to-3D mapping approaches in real-world environments. The 2D-to-3D hybrid approach proposed in this thesis provides an answer to the question whether 3D gaze estimation can be done properly with only a lightweight head-mounted eye tracker, because this problem had only been evaluated with synthetic eye images (Świrski and Dodgson, 2013) as of yet. The composition of 3D gaze targets collected from different depths in the calibration procedure delivered the proof that the parallax error can be significantly reduced, resulting in an improved gaze estimation which outperforms the common 2D-to-2D mapping. Thus, this novel approach is a valid and effective alternative for every eye-tracking-enabled HMD.

- **Analysis** – In terms of eye movement analysis, this thesis made the crucial step from EOG-based to video-based eye movement analysis, presenting a number of novel encoding approaches for blinks, fixations and saccades. These encoding approaches are highly flexible and are built based on a modular principle so that an encoding of a specific eye movement can be arbitrarily extended by the encoding of other eye movements.

  In particular, this work concentrated on the analysis of fixations and demonstrated that velocity-based and dispersion-based fixation detection methods, adopted from remote eye tracking and applied for years in mobile eye tracking, are outdated and unsuitable for mobile settings. To address this significantly challenging technical task, this thesis proposed a novel, robust, and highly efficient fixation detection method which is able to deal with natural user and gaze target motion. The key to solving this problem is not to rely on the analysis of gaze data alone, but to combine real-world features extracted from the eye tracker's scene camera and information gained from a user's gaze behaviour, which are further applied to data-driven deep-trained CNN models.

  The reliable detection of fixations in mobile settings is only a starting point to freeing the detection of other eye movements from the still widely applied but highly inflexible threshold-based approaches.

- **Social Acceptability** – To allow acceptance by the masses, this thesis took the essential step of adjusting the appearance of mobile eye trackers, proposing a novel, unobtrusive, and thus consumer-friendly and attractive design for upcoming head-mounted devices. The fundamental design changes presented with *InvisibleEye* overcome the current antiquated bulky appearance of mobile eye trackers. These consist of the significant reduction of image sensors using lightweight millimetre-size RGB cameras that can be fully embedded into normal glasses frames without occluding a user's field of view. Complementary to the presented hardware prototype, this thesis provides an appealing solution to improve the usability of mobile eye trackers, and thus the social acceptability. Instead of using a marker-based approach to calibrate the eye tracker before each usage, relying on high-resolution eye images, *InvisibleEye* presented a calibration-free, learning-based gaze estimation approach, directly regressing gaze directions from eye images using multiple cameras in parallel. This necessary step compensates for low-resolution sensors but opens up the possibility to convince users of the easy handling of mobile eye trackers without a time-consuming calibration pre-step. With *InvisibleEye* this work finally realised the vision of unobtrusive mobile eye tracking and pervasive attentive user interfaces.

- **Privacy** – This work paved the way for privacy-preserving and privacy-aware eye tracking. To assuage users' and bystanders' concerns about the recording of sensitive scene data or being recorded by the first-person camera of a mobile eye tracking device, this thesis developed *PrivacEye*. It is the first proof-of-concept prototype which is able to automatically de-activate and re-activate a user's first-person scene camera using a physical shutter. Thus, *PrivacEye* is the first prototype that prevents potentially sensitive imagery from being recorded at all, without the need of active user input. But its key component is a method that detects privacy-sensitive situations by leveraging deep scene and eye movement features. The concatenation of scene and eye movement features in this method opens up a new and promising direction for privacy-preserving future work, which will be significantly fostered by the proceeding trend of miniaturisation and integration of eye tracking in head-mounted devices.

- **Novel Applications** – Addressing sensing and analysis challenges and increasing social acceptability and privacy have enabled the development of exciting, novel applications.

  To protect users' privacy, not only on the hardware side using a mechanical shutter, this work also proposed a practical software solution for privacy-aware eye tracking applying differential privacy. This approach adds noise to recorded eye tracking data which minimises an attacker's ability to infer private information from users' gaze behaviour while maintaining data utility for desired applications. Differential privacy was first applied on eye movement data in this work, and proved its effectiveness. It showed its potential when installed on future VR and AR eyewear devices to protect users' private attributes from recorded eye tracking data and laid foundations for future datasets to be released or transmitted fully anonymised.

  Inspired by previous work using EOG or IMU, this thesis presented the first video-based activity recognition approach that relies only on eye movement data. With the decisive step from EOG-based to video-based eye movement analysis, novel encoding options become available. The resulting bag-of-words representation of users' gaze behaviour was combined with a fully unsupervised LDA topic model to discover users' current activities with performance competitive with that of supervised approaches. This application can run purely on the users' side without impeding bystanders' privacy, as it relies only on an eye camera, without the necessity of a scene camera. However, in future information gained from only one source will not be enough. The combination of eye and scene content information is the most relevant step for advanced functionality.

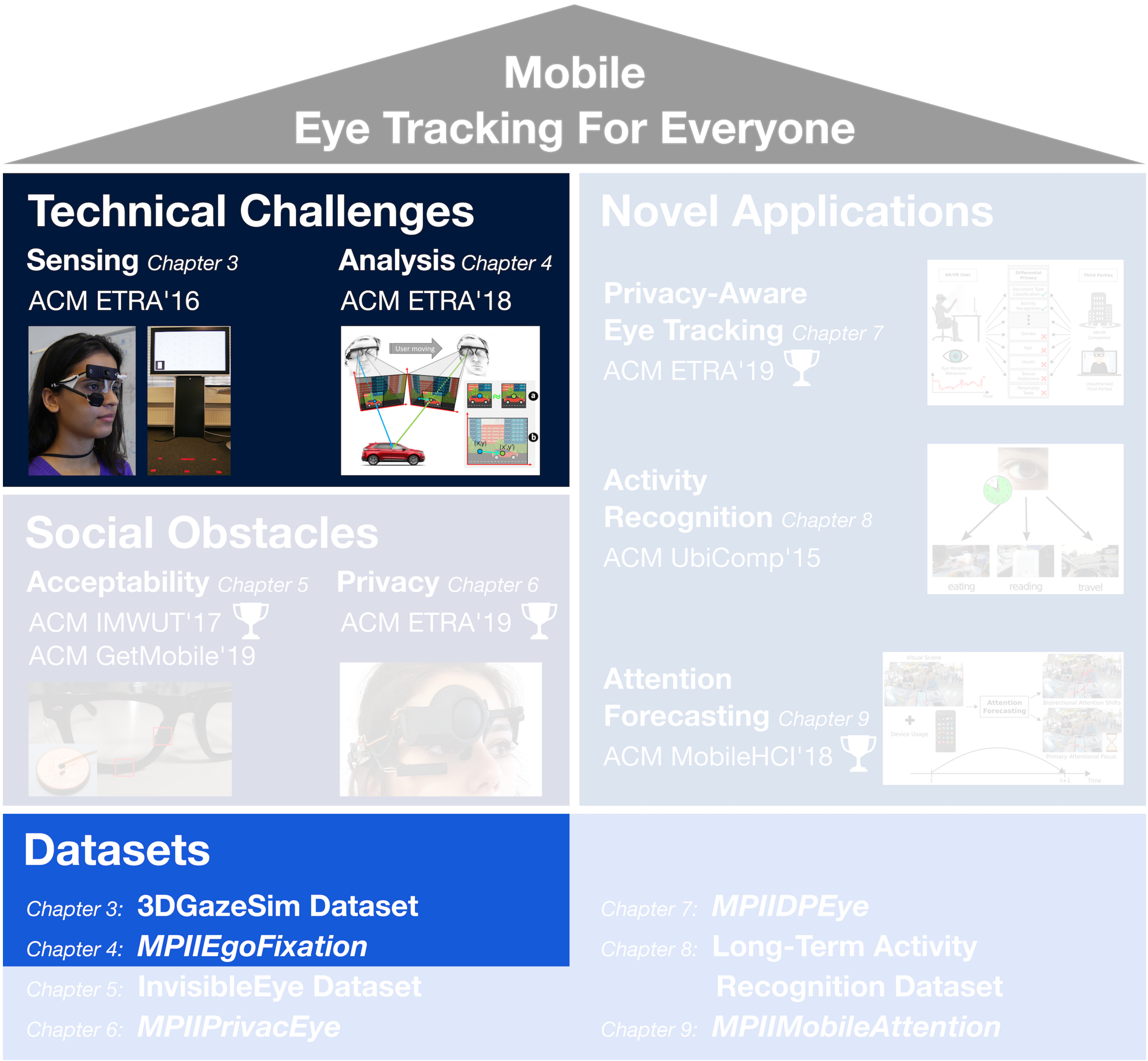
Consequently, this work presents a multi-modal sensing approach to forecast users' temporal attentive behaviour towards mobile devices, which paves the way for a huge variety of essential and promising applications. The approach uses information from users' visual scene as well as device usage to predict bidirectional attention shifts between the mobile device and the environment, as well as the primary attentional focus on the mobile device.

The fusion of different sensors, their miniaturisation, and integration in head-mounted devices like VR or AR headsets will be the fundamental basis to enable the development of novel future applications. This growing versatility and functionality of devices equipped with eye tracking technology is essential to persuade users and to make mobile eye tracking ready for the masses.

- **Datasets** – The basis to evaluate future applications is the time-consuming recording and fine-grained annotation of novel datasets. This work presented the first long-term mobile (∼80 hours) and multi-sensor eye tracking datasets (∼90 hours). In a first stage, these datasets proved the general feasibility of long-term mobile eye tracking recordings, and they displayed the rich amount of information contained in the data in a second stage. In total, this thesis contributes the impressive number of seven novel, publicly available datasets consisting of more than 180 hours of recorded and fully annotated video data. To further substantiate the impact of this thesis, all datasets have been made publicly available.

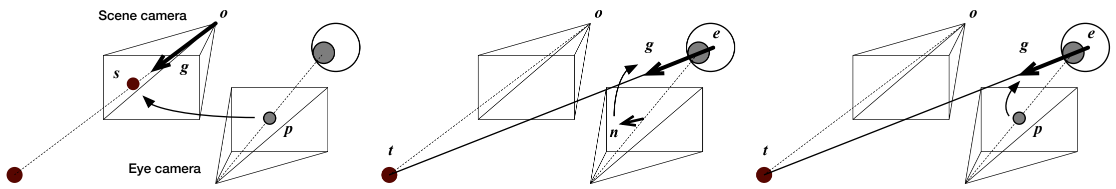
This thesis is intended to inspire and encourage other researchers to use or extend the released datasets to solve upcoming challenges and obstacles, and to develop and evaluate novel, innovative, and exciting applications so that the vision of mobile eye tracking for everyone finally comes true.

Part I


Mobile
Eye Tracking For Everyone
Technical Challenges
Sensing Chapter 3
ACM ETRA'16
Analysis Chapter 4
ACM ETRA'18
User moving
Social Obstacles
Acceptability Chapter 5
ACM IMWUT'17
ACM GetMobile'19
Privacy Chapter 6
ACM ETRA'19
Novel Applications
Privacy-Aware
Eye Tracking Chapter 7
ACM ETRA'19
Activity
Recognition Chapter 8
ACM UbiComp'15
eating
reading
travel
Attention
Forecasting Chapter 9
ACM MobileHCI'18
Attention Forecasting
Datasets
Chapter 3: 3DGazeSim Dataset
Chapter 4: MPIIEgoFixation
Chapter 5: InvisibleEye Dataset
Chapter 6: MPIIPrivacEye
Chapter 7: MPIIDPEye
Chapter 8: Long-Term Activity
Recognition Dataset
Chapter 9: MPIIMobileAttention

3D Gaze Estimation from 2D Pupil Positions on Monocular Head-Mounted Eye Trackers

3D gaze information is important for scene-centric attention analysis, but accurate estimation and analysis of 3D gaze in real-world environments remains challenging. We present a novel 3D gaze estimation method for monocular head-mounted eye trackers. In contrast to previous work, our method does not aim to infer 3D eyeball poses, but directly maps 2D pupil positions to 3D gaze directions in scene camera coordinate space. We first provide a detailed discussion of the 3D gaze estimation task and summarise different methods, including our own. We then evaluate the performance of different 3D gaze estimation approaches using both simulated and real data. Through experimental validation, we demonstrate the effectiveness of our method in reducing parallax error, and we identify research challenges for the design of 3D calibration procedures.



(a) 2D-to-2D mapping (b) 3D-to-3D mapping (c) 2D-to-3D mapping

Figure 3.1: Illustration of the (a) 2D-to-2D, (b) 3D-to-3D, and (c) 2D-to-3D mapping approaches. 3D gaze estimation in wearable settings is a task of inferring 3D gaze vectors in the scene camera coordinate system.

3.1 Introduction

Research on head-mounted eye tracking has traditionally focused on estimating gaze in screen coordinate space, e.g. of a public display. Estimating gaze in scene or world coordinates enables gaze analysis on 3D objects and scenes and has the potential for new applications, such as real-world attention analysis (Bulling, 2016). This approach requires two key components: 3D scene reconstruction and 3D gaze estimation.

In prior work, 3D gaze estimation was approximately addressed as a projection from estimated 2D gaze positions in the scene camera image to the corresponding 3D scene (Munn and Pelz, 2008; Pfeiffer and Renner, 2014; Takemura *et al.*, 2014). However, without proper 3D gaze estimation, gaze mapping suffers from parallax error caused by the offset between the scene camera origin and eyeball position (Mardanbegi and Hansen, 2012; Duchowski *et al.*, 2014). To fully utilise the 3D scene information it is essential to estimate 3D gaze vectors in the scene coordinate system.

While 3D gaze estimation has been widely studied in remote gaze estimation, there have been very few studies in head-mounted eye tracking. This is mainly because 3D gaze estimation typically requires model-based approaches with special hardware, such as multiple IR light sources and/or stereo cameras (Beymer and Flickner, 2003; Nagamatsu *et al.*, 2010). Hence, it remains unclear whether 3D gaze estimation can be done properly only with a lightweight head-mounted eye tracker. Świrski and Dodgson proposed a method to recover 3D eyeball poses from a monocular eye camera (Świrski and Dodgson, 2013). While it can be applied to lightweight mobile eye trackers, their method has been only evaluated with synthetic eye images, and its realistic performance including the eye-to-scene camera mapping accuracy has never been quantified.

We present a novel 3D gaze estimation method for monocular head-mounted eye trackers. Contrary to existing approaches, we formulate 3D gaze estimation as a direct mapping task from 2D pupil positions in the eye camera image to 3D gaze directions in the scene camera. Therefore, for the calibration we collect the 2D pupil positions as well as 3D target points, and finally minimise the distance between the 3D targets and the estimated gaze rays.

The contributions of this chapter are threefold. First, we summarise and analyse different 3D gaze estimation approaches for a head-mounted setup. We discuss potential error sources and technical difficulties in these approaches, and provide clear guidelines for designing lightweight 3D gaze estimation systems. Second, following from this discussion, we propose a novel 3D gaze estimation method. Our method directly maps 2D pupil positions in the eye camera to 3D gaze directions, and does not require 3D observation from the eye camera. Third, we provide a detailed comparison of our method with state-of-the-art methods in terms of 3D gaze estimation accuracy. The open-source simulation environment and the dataset are available at `http://mpii.de/3DGazeSim/` (date: 12.07.2019).

3.2 3D Gaze Estimation

3D gaze estimation is the task of inferring 3D gaze vectors to the target objects in the environment. Gaze vectors in scene camera coordinates can then be intersected with the reconstructed 3D scene. There are three mapping approaches we discuss in this chapter: 2D-to-2D, 3D-to-3D, and our novel 2D-to-3D mapping approach. In this section, we briefly summarise three approaches. For more details, please refer to Section A.1.

3.2.1 2D-to-2D Mapping

Standard 2D gaze estimation methods assume 2D pupil positions $\boldsymbol{p}$ in the eye camera images as input. The task is to find the mapping function from $\boldsymbol{p}$ to 2D gaze positions $\boldsymbol{s}$ in the scene camera images (Figure 3.1a). Given a set of N calibration data items $(\boldsymbol{p}_i, \boldsymbol{s}_i)_{i=1}^{N}$, the mapping function is typically formulated as a polynomial regression. 2D pupil positions are first converted into their polynomial representations $\boldsymbol{q}(\boldsymbol{p})$, and the linear regression weight is obtained via linear regression methods. Following Kassner et al. (Kassner *et al.*, 2014), we did not include cubic terms and used an anisotropic representation as $\boldsymbol{q} = (1, u, v, uv, u^2, v^2, u^2v^2)$ where $\boldsymbol{p} = (u, v)$.

In order to obtain 3D gaze vectors, most of the prior work assumes that the 3D gaze vectors are originating from the origin of the scene camera coordinate system. In this case, estimated 2D gaze positions $\boldsymbol{f}$ can be simply back-projected to 3D vectors $\boldsymbol{g}$ in the scene camera coordinate system. This is equivalent to assuming that the eyeball centre position $\boldsymbol{e}$ is exactly the same as the origin $\boldsymbol{o}$ of the scene camera coordinate system. However, in practice there is always an offset between the scene camera origin and the eyeball position, and this offset causes the parallax error.

3.2.2 3D-to-3D Mapping

If we can estimate a 3D pupil pose (unit normal vector of the pupil disc) from the eye camera as done in Świrski and Dodgson (Świrski and Dodgson, 2013), we can instead take a direct 3D-to-3D mapping approach (Figure 3.1b). Instead of the 2D calibration targets $\boldsymbol{s}$, we assume 3D calibration targets $\boldsymbol{t}$ in this case.

With the calibration data $(\boldsymbol{n}_i, \boldsymbol{t}_i)_{i=1}^{N}$, the task is to find the rotation $\boldsymbol{R}$ and translation $\boldsymbol{T}$ between the scene and eye camera coordinate systems. This can be done by minimising distances between 3D gaze targets $\boldsymbol{t}_i$ and the 3D gaze rays which are rotated and translated to the scene camera coordinate system. In the implementation, we further parameterise the rotation $\boldsymbol{R}$ by a 3D angle vector with the constraint that rotation angles are between $-\pi$ and π, and we initialise $\boldsymbol{R}$ assuming that the eye camera and the scene camera are facing opposite directions.

3.2.3 2D-to-3D Mapping

Estimating 3D pupil pose is not a trivial task in real-world settings. Another potential approach is to directly map 2D pupil positions $\boldsymbol{p}$ to 3D gaze directions $\boldsymbol{g}$ (see Figure 3.1c).

In this case, we need to map the polynomial feature $\boldsymbol{q}$ to unit gaze vectors $\boldsymbol{g}$ originating from an eyeball centre $\boldsymbol{e}$. $\boldsymbol{g}$ can be parameterised in a polar coordinate system, and we assume a linear mapping from the polynomial feature $\boldsymbol{q}$ to the angle vector. The regression weight is obtained by minimising distances between 3D calibration targets $\boldsymbol{t}_i$ and the mapped 3D gaze rays as in the 3D-to-3D approach. In the implementation, we used the same polynomial representation as the 2D-to-2D method to provide a fair comparison with the baseline.

3.3 Data Collection

In order to evaluate the potential and limitations of the introduced mapping approaches, we conducted two studies. First, we used data we obtained from a simulation environment, whereas the second study exploited real-world data collected from 14 participants.

3.3.1 Simulation Data

We first analysed the different mapping approaches in a simulation environment. Our simulation environment is based on a basic model of the human eye consisting of a pair of spheres (Lefohn *et al.*, 2003) and the scene and eye camera models. The eye model and a screenshot of the simulation environment are illustrated in Figure 3.2. We used human average anatomical parameters: $R = 11.5mm$, $r = 7.8mm$, $d = 4.7mm$, and $q = 5.8mm$. The pupil is considered as the centre of the circle which represents the intersection of the two spheres. For both eye and scene cameras, we used the pinhole camera model. Intrinsic parameters were set to values similar to those of the actual eye tracking headset we used in the real-world environment.

One of the key questions about 3D gaze estimation is whether calibration at single depth is sufficient or not. Intuitively, obtaining calibration data at different depths from the scene camera can improve the 3D mapping performance. We set calibration and test plane depths d_c and d_t to 1m, 1.25m, 1.5m, 1.75m, and 2m. At each depth, points are selected from two grids, a 5 by 5 grid which gives us 25 calibration points (blue) and an inner 4 by 4 grid for 16 test points (red) displayed on the white plane of Figure 3.2.

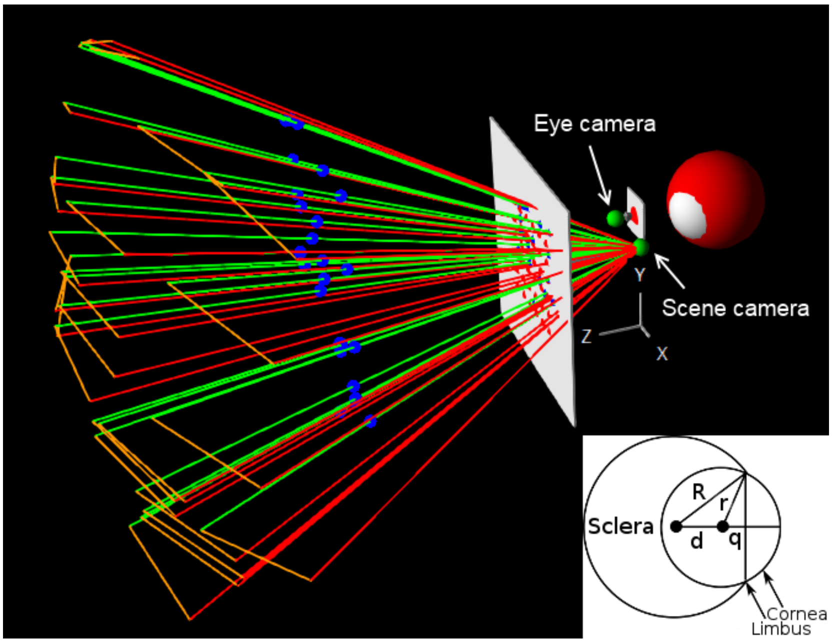


Figure 3.2: 3D eye model and simulation environment with 3D target points given as blue dots. The green and red rays correspond to ground truth and estimated gaze vectors, respectively.

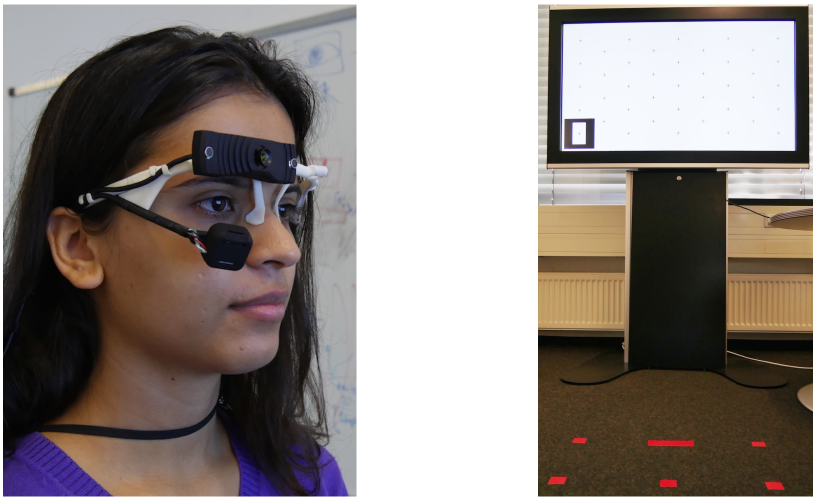
Both of the grids are symmetric with respect to the scene camera's principal axis. From the eye model used, we are able to estimate the corresponding gaze ray.

3.3.2 Real-World Data

We also present evaluation of gaze estimation approaches using a real-world dataset to show the validity of 3D gaze estimation approaches.

Procedure. The recording system consisted of a Lenovo G580 laptop and a Phex Wall 55" display (121.5cm × 68.7cm) with a resolution of 1920 × 1080. Gaze data was collected using a Pupil head-mounted eye tracker connected to the laptop via USB (Kassner *et al.*, 2014) (see Figure 3.3a). The eye tracker has two cameras: one eye camera with a resolution of 640 × 360 pixels recording a video of the right eye from close proximity, as well as an egocentric (scene) camera with a resolution of 1280 × 720 pixels. Both cameras recorded videos at 30 Hz. Pupil positions in the eye camera were detected using the Pupil eye tracker's implementation.

We implemented remote recording software which conducts the calibration and test recordings shown on the display to the participants. As shown in Figure 3.3b, the target markers were designed so that their 3D positions can be obtained using the ArUco library (Garrido-Jurado *et al.*, 2014). Intrinsic parameters of the scene and eye cameras were calibrated before recording, and used for computing 3D fixation target positions $\boldsymbol{t}$ and 3D pupil poses $\boldsymbol{n}$.



(a) Video-based head-mounted eye tracker

(b) Display and distance markers

Figure 3.3: The recording setup consisted of a Lenovo G580 laptop, a Phex Wall 55" display and a Pupil head-mounted eye tracker.

We recruited 14 participants aged between 22 and 29 years. The majority of them had little or no previous experience with eye tracking. Every participant had to perform two recordings, a calibration and a test recording of five different distances from the display. Recording distances were marked by red stripes on the ground (see Figure 3.3b). They were aligned parallel to the display with an initial distance of 1 meter and the following recording distances with a spacing of 25cm (1.0, 1.25, 1.5, 1.75, 2.0). For every participant we recorded 10 videos.

As in the simulation environment, the participants were instructed to look at 25 fixation target points from the grid pattern in Figure 3.3b. After this step the participants had to perform the same procedure again while looking at 16 fixation targets placed on different positions than in the initial calibration to collect the test data for our evaluation part. This procedure was then repeated for the other four mentioned distances. The only restriction we imposed was that the participants should not move their head during the recording.

Error Measurement. Since the ground-truth eyeball position $\boldsymbol{e}$ is not available in the real-world study, we evaluate the estimation accuracy using an angular error observed from the scene camera. For the case where 2D gaze positions are estimated (2D-to-2D mapping), we back-projected the estimated 2D gaze position $\boldsymbol{f}$ into the scene, and directly measured the angle θ between this line and the line from the origin of the scene camera $\boldsymbol{o}$ to the measured fixation target $\boldsymbol{t}$. For the cases where 3D gaze vectors are estimated, we first determined the estimated 3D fixation target position $\boldsymbol{t}'$ assuming the same depth as the ground-truth target $\boldsymbol{t}$. Then the angle between the lines from the origin $\boldsymbol{o}$ was measured.

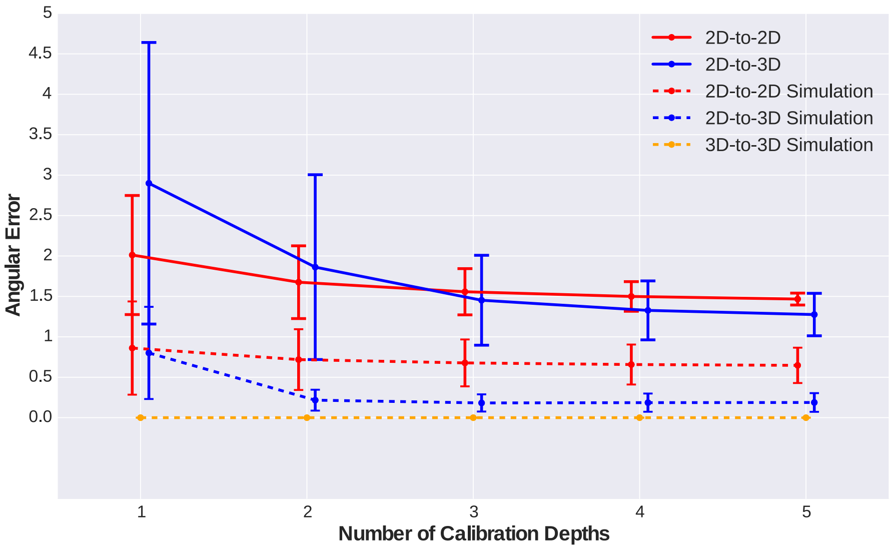


Figure 3.4: Angular error performance over different numbers of calibration depths for 2D-to-2D, 2D-to-3D and 3D-to-3D mapping approaches. Each point corresponds to the mean over all angular error values for each number of calibration depths. The error bars provide the corresponding standard deviations.

3.4 Results

We compared different mapping approaches in Figure 3.4 using an increasing number of calibration depths in both simulation and real-world environments. Each plot corresponds to mean estimation errors of all test planes and all combinations of calibration planes. Angular error is evaluated from the ground-truth eyeball position. It can be seen that in all cases the estimation performance can be improved by taking more calibration planes. Even the 2D-to-2D mapping approach performs slightly better with multiple calibration depths overall in both environments. The 2D-to-3D mapping approach performed better than the 2D-to-2D mapping in all cases in the simulation environment. For the 3D-to-3D mapping approach a parallax error near to zero can be achieved.

Similarly to the simulation case, we first compare the 2D-to-3D mapping with the 2D-to-2D mapping in terms of the influence of different calibration depths displayed as stable lines in Figure 3.4. Since it turned out that the 3D-to-3D mapping on real-world data has more angular error (over 10°) than the 2D-to-3D mapping, we omit the results in the following analysis.

Contrary to the simulation result, with a lower number of calibration depths the 2D-to-2D approach performs better than the 2D-to-3D approach for real-world data. However, with an increasing number of calibration depths, the 2D-to-3D approach outperforms 2D-to-2D comparing the angular error in visual degrees. For five calibration depths we can achieve for the 2D-to-3D case an overall mean of less than 1.3 visual degrees over all test depths and all participants. A more detailed analysis and discussion with corresponding performance plots are available in Section A.2.

3.5 Discussion

We discussed three different approaches for 3D gaze estimation using head-mounted eye trackers. Although it was shown that the 3D-to-3D mapping is not a trivial task, the 2D-to-3D mapping approach was shown to perform better than the standard 2D-to-2D mapping approach using simulation data. One of the key observations from the simulation study is that the 2D-to-3D mapping approach requires at least two calibration depths. Given more than two calibration depths, the 2D-to-3D mapping can significantly reduce the parallax error.

On the real data, we could observe a decreasing error for the 2D-to-3D mapping with an increasing number of calibration depths, and could outperform the 2D-to-2D mapping. However, the performance of the 2D-to-3D mapping became worse than in the simulation environment. Reasons for the different performance of the mapping approaches in the simulation and real-world environment are manifold and reveal their limitations. Our simulation environment considers an ideal setting and does not include noise that occurs in the real world. This noise is mainly produced by potential errors in the pupil and marker detection, as well as head movements of the participants.

In future work it will be important to investigate how the 3D-to-3D mapping approach can work in practice. The fundamental difference from the 2D-to-3D mapping is that the mapping function has to explicitly handle the rotation between eye and scene camera coordinate systems. In addition to the fundamental estimation inaccuracy of the 3D pupil pose estimation technique without modelling real-world factors such as corneal refraction, we did not consider the difference between optical and visual axes. A more appropriate mapping function could be a potential solution for the 3D-to-3D mapping, and another option could be to use more general regression techniques considering the 2D-to-3D results.

Throughout the experimental validation, this research also illustrated the fundamental difficulty of the 3D gaze estimation task. It has been shown that the design of the calibration procedure is also quite important, and it is essential to address the issue from the standpoint of both calibration design and mapping formulation. Since the importance of different calibration depths has been shown, the design of automatic calibration procedure, e.g., how to obtain calibration data at different depths using only digital displays, is another important HCI research issue.

Finally, it is also important to combine the 3D gaze estimation approach with 3D scene reconstruction methods and evaluate the overall performance of 3D gaze mapping. In this sense, it is also necessary to evaluate performance with respect to scene reconstruction error.

3.6 Conclusion

In this chapter, we provided an extensive discussion on different approaches for 3D gaze estimation using head-mounted eye trackers. In addition to the standard 2D-to-2D mapping approach, we discussed two potential 3D mapping approaches using either 3D or 2D observation from the eye camera. We conducted a detailed analysis of 3D gaze estimation approaches using both simulation and real data.

Experimental results showed the advantage of the proposed 2D-to-3D estimation methods, but its complexity and technical challenges were also revealed. Together with the dataset and simulation environment, this study would provide a solid basis for future research on 3D gaze estimation with lightweight head-mounted devices.

Fixation Detection for Head-Mounted Eye Tracking Based on Visual Similarity of Gaze Targets

Fixations are widely analysed in human vision, gaze-based interaction, and experimental psychology research. However, robust fixation detection in mobile settings is profoundly challenging given the prevalence of user and gaze target motion. These movements feign a shift in gaze estimates in the frame of reference defined by the eye tracker's scene camera. To address this challenge, we present a novel fixation detection method for head-mounted eye trackers. Our method exploits that, independent of user or gaze target motion, target appearance remains about the same during a fixation. It extracts image information from small regions around the current gaze position and analyses the appearance similarity of these gaze patches across video frames to detect fixations. We evaluate our method using fine-grained fixation annotations on a five-participant indoor dataset (*MPIIEgoFixation*) with more than 2,300 fixations in total. Our method outperforms commonly used velocity- and dispersion-based algorithms, which highlights its significant potential to analyse scene image information for eye movement detection.

4.1 Introduction

Fixations are one of the most informative and thus important characteristics of human gaze behaviour. Given the strong link between fixations and overt visual attention, human fixations have been widely studied in experimental psychology, such as in the context of mind wandering (Faber *et al.*, 2017), reading comprehension (Li *et al.*, 2016), or face processing (Dalton *et al.*, 2005). Fixations have also been used to understand users' visual attention (Nguyen and Liu, 2016), to assess on-line learning (D'Mello *et al.*, 2012) or to enhance the awareness in computer-mediated communication (Higuch *et al.*, 2016). Recent efforts have investigated using information on fixations to user behaviour modelling (Bulling *et al.*, 2011b; Bulling and Zander, 2014; Steil and Bulling, 2015) and personality traits (Hoppe *et al.*, 2015, 2018). The development of methods to automatically detect fixations in continuous gaze data has consequently emerged as an important and highly active area of research (Salvucci and Goldberg, 2000; Hessels *et al.*, 2017). With head-mounted eye trackers becoming ever more lightweight, accurate, and affordable (Kassner *et al.*, 2014; Tonsen *et al.*, 2017), fixation detection is also becoming increasingly important for mobile settings (Kurzhals *et al.*, 2017).

Fixation detection methods can be broadly classified as dispersion- or velocity-based (Holmqvist *et al.*, 2011) as well as data-driven (Urruty *et al.*, 2007). While dispersion-based methods analyse the spatial scattering of gaze estimates within a certain time window, velocity-based methods detect fixations by analysing point-to-point velocities of the gaze estimates. A key property of all of these methods is that they rely solely on gaze data, i.e. they typically do not take any other information into account, such as the target being looked at. This approach works well for remote eye trackers used in stationary settings in which the estimated gaze is analysed within a fixed frame of reference, i.e. the screen coordinate system.

In contrast, fixation detection for head-mounted eye trackers and mobile settings is significantly more challenging. Gaze estimates are typically given in the eye tracker's scene camera coordinate system but this frame of reference changes constantly with respect to the world coordinate system as the wearer moves around or turns his head while looking at a target (see Figure 4.1). As a result, gaze estimates during a fixation seem to shift within the scene camera coordinate system, resulting in failures of fixation detection methods that rely solely on gaze information. Maintaining gaze on a particular real-world target consequently involves a complex combination of fixations, smooth pursuit, and vestibulo-ocular reflex movements. In this chapter we use the term *fixation* to jointly refer to users' *visual focus of attention* (Massé *et al.*, 2017) on a gaze target irrespective of scene and head motion.

To the best of our knowledge, we are the first to address the challenging task of fixation detection for head-mounted eye tracking. The specific contributions of this chapter are three-fold. First, we propose a novel fixation detection method that is robust to user and gaze target movements prevalent in mobile everyday settings. Our method leverages visual information of the scene camera image and exploits that, independent of user or gaze target motion, target appearance remains about the same during a fixation. Specifically, our method considers image information from small

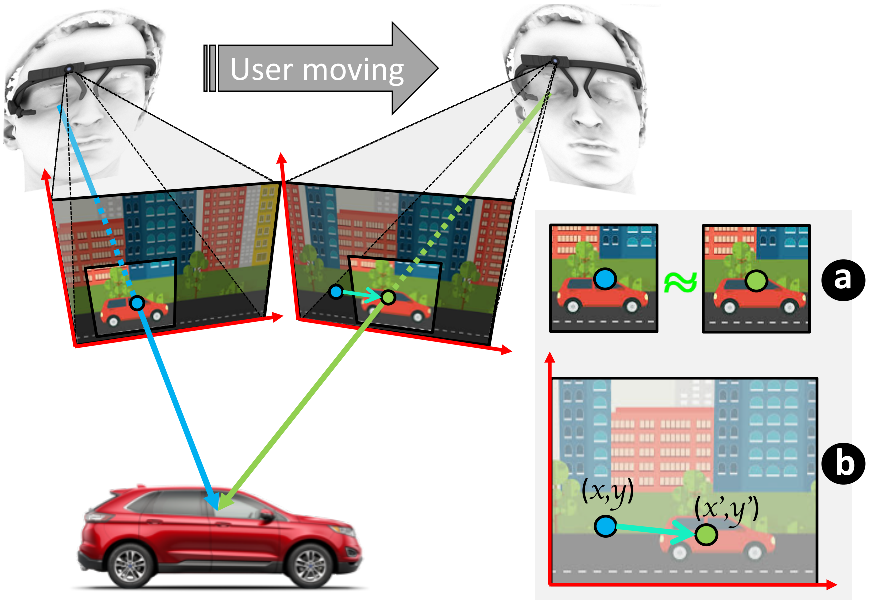


Figure 4.1: (a) Our method exploits that, independent of user or gaze target motion prevalent in mobile settings, target appearance remains about the same during a fixation. To detect fixations, it analyses the visual similarity of small patches around each gaze estimate. (b) Existing methods that only use gaze estimates face challenges due to these estimates shifting in the scene camera coordinate system.

regions around the current gaze position and analyses the appearance similarity of these *gaze patches* across video frames using a state-of-the-art deep convolutional image patch similarity network (Zagoruyko and Komodakis, 2015). Second, we annotate a subset of a recent mobile eye tracking dataset (Sugano and Bulling, 2015) with fine-grained fixation annotations – the first of its kind with annotations at the individual video frame level. Our *MPIIEgoFixation* dataset is publicly available at `https://www.mpi-inf.mpg.de/MPIIEgoFixation/` (date: 12.07.2019). Third, through experimental evaluations on this dataset, we show that our method outperforms widely used, state-of-the-art dispersion-based and velocity-based methods for fixation detection.

4.2 Related Work

The work of this chapter is related to previous works on 1) fixations in mobile settings, 2) computational methods for fixation detection, and 3) applications that used gaze patches.

4.2.1 Fixations in Mobile Settings

With the proliferation of head-mounted eye trackers, an increasing number of studies have been conducted in mobile settings. Fixation behaviours together with other eye movement characteristics have been exploited for activity recognition (Bulling *et al.*, 2011b; Steil and Bulling, 2015). Spatial-temporal patches around fixations have been used to capture the joint visual attention of multiple users (Kera *et al.*, 2016; Huang *et al.*, 2017). Visualising the fixation location has been shown to be effective in enhancing situation-awareness for remote collaboration in mobile settings (Higuch *et al.*, 2016). Researchers have also investigated fixation-based visualisation methods to facilitate egocentric video understanding (Blascheck *et al.*, 2016), user interest analysis (Kurzhals *et al.*, 2017), or video summarisation (Xu *et al.*, 2015). Despite the significant potential and ever-increasing interest in head-mounted eye tracking, works have up to now used fixation detection methods originally developed for remote eye trackers and stationary settings. To the best of our knowledge, we now present the first method specifically geared to mobile settings for tracking users' fixations without a fixed frame of reference, only using the similarity of the gaze patches.

4.2.2 Fixation Detection Methods

Existing fixation detection methods can be categorised into velocity-based, dispersion-based, and data-driven approaches, the first being the most widely used (Andersson *et al.*, 2017). These methods have often been used to discriminate fixation from smooth pursuit (eye tracing a moving target) and saccadic movements (shifting gaze between one fixation and another). Since fixations, smooth pursuits, and saccades are characterised by different velocities of eye movements, velocity-based methods have usually defined a velocity threshold to detect fixations from saccades (Salvucci and Goldberg, 2000), where eye movements with a velocity below the threshold are classified as fixations and above as saccades. If needed, an additional threshold is used to discriminate smooth pursuit from saccades (Ferrera, 2000; Komogortsev and Karpov, 2013). Dispersion-based algorithms assume that gaze estimates belonging to a fixation should locate in a cluster (Salvucci and Goldberg, 2000; Blignaut, 2009; Holmqvist *et al.*, 2011). Therefore, these algorithms measured the degree of gaze estimates' scattering to identify fixations. A number of recent research has applied data-driven approaches to improve eye movement detection, including smooth pursuits (Vidal *et al.*, 2012a) and fixations. For fixations, prior works have proposed the use of projection clustering (Urruty *et al.*, 2007), principle component analysis (Kasneci *et al.*, 2015), eigenvector analyses (Berg *et al.*, 2009), Bayesian decision theory (Santini *et al.*, 2016), or detailed geometric properties of signal components (Vidal *et al.*, 2012a). Only few previous works have addressed the challenging task of discriminating between multiple eye movement types at once (Hoppe and Bulling, 2016; Zemblys *et al.*, 2017). However, all of these methods have relied on the gaze estimates alone to identify fixations, regardless of the visual information available on the gaze targets. Please note Kinsman has pointed out that regular eye movement detectors are unsuitable for mobile eye tracking scenarios (Kinsman *et al.*, 2012) and improved the velocity-based approach (Pontillo *et al.*, 2010) to compensate ego-motion from scene motion using Fast Fourier Transformation, which could be much more computationally expensive than our method.

4.2.3 Applications Using Gaze Patches

Gaze patches have been analysed in different applications. For instance, Shiga et al. extracted visual features from gaze patches for activity recognition of the wearer (Shiga *et al.*, 2014) and Sattar et al. used gaze patches to predict the category and attributes of targets during visual search (Sattar *et al.*, 2015, 2017a). Another line of works exploited gaze patches for eye tracking data visualisation as well as video summarisation and segmentation. For example, Tsang et al. created a tree structure of gaze patches to visualise sequential fixation patterns (Tsang *et al.*, 2010). Pontillo et al. presented an interface with visualisation of gaze patches to facilitate the semantic labelling of data (Pontillo *et al.*, 2010). Kinsman et al. performed a hierarchical image clustering of gaze patches so as to accelerate the analysis of eye tracking data (Kinsman *et al.*, 2010). Similarly, Kurzhals et al. represented a video by gaze patches to show temporal changes in viewing behaviour (Kurzhals *et al.*, 2016a,b, 2017). However, all of these studies detected fixations using conventional techniques and analysed gaze patches of these detected fixations. Anantrasirichai et al. trained an SVM classifier to identify fixations for low-sample-rate mobile eye trackers based only on means and variances of CNN layer activations, thus much of the detailed spatial information was not used (Anantrasirichai *et al.*, 2016). In contrast, we are the first to propose and demonstrate a gaze patch approach for fixation detection directly, without model training and eye movement feature extraction.

4.3 Detecting Fixations in Mobile Settings

As mentioned before, fixation detection in gaze data recorded using head-mounted eye trackers faces a number of unique challenges compared to remote eye tracking. Gaze estimates are typically represented by a 2D coordinate in the screen coordinate system. Consequently, dispersion-based methods can detect fixations by measuring the spatial scattering of gaze estimates over a certain time window. That is, a new fixation occurs when the recent gaze estimates are too far away from the previous location. Similarly, the velocity-based method detects the end of a fixation when there is a large location change of gaze estimates over a certain time interval.

A key requirement for the current fixation detection methods is that they require a fixed frame of reference for the gaze estimates, i.e. the screen coordinate system in the case of stationary eye trackers. However, mobile settings are characterised by their naturalness and mobility. Gaze estimates normally refer to the egocentric camera coordinate system, which moves along with the wearer's head and body motion in natural recording. As a result, gaze estimates in the egocentric camera coordinate vary when the head moves, even though the visual attention of the wearer remains fixed on an object. Thus, we exploit the visual similarity of the gaze target.

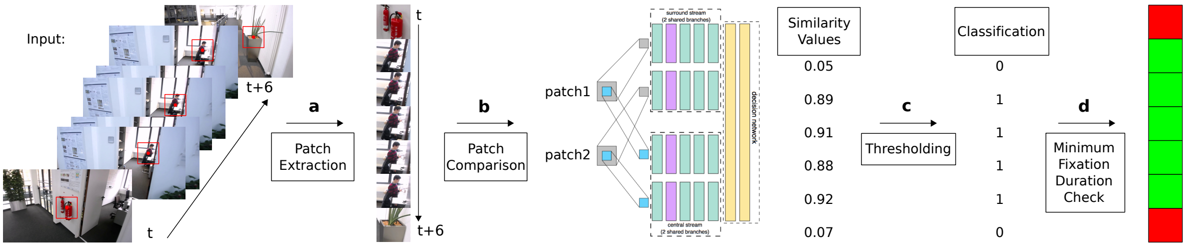


Figure 4.2: Overview of our method. Inputs to our method are scene camera frames with corresponding gaze estimates. First, our method (a) extracts gaze patches around the gaze estimates and (b) then computes similarity values with a state-of-the-art deep convolutional image patch similarity network (Zagoruyko and Komodakis, 2015). (c) In the next step, the similarity values are thresholded to classify patch pairs into fixation candidates. (d) Finally, fixation candidates are checked for a minimum length (Irwin, 1992).

4.3.1 Patch-Based Similarity

The core idea of our method is based on the observation that the appearance of gaze target stays similar regardless of head motion. Therefore, given the inputs of egocentric video and gaze estimates from the head-mounted eye tracker, our method compares the sequential gaze patch information around each gaze estimate to determine fixations (see Figure 4.2). Specifically, our method takes the egocentric video and the corresponding sequence of gaze estimates as input. It first extracts gaze patches with the gaze estimate as centre in each video frame and feeds each pair of gaze patches from consecutive frames to a CNN network that measures the patch similarity. We then determine the fixation segments based on the sequence of similarity measurement. Being independent of the frame of reference, our method can be robust to head motion in mobile settings and thus address the shortcomings of existing fixation detection methods. The following subsections detail each step of our method.

Extracting Gaze Patches from Video Stream. In the first step, our method extracts a gaze patch from each frame in the egocentric video, using the location of a gaze estimate as the patch centre. The egocentric videos we use in this chapter have a resolution of 1280 × 720 pixels, which covers 78.44 horizontal and 44.12 vertical visual degree. The size of a gaze patch is set to 200 × 200 pixels. Prior studies on video summarisation have extracted patches of 100 × 100 pixels, which corresponds to the size of fovea (Kurzhals *et al.*, 2017). In contrast to their purpose of scene understanding, gaze patches in our study are used to represent the human visual focus of attention in fixations. To simulate the spotlight effect of fixations (Eriksen and Hoffman, 1972) that a human has clearer vision in the focus and more blurry vision in the peripheral area, we exploit a larger size of patch (200 × 200 pixels) for similarity comparison. To this end, the patch comparison we use focuses more on the central region and less on the fringe area. In accordance with the size of fovea suggested by Kurzhals et al. (Kurzhals *et al.*, 2017), the central region in our gaze patch is 100 × 100 pixels. If the gaze patch does not fit into the camera's field of view, the gaze patch is cut so that it only covers the scene content until the border. Please note that we discard scene frames with no valid gaze estimate, as the eye tracking would fail for these.

Computing the Similarity of Gaze Patches. Next, we compute the similarity between gaze patches in each pair of consecutive frames. To account for the spotlight effect in patch similarity comparison, we adopt the convolutional neural network (2ch2stream) by Zagoruyko et al. (Zagoruyko and Komodakis, 2015) that heightens the importance of the patch central region in comparison. More specifically, this network uses a two channel structure, one of which processes the holistic patch information and the other of which analyses only the central region. This network provides unbounded similarity values $(-1, \infty)$, and it is trained on the Notredame dataset (Winder and Brown, 2007). In practice, we resize the gaze patches from 200×200 pixels captured from the egocentric video (1280×720) to 64×64 pixels and feed them into 2ch2stream.

Determining Fixation from Patch Similarity. Once we obtain the similarity sequence of patch pairs given by 2ch2stream, we identify fixations using a light-weight method. Specifically, we use a thresholding method to determine whether consecutive patches belong to the same fixation segment. If the similarity of consecutive patches is higher than the *similarity threshold*, their corresponding time periods are grouped together. This process groups similar sequential patches together, and each group of patches corresponds to one fixation. Finally, we run a duration validation to verify that each resulting fixation should be at least 150 ms (cf. (Irwin, 1992)).

4.4 Dataset

We have evaluated our method on a recent mobile eye tracking dataset (Sugano and Bulling, 2015). This dataset is particularly suitable because participants walked around throughout the recording period. Walking leads to a large amount of head motion and scene dynamics, which is both challenging and interesting for our detection task. Since the dataset was not yet publicly available, we requested it directly from the authors.

The eye tracking headset (Pupil (Kassner *et al.*, 2014)) featured a 720p world camera as well as an infrared eye camera equipped on an adjustable camera arm. Both cameras recorded at 30 Hz. Egocentric videos were recorded using the world camera and synchronised via hardware timestamps. Gaze estimates were given in the dataset.

4.4.1 Data Annotation

Given the significant amount of work and cost of fine-grained fixation annotation, we used only a subset from five participants (four males, one female, all ages 20–33). This subset contains five videos, each lasting five minutes (i.e. 9,000 frames each). We asked one annotator to annotate fixations frame-by-frame for all recordings using Advene (Aubert *et al.*, 2012). Each frame was assigned a fixation ID, so that frames belonging to the same fixation had the same ID. We instructed the annotator to start a new fixation

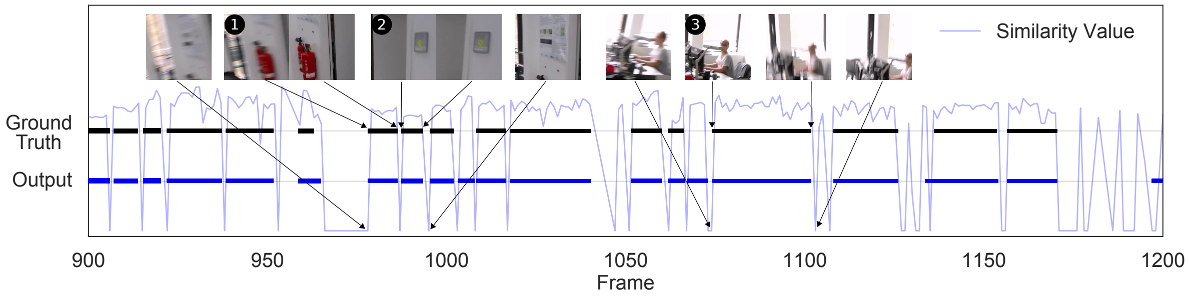


Figure 4.3: Example sequence of similarity values calculated by the deep convolutional image patch similarity network, ground truth fixation events (black), and detected fixations (blue). The example patches are from two short fixations (1), (2) and a longer fixation (3). We see that the visual content of gaze patches, even shortly before or after a fixation, differs considerably from those within the fixation.

segment after an observable gaze shift and a change of gaze target. Similarly, a fixation segment should end when the patch content changes noticeably, even though the position of the gaze point might remain in the same position in the scene video. In addition, if a fixation segment lasted for less than five consecutive frames (i.e. 150 ms), it was to be discarded. During the annotation, the gaze patch as well as the scene video superimposed with gaze points were shown to the annotator. The annotator was allowed to scroll back and forth along the time line to mark and correct the fixation annotation.

An example sequence containing the annotated ground truth and detected fixations based on the corresponding similarity values is shown in Figure 4.3. The figure shows example gaze patches from two short and a longer fixation as well as gaze patches before and after a detected fixation. We see that the visual content of gaze patches, even shortly before or after a fixation, differs considerably from those within the fixation.

4.4.2 Dataset Statistics

To better understand the fixation behaviours in mobile settings, we computed fixation statistics based on ground truth annotation. We also measured head motion by calculating optical flow within the boundary region (100 pixels) of the egocentric videos. We empirically set a flow threshold of 2° to capture the large head motion. Similarly, we defined a visual angular threshold of 0.5° to capture large gaze shifts from the sequence of gaze estimates.

We see that almost three quarters of the time (74%), eyes were in fixation across different participants. In addition, large head motion and gaze shifts occurred about 85% and 80% of the time during these fixation segments, respectively. These numbers indicate that fixations and head motion were pervasive in natural mobile recordings. More importantly, they suggest that the reliability of conventional fixation detection that relied on a fixed coordinate system should be questioned for a clear majority of the time. Our experimental evaluation provides a more in-depth performance comparison of our method against different fixation detection methods.

4.5 Evaluation

In this section, we compare our proposed method against commonly used dispersion-based and velocity-based methods. As for the dispersion-based method, we have adopted the implementation available in Pupil (Kassner *et al.*, 2014). The method uses a dispersion threshold to identify fixations as groups of gaze estimates that locate closely in the egocentric camera coordinate system. For the velocity-based method, we have reimplemented Salvucci and Goldbergs's velocity-threshold identification algorithm (Salvucci and Goldberg, 2000). The method uses a threshold to segment fixations when the velocity of the gaze estimated point changes rapidly. Given that our method also uses a similarity threshold for fixation detection, we evaluate the performance of our method against the dispersion- and velocity-based methods for increasing thresholds, respectively. Please note, we followed the practice of defining the minimal duration of fixation as 150 ms (Irwin, 1992), which has been used consistently across the different methods.

4.5.1 Evaluation Metrics

To provide a thorough evaluation on the performance of our fixation detection, we break down the errors in fixation detection events and analyse the underlying issues of the proposed method against the conventional fixation detection methods. We use the evaluation metrics originally developed by Ward et al. for fine-grained analysis of activity recognition systems. A comprehensive explanation of the different evaluation metrics, i.e. their meaning and how they are calculated, is beyond the scope of this thesis. We refer the interested reader to the original paper (Ward *et al.*, 2006). In a nutshell, in addition to the *Correctly classified* (C) fixation events, we have also studied the errors from three main perspectives, which we briefly discuss as follows:

1. *Deletion* (D) and *insertion* (I'): Both belong to the classical errors in event detection. In our case, a deletion error indicates the failure to detect a fixation, while an insertion means a fixation is detected where there is none in the ground truth.

2. *Fragmentation* (F) and *merge* (M'): These are associated with sensitivity of event segmentation. A fragmentation error describes a single fixation in ground truth being detected as multiple ones. In contrast, a merge error depicts multiple fixations in ground truth being recognised as being one by the method.

3. *Overfill* (O) and *Underfill* (U): These errors are related to the erroneous timing of fixation detection. An overfill error denotes that the identified fixation covers too much time compared to the ground truth. As the opposite, an underfill indicates that the detected fixation fails to cover parts of the ground truth.

To better describe the fragmentation and merge errors, we further refer to a "fragmenting" output (F') as an *output*, i.e. the identified fixation, that belongs to one of the detected fragments of a large ground truth fixation, and a "merging" output (M') as a large identified fixation that covers multiple ground truth fixations. In other words, F' and M' are counted from the output side, while F and M are counted from the ground

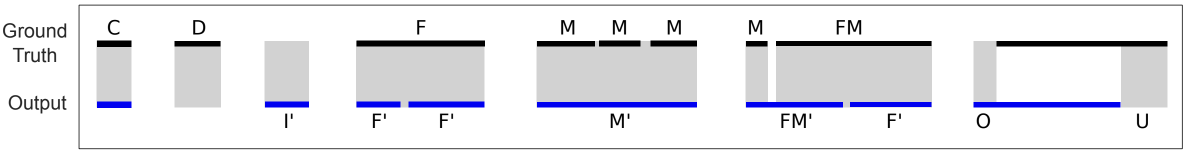


Figure 4.4: Example event based errors in continuous detection of fixations. The metrics include correctly classified events (C), overfill (O) and underfill (U) errors, deletion (D) and insertion (I'), as well as merge (M, M') and fragmentation (F, F') errors, and their interplay (FM, FM').

truth. We also group events that are both, fragmented and merged, as FM; similarly, an output event that is both fragmenting and merging as FM'. An example overview of all event-based error cases is shown in Figure 4.4.

As in event detection, the most important implication often comes from the number of correctly classified events (C) as well as the over- and underestimated events, i.e. insertion (I') and deletion (D). We therefore adapt a unified metric (CDI') (Bulling *et al.*, 2012) to assess these three important aspects:

$$CDI' = C - D - I' \tag{4.1}$$

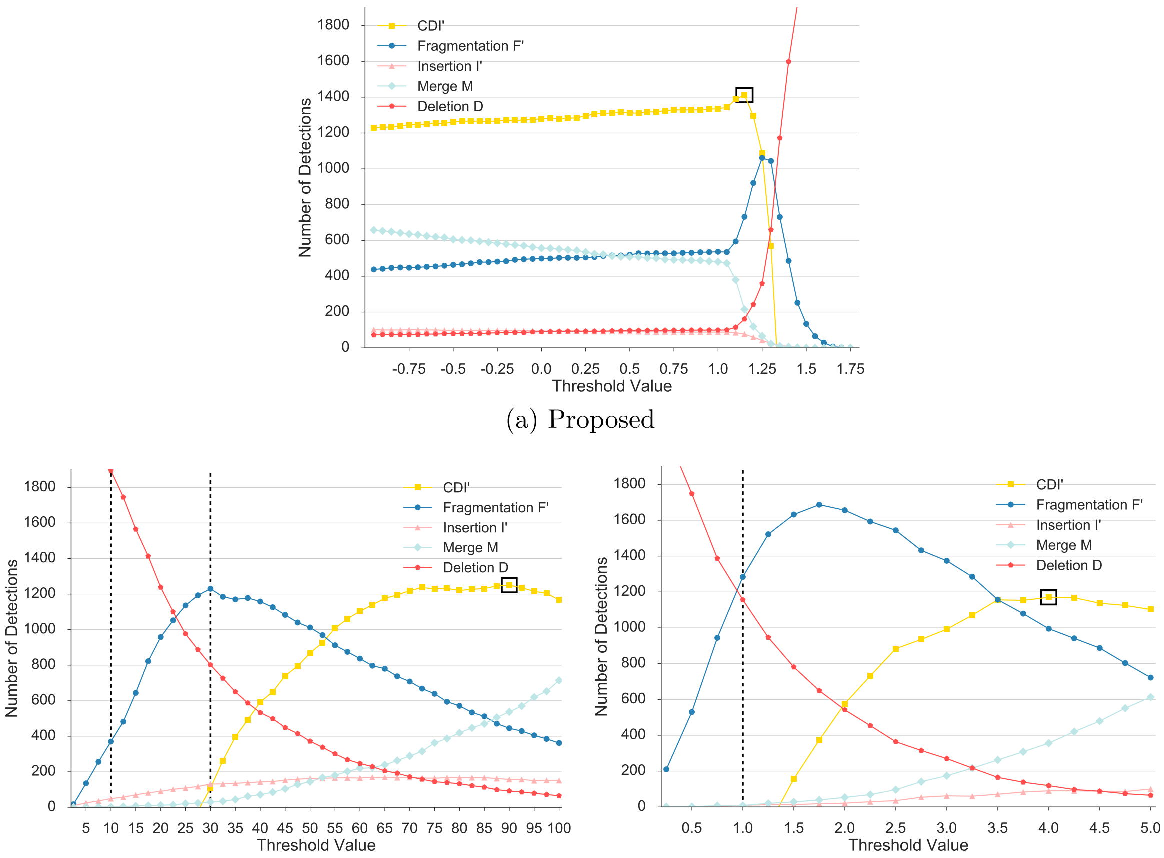
Using the unified and the individual measurements as performance metrics for fixation detection not only sheds light on how a fixation has been correctly detected as an event, but also endows us with a more in-depth understanding of the detection reliability of event characteristics, such as detection delay and duration error.

4.5.2 Fixation Detection Performance

Our evaluation begins with an overall fixation detection performance with respect to different important event-based metrics, including the unified metric CDI', insertion (I'), deletion (D) as well as fragmenting output (F') and merge (M) of ground truth. There are interesting findings when we evaluate the performance change for increasing thresholds for each method. The performance of the interesting metrics are selected and shown in Figure 4.5.

First, comparing across the methods, we see that our method achieves the highest score of the unified metric. It reaches approximately 1,400 for CDI', while the numbers of the velocity- and dispersion-based approaches are around 1,200. Although the optimal thresholds (shown in black squares) for conventional techniques also lead to a high CDI' number, these thresholds are surprisingly large compared to the suggested values (represented in the black vertical lines) in traditional stationary settings. Interestingly and as expected, the commonly used velocity and dispersion thresholds (Eriksen and Hoffman, 1972; Holmqvist *et al.*, 2011) correspond to only poor performances in mobile settings, which are generally associated with a large number of deletion (D) and fragmenting output (F'), and more importantly, a very low number of the unified metric (CDI').

Most interestingly, we see that our method performs robustly for the unified metric (CDI') as well as for individual metrics. As the similarity threshold increases from



(a) Proposed

(b) Velocity-based

(c) Dispersion-based

Figure 4.5: Performance of fixation detection of our proposed method as well as the velocity- and dispersion-based methods over sweeps of their threshold parameters. Black dashed lines indicate the thresholds recommended for the velocity-based (10°/sec and 30°/sec (Holmqvist *et al.*, 2011)) and dispersion-based methods (1° (Eriksen and Hoffman, 1972)). Black squares mark the best performing threshold for each method.

-0.95 to 1.15, CDI' rises steadily, and the rest of individual metric stays stable without significant variations. Furthermore, there is a very wide range of acceptable thresholds for our method. In contrast, performance of the velocity- and dispersion-based counterparts changes considerably with their thresholds.

It is also interesting to note that the behaviours of all the thresholding methods toward the change in threshold are in good agreement. In particular for the velocity- and dispersion-based methods, almost all the curves have similar trends and shapes. That is, a threshold that is over restrictive for fixation detection gives a high number of deletions (D) and a mounting number of fragmenting output (F'). On the other hand, a threshold that is over loose for fixation detection yields the growth of merge error (M). In contrast to the robustness of our method, conventional techniques fail to present a wide range of acceptable thresholds that can lead to overall good performance.

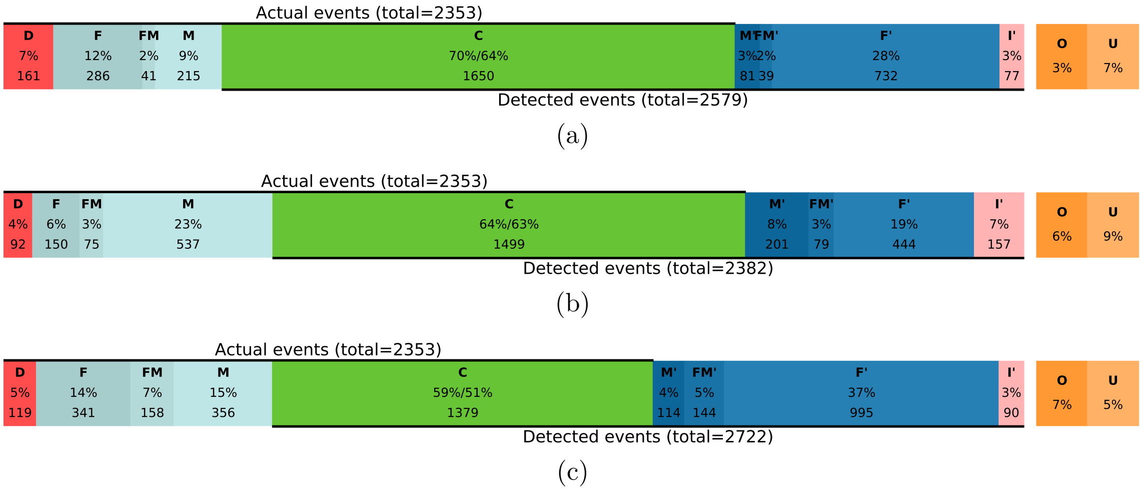


Figure 4.6: Event analysis diagram (EAD) for (a) our proposed patch-based, (b) the velocity-based, and (c) the dispersion-based fixation detection method for the best-performing thresholds shown in Figure 4.5. The EAD shows an overview of the typical errors occurring in continuous event detection, i.e. the number of correct detections (C), merges (M, M'), fragmentations (F, F'), deletions (D), and insertions (I'). The corresponding overfills (O) and underfill (U) errors are shown on the right.

4.5.3 Influence of Key Parameters on Performance

In addition to the previous discussion on how the important CDI' performance varies for increasing thresholds, respectively, this section scrutinises all types of fixation detection errors, under the optimal parameter with respect to CDI' for each method.

Figure 4.6 shows the event analysis diagram (EAD) of fixation detection results of our method, velocity-, and dispersion-based methods. Starting from the most important metrics, we see that the number of correctly detected (C) fixation of our method (1,650) clearly exceeds that of the velocity- (1,499) and dispersion-based (1,379) methods. For insertion error (I'), our method (77) can also outperform its counterparts (157 and 90, respectively) by sacrificing a marginal performance decrease of deletion error (D).

As regards the fragmentation error from both sides of ground truth (F) and output (F'), the velocity-based method gives the best result. In contrast, the velocity-based method performs worst in terms of merge error. This is quite intuitive, as large fragmentation error tends to correlate with small merge error, and vice versa. It is encouraging that our method gives the minimal overall fragmentation and merge errors (F+FM+M+M'+FM'+F'=1394), compared to the velocity-based (1,486) and dispersion-based (2,108) methods. With respect to the timing errors, we see that our method results in the lowest overfill error (3%) and a moderate underfill error (7%).

In conclusion, the proposed method is able to precisely identify the majority of ground truth annotated fixations, with an overall minimal number of fragmentation and merge cases and an acceptable number of timing errors.

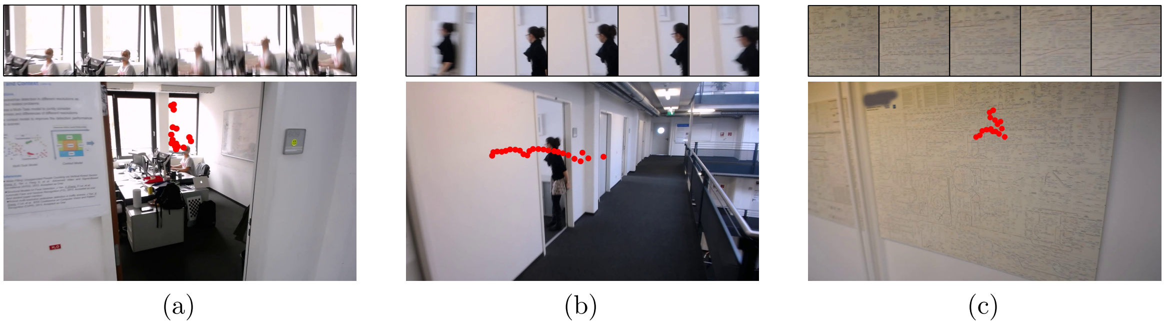


Figure 4.7: Example eye tracker scene images with gaze estimates (in red) and corresponding sequences of gaze patches (top) for cases in which our method successfully detected fixations while the conventional methods failed. Our method robustly deals with (a) vertical head motion during nodding, (b) horizontal head motion that follows a target of interest while walking down a corridor, and (c) compensating head movements while walking towards an object.

4.5.4 Example Detections

One important feature of our proposed approach lies in that it can be robust to blurry image inputs, which are a common problem when using egocentric scene cameras. For example, participants were mobile most of the time in our evaluation dataset. Blurry images occurred frequently when users directed their gaze through intentional head or body motion as well as when users fixated but compensated head motion through eye movement.

The examples in Figure 4.7 show detection cases where our method can successfully identify fixations while conventional methods fail. In order to maintain the privacy rights, we have blurred observable logos and show only blurry faces of people in the scene. In the image sequence of Figure 4.7a, a participant is interacting with another person and nodding, resulting in a group of widely scattered gaze estimates. However, our method can identify the fixation, as the gaze target person remains in gaze patches throughout the process. Figure 4.7b shows a sequence where the participant is walking along the corridor while fixating on a girl leaning against the door frame. Since the participant is moving forward with his head turning left to follow the girl, the path of gaze estimates appears in a line. Conventional fixation detection methods in this case would not detect the fixation due to the obvious shift of gaze estimates. In Figure 4.7c, the participant is moving closer to a poster of chemical formulas and shaking his head at the same time. The head motion is so large that conventional methods fail. Although the image content looks similar and blurry, our proposed method is still able to detect the fixation correctly based on the visual similarity of gaze patches.

4.6 Discussion

This study points out an important but overlooked issue of fixation detection in mobile settings. Since the coordinate system for mobile gaze estimates often moves during natural head motion, eye fixations no longer correspond to a fixed coordinate system of gaze estimates, as assumed by the existing methods. This change of setting hampers the velocity- and dispersion-based fixation detection methods. We are the first to address the challenges of fixation detection in mobile settings by exploiting the visual similarity of gaze targets. We also provide the first mobile dataset with fine-grained fixation annotation for the purpose of this line of studies (*MPIIEgoFixation*). In addition, we have suggested appropriate evaluation metrics for fixation event detection and have conducted an in-depth evaluation of our method against the existing widely used counterparts in mobile settings.

It is encouraging to see that our method can be robust to head motions. It outperforms the velocity- and dispersion-based methods with respect to a number of major metrics for fixation event detection, such as correctly detecting events, insertion errors, merge errors, and overfills. The slightly higher number of deletion errors of our proposed method in comparison to the velocity- and dispersion-based approaches is a side effect of optimising for the CDI' score. There is a general trade-off between deletion and merge errors that can be determined depending on the particular application. In our method, a higher threshold leads to a sharper cut between frames that belong to a fixation or not, whereas an increasing threshold for the velocity- and dispersion-based approaches makes these approaches more greedy so that the deletion errors transit to an increasing number of merge errors that result in higher overfill errors, whereas our proposed method suffers from increasing underfill errors. Our experimental results also reveal that our method is much more robust to the parameter value, compared to the conventional techniques.

Given the advance of mobile eye tracking and the emerging attention to mobile computing, we believe that our method can open up numerous opportunities for application studies as well as follow-up gaze behaviour research. Regarding the commercial potential and application studies, our study meets the need of the recent exploding interest in augmented reality research and user experience studies. Our method requires only low computational cost, thus it is suitable for mobile and portable devices. As regards the gaze behaviour research, this study sheds light on a proper fixation detection method in mobile settings and provides guidance for appropriate evaluation metrics.

As the very first step in addressing the challenge of mobile fixation detection, we propose a simple yet effective method and have made a considerable effort in annotation and evaluation. We have conducted extensive evaluation on our *MPIIEgoFixation* dataset with fine-grained fixation annotation. Although this dataset contains only five participants, we have annotations of over 2,300 fixations and more than 40,000 frames, which are sufficient to properly evaluate our method.

Given that the goal of this chapter is to study the detection of fixations in mobile settings, we focused on cases where participants are on the move. In future work we will evaluate our approach on a novel dataset covering both mobile and stationary settings.

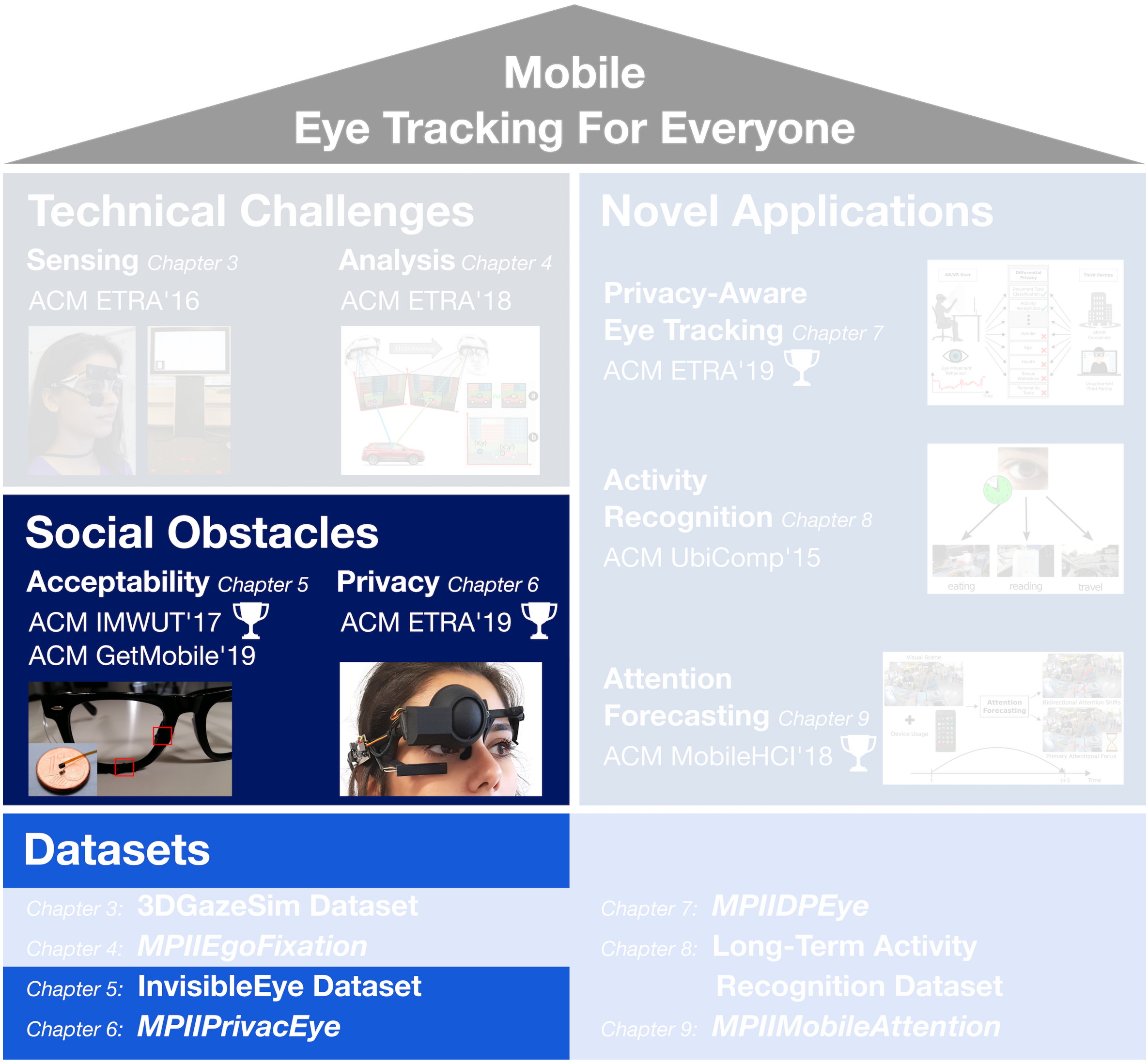
We will also extend our patch-based method by training an end-to-end neutral network to incorporate additional visual information such as scene dynamics in a joint framework.

Besides, not taking eye motion as input increases the difficulty of fixation detection when gaze targets share very similar textures or completely homogeneous appearances, though this only happened rarely in our dataset. To address this, we plan to experiment with an adaptive threshold based on the visual variability of the scene and gaze patch.

4.7 Conclusion

In this chapter we have presented a novel fixation detection method for head-mounted eye trackers. Our method analyses the image appearance in small regions around the current gaze position, which, independent of user or gaze target motion, remains about the same during a fixation. We have evaluated our method on a novel, fine-grained annotated five-participant indoor dataset *MPIIEgoFixation* with more than 2,300 fixations in total. We have shown that our method outperforms commonly used velocity- and dispersion-based algorithms, particularly with respect to the total number of correctly detected fixations as well as insertion and merge event errors. These results are promising and highlight the significant potential of analysing scene image information for eye movement detection – particularly given the emergence of head-mounted eye tracking and, with it, the increasing need for robust and accurate gaze behaviour analysis methods.

Part II



Mobile
Eye Tracking For Everyone

Technical Challenges

Sensing *Chapter 3*
ACM ETRA'16

Analysis *Chapter 4*
ACM ETRA'18

Social Obstacles

Acceptability *Chapter 5*
ACM IMWUT'17 🏆
ACM GetMobile'19

Privacy *Chapter 6*
ACM ETRA'19 🏆

Novel Applications

Privacy-Aware Eye Tracking *Chapter 7*
ACM ETRA'19 🏆

Activity Recognition *Chapter 8*
ACM UbiComp'15

Attention Forecasting *Chapter 9*
ACM MobileHCI'18 🏆

Datasets

Chapter 3: 3DGazeSim Dataset
Chapter 4: ***MPIIEgoFixation***
Chapter 5: InvisibleEye Dataset
Chapter 6: ***MPIIPrivacEye***
Chapter 7: ***MPIIDPEye***
Chapter 8: Long-Term Activity Recognition Dataset
Chapter 9: ***MPIIMobileAttention***

5 InvisibleEye: Mobile Eye Tracking Using Multiple Low-Resolution Cameras and Learning-Based Gaze Estimation

Analysis of everyday human gaze behaviour has significant potential for ubiquitous computing, as evidenced by a large body of work in gaze-based human-computer interaction, attentive user interfaces, and eye-based user modelling. However, current mobile eye trackers are still obtrusive, which not only makes them uncomfortable to wear and socially unacceptable in daily life, but also prevents them from being widely adopted in the social and behavioural sciences. To address these challenges we present *InvisibleEye*, a novel approach for mobile eye tracking that uses millimetre-size RGB cameras that can be fully embedded into normal glasses frames. To compensate for the cameras' low image resolution of only a few pixels, our approach uses multiple cameras to capture different views of the eye, as well as learning-based gaze estimation to directly regress from eye images to gaze directions. We prototypically implement our system and characterise its performance on three large-scale, increasingly realistic, and thus challenging datasets: 1) eye images synthesised using a recent computer graphics eye region model, 2) real eye images recorded of 17 participants under controlled lighting, and 3) eye images recorded of four participants over the course of four recording sessions in a mobile setting. We show that *InvisibleEye* achieves a top person-specific gaze estimation accuracy of 1.79° using three cameras with a resolution of only 5×5 pixels. Our evaluations not only demonstrate the feasibility of this novel approach but, more importantly, underline its significant potential for finally realising the vision of invisible mobile eye tracking and pervasive attentive user interfaces.

5.1 Introduction

Human gaze has a long history as a means for hands-free interaction with ubiquitous computing systems and has, more recently, also been shown to be a rich source of information about the user (Bulling *et al.*, 2011a; Bulling and Zander, 2014; Majaranta and Bulling, 2014). Prior work has demonstrated that gaze can be used for fast, accurate, and natural interaction with both ambient (Stellmach and Dachselt, 2013; Vidal *et al.*, 2013; Turner *et al.*, 2014; Zhang *et al.*, 2014; Lander *et al.*, 2015) and body-worn displays, including smartwatches (Akkil *et al.*, 2015; Esteves *et al.*, 2015). Eye movements are closely linked to everyday human behaviour and cognition and can therefore be used for computational user modelling, such as for eye-based recognition of daily activities (Bulling *et al.*, 2008b, 2009b), visual memory recall (Bulling and Roggen, 2011), visual search targets (Zelinsky *et al.*, 2013; Sattar *et al.*, 2015, 2017b), and intents (Bednarik *et al.*, 2012), or personality traits (Hoppe *et al.*, 2015) – including analyses over long periods of time for lifelogging applications (Bulling *et al.*, 2013; Steil and Bulling, 2015). Interest in gaze has been fuelled by recent technical advances and significant reductions in the cost of mobile eye trackers that can be worn in daily life and thus provide access to users' everyday gaze behaviour (Bulling and Gellersen, 2010).

However, despite its appeal, mobile eye tracking still suffers from several fundamental usability problems. First, current mobile trackers are still rather uncomfortable to wear, especially during long-term recordings. The main reason for this is high-quality imaging sensors that are large and thus often occlude the user's field of view. In addition, the sensors themselves as well as the additional electronics and wiring required to operate them makes current headsets heavy and cause discomfort or even pain. Second, current mobile eye trackers limit users' mobility given that they require a wired connection to a recording computer both as a power supply and for real-time image processing (often in the form of a laptop worn in a backpack). Eye trackers that do not require a wired connection instead store data on the device itself but, on the downside, are not well-suited for real-time applications. In addition, tetherless headsets require a battery, which adds to their weight and further limits their recording time. Finally, the obtrusive design of current eye trackers leads to low social acceptance and unnatural behaviour of both the wearer and people they interact with (Risko and Kingstone, 2011; Nasiopoulos *et al.*, 2015), thus fundamentally limiting the practical usefulness of mobile eye tracking as a tool in the social and behavioural sciences.

To address these issues, we argue that it is ultimately necessary to fully integrate eye tracking into regular glasses, i.e. to effectively make eye tracking visually and physically unnoticeable to both the wearer and others. We believe that a key requirement for such unnoticeable (*invisible*) integration is to reduce the size of an eye tracker's core component: the imaging sensors. Smaller sensors would not only significantly reduce the device's weight but could also be positioned in the visual periphery to avoid occlusions within the users' field of view. In addition, the low resolution common to these sensors generates significantly less data that could more easily be processed on the device itself, stored, or transmitted wirelessly, thus removing the need for a separate recording device altogether. Finally, the reduced computation required to process low-resolution images

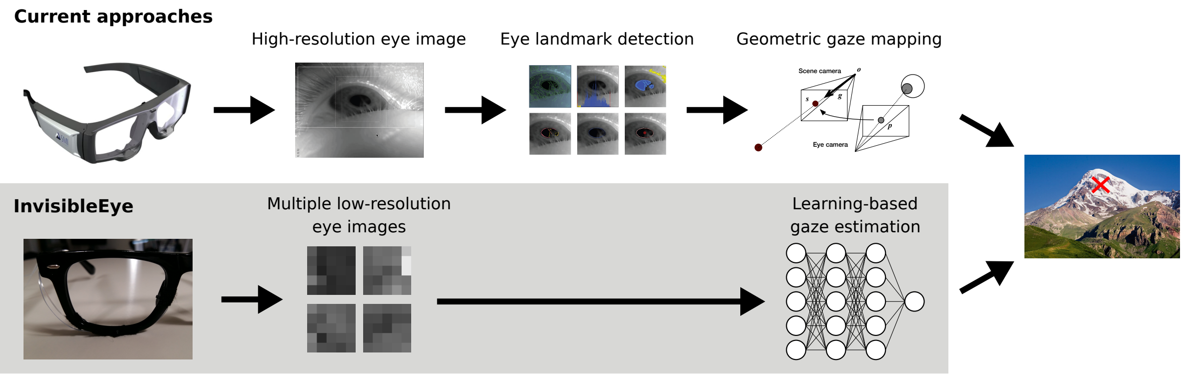


Figure 5.1: (top) Classic approaches require high-resolution imaging sensors, resulting in rather bulky and obtrusive headsets, as well as hand-optimised algorithms for eye landmark detection and geometric gaze mapping. (bottom) *InvisibleEye* is a novel approach for mobile eye tracking that uses millimetre-size RGB cameras that can be fully embedded into normal glasses frames. To compensate for the cameras' low image resolution of only a few pixels, our approach uses multiple cameras in parallel and learning-based gaze estimation to regress to gaze position in the scene camera coordinate system (red cross).

decreases the load on the processor, which in turn could help to extend the recording time, which is limited to a few hours for current mobile eye trackers.

As a first step towards realising the above vision, we present *InvisibleEye*, a novel mobile eye tracker that uses millimetre-size imaging sensors with a resolution of only a few pixels that can be fully embedded into a normal glasses frame (see Figure 5.1). Traditional image processing and computer vision methods for eye landmark detection (most importantly pupil and pupil centre) and gaze estimation in mobile eye trackers require high-quality eye region images and are thus not suitable for such low-resolution sensors. Inspired by recent advances in remote gaze estimation in computer vision (Zhang *et al.*, 2015, 2017b), we instead propose a learning-based approach that does not require robust detection of eye landmarks but directly regresses from low-resolution eye images to 3D gaze directions. To compensate for the low resolution of each individual imaging sensor, and thus to improve overall gaze estimation accuracy, *InvisibleEye* uses multiple sensors positioned around the eye in parallel. In the work of this chapter we learn a person-specific model for each user using training data recorded beforehand. Calibration-free (person-independent) gaze estimation is an open research challenge and an important direction for future work. We evaluate *InvisibleEye* on three large-scale, increasingly realistic datasets: 1) 200,000 eye images synthesised using a recent computer graphics method (Wood *et al.*, 2015), which allows us to explore the influence of the number of cameras, camera positioning, and image resolution on gaze estimation performance in a principled way, 2) 280,000 real eye images recorded with a first prototype implementation in a laboratory setting with controlled lighting during a calibration-like procedure, and 3) 240,000 real eye images recorded using a second prototype over the course of four recording sessions in a mobile setting in which four participants gazed at a physical targets from various angles. The second dataset is publicly available at `hhttp://www.mpi-inf.mpg.de/invisibleeye/` (date: 12.07.2019). We demonstrate that our approach can achieve a person-specific gaze estimation accuracy of 1.79° in the mobile setting using three cameras with an image resolution of only 5 × 5 pixels.

The specific contributions of this chapter are three-fold: First, we propose a novel approach for mobile eye tracking that leverages multiple tiny, low-resolution cameras that can be fully and thus invisibly integrated into a normal glasses frame. Second, we introduce a first-of-its-kind dataset of 280,000 close-up eye images that have been captured from multiple views and that are annotated with corresponding ground-truth gaze directions in both a stationary controlled and mobile setting. Third, we present extensive evaluations of two prototypical implementations of our approach on these datasets plus synthetic data and characterise their performance across key design parameters including image resolution, number of cameras, and camera angle and positioning.

5.2 Related Work

The work of this chapter is related to previous works on 1) mobile eye tracking, 2) gaze estimation using multiple cameras, and 3) datasets for the development and evaluation of gaze estimation algorithms.

5.2.1 Mobile Eye Tracking

Many approaches for mobile eye tracking have been explored in the past, including some at low cost (San Agustin *et al.*, 2010b; Noris *et al.*, 2011; Kassner *et al.*, 2014; Kim *et al.*, 2014). The traditional computational pipeline for mobile gaze estimation involves 1) eye landmark detection, in particular detecting the pupil centre, and ellipse fitting either using special-purpose image processing techniques (Li *et al.*, 2005; Li and Parkhurst, 2006; Long *et al.*, 2007; Świrski *et al.*, 2012; Fuhl *et al.*, 2015; Javadi *et al.*, 2015; Fuhl *et al.*, 2016c) or machine learning (Fuhl *et al.*, 2016b), and 2) gaze mapping, traditionally using a geometric eye model (Tsukada *et al.*, 2011; Pires *et al.*, 2013; Świrski and Dodgson, 2013; Plopski *et al.*, 2015) or, more recently, by directly mapping 2D pupil positions to 3D gaze directions (Mansouryar *et al.*, 2016). Instead of using two cameras, Nakazawa and Nitschke relied on only an eye camera and proposed a geometric approach to estimate gaze using corneal imaging (Nakazawa and Nitschke, 2012). All of these video-based methods rely on high-quality eye images and cameras, and therefore all suffer from the disadvantages discussed in the introduction.

Although a large body of works investigated learning-based gaze estimation, they mostly focused on remote settings, i.e. settings in which the camera is placed in front of the user, for example under a display (Sugano *et al.*, 2014; Lu *et al.*, 2014; Zhang *et al.*, 2015; Krafka *et al.*, 2016). More closely related to ours is the work by Mayberry et al., who used a subset of pixels from an eye image to estimate gaze direction with an accuracy of up to 3° (Mayberry *et al.*, 2014). However, they still assumed high-resolution eye images as input, and did not fully explore the potential of the learning-based approach, in particular in terms of input image resolution. Although Abdulin et al. investigated the impact of image resolution of an eye camera and found that the iris-diameter resolution should be at least 50 pixels for model-based approaches (Abdulin *et al.*, 2016), the minimum image resolution for learning-based approaches has not been fully investigated in prior work. In contrast, the work of this chapter is first to utilise multiple low-resolution eye cameras that can be fully embedded into an ordinary glasses frame in combination with a learning-based gaze estimation method.

In an attempt to further integrate mobile eye tracking, a smaller number of works investigated alternative measurement techniques, such as electro-oculography (EOG). EOG involves attaching electrodes on the skin around the eyes to measure the electric potential differences caused by eye movements. While EOG is computationally lightweight compared to video-based approaches, and thus promises full and low-power integration (Manabe and Fukumoto, 2006; Bulling *et al.*, 2008a, 2009a), due to drift and a low signal-to-noise ratio EOG is only suited for measuring relative movement of the eye. Borsato et al. instead used the sensor of a computer mouse to track the episcleral surface of the eye (the white part of the eye) using optic flow (Borsato and Morimoto, 2016). Using this approach they reported an accuracy of 2.1° of error at a 1 kHz sampling rate. However, the tracking was lost during every blink and the system had to be recalibrated each time, rendering it impractical for actual use. A few other works explored the use of phototransistors for mobile eye tracking that can, potentially, be fully integrated into a glasses frame. For example, Ishiguro et al. used infrared illumination in combination with four infrared sensitive phototransistors attached to a glasses frame to record relative movement of the eyes (Ishiguro *et al.*, 2010). Their use of phototransistors allowed for a fairly compact, occlusion-free, and low-power design but the proposed system was only evaluated in a usability study without a quantitative analysis. With the goal of obtaining actual gaze estimates, Topal et al. used up to six infrared sensitive phototransistors per eye and trained a support vector machine to regress the gaze point from the signals achieving an average angular error of about 0.93° (Topal *et al.*, 2014). However, their evaluation was also limited to a constrained laboratory setting.

5.2.2 Multi-Camera Gaze Estimation

Several previous works investigated the use of multiple cameras for head pose estimation as a proxy to gaze, or gaze estimation directly. For example, Voit and Stiefelhagen equipped a room with multiple cameras to track horizontal head orientation of multiple users and, eventually, estimate who was looking at whom (Voit and Stiefelhagen, 2006). As a follow up work of (Ruddarraju *et al.*, 2003a), Ruddarraju et al. presented a method for detecting gaze in interaction (Ruddarraju *et al.*, 2003b). Head pose was used to estimate a user's eye gaze and to measure if a user was looking at a previously defined region of interest. Utsumi et al. estimated users' head pose to choose the best out of multiple remote cameras positioned around the user to estimate gaze (Utsumi *et al.*, 2012). Arar et al. proposed a general framework for gaze estimation using multiple cameras placed around a computer screen by computing a weighted average of the estimations of each individual camera (Arar *et al.*, 2015). While all of these works explored multi-camera gaze estimation in remote settings, also using learning-based methods, the work of this chapter is first to explore this approach for mobile eye tracking.

5.2.3 Gaze Estimation Datasets

In computer vision, but increasingly also in other fields, the availability of large-scale, annotated datasets to develop and evaluate learning-based methods has emerged as a critical requirement. Consequently, recent years have seen an increasing number of datasets being published, including for mobile gaze estimation. Swirski et al. presented a small dataset of 600 eye images recorded with a head-mounted camera, but the dataset only covered a single camera view and offered no variability in terms of participants or lighting conditions (Świrski *et al.*, 2012). Tonsen et al. and Fuhl et al. provided large and challenging datasets with a lot of variability in personal appearance and illumination conditions but they, too, only included single-view recordings of one eye (Fuhl *et al.*, 2015; Tonsen *et al.*, 2016). While an ever-increasing number of datasets have been proposed, all of them target the tasks of pupil detection and ellipse fitting. To the best of our knowledge, none of the existing datasets offers ground truth gaze directions in addition to the eye images, thus limiting their use for developing and evaluating mobile gaze estimation pipelines. In contrast, we present the first-of-its-kind large-scale dataset of eye images that have been captured from multiple views and that are annotated with corresponding ground-truth gaze directions in both a stationary controlled and mobile everyday settings.

With the goal of reducing the time and effort required to record and annotate gaze estimation datasets, a relatively new line of work is exploring means to instead render highly realistic and perfectly annotated eye images using computer graphics techniques. Two representatives of this line of work are the methods by Swirski and Dodgson (Świrski and Dodgson, 2014) as well as SynthesEyes and UnityEyes by Wood et al. (Wood *et al.*, 2015, 2016a), the latter of which was more recently extended into a fully morphable 3D eye region model (Wood *et al.*, 2016a). While both methods allow synthesis of annotated eye images for different camera positions, they differ in that (Wood *et al.*, 2016a) uses a more realistic eye region model and can simulate different lighting conditions. We therefore opted to use UnityEyes for part of our evaluation.

5.3 Multi-View Low-Resolution Mobile Eye Tracking

The goal of this chapter is to design a fully-integrated, *invisible* eye tracking device. As illustrated in Figure 5.1, our proposed system consists of eye cameras fully embedded into ordinary eyeglasses. While the scene camera is still expected to have higher resolution, the eye cameras are expected to be built with tiny low-resolution imaging sensors. Since the use of low-resolution and low-quality eye images leads to a fundamental difficulty in employing the conventional mobile eye tracking approaches through, e.g., eye landmark detection, we further propose to take a machine learning-based approach for gaze estimation. Here, the specific technical challenges are: 1) whether such tiny

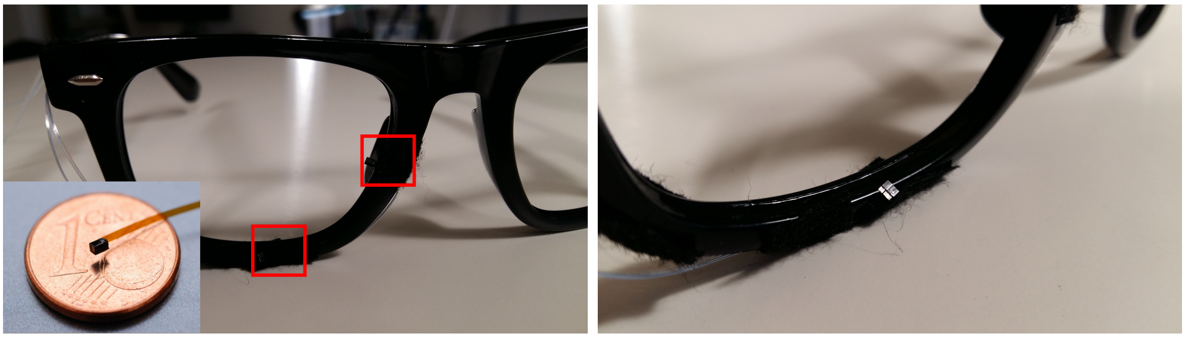


Figure 5.2: Fully integrated version of *InvisibleEye* consisting of multiple, millimetre-size Awiba NanEye RGB cameras (marked in red) that are invisibly integrated into an off-the-shelf glasses frame. For our evaluations we developed two other prototypes to be able to characterise performance across key design parameters, including image resolution and number of cameras, as well as camera angle and positioning, and to compare with a state-of-the-art (high-resolution) mobile eye tracker.

imaging sensors are available, and 2) what is the minimum image quality and resolution, as well as the minimum number of sensors, required for mobile learning-based gaze estimation. Considering previous works that have used individual photo transistors for gaze estimation (Ishiguro *et al.*, 2010; Topal *et al.*, 2014), in this chapter we explore eye image resolutions as low as 1 × 1 pixels.

In terms of sensor footprint, millimetre-size RGB cameras are available on the market mainly for medical imaging purposes such as endoscopy. Figure 5.2 shows a fully integrated prototype of our proposed system using an off-the-shelf glasses frame and medical-purpose millimetre-size cameras. In this prototype and one of the following experiments, we used the Awaiba NanEye camera which has a footprint of only 1 × 1 mm (Sousa *et al.*, 2017). As can be seen, this hardware concept using tiny eye cameras enables extremely unobtrusive design. In addition, to compensate for both low image quality and limited visibility of non-adjustable embedded cameras, we further propose to use multiple low-resolution eye images as input to the gaze estimation pipeline.

5.3.1 Neural Network for Multi-View Gaze Estimation

As discussed above, we propose to take a machine learning-based gaze estimation approach. Using a set of training (calibration) eye images associated with ground-truth gaze positions in the scene camera, our system trains a gaze estimation function that can directly output gaze positions from arbitrary input eye images. Prior work on remote appearance-based estimation already demonstrated that, in the ideal case, only a 15-dimensional feature representation (eye image of 3 × 5 pixels) is sufficient to achieve less than one degree of accuracy (Lu *et al.*, 2014). Inspired by such prior attempts, in this chapter we examine the machine learning pipeline assuming low-resolution cases.

We use an artificial neural network as illustrated in Figure 5.3 to learn a mapping from low-resolution eye images to gaze positions. We assume the existence of training (calibration) data from the target user, and train a person-specific mapping function

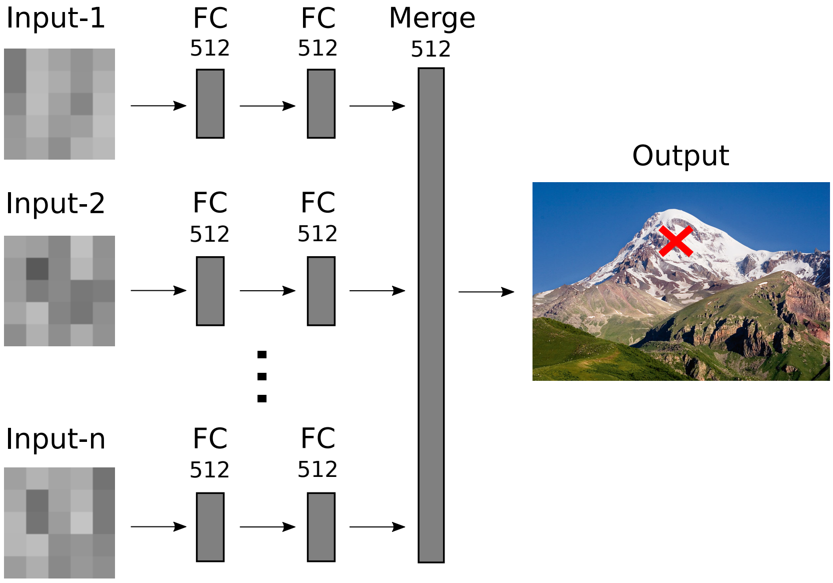


Figure 5.3: Overview of the neural network used in this chapter for learning-based gaze estimation. The network takes multiple low-resolution eye images as input. Each image is encoded using two fully connected (FC) layers and image-specific representations are then merged to jointly predict gaze direction in scene camera coordinates (red cross).

for each user. Unlike prior work (Baluja and Pomerleau, 1994), our method takes multiple eye images obtained from the tiny wearable eye cameras and learn a joint mapping function from all eye images. While there is a trade-off between the depth and performance of the neural network, our proposed network architecture is designed to be sufficiently shallow to reduce training time and inference time at run-time. Separate stacks of two fully connected layers with 512 hidden units and ReLU activation take raster-scanned image vectors from each of the N eye cameras as input. The outputs of those stacks are merged in another fully connected layer with 512 hidden units, and the output is predicted by a linear regression layer. The network is trained to jointly predict the x- and y-coordinate of the gaze positions, and the loss function is defined as the mean absolute distance between the predicted and ground-truth values. We implemented the network using Keras (Chollet *et al.*, 2015) with the Tensorflow (Abadi *et al.*, 2016) backend and chose the Adagrad algorithm (Duchi *et al.*, 2011) as optimiser with a learning rate of $lr = 0.005$. We trained our models on a modern i7-6850K CPU, on which training until convergence took about 1-2 minutes in all cases. At test time, we achieved ~700 frames per second (fps) on the same CPU using a single core. When using a Nvidia GeForce GTX 1080 Ti GPU we achieved up to ~850 fps. For comparison, for the gaze estimation pipeline of Pupil Labs (Kassner *et al.*, 2014), a commercial, state-of-the art mobile eye tracker, we achieved only ~270 fps. These results indicate the significantly smaller amount of computation required for *InvisibleEye* and thus its potential for mobile and embedded platforms that have only limited computational power.

5.4 Experiments

To systematically explore the feasibility and performance of *InvisibleEye* we conducted a series of experiments on three large-scale and increasingly difficult datasets, two of which we collected specifically for the purpose of the work in this chapter. Experiment 1 was conducted in an idealised setting using synthesised eye images. Synthesising the eye images allowed us to use an arbitrary number of “virtual” cameras in different positions, which would not be possible when recording with real cameras. For Experiment 2 we implemented a first prototype to record real data in a constraint environment. This allowed us to control several of the parameters that make mobile gaze estimation difficult, in particular slippage of the headgear or changes in lighting conditions. Experiment 3 evaluated the performance of *InvisibleEye* in a challenging mobile real-world setting using a second prototype. It is important to note that, in all experiments that follow, the network was trained in a person-dependent fashion, i.e., trained for each user individually with person-specific training data. In the following, we report on each of these experiments in turn.

5.4.1 Experiment 1: Evaluation on Synthetic Images

Before constructing the first hardware prototype for *InvisibleEye*, we opted to investigate the design space using synthetic eye image data. The goal of Experiment 1 on these synthetic images was to evaluate the minimum number and positions of cameras.

Data Synthesis. The dataset for Experiment 1 was generated using UnityEyes, a computer graphics eye region model to synthesise highly-realistic and perfectly annotated eye region images (Wood *et al.*, 2016a). UnityEyes combines a novel generative 3D model of the human eye region with a real-time rendering framework. The model is based on high-resolution 3D face scans and uses real-time approximations for complex eyeball materials and structures as well as anatomically inspired procedural geometry methods for eyelid animation. Using UnityEyes, we synthesised images for five different eye regions as illustrated in Figure 5.4. We used a uniform 5×5 grid of camera angles to synthesise the images (see Figure 5.6a). The used camera angles span the full range of angles UnityEyes is capable of synthesising, which is a frontal view as one extreme, and views that are increasingly bottom-up or from the side. Top-down views were largely occluded by the ridge bone and were therefore not considered here. For each combination of eye region, camera angle, and lighting condition, we recorded a set of 1,600 different eyeball poses, corresponding to a uniform 40×40 grid of gaze angles. The step size in this grid was 1°, so the dataset covers a horizontal and vertical field of view of 40°. Each set was randomly split into a set of 1,280 training images and 320 test images. The images produced by UnityEyes are of high resolution and we therefore down-sampled them to resolutions below 20×20 pixels to simulate the images a low-quality sensor would yield. We also converted them to grayscale to further lower their dimension.

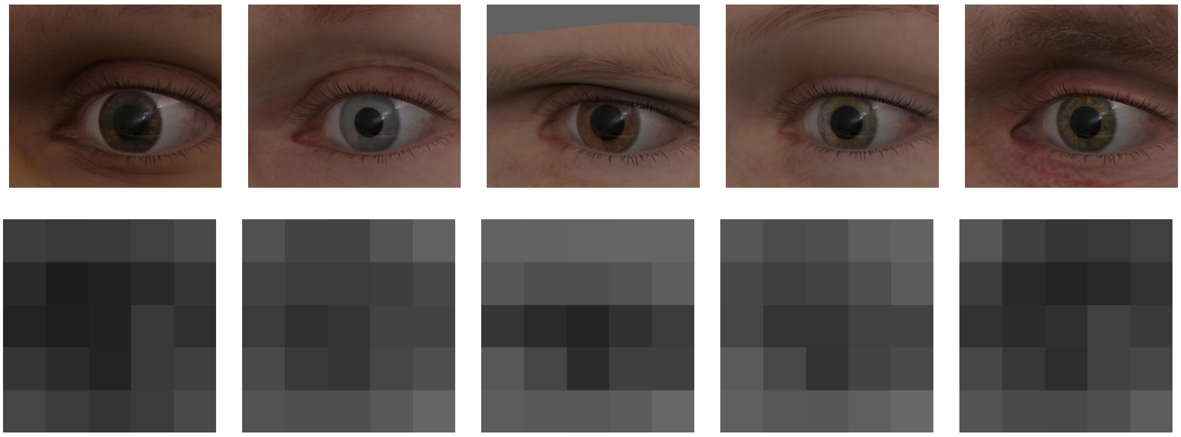
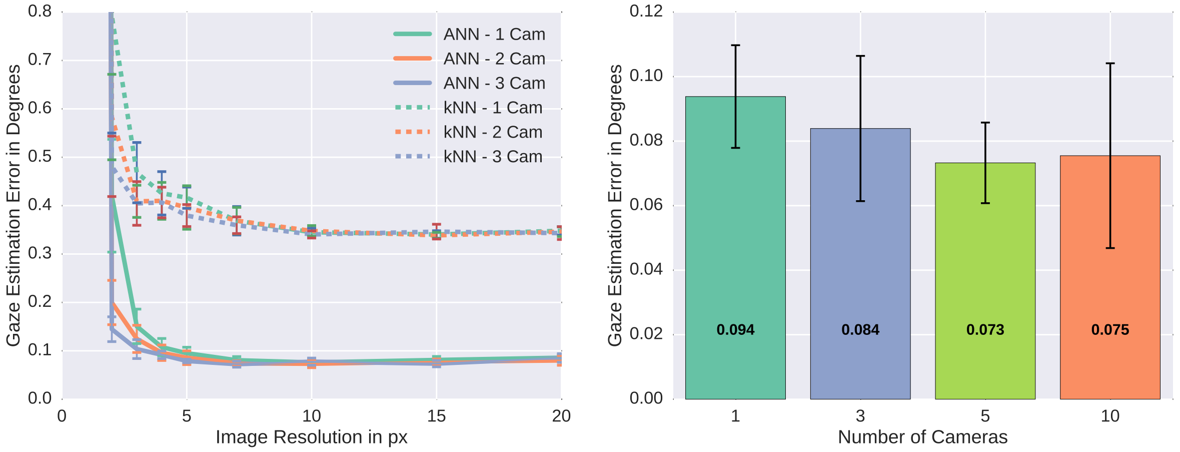


Figure 5.4: (top row) Sample eye images from the original UnityEyes dataset (Wood *et al.*, 2016a) and the corresponding low-resolution grey-scale images (bottom row) that were used as input to the learning-based gaze estimation method.

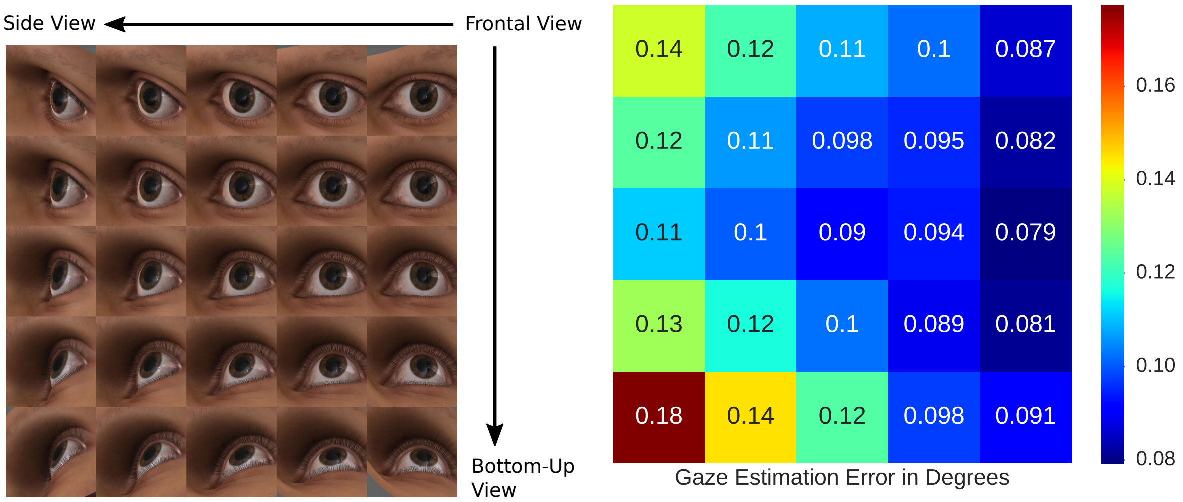


(a)

(b)

Figure 5.5: (a) Average gaze estimation error for different image resolutions for a k-Nearest-Neighbours approach and our suggested neural network approach. (b) Gaze estimation error for different numbers of cameras at 5-pixel resolution. Each bar corresponds to the error of the best combination of x cameras out of 5 randomly selected sets.

Results. To investigate the difficulty of estimating gaze with extremely low resolution images and the capabilities of *InvisibleEye* at this task, we trained different neural networks for different image resolutions. For a baseline comparison we also computed results using k-Nearest-Neighbours (kNN) with $k = 5$. Furthermore, we evaluated all approaches for different numbers of cameras. For the kNN approach we concatenated the corresponding images of different cameras before training. The results of this series of experiments are summarised in Figure 5.5a. As can be seen from the figure, both kNN and our approach achieve very low gaze estimation error. For example, at 10×10 image resolution, using a single camera, kNN achieves 0.345° error and the neural network



0.14	0.12	0.11	0.1	0.087
0.12	0.11	0.098	0.095	0.082
0.11	0.1	0.09	0.094	0.079
0.13	0.12	0.1	0.089	0.081
0.18	0.14	0.12	0.098	0.091

(a) (b)

Figure 5.6: (a) The use of synthetic images allows us to explore a wide range of camera angles (25 in the work of this chapter) in an efficient and principled manner. (b) Average gaze estimation error in degrees when evaluating for each individual angle.

achieved 0.078°. One can also see that there is little benefit in increasing the resolution of the input images. For both approaches, however, the figure also shows that the addition of cameras to the system helps to improve the results, especially for very low resolutions. At 3×3-pixel image resolution, for example, the result of the neural network improves from 0.15° error to 0.12° and 0.1° error for two and three cameras respectively, which is an improvement of 20% and 33%. Figure 5.5b shows the results for even higher numbers of cameras. As one can see, additional cameras help to improve performance slightly, but beyond four to five cameras the error does not significantly decrease any further.

Besides choosing the right number of cameras, another important parameter is the positioning of those cameras. One would like them to have an informative view but not to occlude the user's field of view. Figure 5.6b shows the error when using every available camera individually. As one can see, frontal views of the eye yield the lowest error, while bottom-up views are superior to side views. The worst result is achieved with the highly off-axis view from the very bottom and on the far side.

Discussion. Although the results achieved on synthetic data do not directly translate to the real world, since the gaze estimation task is a lot easier without real-world noise, the first set of experiments clearly demonstrates that mobile gaze estimation does not necessarily require high-resolution images. Further, we found that using multiple cameras can improve performance, but more than three to four cameras are unlikely to yield significant improvements. These results thus serve as important guidelines for designing *InvisibleEye* prototypes, which will be discussed in the following sections. We also found that frontal views of the eye yield the best results. We believe this is because frontal views have the least occluded view of the pupil and iris (e.g. with respect to the eyelashes), resulting in more distinct features for gaze estimation. However, since one of the key attributes of *InvisibleEye* should be that its cameras are in non-occluding positions, frontal views are not an option in practice. Since bottom-up views and pure side views were the next best options according to these first experiments, we positioned the cameras in corresponding positions in our prototypes.

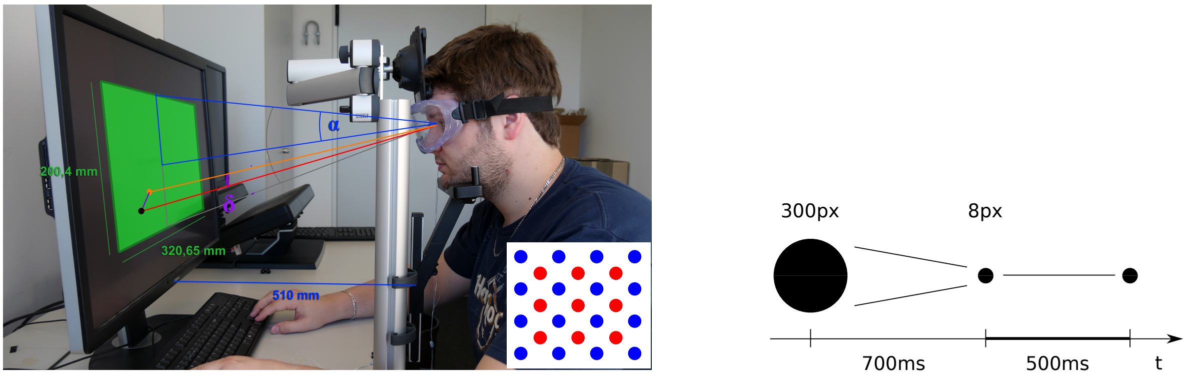
5.4.2 Experiment 2: Evaluation in a Controlled Laboratory Setting

Based on our experiments on synthetic eye images, we built a hardware prototype of *InvisibleEye* to evaluate its performance on real images. We conducted the second experiment using this prototype in a controlled laboratory environment. As discussed earlier, we used Awaiba NanEye cameras to achieve the small footprint of 1 × 1 mm. The NanEye cameras have an image resolution of 250 × 250 pixels and can capture images at 44 frames per second. Although the form factor of this medium-resolution camera is already sufficient to realise fully invisible mobile eye tracking (see Figure 5.2), we wanted to explore even lower image resolutions, i.e. below 20 × 20 pixels, which also promises further decreased bandwidth and computational requirements. We therefore opted to simulate this setting by artificially degrading the image resolution further.

The prototype was built by attaching four NanEye cameras to a pair of safety glasses. The NanEye cameras are very fragile and, since they are so small, also difficult to work with. We therefore opted to use safety glasses as the basis of our prototype, because it allowed us to carefully attach the cameras to the glass. The number of cameras and their positioning was motivated by the results of Experiment 1, i.e. two cameras were positioned with bottom-up views of the eye and one camera each was positioned on the far left and right side of the eye. The cameras were attached using "Blu-Tack", a reusable putty-like pressure-sensitive adhesive. Since we attached the cameras to a pair of panoramic safety glasses, their angles are similar to what they would be in a regular glasses frame. The main difference in the angles is, that they are further away from the eye than they would be in a regular frame. We compensated for this by cropping the image by 25% from the center in each direction, which has a similar effect on the image as moving the camera closer to the eye while reducing the resolution.

Data Collection. We used this first hardware prototype to record a dataset of more than 280,000 close-up eye images with ground truth annotation of the gaze location. Figure 5.8 shows a few example images indicating the positional differences between the cameras and the impacts of cropping and down-sampling the images. A total of 17 participants were recorded, covering a wide range of appearances:

- **Gender**: Five (29%) female and 12 (71%) male
- **Nationality**: Seven (41%) German, seven (41%) Indian, one (6%) Bangladeshi, one (6%) Iranian, and one (6%) Greek
- **Eye Colour**: 12 (70%) brown, four (23%) blue, and one (5%) green
- **Glasses**: Four participants (23%) wore regular glasses and one (6%) wore contact lenses



(a) (b)

Figure 5.7: (a) Overview of the recording setup used in Experiment 2 with a participant wearing the first prototype and resting his head on a chin rest. Ground truth gaze targets (marked in black) were shown in the central area of the screen covering $2 \cdot \alpha$ of the participant's visual field (marked in green). The angular error δ (purple) could then be calculated as the distance between true and predicted gaze targets (marked in orange). On-screen gaze targets were distributed in a grid and split into training (blue) and test data (red). (b) To increase ground truth accuracy, gaze targets were shown with a shrinking animation for 700 ms, and then for another 500 ms at the smallest size. Data was only recorded during the latter 500 ms.

For each participant, two sets of data were recorded: one set of of training data and a separate set of test data. For each set, a series of gaze targets was shown on a display that participants were instructed to look at. For both training and test data the gaze targets covered a uniform grid in a random order, where the grid corresponding to the test data was positioned to lie in between the training points (see Figure 5.7a). Since the NanEye cameras record at about 44 fps, we gathered approximately 22 frames per camera and gaze target. The training data was recorded using a uniform 24 × 17 grid of points, with an angular distance in gaze angle of 1.45° horizontally and 1.30° vertically between the points. In total the training set contained about 8,800 images per camera and participant. The test set's points belonged to a 23 × 16 grid of points and it contains about 8,000 images per camera and participant. This way, the gaze targets covered a field of view of 35° horizontally and 22° vertically.

The recording procedure was split into two parts for training and test data. For both parts, participants were instructed to put on the prototype and rest their head on a chin rest positioned exactly 510 mm in front of a display. The display was a 30-inch LED monitor with a pixel pitch of 0.25 mm and viewable image dimensions of 641.3 × 400.8 mm, set to 2560 × 1600-pixel resolution. On the display, the grid of gaze targets was shown, which the participants were instructed to look at. Each point appeared as a big circle 300 pixels in diameter and shrunk to a circle of 8 pixels diameter over the course of 700 ms. The small circle was then displayed for another 500 ms, until the display of the next point started. Data was only recorded during the latter 500 ms, i.e. while the small circle was shown (see Figure 5.7a). It is important to note

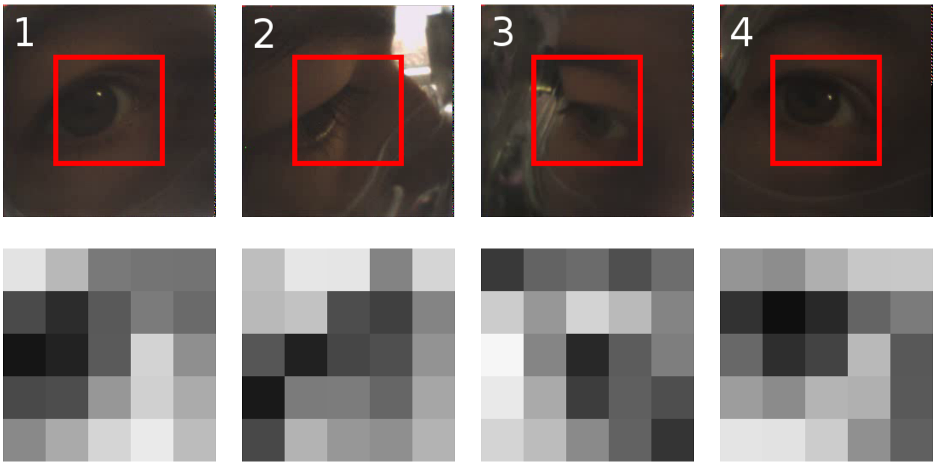


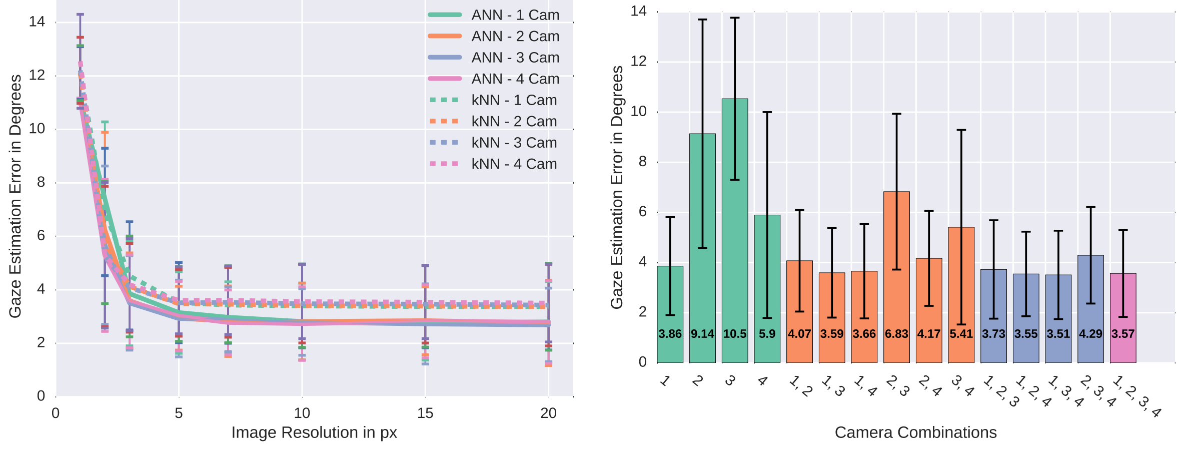
Figure 5.8: (top) Sample eye images from one participant recorded using four NanEye cameras. We identify each of the cameras by the number at the top-left. (bottom) Corresponding cropped low-resolution versions of these images.

that the chin rest did not fully restrain participants and we noticed that their head sometimes moved noticeably, thus resulting in a certain amount of label noise. Using the shrinking animation for the circle helps the participants to locate the circle on the screen and gives them time to relocate their gaze. Similar to (Krafka *et al.*, 2016), we also showed an “L” or an “R” in between every 20th pair of points in the sequence. The letter was displayed for 500 ms at the position of the last point. Participants were asked to confirm the letter they had seen by pressing the corresponding left or right arrow-key. This was done to ensure participants focused on the gaze targets and task at hand throughout the recording.

The data is publicly available at `http://www.mpi-inf.mpg.de/invisibleeye/` (date: 12.07.2019).

Results. We again computed the performance of *InvisibleEye* for different resolutions and camera combinations. Figure 5.9a shows the performance for different resolutions and up to four cameras. Compared to the synthetic case, one can see that the gaze estimation error is now considerably higher but still follows a similar distribution as before. Specifically, for resolutions above 5×5 pixels, the error remains stable with for example 2.9° error for the ANN and 3.52° error for kNN with one camera at exactly 5×5 pixels. These error values are in a range that is low enough for many practical applications like activity recognition (Bulling *et al.*, 2011b) or attention analysis (Sugano *et al.*, 2016). However, if we consider Figure 5.9b we can see that additional cameras do not help for every combination of cameras. Instead, combining cameras that perform worse individually achieves the biggest increase in performance.

Discussion. We have seen that even for very low image resolutions of only 3×3 pixels, *InvisibleEye* is capable of estimating gaze at a low error of 3.86° with a single camera and 3.57° when combining four cameras. This shows that gaze estimation at these low



(a)

(b)

Figure 5.9: (a) Average gaze estimation error for different image resolutions for a k-Nearest-Neighbours approach and our suggested neural network approach on the controlled laboratory data. (b) Average gaze estimation error for different numbers of cameras at 3-pixel resolution. Please refer to Figure 5.8 for the camera-label assignment.

resolutions is possible with real-world data at an accuracy that is practically relevant. These error values further represent an upper bound to what *InvisibleEye* can achieve in this setting, due to the label noise in the data. In the following experiment we will see that, although we move into a more difficult setting, the achieved errors will be even lower since we do not have as much label noise in the data.

Furthermore, the results suggest that combining multiple cameras does not yield a benefit in every case but can improve performance markedly when combining cameras that perform badly individually. Since, in practice, one will always have design constraints on the hardware and a different fit of the device on every user, one runs the risk of positioning the cameras badly for at least some participants. The possibility of combining the information from multiple bad cameras is therefore highly relevant in practice.

5.4.3 Experiment 3: Evaluation in a Mobile Setting

In the controlled setting, we assumed a display at a fixed distance in front of the user and predicted gaze in the screen coordinate system. In practice, however, we want to allow users to move around freely and still be able to track gaze on all kinds of objects, not only displays. Bridging this gap between the controlled laboratory setting and the real world requires adding a scene camera to the system that records the user's field of view and allows us to estimate gaze in scene camera coordinates.

We built a second hardware prototype featuring such a scene camera to test *InvisibleEye* in a mobile setting. We also explicitly allowed gaze targets at arbitrary depths. The depth at which a gaze target lies directly correlates with the location of the target projected into the camera image. From only the view of one eye, this location in the image is, however, in general not inferable. If, for example, the target is moved

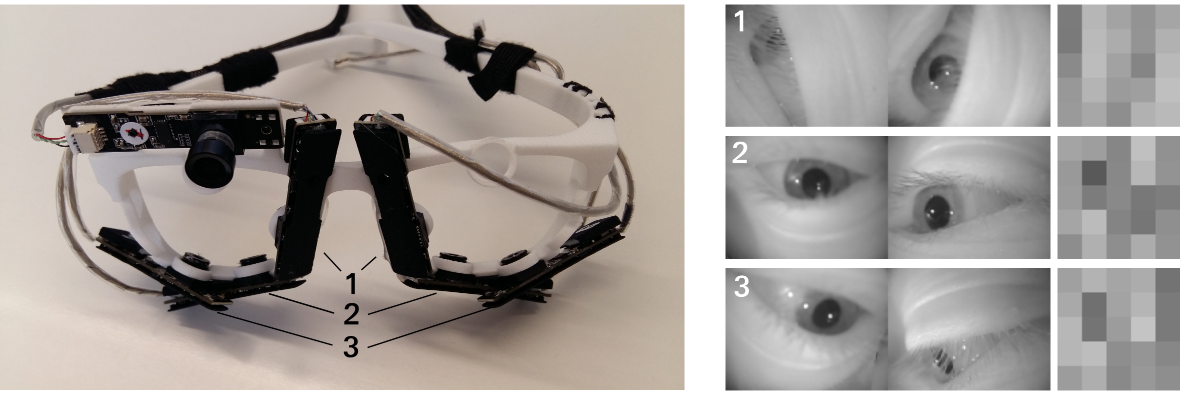


Figure 5.10: (left) Second prototype consisting of a custom 3D-printed glasses frame that can hold up to six Pupil Labs (Kassner *et al.*, 2014) eye cameras with an additional scene camera. (right) Sample images recorded with the prototype with original image on the left and corresponding low-resolution counterparts on the right. We identify camera pairs by the number in the top-left corner.

along the gaze ray projected from the recorded eye into the world, the image of the eye will not change at all if it keeps gazing at the target, while the location of the target in the scene camera image might change considerably (Barz *et al.*, 2016). It is therefore necessary to use views from both eyes to resolve this ambiguity, which we do by using symmetric pairs of cameras recording both eyes. Further, we explicitly allow slippage of the headset, which is a problem frequently occurring in practice.

For this second prototype we decided against using NanEye cameras mainly because comparison with state-of-the-art mobile gaze estimation methods is impossible due to the lower image resolution. We instead used Pupil Labs cameras (Kassner *et al.*, 2014) to record the eyes and the scene using a custom-built, 3D printed frame (see Figure 5.10). Note that, unlike NanEye cameras, the Pupil Labs eye cameras record infrared images of the eye similarly as most cameras in commercially available eye trackers. The field of view of the scene camera was approximately 80° × 60°. Please note that although these cameras are slightly bigger, they are now located directly in the frame of a pair of glasses, i.e. their viewing angles are exactly as they should be.

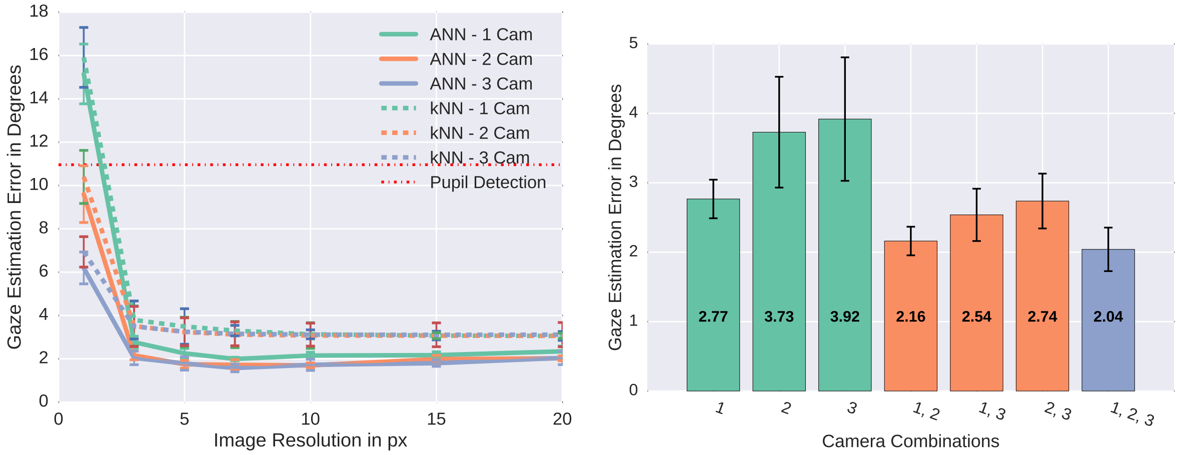
Data Collection. Using this prototype, we recorded another dataset of 240,000 eye images with four participants (four male, aged between 24 and 38 years). To record gaze data at varying distances in a mobile setting, a calibration marker was attached to a wall in front of the participants. Participants were asked to position themselves at an arbitrary distance of up to 3 meters in front of the marker and to perform a series of head movements while gazing at the marker. The head movements consisted of continuously moving the head upwards and downwards while rotating it from the far left to the far right within approximately 10 seconds. Participants were asked to perform the movement such that the marker would move to the edge of their field of view but always remain visible, so they could gaze at it. After performing the head movements, participants were asked to position themselves at a new randomly selected distance for another recording. We repeated this procedure for the whole duration of the recording session. Additionally, to simulate slippage of the headset that is pervasive in mobile settings (Sugano and Bulling, 2015), participants were asked to take off the headset and to put it back on after every 6th recording. Each recording session lasted for about 15 minutes and every participant performed a total of four sessions. This way we were able to efficiently gather images for gaze angles of a large field of view of roughly 70° × 60°.

Each eye image was automatically labelled with the position of the calibration marker in the scene camera. All cameras were set to record images of 640×480 pixels resolution at 120 Hz. Per session, approximately 30,000 images were recorded by each camera. To reduce the required time for training, we reduced the training set to a random subset of 15,000 images. The data of the first three sessions was used as training data, while the data of the fourth session was used for testing. Given that the data was recorded indoors, the images recorded by the infrared cameras were not subject to any significant changes in lighting conditions.

As before, the images we recorded with this second prototype were of much higher quality than what we required for *InvisibleEye*. We therefore down-sampled the images to a lower resolution. We did not crop the images this time because the cameras were sufficiently close to the eye in this second prototype. The images recorded from each camera pair, i.e. one camera from the left side and its symmetrical counterpart from the right side, were concatenated before the process. Sample images and corresponding low-resolution versions are shown in Figure 5.10. For gaze estimation, we used the same neural network architecture as before.

Results. Because the data recorded with our prototype for the mobile setting was recorded with high-quality cameras, we first computed a baseline performance using a state-of-the-art gaze estimation approach based on pupil detection on the original high-resolution images. For this we used the previously mentioned publicly available pipeline from Pupil Labs (Kassner *et al.*, 2014). We randomly picked 200 images out of the first 1,650 images recorded for every participant as calibration data. Due to the continuous and random head movement during recording sessions, the first 1,650 images already cover the entire field of view of the participant and thus represent a realistic set of calibration images. We picked only images recorded by the left camera of pair number two, since this camera position is the closet to that of traditional eye trackers. After detecting the pupil positions in all calibration images, the next step in the Pupil Labs pipeline is to fit a 7th order polynomial to map the pupil positions to the ground truth gaze positions. Using this polynomial, we then estimated gaze positions for all other images from the same participant using the detected pupil position in each image. This baseline method achieved an error of 10.96°. This high error is due to the strong slippage of the headset that is present in the data but not being compensated for. By comparing the positions of one eye corner in a random subset of images, which can be interpreted as an estimate of this slippage, we found that the average distance to the centroid of all eye corner positions was 36.3 pixels.

Similar as before, we evaluated the average gaze estimation performance of *InvisibleEye* for increasingly lower resolutions as well as the number of used cameras.



(a)

(b)

Figure 5.11: (a) Average gaze estimation error for different image resolutions and number of cameras. Also shown is the performance of a state-of-the-art (high-resolution) mobile eye tracking method. (b) Average gaze estimation error for every possible combination of cameras averaged across all participants for an image resolution of 3×3 pixels. Please refer to Figure 5.10 for the camera-label assignment.

The results of this analysis are shown in Figure 5.11a. As we can see, the curves look similar to corresponding ones in the constraint setting, i.e. for resolutions larger than 5×5 pixels the performance remains stable, whereas it drops for lower resolutions. At 5×5-pixel resolution the average error when using all three camera pairs was 1.79°. The average error when using only a single camera was 2.25°. Thus, the use of multiple views of the eye led to a performance increase of approximately 20%. Figure 5.11b shows the average performance of every possible combination of camera pairs across all participants. Here we can see that, in all cases, the addition of a second camera pair improved the results on average.

Discussion. Experiment 3 has shown that for images recorded using cameras positioned around the frame, even if using high-resolution images classical approaches based on pupil detection perform badly. We showed that in contrast, for the same camera positions, *InvisibleEye* achieves a better performance, even for image resolutions as low as 3×3 pixels (corresponding to an error of 2.04°) using three cameras. This result shows that *InvisibleEye* is a viable option even in difficult settings. In this setting we have also seen that adding more cameras can improve performance. This might indicate that the apparent camera angles are difficult enough by themselves and that they can complement each other well, as was the case for cameras 2 and 3 in the controlled laboratory setting.

5.5 Discussion

In this chapter we introduced *InvisibleEye*, a new approach that addresses several key challenges of current mobile eye trackers. The key novelty of our approach is the combination of small and low-quality cameras with an image resolution as low as 3×3 pixels with a method for learning-based gaze estimation. Our experiments show that despite the very low image resolution, *InvisibleEye* can still achieve an accuracy of 2.25° at 5×5-pixel image resolution when using a single pair of cameras in a mobile setting. We have also shown that using three pairs of cameras capturing different views of the eye can further improve performance to 1.79°. The hardware requirements for an embedded system to run *InvisibleEye* at test time are also very low. While model training might be feasible on a mobile device, it could be outsourced to a standard desktop machine or a cloud service too, making *InvisibleEye* easy to deploy in practice. These findings are highly encouraging given that they not only demonstrate the feasibility of our approach but, more importantly, underline its potential for finally realising the vision of invisible mobile eye tracking. Despite these promising results, our *InvisibleEye* prototypes still have several limitations. First, all evaluations shown here are based on person-specific training, i.e. every user needs to record training data with the device prior to first use. It is important to note, however, that while highly undesirable from a usability point of view, the requirement for person-specific training or calibration also applies to state-of-the-art mobile eye trackers that use a classic gaze estimation method and person-specific calibration. Nonetheless, the amount of person-specific training data currently required for our method is still significantly larger than the one for standard calibration approaches. Methods from transfer learning could, for example, be used to reduce the amount of required training data, and it is also promising to investigate implicit calibration approaches.

The ability to robustly estimate gaze across the large variability in eye appearances of different people is a significantly more challenging task and thus represents the most important direction for future work. A less challenging yet still highly practical solution could be eye tracker self-calibration in which gaze positions are inferred, for example, from saliency maps calculated from the scene camera images (Sugano and Bulling, 2015). This has the potential to allow the user to gather training data naturally just by wearing the device for an extended amount of time, thereby continuously improving performance during everyday use.

Second, in this chapter we have not yet evaluated the performance of *InvisibleEye* in an outdoor environment, nor during long-term recordings. Usually, mobile eye tracking systems perform a lot worse outdoors because the sun can create intense reflections and shadows on the eye image (Tonsen *et al.*, 2016). It remains to be explored if a learning-based approach can improve the robustness in such challenging environments.

Finally, while the two prototype systems of *InvisibleEye* that we have built were sufficient to investigate its performance in both stationary controlled and mobile real-world settings, a fully integrated mobile eye tracker that can be used robustly in daily life is still highly desirable. Currently, such full integration is not possible with the NanEye cameras used in the work of this chapter, given that they have to be connected to a desktop computer using a special-purpose USB breakout board. The cameras do use a standard video interface, however, which makes us confident that fully embedded integration of both hardware and software will soon be feasible.

5.6 Conclusion

In this chapter we presented *InvisibleEye* – a novel approach that, in contrast to a long line of work on mobile eye tracking, relies on tiny cameras that can be nearly invisibly integrated into a normal glasses frame. To compensate for the cameras' low image resolution of only a few pixels, we showed how to combine multiple of them using a learning-based gaze estimation method that directly regresses from eye images to gaze directions. We evaluated our system on three increasingly challenging datasets to study its performance across key design parameters including image resolution, number of cameras, as well as camera angle and positioning. Our approach achieved a person-specific gaze estimation accuracy of 1.79° using three cameras with a resolution of only 5×5 pixels. These findings are promising and not only underline the potential of this new approach but mark an important step towards realising the vision of fully unobtrusive, comfortable, and socially acceptable mobile eye tracking.

6 PrivacEye: Privacy-Preserving Head-Mounted Eye Tracking Using Egocentric Scene Image and Eye Movement Features

Eyewear devices, such as augmented reality displays, increasingly integrate eye tracking, but the first-person camera required to map a user's gaze to the visual scene can pose a significant threat to user and bystander privacy. We present *PrivacEye*, a method to detect privacy-sensitive everyday situations and automatically enable and disable the eye tracker's first-person camera using a mechanical shutter. To close the shutter in privacy-sensitive situations, the method uses a deep representation of the first-person video combined with rich features that encode users' eye movements. To open the shutter without visual input, *PrivacEye* detects changes in users' eye movements alone to gauge changes in the "privacy level" of the current situation. We evaluate our method on a first-person video dataset recorded in daily life situations of 17 participants, annotated by themselves for privacy sensitivity, and show that our method is effective in preserving privacy in this challenging setting.

6.1 Introduction

Eyewear devices, such as head-mounted displays or augmented reality glasses, have recently emerged as a new research platform in fields such as human-computer interaction, computer vision, or the behavioural and social sciences (Bulling and Kunze, 2016). An ever-increasing number of these devices integrate eye tracking to analyse attention allocation (Eriksen and Yeh, 1985; Sugano *et al.*, 2016), for computational user modelling (Fischer, 2001; Itti and Koch, 2001), or hands-free interaction (Hansen *et al.*, 2003; Vertegaal *et al.*, 2003). Head-mounted eye tracking typically requires two cameras: An eye camera that records a close-up video of the eye and a high-resolution first-person (scene) camera to map gaze estimates to the real-world scene (Kassner *et al.*, 2014). The scene camera poses a serious privacy risk as it may record sensitive personal information, such as login credentials, banking information, or text messages, as well as infringe on the privacy of bystanders (Perez *et al.*, 2017). Privacy risks intensify with the unobtrusive integration of eye tracking in ordinary glasses frames (Tonsen *et al.*, 2017).

In the area of first-person vision, prior work identified strategies of self-censorship (Koelle *et al.*, 2017) that, however, are prone to (human) misinterpretations and forgetfulness, or the accidental neglect of social norms and legal regulations. In consequence, user experience and comfort are decreased and the user's mental and emotional load increases, while sensitive personal information can still be accidentally disclosed. Other works therefore investigated alternative solutions, such as communicating a bystander's privacy preferences using short-range wireless radio (Aditya *et al.*, 2016), visual markers (Schiff *et al.*, 2007), or techniques to compromise recordings (Harvey, 2012; Truong *et al.*, 2005). However, all of these methods require bystanders to take action themselves to protect their privacy. None of these works addressed the problem at its source, i.e. the scene camera, nor did they offer a means to protect the privacy of both the wearer and potential bystanders.

To address this limitation, we propose *PrivacEye*, the first method for privacy-preserving head-mounted eye tracking (see Figure 6.1). The key idea and core novelty of our method is to detect users' transitions into and out of privacy-sensitive everyday situations by leveraging both cameras available on these trackers. If a privacy-sensitive situation is detected, the scene camera is occluded by a physical shutter. Our design choice to use a non-spoofable physical shutter, which closes for some time and therefore provides feedback to bystanders, is substantiated by Koelle et al., who highlight an increased trustworthiness over LED lights on the camera or pure software solutions (Koelle *et al.*, 2018b). While this approach is secure and visible to bystanders, it prohibits visual input from the scene. Thus, our method analyses changes in the users' eye movement behaviour alone to detect if they exit a privacy-sensitive situation and then reopens the camera shutter. A naive, vision-only system could reopen the shutter at regular intervals, e.g. every 30 seconds, to detect whether the current situation is still privacy-sensitive. However, this approach may negatively affect perceived reliability and increase mistrust in the system. Thus, our eye tracking approach promises significant advantages over a purely interval-based approach in terms of user experience and perceived trustworthiness.

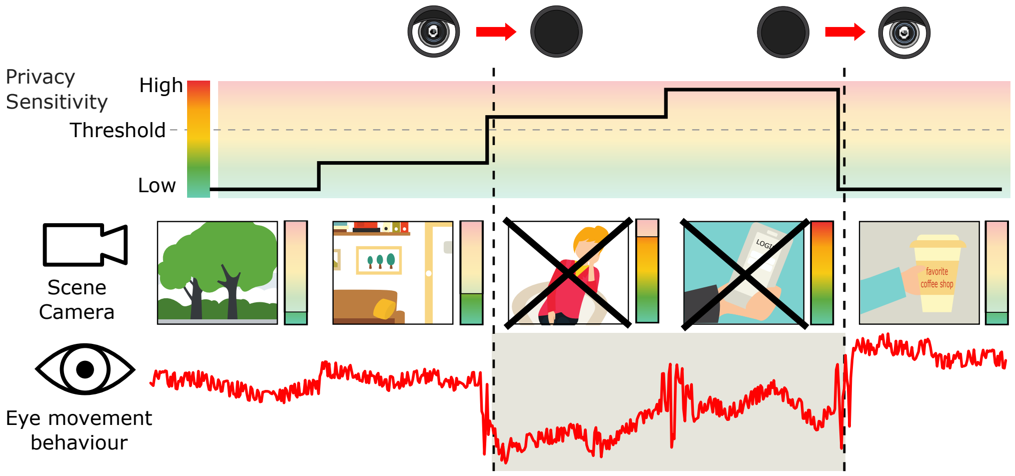


Figure 6.1: Our method uses a mechanical camera shutter (top) to preserve users' and bystanders' privacy with head-mounted eye trackers. Privacy-sensitive situations are detected by combining deep scene image and eye movement features (middle) while changes in eye movement behaviour alone trigger the reopening of the camera shutter (bottom).

Our approach is motivated by prior work that demonstrates that eye movements are a rich source of information on a user's everyday activities (Bulling *et al.*, 2011b; Steil and Bulling, 2015), social interactions and current environment (Bulling *et al.*, 2013), or even a user's personality traits (Hoppe *et al.*, 2018). In addition, prior work showed that perceived privacy sensitivity is related to a user's location and activity (Hoyle *et al.*, 2015). We therefore hypothesise that *privacy sensitivity* transitively informs a user's *eye movements*. We are the first to confirm this transitivity, which results as a reasoned deduction from prior work.

The specific contributions of this chapter are three-fold: First, we present *PrivacEye*, the first method that combines the analysis of egocentric scene image features with eye movement analysis to enable context-specific, privacy-preserving de-activation and re-activation of a head-mounted eye tracker's scene camera. As such, we show a previously unconfirmed transitive relationship over the users' eye movements, their current activity and environment, as well as the perceived privacy sensitivity of the situation they are in. Second, we evaluate our method on a dataset of real-world mobile interactions and eye movement data, fully annotated with locations, activities, and privacy sensitivity levels of 17 participants. Third, we provide qualitative insights on the perceived social acceptability, trustworthiness, and desirability of *PrivacEye*, based on semi-structured interviews, using a fully functional prototype.

6.2 Related Work

Research on eye tracking privacy is sparse. Thus, the work in this chapter mostly relates to previous works on (1) privacy concerns with first-person cameras and (2) privacy enhancing methods for (wearable) cameras.

6.2.1 Privacy Concerns – First-Person Cameras

First-person cameras are well-suited for continuous and unobtrusive recordings, which causes them to be perceived as unsettling by bystanders (Denning *et al.*, 2014). Both users' and bystanders' privacy concerns and attitudes towards head-mounted devices with integrated cameras were found to be affected by context, situation, usage intentions (Koelle *et al.*, 2015), and user group (Profita *et al.*, 2016). Hoyle et al. showed that the presence and the number of people in a picture, specific objects (e.g., computer displays, ATM cards, physical documents), location, and activity affected whether lifeloggers deemed an image "shareable" (Hoyle *et al.*, 2014). They also highlighted the need for automatic privacy-preserving mechanisms to detect those elements, as individual sharing decisions are likely to be context-dependent and subjective. Their results were partly confirmed by Price et al., who, however, found no significant differences in sharing when a screen was present (Price *et al.*, 2017). Chowdhury et al. found that whether lifelogging imagery is suitable for sharing is (in addition to content, scenario, and location) mainly determined by its sensitivity (Chowdhury *et al.*, 2016). Ferdous et al. proposed a set of guidelines that, among others, include semi-automatic procedures to determine the sensitivity of captured images according to user-provided preferences (Ferdous *et al.*, 2017). All highlight the privacy sensitivity of first-person recordings and the importance of protecting user and bystander privacy.

6.2.2 Enhancing Privacy of First-Person Cameras

To increase the privacy of first-person cameras for bystanders, researchers have suggested communicating their privacy preferences to nearby capture devices using wireless connections as well as mobile or wearable interfaces (Krombholz *et al.*, 2015). Others have suggested preventing unauthorised recordings by compromising the recorded imagery, e.g., using infrared light signals (Harvey, 2010; Yamada *et al.*, 2013) or disturbing face recognition (Harvey, 2012). In contrast to our approach, these techniques all require the bystander to take action, which might be impractical due to costs and efforts (Denning *et al.*, 2014).

A potential remedy are automatic, or semi-automatic approaches, such as *PlaceAvoider*, a technique that allows users to "blacklist" sensitive spaces, e.g., bedroom or bathroom (Templeman *et al.*, 2014). Similarly, *ScreenAvoider* allowed users to control the disclosure of images of computer screens showing potentially private content (Korayem *et al.*, 2016). Erickson et al. proposed a method to identify security risks, such as ATMs, keyboards, and credit cards, in images captured by first-person wearable devices (Erickson *et al.*, 2014). However, instead of assessing the whole scene in terms of privacy sensitivity, their systems only detected individual sensitive objects. Raval et al. presented *MarkIt*, a computer vision-based privacy marker framework that allowed users to use self-defined bounding boxes and hand-gestures to restrict visibility of content on two dimensional surfaces (e.g. white boards) or sensitive real-world objects (Raval *et al.*, 2014). *iPrivacy* automatically detects privacy-sensitive objects from social images users are willing to share using deep multi-task learning (Yu *et al.*, 2017). It warns the image owners what objects in the images need to be protected before sharing and recommends privacy settings.

While all of these methods improved privacy, they either only did so post-hoc, i.e. after images had already been captured, or they required active user input. In contrast, our approach aims to prevent potentially sensitive imagery from being recorded at all, automatically in the background, i.e. without engaging the user. Unlike current computer vision based approaches that work in image space, e.g. by masking objects or faces (Yamada *et al.*, 2013; Raval *et al.*, 2014; Shu *et al.*, 2016), restricting access (Korayem *et al.*, 2016), or deleting recorded images post-hoc (Templeman *et al.*, 2014), we de-activate the camera completely using a mechanical shutter and also signal this to bystanders. Our approach is the first to employ eye movement analysis for camera re-activation that, unlike other sensing techniques (e.g., microphones, infrared cameras), does not compromise the privacy of potential bystanders.

6.3 Design Rationale

PrivacEye's design rationale is based on user and bystander goals and expectations. In this section, we outline how *PrivacEye*'s design contributes to avoiding erroneous disclosure of sensitive information, so-called misclosures (User Goal 1), and social friction (User Goal 2), and detail on three resultant design requirements.

6.3.1 Goals and Expectations

Avoid Misclosure of Sensitive Data. A user wearing smart glasses with an integrated camera would typically do so to make use of a particular functionality, e.g., visual navigation. However, the device's "always-on" characteristic causes it to capture more than originally intended. A navigation aid would require capturing certain landmarks for tracking and localisation. In addition, unintended imagery and potentially sensitive data is captured. Ideally, to prevent misclosures (Caine, 2009), sensitive data should not be captured. However, requiring the user to constantly monitor her actions and environment for potential sensitive information (and then de-activate the camera manually) might increase the workload and cause stress. As users might be forgetful, misinterpret situations, or overlook privacy-sensitive items, automatic support from the system would be desirable from a user's perspective.

Avoid Social Friction. The smart glasses recording capabilities may cause social friction if they do not provide a clear indication whether the camera is on or off: Bystanders might even perceive device usage as a privacy threat when the camera is turned off (Koelle *et al.*, 2015, 2018b). In consequence, they feel uncomfortable around such devices (Bohn *et al.*, 2005; Denning *et al.*, 2014; Ens *et al.*, 2015; Koelle *et al.*, 2015). Similarly, user experience is impaired when device users feel a need for justification as they could be accused of taking surreptitious pictures (Häkkilä *et al.*, 2015; Koelle *et al.*, 2018b).

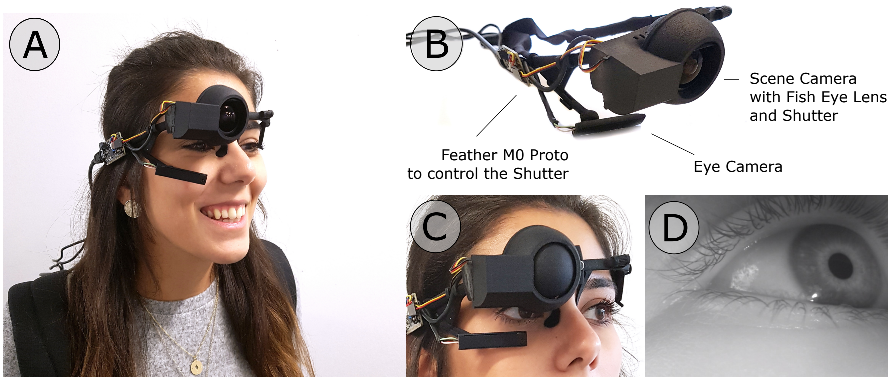

A
B
Scene Camera with Fish Eye Lens and Shutter
Feather M0 Proto to control the Shutter
Eye Camera
C
D

Figure 6.2: *PrivacEye* prototype with labelled components (B) and worn by a user with a USB-connected laptop in a backpack (A). Detection of privacy-sensitive situations using computer vision closes the camera shutter (C), which is reopened based on a change in the privacy detected level in a user's eye movements (D).

6.3.2 Design Requirements

As a consequence of these user goals there are three essential design requirements that *PrivacEye* addresses: (1) The user can make use of the camera-based functionality without the risk of misclosures or leakage of sensitive information. (2) The system pro-actively reacts to the presence or absence of potentially privacy-sensitive situations and objects. (3) The camera device communicates the recording status clearly to both user and bystander.

6.4 PrivacEye Prototype

Our fully functional *PrivacEye* prototype, shown in Figure 6.2, is based on the Pupil head-mounted eye tracker (Kassner *et al.*, 2014) and features one 640 × 480 pixel camera (the so-called "eye camera") that records the right eye from close proximity (30 fps), and a second camera (1280 × 720 pixels, 24 fps) to record a user's environment (the so-called "scene camera"). The first-person camera is equipped with a fish eye lens with a 175° field of view and can be closed with a mechanical shutter. The shutter comprises a servo motor and a custom-made 3D-printed casing, including a mechanical lid to occlude the camera's lens. The motor and the lid are operated via a micro controller, namely a Feather M0 Proto. Both cameras and the micro controller were connected to a laptop via USB. *PrivacEye* further consists of two main software components: (1) detection of privacy-sensitive situations to close the mechanical camera shutter and (2) detection of changes in user's eye movements that are likely to indicate suitable points in time for reopening the camera shutter.

6.4.1 Detection of Privacy-Sensitive Situations

The approaches for detecting privacy-sensitive situations we evaluated are (1) *CNN-Direct*, (2) *SVM-Eye*, and (3) *SVM-Combined*.

CNN-Direct. Inspired by prior work on predicting privacy-sensitive pictures posted in social networks (Orekondy *et al.*, 2017), we used a pre-trained GoogleNet, a 22-layer deep convolutional neural network (Szegedy *et al.*, 2015). We adapted the original GoogleNet model for our specific prediction task by adding two additional fully connected (FC) layers. The first layer was used to reduce the feature dimensionality from 1024 to 68 and the second one, a Softmax layer, to calculate the prediction scores. Output of our model was a score for each first-person image indicating whether the situation visible in that image was privacy-sensitive or not. The cross-entropy loss was used to train the model. The full network architecture is included in Appendix B (see Figure B.2).

SVM-Eye. Given that eye movements are independent from the scene camera's shutter status, they can be used to (1) detect privacy-sensitive situations while the camera shutter is open and (2) detect changes in the subjective privacy level while the camera shutter is closed. The goal of this second component is to instead detect changes in a user's eye movements that are likely linked to changes in the privacy sensitivity of the current situation and thereby to keep the number of times the shutter is reopened as low as possible. To detect privacy-sensitive situations and changes, we trained SVM classifiers (kernel = rbf, $C = 1$) with characteristic eye movement features, which we extracted using only the eye camera video data. We extracted a total of 52 eye movement features, covering fixations, saccades, blinks, and pupil diameter (see Table B.2 in Appendix B for a list and description of the features). Similar to (Bulling *et al.*, 2011b), each saccade is encoded as a character forming words of length n (wordbook). We extracted these features using a sliding window of 30 seconds (step size of 1 sec).

SVM-Combined. A third approach for the detection of privacy-sensitive situations is a hybrid method. We trained SVM classifiers using the extracted eye movement features (52) and combined them with CNN features (68) from the scene image, which we extracted from the first fully connected layer of our trained CNN model, creating 120 feature large samples. With the concatenation of eye movement and scene features, we are able to extend the information from the two previous approaches during recording phases where the camera shutter is open.

6.5 Experiments

We evaluated the different approaches on their own and in combination in a realistic temporal sequential analysis trained in a person-specific (leave-one-recording-out) and person-independent (leave-one-person-out) manner. We assume that the camera shutter is open at start up. If no privacy-sensitive situation is detected, the camera shutter remains open and the current situation is rated "non-sensitive", otherwise, the camera shutter is closed and the current situation is rated "privacy-sensitive". Finally, we analysed error cases and investigated the performance of *PrivacEye* in different environments and activities.

6.5.1 Dataset

While an ever-increasing number of eye movement datasets have been published in recent years (see (Bulling *et al.*, 2011b, 2012; Steil and Bulling, 2015; Sugano and Bulling, 2015; Hoppe *et al.*, 2018) for examples), none of them focused on privacy-related attributes. We therefore make resource to a previously recorded dataset (Steil *et al.*, 2018b). The dataset of Steil et al. contains more than 90 hours of data recorded continuously from 20 participants (six females, aged 22-31) over more than four hours each. Participants were students with different backgrounds and subjects with normal or corrected-to-normal vision. During the recordings, participants roamed a university campus and performed their everyday activities, such as meeting people, eating, or working as they normally would on any day at the university. To obtain some data from multiple, and thus also "privacy-sensitive", places on the university campus, participants were asked to not stay in one place for more than 30 minutes. Participants were further asked to stop the recording after about one and a half hours so that the laptop's battery packs could be changed and the eye tracker re-calibrated. This yielded three recordings of about 1.5 hours per participant. Participants regularly interacted with a mobile phone provided to them and were also encouraged to use their own laptop, desktop computer, or music player if desired. The dataset thus covers a rich set of representative real-world situations, including sensitive environments and tasks. The data collection was performed with the same equipment as shown in Figure 6.2 excluding the camera shutter.

6.5.2 Data Annotation

The dataset was fully annotated by the participants themselves with continuous annotations of location, activity, scene content, and subjective privacy sensitivity level. 17 out of the 20 participants finished the annotation of their own recording resulting in about 70 hours of annotated video data. They again gave informed consent and completed a questionnaire on demographics, social media experience and sharing behaviour (based on Hoyle et al. (Hoyle *et al.*, 2014)), general privacy attitudes, as well as other-contingent privacy (Baruh and Cemalcılar, 2014) and respect for bystander privacy (Price *et al.*, 2017). General privacy attitudes were assessed using the *Privacy Attitudes Questionnaire* (PAQ), a modified Westin Scale (Westin, 2003) as used by (Caine, 2009; Price *et al.*, 2017).

Annotations were performed using Advene (Aubert *et al.*, 2012). Participants were asked to annotate continuous video segments showing the same situation, environment, or activity. They could also introduce new segments in case a privacy-relevant feature in the scene changed, e.g., when a participant switched to a sensitive app on the mobile phone. Participants were asked to annotate each of these segments according to the annotation scheme (see Table B.1 in Appendix B). Privacy sensitivity was rated on a 7-point Likert scale ranging from 1 (fully inappropriate) to 7 (fully appropriate). As we expected our participants to have difficulties understanding the concept of "privacy sensitivity", we rephrased it for the annotation to "How appropriate is it that a camera is in the scene?". Figure 6.3 visualises the labelled privacy sensitivity levels for each participant. Based

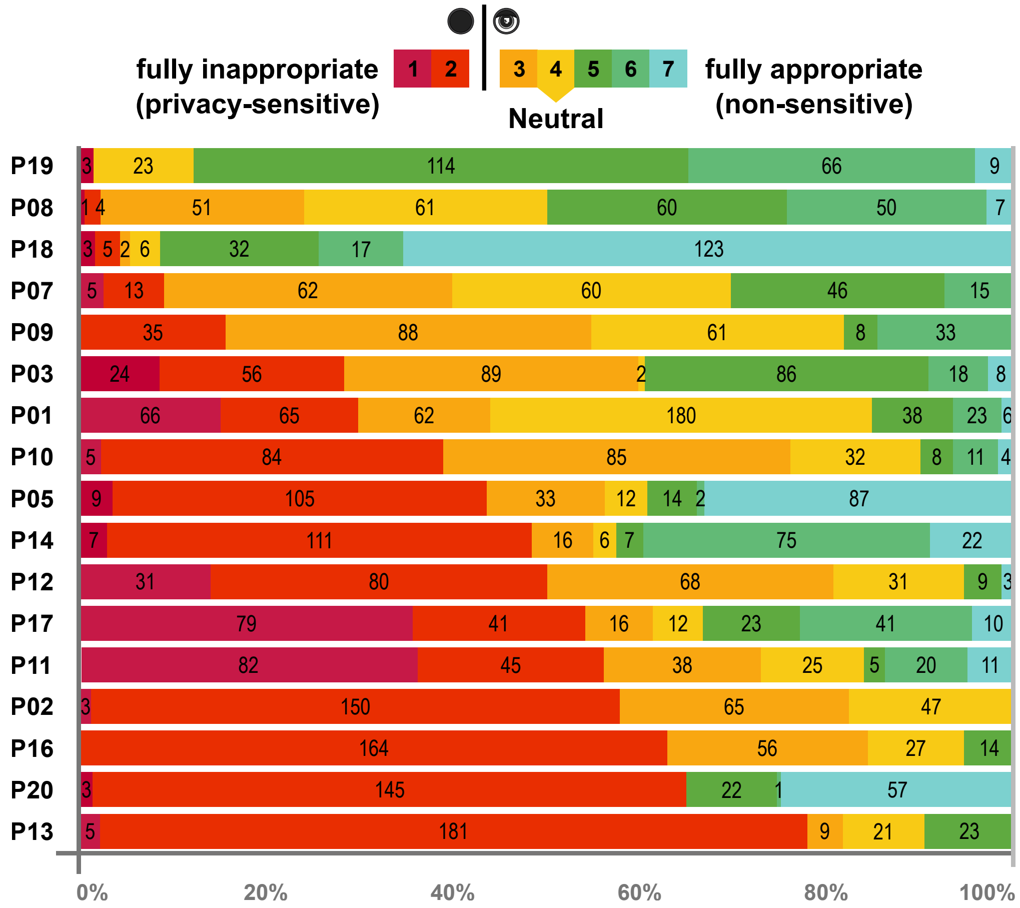


Figure 6.3: Privacy sensitivity levels rated on a 7-pt Likert scale from 1: fully inappropriate (i.e. privacy-sensitive) to 7: fully appropriate (i.e. non-sensitive). Distribution in labelled minutes/level per participant, sorted according to a “cut-off” between closed shutter (level 1-2) and open shutter (level 3-7). In practice, the “cut-off” level could be chosen according to individual ratings as measured by PAQ.

on the latter distribution, we pooled ratings of 1 and 2 in the class “privacy-sensitive”, and all others in the class “non-sensitive”. A consumer system would provide the option to choose this “cut-off”. We will use these two classes for all evaluations and discussions that follow in order to show the effectiveness of our proof-of-concept system. The dataset is available at `https://www.mpi-inf.mpg.de/MPIIPrivacEye/` (date: 12.07.2019).

6.5.3 Sequential Analysis

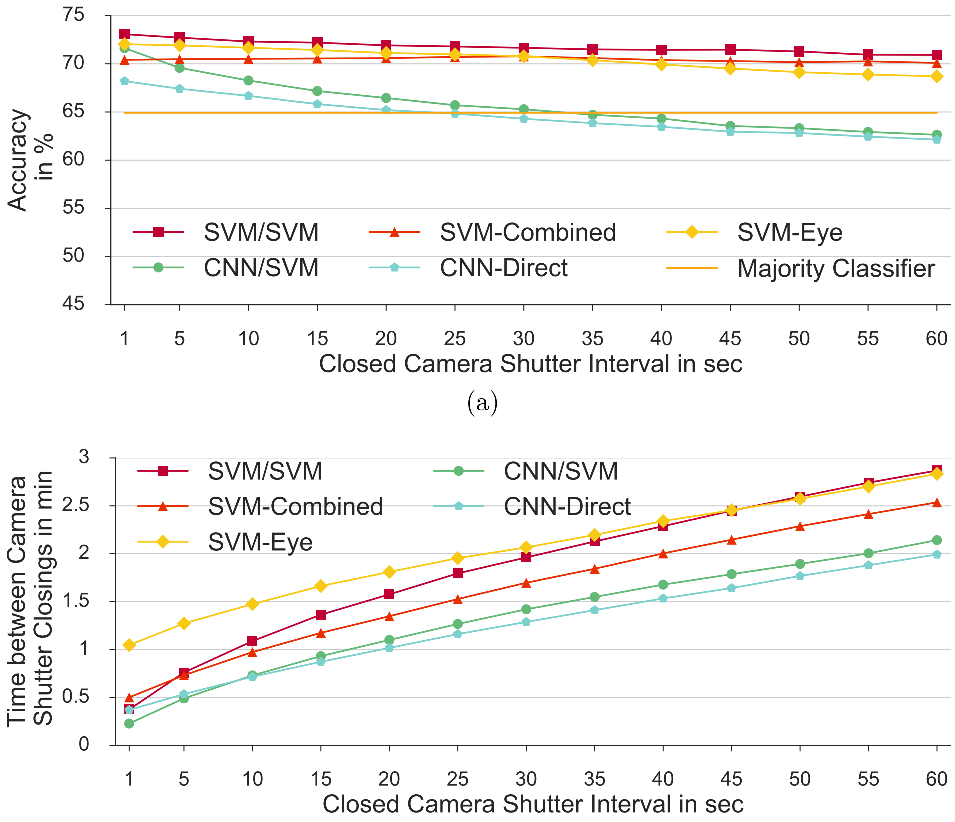
To evaluate *PrivacEye*, we applied the three proposed approaches separately as well as in combination in a realistic temporal sequential analysis, evaluating the system as a whole within person-specific (leave-one-recording-out) and person-independent (leave-one-person-out) cross validation schemes. Independent of CNN or SVM approaches, we first trained and then tested in a person-specific fashion. That is, we trained on two of the three recordings of each participant and tested on the remaining one – iteratively over all combinations and averaging the performance results in the end. For the leave-one-person-out cross validation, we trained on the data of 16 participants and tested on the remaining one. *SVM-Eye* is the only one of the three proposed approaches that allows *PrivacEye* to be functional when no scene imagery is available, i.e., when the shutter is closed. Additionally, it can be applied when the shutter is open thus serving both software components of *PrivacEye*. While the camera shutter is not closed, i.e., scene imagery is available, *CNN-Direct* or *SVM-Combined* can be applied. To provide a comprehensive picture, we then analysed the combinations *CNN-Direct* + *SVM-Eye* (*CNN/SVM*) and *SVM-Combined* + *SVM-Eye* (*SVM/SVM*). The first approach is applied when the camera shutter is open and *SVM-Eye* only when the shutter is closed. For the sake of completeness, we also evaluated *SVM-Combined* and *CNN-Direct* on the whole dataset. However, these two methods represent hypothetical best-case scenarios in which eye and scene features are always available. As this is in practice not possible, they have to be viewed as an upper-bound baseline. For evaluation purposes, we apply the proposed approaches within a step size of one second in a sequential manner. The previously predicted camera shutter position (open or close) decides which approach is applied for the prediction of the current state to achieve realistic results. We use $Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$, where TP, FP, TN, and FN count sample-based true positives, false positives, true negatives, and false negatives, as performance indicator.

CNN-Direct. For training the CNN, which classifies a given scene image directly as privacy-sensitive or non-sensitive, we split the data from each participant into segments. Each change in environment, activity, or the annotated privacy sensitivity level starts a new segment. We used one random image per segment for training.

SVM-Eye and SVM-Combined. The SVM classifiers use only eye movement features (*SVM-Eye*) or the combination of eye movement and CNN features (*SVM-Combined*). We standardised the training data (zero mean, unit variance) for the person-specific and leave-one-person-out cross validation before training the classifiers, and used the same parameters for the test data.

6.5.4 Results

With potential usability implications in mind, we evaluate performance over a range of closed camera shutter intervals. If a privacy-sensitive situation is detected from the *CNN-Direct* or *SVM-Combined* approach, the camera shutter is kept closed for an interval between 1 and 60 seconds. If *SVM-Eye* is applied and no privacy change is detected, the shutter remains closed. In a practical application, users build more trust when the camera shutter remains closed, at least for a sufficient amount of time, to guarantee the protection of privacy-sensitive scene content when such a situation is detected (Koelle *et al.*, 2018b). We also evaluated *CNN-Direct* and *SVM-Combined* on the whole recording as hypothetical best-case scenarios. However, comparing their performance against the combinations *SVM/SVM* and *CNN/SVM* illustrate the performance improvement using *SVM-Eye* when the camera shutter is closed.

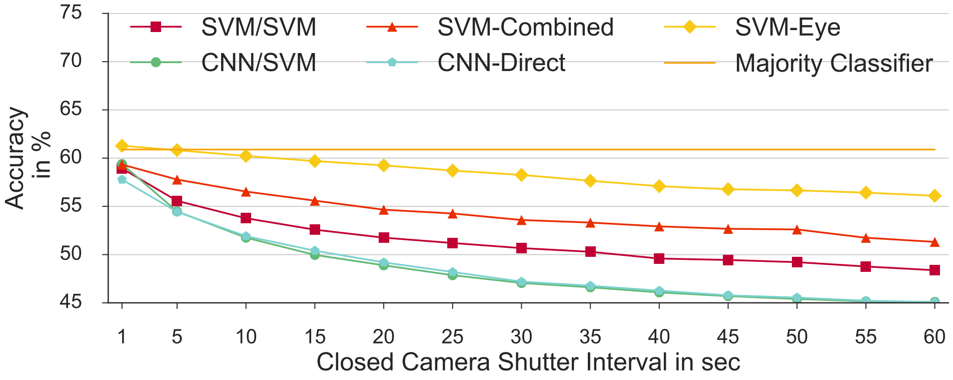


(a)

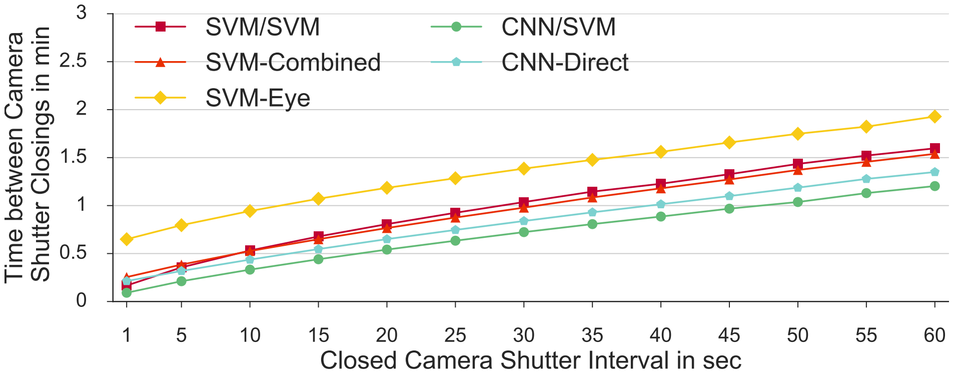
(b)

Figure 6.4: Person-specific leave-one-recording-out evaluation showing the achieved accuracy (a) and the time between camera shutter closings (b) across different closed camera shutter intervals.

Person-Specific (Leave-One-Recording-Out) Evaluation. Figure 6.4a shows the person-specific accuracy performance of *PrivacEye* against increasing camera shutter closing time for two combinations *CNN/SVM* and *SVM/SVM*, and *SVM-Eye*, which can be applied independent of the camera shutter status. Besides *CNN-Direct* and *SVM-Combined*, the majority class classifier serves as a baseline, predicting the majority class from the training set. The results reveal that all trained approaches and combinations perform above the majority class classifier. However, we can see that *CNN-Direct* and its combination with *SVM-Eye* (*CNN/SVM*) perform below the other approaches and below the majority class classifier for longer closed camera shutter intervals. *SVM-Eye* and *SVM-Combined* perform quite robustly, around 70% accuracy, while *SVM-Eye* performs better for shorter intervals and *SVM-Combined* for longer intervals. The interplay approach *SVM/SVM*, which we would include in our prototype, exceeds 73% with a closed camera shutter interval of one second and outperforms all other combinations in terms of accuracy in all other intervals. One reason for the performance improvement of *SVM/SVM* in comparison to its single components is that *SVM-Combined* performs better for the detection of privacy-sensitive situations when the camera shutter is open while *SVM-Eye* performs better for preserving privacy-sensitive situations so



(a)



(b)

Figure 6.5: Person-independent leave-one-person-out evaluation showing the accuracy results (a) and the time between closing the camera shutter (b) across different closed camera shutter intervals.

that the camera shutter remains closed. Another aim of our proposed approach is the reduction of opening and closing events during a recording to strengthen reliability and trustworthiness. A comparison of Figure 6.4a and Figure 6.4b renders a clear trade-off between accuracy performance and time between camera shutter closing instances. For very short camera shutter closing times the *SVM-Eye* approach, which only relies on eye movement features from the eye camera, shows the best performance, whereas for longer camera shutter closing times, the combination *SVM/SVM* shows better accuracy with a comparable amount of time between camera shutter closing instances. However, the current approaches are actually not able to reach the averaged ground truth of about 8.2 minutes between camera shutter closings.

Person-Independent (Leave-One-Person-Out) Evaluation. The more challenging task, which assumes that privacy-sensitivity could generalise over multiple participants, is given in the person-independent leave-one-person-out cross validation of Figure 6.5a. Similar to the person-specific evaluation, *CNN-Direct* and *CNN/SVM* perform worse than the other approaches. Here, *SVM-Eye* outperforms *SVM-Combined* and *SVM/SVM*. However, none of the approaches are able to outperform the majority

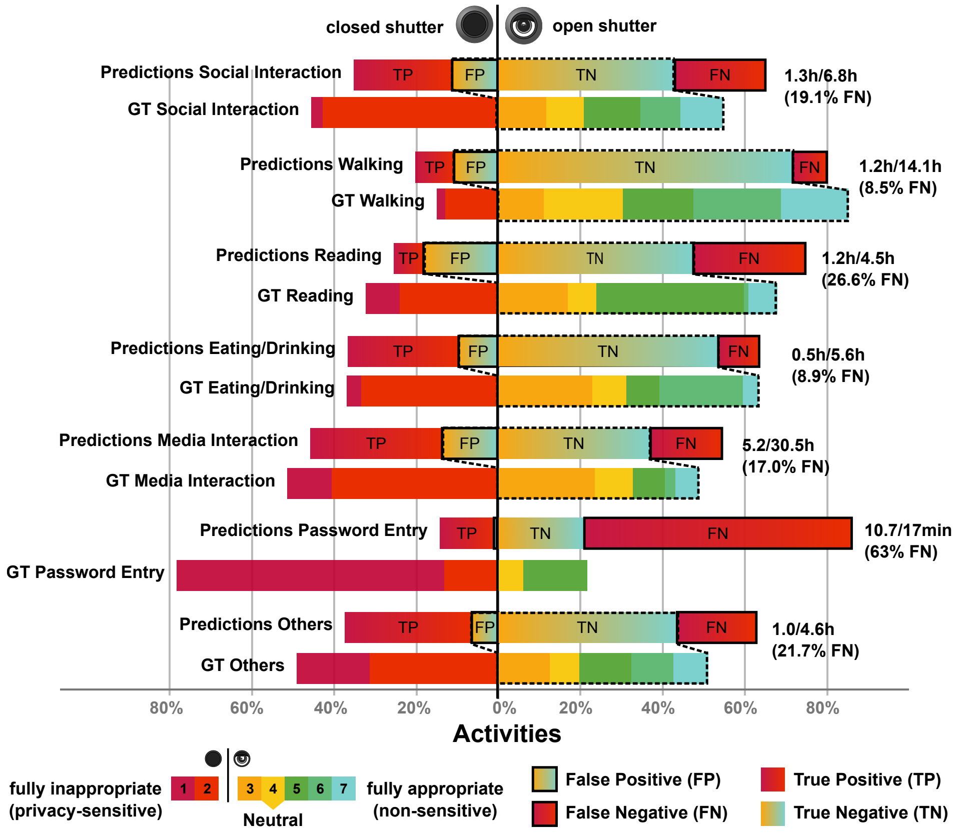


Figure 6.6: Error case analysis for different activities showing the "cut-off" between closed shutter (left, *privacy-sensitive*) and open shutter (right, *non-sensitive*) with *PrivacEye* prediction and the corresponding ground truth (GT). False positives (FP) are *non-sensitive* but protected (closed shutter), false negatives (FN) are *privacy-sensitive* but unprotected (open shutter).

classifier. These results show that eye movement features generalise better over multiple participants to detect privacy-sensitive situations than scene image information. Comparing the number of minutes between camera shutter closing events of person-specific and leave-one-person-out in Figure 6.4b and Figure 6.5b, the person-specific approach outperforms the person-independent leave-one-person-out evaluation scheme for each approach. This shows that privacy sensitivity does not fully generalise, and consumer systems would require a person-specific calibration and online learning.

6.5.5 Error Case Analysis

For *PrivacEye*, it is not only important to detect the privacy-sensitive situations (TP), but equally important to detect non-sensitive situations (TN), which are relevant to grant a good user experience. Our results suggest that the combination *SVM/SVM* performs best for the person-specific case. For this setting we carry out a detailed error case analysis of our system for the participants' different activities. For the activities outlined in Figure 6.6, *PrivacEye* works best while eating/drinking and in media interactions.

Also, the results are promising for detecting social interactions. The performance for password entry, however, is still limited. Although the results show that it is possible to detect password entry, the amount of false negatives (FN) is high compared to other activities. This is likely caused by the dataset's under-representation of this activity, which characteristically lasts only a few seconds. Future work might be able to eliminate this by specifically training for password and PIN entry, which will enable the classifier to better distinguish between PIN entry and, e.g., reading. In Section B.4 we provide an in-depth error case analysis to further investigate error cases in different environments.

6.6 User Feedback

Collecting initial subjective feedback during early stages of system development allows us to put research concepts in a broader context and helps to shape hypotheses for future quantitative user studies. In this section, we report on a set of semi-structured one-to-one interviews on the use of head-worn augmented reality displays in general, and our interaction design and prototype in particular. To obtain the user feedback, we recruited 12 new and distinct participants (six females), aged 21 to 31 years (M = 24, SD = 3) from the local student population. They were enrolled in seven highly diverse majors, ranging from computer science and biology to special needs education. We decided to recruit students, given that we believe they and their peers are potential users of a future implementation of our prototype. We acknowledge that this sample, consisting of rather well educated young adults (with six of them having obtained a Bachelor's degree), is not representative for the general population. Interviews lasted about half an hour and participants received a 5 Euro Amazon voucher. We provide a detailed interview protocol in Section B.5. The semi-structured interviews were audio recorded and transcribed for later analysis. Subsequently, qualitative analysis was performed following inductive category development (Mayring, 2014). Key motives and reoccurring themes were extracted and are presented in this section, where we link back to *PrivacEye*'s design and discuss implications for future work.

6.6.1 User Views on Transparency

Making it transparent (using the 3D-printed shutter), whether the camera was turned on or off, was valued by all participants. Seven participants found the integrated shutter increased perceived safety in contrast to current smart glasses; only few participants stated that they made no difference between the shutter and other visual feedback mechanisms, e.g. LEDs (n = 2). Several participants noted that the physical coverage increased trustworthiness because it made the system more robust against hackers (*concerns:hacking*, n = 3) than LEDs. Concluding, the usage of physical occlusion could increase perceived safety and, thus, could be considered an option for future designs. Participants even noted that the usage of the shutter as reassuring as pasting up a laptop camera (*laptop comparison*, n = 4), which is common practice.

6.6.2 User Views on Trustworthiness

In contrast, participants also expressed technology scepticism, particularly that the system might secretly record audio (*concerns:audio*, n = 5) or malfunction (*concerns:malfunction*, n = 4). With the increasing power of deep neural networks malfunctions, system failures, or inaccuracies will be addressable in the future, interaction designers will have to address this fear of "being invisibly audio-recorded". A lack of knowledge about eye tracking on both the user's and the bystander's side might even back this misconception. Therefore, future systems using eye tracking for context recognition will have to clearly communicate their modus operandi.

6.6.3 Perceived Privacy of Eye Tracking

The majority of participants claimed to have no privacy concerns about smart glasses with integrated eye tracking functionality: *"I do see no threat to my privacy or the like from tracking my eye movements; this [the eye tracking] would rather be something which could offer a certain comfort."* (P11) Only two participants expressed concerns about their privacy, e.g., due to fearing eye-based emotion recognition (P3). One was uncodeable. This underlines our assumption that eye tracking promises privacy-preserving and socially acceptable sensing in head-mounted augmented reality devices and, thus, should be further explored.

6.6.4 Desired Level of Control

Participants were encouraged to elaborate on whether the recording status should be user-controlled or system-controlled. P10 notes: *"I'd prefer if it was automatic, because if it is not automatic, then the wearer can forget to do that [de-activating the camera]. Or maybe he will say 'Oh, I do not want to do that' and then [...] that leads to a conflict. So better is automatic, to avoid questions."* Four other participants also preferred the camera to be solely controlled by the system (*control:automatic*, n = 4). Their preference is motivated by user forgetfulness (n = 5), and potential non-compliance of users (in the bystander use case, n = 1). Only two participants expressed a preference for sole (*control:manual*) control, due to an expected lack of system reliability, and technical feasibility. Two responses were uncodable. All other participants requested to implement manual confirmation of camera de-activation/re-activation or manual operation as alternative modes (*control:mixed*, n = 4), i.e., they like to feel in control. To meet these user expectations, future interaction designs would have to find an adequate mix of user control and automatic support through the system; for example, by enabling users to explicitly record sensitive information (e.g. in cases of emergency) or label seemingly non-sensitive situations "confidential".

6.7 Discussion

We discuss *PrivacEye* in light of the aforementioned design and user requirements and results of the technical evaluation.

6.7.1 Privacy-Preserving Device Behaviour

Design Requirements 1 and *2* demand privacy-preserving device behaviour. With *PrivacEye*, we have presented a computer vision routine that analyses all imagery obtained from the scene camera, combined with eye movement features with regard to privacy sensitivity and, in case a situation requires protection, the ability to de-activate the scene camera and close the system's camera shutter. This approach prevents both accidental misclosure and malicious procurance (e.g. hacking) of sensitive data, as has been positively highlighted by our interview participants. However, closing the shutter comes at the cost of having the scene camera unavailable for sensing after it has been de-activated. *PrivacEye* solves this problem by using a second eye camera that allows us, in contrast to prior work, to locate all required sensing hardware on the user's side. With *PrivacEye* we have provided proof-of-concept that context-dependent re-activation of a first-person scene camera is feasible using only eye movement data. Future work will be able to build upon these findings and further explore eye tracking as a sensor for privacy-enhancing technologies. Furthermore, our results provide first prove that there is indeed a transitive relationship over privacy sensitivity and a user's eye movements.

6.7.2 Defining Privacy Sensitivity

Prior work indicates that the presence of a camera may be perceived appropriate or inappropriate depending on social context, location, or activity (Hoyle *et al.*, 2014, 2015; Price *et al.*, 2017). However, related work does, to the best of our knowledge, not provide any insights on eye tracking data in this context. For this reason, we run a dedicated data collection and ground truth annotation. Designing a practicable data collection experiment requires the overall time spent by a participant for data recording and annotation to be reduced to a reasonable amount. Hence, we made use of an already collected dataset, and re-invited the participants only for the annotation task. While the pre-existing dataset provided a rich diversity of privacy-sensitive locations and objects, including smart phone interaction, and realistically depicts everyday student life, it is most likely not applicable to other contexts, e.g., industrial work or medical scenarios.

For *PrivacEye*, we rely on a 17-participant-large, ground truth annotated dataset with highly realistic training data. Thus, the collected training data cannot be fully generalised, e.g., to other regions or age groups. On the plus side, however, this data already demonstrates that in a future real-world application, sensitivity ratings may vary largely between otherwise similar participants. This might also be affected by their (supposedly) highly individual definition of "privacy". Consequently, a future consumer system should be pre-trained and then adapted online, based on personalised retraining after user feedback. In addition, users should be enabled to select their individual “cut-off”, i.e., the level from which a recording is blocked, which was set to “2” for *PrivacEye*. Future users of consumer devices might choose more rigorous or relaxed “cut-off” levels depending on their personal preference. Initial user feedback also indicated that an interaction design that combines automatic, software-controlled de- and re-activation, with conscious control of the camera by the user, could be beneficial.

6.7.3 Eye Tracking for Privacy-Enhancement

Eye tracking is advantageous for bystander privacy given that it only senses users and their eye movements. In contrast to, e.g., microphones or infrared sensing, it senses a bystander and/or an environment only indirectly via the user’s eye motion or reflections. Furthermore, eye tracking allows for implicit interaction and is non-invasive, and we expect it to become integrated into commercially available smart glasses in the near future. On the other hand, as noted by Liebling and Preibusch (Liebling and Preibusch, 2014; Preibusch, 2014), eye tracking data is a scare resource, which can be used to identify user attributes like age, gender, health, or user’s current task. For this reason, the collection and use of eye tracking data could be perceived as a potential threat to user privacy. However, our interviews showed that eye tracking was not perceived as problematic by a large majority of our participants. Nevertheless, eye tracking data must be protected by appropriate privacy policies and data hygiene.

To use our proposed hardware prototype in a real-world scenario, data sampling and analysis need to run on a mobile phone. The CNN feature extraction is currently the biggest computational bottleneck, but could be implemented in hardware to allow for real-time operation (cf. Qualcom’s Snapdragon 845). Further, we believe that a consumer system should provide an accuracy >90% which could be achieved using additional sensors such as GPS or inertial tracking. However, presenting the first approach for automatic de- and re-activation of a first-person camera that achieves ∼73% with competitive performance to *ScreenAvoider* (54.2 - 77.7%) (Korayem *et al.*, 2014) and *iPrivacy* (∼75%) (Yu *et al.*, 2017), which are restricted to scene content protection and post-hoc privacy protection, we provide a solid basis for follow up work. We note that a generalised person-independent model for privacy sensitivity protection is desirable. For the work in this chapter only the participants themselves labelled their own data. Aggregated labels of multiple annotators would result in a more consistent and generalisable “consensus” model and improve test accuracy, but would dilute the measure of perceived privacy sensitivity, which is highly subjective (Price *et al.*, 2017). Specifically, similar activities and environments were judged differently by the individual participants, as seen in Figure 6.3. The availability of this information is a core contribution of our dataset.

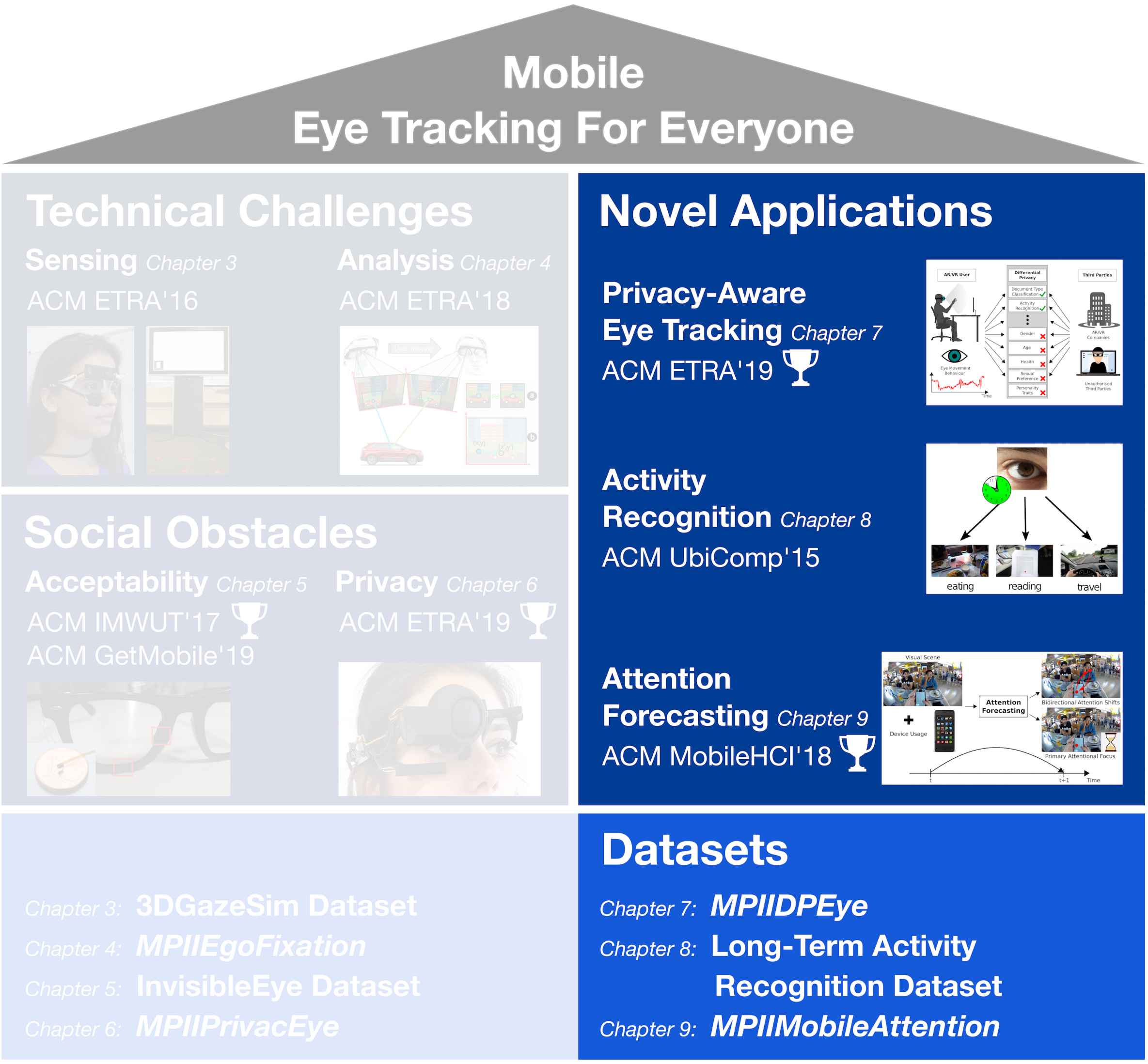
6.7.4 Communicating Privacy Protection

The interaction design of *PrivacEye* tackles *Design Requirement 3* using a non-transparent shutter. Ens et al. (Ens *et al.*, 2015) reported that the majority of their participants expected to feel more comfortable around a wearable camera device if it clearly indicated to be turned on or off. Hence, our proposed interaction design aims to improve a bystander's awareness of the recording status by employing an *eye metaphor*. Our prototype implements the "eye lid" as a retractable shutter made from non-transparent material: open when the camera is active, closed when the camera is inactive. Thus, the metaphor mimics "being watched" by the camera. The "eye lid" shutter ensures that bystanders can comprehend the recording status without prior knowledge, as eye metaphors have been widely employed for interaction design, e.g., to distinguish visibility or information disclosure (Pousman *et al.*, 2004; Schlegel *et al.*, 2011; Motti and Caine, 2016) or to signal user attention (Chan and Minamizawa, 2017). Furthermore, in contrast to visual status indicators, such as point lights (LEDs), physical occlusion is non-spoofable (cf. (Denning *et al.*, 2014; Portnoff *et al.*, 2015)). This concept has been highly appreciated during our interviews, which is why we would recommend adopting it for future hardware designs.

6.8 Conclusion

In this chapter, we have proposed *PrivacEye*, a method that combines first-person computer vision with eye movement analysis to enable context-specific, privacy-preserving de-activation and re-activation of a head-mounted eye tracker's scene camera. We have evaluated our method quantitatively on a 17-participant dataset of fully annotated everyday behaviour as well as qualitatively, by collecting subjective user feedback from 12 potential future users. To the best of our knowledge, our method is the first of its kind and prevents potentially sensitive imagery from being recorded at all, without the need for active user input. As such, we believe the method opens up a new and promising direction for future work in head-mounted eye tracking, the importance of which will only increase with further miniaturisation and integration of eye tracking in head-worn devices or even in normal glasses frames.

Part III


Mobile
Eye Tracking For Everyone
Technical Challenges
Sensing Chapter 3
ACM ETRA'16
Analysis Chapter 4
ACM ETRA'18
Social Obstacles
Acceptability Chapter 5
ACM IMWUT'17
ACM GetMobile'19
Privacy Chapter 6
ACM ETRA'19
Novel Applications
Privacy-Aware
Eye Tracking Chapter 7
ACM ETRA'19
Activity
Recognition Chapter 8
ACM UbiComp'15
eating
reading
travel
Attention
Forecasting Chapter 9
ACM MobileHCI'18
Visual Scene
Attention Forecasting
Device Usage
Bidirectional Attention Shifts
Primary Attentional Focus
Time
Chapter 3: 3DGazeSim Dataset
Chapter 4: MPIIEgoFixation
Chapter 5: InvisibleEye Dataset
Chapter 6: MPIIPrivacEye
Datasets
Chapter 7: MPIIDPEye
Chapter 8: Long-Term Activity Recognition Dataset
Chapter 9: MPIIMobileAttention

7

Privacy-Aware Eye Tracking Using Differential Privacy

With eye tracking being increasingly integrated into virtual and augmented reality (VR/AR) head-mounted displays, preserving users' privacy is an ever more important, yet under-explored, topic in the eye tracking community. We report a large-scale online survey (N=124) on privacy aspects of eye tracking that provides the first comprehensive account of with whom, for which services, and to what extent users are willing to share their gaze data. Using these insights, we design a privacy-aware VR interface that uses differential privacy, which we evaluate on a new 20-participant dataset for two privacy sensitive tasks: We show that our method can prevent user re-identification and protect gender information while maintaining high performance for gaze-based document type classification. Our results highlight the privacy challenges particular to gaze data and demonstrate that differential privacy is a potential means to address them. Thus, the work of this chapter lays important foundations for future research on privacy-aware gaze interfaces.

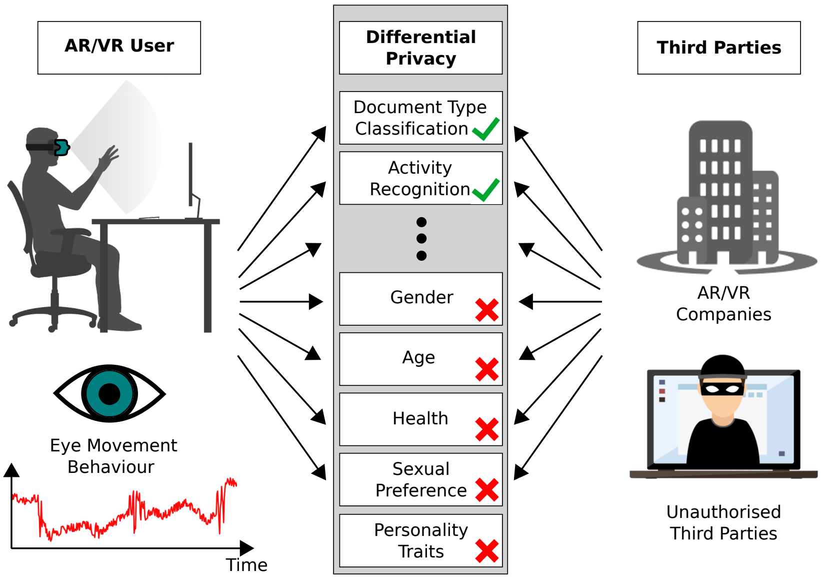


Figure 7.1: Using differential privacy prevents third parties, like companies or hackers, from deriving private attributes from a user's eye movement behaviour while maintaining the data utility for non-private information.

7.1 Introduction

With eye tracking becoming pervasive (Bulling and Gellersen, 2010; Tonsen *et al.*, 2017), preserving users' privacy has emerged as an important topic in the eye tracking, eye movement analysis, and gaze interaction research communities. Privacy is particularly important in this context given the rich information content available in human eye movements (Bulling *et al.*, 2011a), on one hand, and the rapidly increasing capabilities of interactive systems to sense, analyse, and exploit this information in everyday life (Hansen *et al.*, 2003; Vertegaal *et al.*, 2003; Stellmach and Dachselt, 2012) on the other. The eyes are more privacy-sensitive than other input modalities: They are typically not consciously controlled; they can reveal unique private information, such as personal preferences, goals, or intentions. Moreover, eye movements are difficult to remember, let alone reconstruct in detail, in retrospect, and hence do not easily allow users to "learn from their mistakes", i.e. to reflect on their past and change their future privacy-related behaviour.

These unique properties and rapid technological advances call for new research on next-generation eye tracking systems that are *privacy-aware*, i.e. that preserve users' privacy in all interactions they perform with other humans or computing systems in everyday life. However, *privacy-aware eye tracking* remains under-investigated as of yet (Liebling and Preibusch, 2014).

The lack of research on privacy-aware eye tracking results in two major limitations: First, there is a lack of even basic understanding of users' privacy concerns with eye tracking in general and eye movement analysis in particular. Second, there is a lack of eye tracking methods to preserve users' privacy, corresponding systems, and user interfaces that implement (and hence permit the evaluation of) these methods with end users. This chapter aims to address both limitations and, as such, make the first crucial step towards a new generation of eye tracking systems that respect and actively protect private information that can be inferred from the eyes.

The work of this chapter first contributes a large-scale online survey on privacy aspects of eye tracking and eye movement analysis. The survey provides the first comprehensive account of with whom, for which services, and to what extent users are willing to share their eye movement data. The survey data is available at `https://www.mpi-inf.mpg.de/MPIIDPEye/` (date: 12.07.2019). Informed by the survey, *we further contribute the first method to protect users' privacy in eye tracking based on differential privacy (DP)*, a well-studied framework in the privacy research community. In a nutshell, DP adds noise to the data so as to minimise chances to infer privacy-sensitive information or to (re-)identify a user while, at the same time, still allow use of the data for desired applications (the so-called utility task), such as activity recognition or document type classification (see Figure 8.1). We illustrate the use of differential privacy for a sample virtual reality (VR) gaze interface. We opted for a VR interface given that eye tracking will be readily integrated into upcoming VR head-mounted displays, and hence, given the significant and imminent threat potential (Adams *et al.*, 2018): Eye movement data may soon be collected at scale on these devices, recorded in the background without the user noticing, or even transferred to hardware manufacturers.

7.2 Related Work

We discuss previous works on 1) information available in eye movements, 2) eye movements as a biometric, and 3) differential privacy.

7.2.1 Information Available in Eye Movements

A large body of work across different research fields has demonstrated the rich information content available in human eye movements. Pupil size is related to a person's interest in a scene (Hess and Polt, 1960) and can be used to measure cognitive load (Matthews *et al.*, 1991). Other works have shown that eye movements are closely linked to mental disorders, such as Alzheimer's (Hutton *et al.*, 1984), Parkinson's (Kuechenmeister *et al.*, 1977), or schizophrenia (Holzman *et al.*, 1974). More recent work in HCI has demonstrated the use of eye movement analysis for human activity recognition (Bulling *et al.*, 2013; Steil and Bulling, 2015) as well as to infer a user's cognitive state (Bulling and Zander, 2014; Faber *et al.*, 2017) or personality traits (Hoppe *et al.*, 2018). More closely related to the work of this chapter, several researchers have shown that gender and age can be inferred from eye movements, e.g. by analysing the spatial distribution of gaze on images like faces (Cantoni *et al.*, 2015; Sammaknejad *et al.*, 2017).

All of these works underline the significant potential of eye movement analysis for a range of future applications, some of which may soon become a reality, for example, with the advent of eye tracking-equipped virtual and augmented reality head-mounted displays. Despite the benefits of these future applications, the wide availability of eye tracking will also pose significant privacy risks that remain under-explored in the eye tracking community.

7.2.2 Eye Movements as a Biometric

Eye movement biometrics has emerged as a promising approach to user authentication (Kasprowski and Ober, 2003). While first works required a point stimulus that users were instructed to follow with their eyes (Kasprowski, 2004; Kasprowski and Ober, 2005), later ones explored static points (Bednarik *et al.*, 2005) or images (Maeder and Fookes, 2003). Kinnunen et al. presented the first method for "task-independent" person authentication using eye movements (Kinnunen *et al.*, 2010). Komogortsev et al. proposed the first attempt to model eye movements for authentication using an Oculomotor Plant Mathematical Model (Komogortsev *et al.*, 2010; Komogortsev and Holland, 2013). Eberz et al. presented a biometric based on eye movement patterns. They used 20 features that allowed them to reliably distinguish and authenticate users across a variety of real-world tasks, including reading, writing, web browsing, and watching videos on a desktop screen (Eberz *et al.*, 2016). Zhang et al. used eye movements to continuously authenticate the wearer of a VR headset by showing different visual stimuli (Zhang *et al.*, 2018b).

While an ever-growing body of research explores eye movements as a promising modality for privacy applications and user authentication, we are the first to practically explore eye movements recorded using eye tracking as a potential threat to users' privacy.

7.2.3 Differential Privacy

Differential privacy has been studied in privacy research for more than a decade in terms of its theoretical foundations and its practical applications to different data types, such as location (Pyrgelis *et al.*, 2017), biomedical data (Saleheen *et al.*, 2016), or continuous time series data (Fan and Xiong, 2012). We refer the reader to (Zhu *et al.*, 2017) for a survey. A key challenge in differential privacy is to find the right trade-off between privacy and utility, that is, the right amount of random noise to "hide" an individual without hampering data utility. Fredrikson et al. demonstrated how important it is to balance privacy and utility (Fredrikson *et al.*, 2014). They observed that either privacy was not preserved or that utility suffered, leading to increased health risks for the patients from unsuitable drug dosage. A good privacy-utility trade-off is possible if privacy mechanisms are tailored towards a specific use case (Fan and Xiong, 2012; Pyrgelis *et al.*, 2017). While differential privacy has a long history in privacy research, to the best of our knowledge, we are the first to apply this framework to eye tracking data.

7.3 Privacy Concerns in Eye Tracking

We conducted a large-scale online survey to shed light on users' privacy concerns related to eye tracking technology and the information that can be inferred from eye movement data. We advertised our survey on social platforms (Facebook, WeChat) and local mailing lists for study announcements. The survey opened with general questions about eye tracking and VR technologies; continued with questions about future use and applications, data sharing and privacy (especially regarding with whom users are willing to share their data); and concluded with questions about the participants' willingness to share different eye movement representations. Participants answered each question on a 7-point Likert scale (1: Strongly disagree to 7: Strongly agree). To simplify the analysis, we merged scores 1 to 3 to "Disagree" and 5 to 7 to "Agree".

The survey took about 20 minutes to complete, was set up as a Google Form, and was split into the parts described above. Our design ensured that participants without pre-knowledge of eye tracking and VR technology could participate as well: We provided a slide show containing information about eye tracking in general, and in VR devices specifically, and introduced the different forms of data representation, showing example images or explanatory texts.

In our survey, 124 people (81 male, 39 female, 4 did not tick the gender box) participated, aged 21 to 66 (mean = 28.07, std = 5.89). The participants were from all over the world, coming from 29 different countries (Germany: 39%, India: 12%, Pakistan: 6%, Italy: 6%, China: 5%, USA: 3%). Sixty-seven percent of them had a graduate university degree (master's or PhD), and 22% had an undergraduate university degree (bachelor's). Fifty-one percent were students of a variety of subjects (law, language science, computer science, psychology, etc.); 34% were scientists and researchers, IT professionals (7%), or had business administration jobs (2%). Since the topic of the survey was in the title of posts and emails, most likely people inherently interested in the topic participated. The majority were young, educated people with a technical background the exact group of people most likely to experience AR/VR technology (73%) in contrast to, for example, older generations.

Given the breadth of results, we highlight key insights most relevant for this chapter. We found nearly all answers for the provided questions to be significantly different from an equal distribution tested with Pearson's chi-squared test ($p < 0.001$, dof = 6). Additionally, we calculated the skewness and observed that the majority of questions show a significant difference to the corresponding normal distribution ($p < 0.1$). Detailed numbers, plots, significance and skewness test results can be found in Section C.2 and Section C.3 (see `https://www.mpi-inf.mpg.de/MPIIDPEye/` (date: 12.07.2019)).

Services and Attributes. In the first part of our survey, we asked participants for which services they would share their eye tracking data and presented both currently available and potential future services as answer options. As we can see from Figure 7.2, more than 80% of all participants agreed to share their eye tracking data for (early) detection of diseases like Alzheimer's or Parkinson's. Likewise, the majority agreed to share their data for hands-free VR and user interface interaction. Similar results can be observed for learning and reading skill detection as well as for stress level monitoring. However, for improved gaze target recognition, website content, and activity recognition, we observe two peaks. A clear majority is unwilling to share data with shopping assistance and interest detection services.

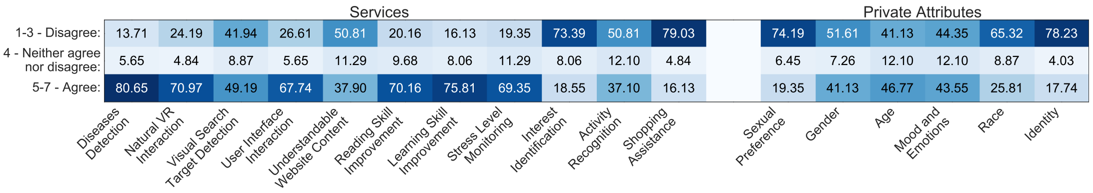


Figure 7.2: Survey results (Services and Attributes): With which services would you agree to share your eye tracking data (Services)?; Would you agree to private attributes being inferred by these services (Private Attributes)?

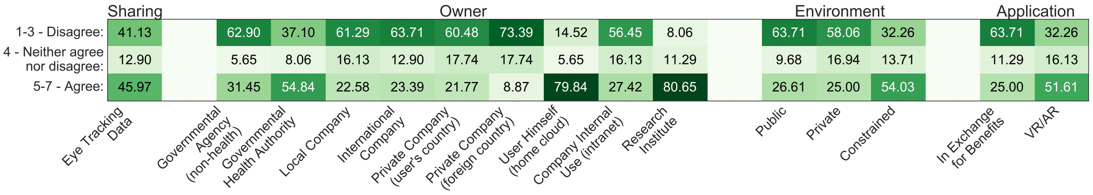


Figure 7.3: Survey results (Whom and Where): Would you agree to share your eye tracking data in general (Sharing); with whom (Owner); where (Environment); in exchange for benefits or for VR/AR usage (Application)?

Our next set of questions indicated the fact that services could be able to infer private attributes from their data, and we asked whether participants would still want to share their eye tracking data. We clearly observed that if the attributes of sexual preference, gender, race, and identity can be inferred, a majority do not want to share their data. It was only for age and emotion detection that we identified two different interest groups that either agree with or object to sharing their data.

Whom and Where. In the second part, of our survey we asked participants whether they would share eye tracking data in general, and with whom. Moreover, we were interested in whether the environment has an influence on their sharing behaviour (see Figure 7.3). Finally, we wanted to know whether the sharing behaviour is different if participants get benefits (not specified) in exchange for their data or if the data is collected during VR/AR usage in general.

The answers as to whether participants would share their eye tracking data in general do not show a clear tendency; the participants' opinions are split in two groups ($\chi^2(\text{dof} = 6) = 32.25$, $p = 1.46 \times 10^{-6}$). Next, we asked more specifically whether participants would share their data if it were later owned and operated by one of the given "owner" options in Figure 7.3. According to their answers, participants would only share their data if the co-owner is a governmental health-agency; they do not trust local and international companies, or company internal use. However, participants would also share their data for research purposes, which is not surprising given that 67% of participants have a graduate university degree and trust in research institutes. Participants would not agree to share their data in public, nor in private environments, but they would agree to constrained environments. Furthermore, the participants object to sharing their data for any kind of benefit, but would agree when their eye tracking data was collected in VR/AR ($\chi^2(\text{dof} = 6) = 26.72$, $p = 0.00016$).

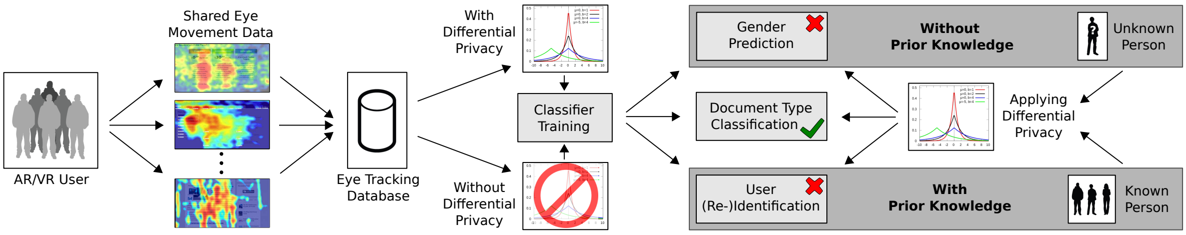


Figure 7.4: (Left) Our method assumes that AR/VR users share their eye tracking data and privacy-sensitive information with a third party, which is able to train classifiers with or without differentially private data to infer private attributes of an unknown (without prior knowledge) or a known (with prior knowledge) person; (Right) Applying differential privacy to test data prevents private information inference (gender, user (re-)identification) but maintains data utility (document type classification).

Data Representation. In the final part of the survey, we asked participants in what form they would agree to share their data. We discriminate 12 different representations, ranging from raw eye tracking, to heatmaps, to aggregated features (see Figure C.1 in Appendix C). Additionally, we were interested in whether their sharing behaviour changes if the data is first anonymised. Information which provides gaze information, like fixations, or scan path information on a surface would mostly not be shared. Participants largely agree to share their eye tracking data as statistical features, and especially aggregated features. This is why we focus in our study on the aggregated feature representation to apply differential privacy. Our survey shows a clear increase in participants willing to share their data in anonymised form.

7.4 Privacy-Preserving Eye Tracking

The findings from our survey underline the urgent need to develop *privacy-aware eye tracking systems* – systems that provide a formal guarantee to protect the privacy of their users. Additionally, it is important not to forget that eye movement data typically also serves a desired task – a so-called *utility*. For example, eye movement data may be used in a reading assistant to detect the documents a user is reading (Kunze *et al.*, 2013c) or to automatically estimate how many words a user reads per day (Kunze *et al.*, 2013b, 2015). Therefore, it is important to ensure that any privacy-preserving method does not render the utility dysfunctional, i.e. that the performance on the utility task will not drop too far. The key challenge can thus be described as *ensuring privacy without impeding utility.*

We assume in the following that multiple users share their eye tracking data in the form of aggregated features. The resulting eye tracking database is visualised in the left part of Figure 7.4. This database can be downloaded both for legitimate use cases as well as for infringing on users' privacy, for example, to train classifiers for various tasks. Therefore, our proposed privacy mechanism is applied prior to the release by a trusted curator.

7.4.1 Threat Models

We have identified two attack vectors on users' privacy in the context of eye tracking that we formalise in two threat models. They differ in their assumption about the attackers' prior knowledge about their target (see the right part of Figure 7.4).

Without Prior Knowledge. In the first threat model, we assume that an attacker has no prior knowledge about the target and wants to infer a private attribute; we focus on gender in our example study. The attacker can only rely on a training dataset from multiple participants different from the target. This data can be gathered by companies or game developers we share our data with in exchange for a specific service. Some users might opt in to share their data with a third party to receive personalised advertisements, or they might create a user account to remove advertisements. These companies with eye tracking data can misuse the data, forward it to third parties or get hacked by external attackers. Another source for attackers to get eye tracking datasets is publicly available datasets generated for research purposes. Concretely, VR glasses are offered in gaming centres and used by multiple visitors, which we refer to as the one-device-multiple-users scenario. An attacker with access to the eye tracking data might be interested in inferring the gender of the current user to show gender-specific advertisements.

With Prior Knowledge. The second threat model assumes that the attacker has already gathered prior knowledge about the target. Observing further eye tracking data, the attacker wants to re-identify the target to inspect the target's habits. Concretely, the target might be using different user accounts or even different devices for work and leisure time (a one-user-multiple-devices scenario). We assume the attacker is able to link the target's work data to the target's identity and now wants to identify the target's data from his/her leisure activities. Again, the attacker could be a VR/AR company exploiting their data to check whether a device is only used by one person, or re-identify a user automatically to adapt device settings. Moreover, data could be released intentionally to a third party for money or unintentionally through a hack.

7.4.2 Differential Privacy for Eye Tracking

We propose to mitigate the privacy threats emerging from our two threat models using *differential privacy*, a well-known framework from privacy research (Dwork *et al.*, 2014). Differential privacy guarantees that the answer of the privacy-preserving mechanism does not depend on whether a single user contributed his/her data or not; hence, there is no way to infer further information about this user. Concretely, the answer to the question "What is the average fixation rate when reading a text?" should be almost the same, whether or not a specific user, say, Alice, has contributed her data to our database of fixation rates. We denote a differentially private mechanism by $\mathcal{M}$ and refer to Alice's data as a single data element in the database D. Typically, $\mathcal{M}$ adds random noise to "hide" each data element, which we will formalise in the following.

Definition 1 (ϵ-Differential Privacy (Dwork *et al.*, 2006)). *A mechanism $\mathcal{M}$ provides ϵ-differential privacy if for all databases D, D' that differ in at most one element and for every $S \subseteq Range(\mathcal{M})$, we have*

$$Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \cdot Pr[\mathcal{M}(D') \in S]. \tag{7.1}$$

Differential privacy allows computing an arbitrary function g over the database, i.e. $g : \mathcal{R}^* \mapsto \mathcal{R}^d$, where d denotes the dimensionality of the output of g. For our running example, g would compute the average and output one number, hence $d = 1$. Similarly, we could define g to average over 30-second windows of fixation data and then output a vector of length d.

How much noise we have to add depends on the variance of the data between two arbitrary elements. Formally:

Definition 2 (L_1 Sensitivity (Dwork *et al.*, 2006)). *For all functions $g : \mathcal{R}^* \mapsto \mathcal{R}^d$, the L_1 sensitivity is the smallest number Δ_g s.th. for all databases D, D' differing in one element, we have*

$$||g(D) - g(D')||_{L_1} \leq \Delta_g. \tag{7.2}$$

Intuitively, the sensitivity captures the maximal influence Alice's data could have on the answer to our query. In the worst case, for her privacy, Alice's data is an outlier, e.g. Alice is a very slow reader compared to all other participants. Even in this case, the difference between Alice's data and any other entry in the database must be smaller than or equal to the sensitivity. The noise to "hide" Alice's contribution is scaled to this worst case, ensuring Alice's privacy.

Next, we formalise the exponential mechanism that is one way to generate differentially private data:

Definition 3 (Exponential Mechanism (Dwork *et al.*, 2014)). *The exponential mechanism selects and outputs an element $r \in \mathcal{R}$ in the range of permissible output elements with probability equal to (written: $r \sim$)*

$$r \sim exp(\frac{\epsilon \cdot u(x,r)}{2\Delta_u}) \tag{7.3}$$

where u is a utility function judging the quality of r with respect to the original data element x.

In order to apply the exponential mechanism to our example database of fixation durations, we would first need to define a utility function u and the set of permissible outputs. Valid answers to the query "What are the average fixation rates when reading a text, sampled at 30 second windows?" are vectors of length d containing real-numbered entries; thus, $\mathcal{R} = \mathbb{R}^d_{\geq 0}$. The utility function u is a measure of quality for the output r with respect to the original data entry x. The exponential mechanism ensures that high-quality outputs r are generated exponentially more often than low-quality r.

Finally, we state one theorem that allows combining several differentially private mechanisms into one.

Theorem 1 (Composition Theorem (Dwork *et al.*, 2006)). *Let $\mathcal{M}_1, ..., \mathcal{M}_k$ be a fixed sequence of mechanisms, where each mechanism $\mathcal{M}_i$ is ϵ_i-differentially private. Then, their joint output $\mathcal{M}(D) = (\mathcal{M}_1(D), ..., \mathcal{M}_k(D))$ is ϵ-differentially private for $\epsilon = \sum_{i=1}^{k} \epsilon_i$.*

7.4.3 Implementing Differential Privacy

Our dataset contains data from n participants, which we refer to as $p_1, ..., p_n$. For each participant, we measure m features, $f_1, ..., f_m$ at different points in time. In summary, p_{1,f_7,t_5} denotes the value of the 7th feature at time point 5 of participant 1, and the vector $(p_{1,f_7,t_0}, ..., p_{1,f_7,t_{max,1}})$ contains all measurements of feature 7 for participant 1. Notice that the data entries available may have different lengths, i.e. $t_{max,1}$, the last time point of participant 1, may be different from another participant's last time point, e.g. $t_{max,2}$.

The sensitivity for our mechanism then depends on the range of the features, which is different across our m features. For example, feature f_{15} is the fixation duration in our dataset, and it has an estimated range of [0.11, 2.75] seconds, while f_{22}, which describes the pupil diameter size, has an estimated range of [21.9, 133.9] pixels. Therefore, we derive one privacy mechanism $\mathcal{M}_{f_i}$ for each feature separately and use the composition theorem (Theorem 1) to combine the m mechanisms into our final mechanism. The exponential mechanism requires a utility function u. We choose the L_1 distance for simplicity of the derivation:

$$u(p_{f_i}, r) = \sum_{j=1}^{t_{max,p}} |p_{f_i,j} - r_j| \tag{7.4}$$

According to Definition 2, the sensitivity Δ_{u,f_i} is

$$\Delta_{u,f_i} = \underset{p_{f_i},q_{f_i}}{max} ||(p_{f_i,t_0}, ..., p_{f_i,t_{max,p}}) - (q_{f_i,t_0}, ..., q_{f_i,t_{max,q}})||_{L_1} \tag{7.5}$$

i.e. the maximal difference between the data vectors of two arbitrary participants p and q for the i-th feature. Next, we unify the length by padding the data vector with the shorter length. Let $tmax$ be the maximal length: $tmax = max(t_{max,p}, t_{max,q})$. Using this and the definition of the L_1 norm:

$$\Delta_{u,f_i} \leq \underset{p_{f_i},q_{f_i}}{max} \sum_{j=1}^{tmax} |p_{f_i,t_j} - q_{f_i,t_j}| = tmax \cdot \delta_i \tag{7.6}$$

In the last step, we used the fact that we can derive the range δ_i of feature f_i, either estimated from the data or by theoretic constraints.

We rely on the exponential mechanism (see Definition 3) to obtain a vector r that is differentially private for each participant p and feature f_i:

$$r \sim exp(\frac{\epsilon_i u(p_{f_i}, r)}{2\Delta_{u,f_i}}) \overset{\text{Eq. 7.4}}{=} exp(\frac{\epsilon_i \sum_{j=1}^{t_{max,p}} |p_{f_i,j} - r_j|}{2 \cdot tmax \cdot \delta_i}) \tag{7.7}$$

To increase readability, we define $\lambda_i = \frac{\epsilon_i}{2 \cdot tmax \cdot \delta_i}$, which is constant once i and ϵ_i are fixed. We generate such a vector r from the exponential distribution by first sampling a random scalar y from the exponential distribution with location 0 and scale parameter $\frac{1}{\lambda_i}$. We derive our differentially private vector r from y as follows:

$$y = exp(\lambda_i \cdot \sum_{j=1}^{t_{max,p}} |p_{f_i,j} - r_j|) \Leftrightarrow \frac{log_e(y)}{\lambda_i} = \sum_{j=1}^{t_{max,p}} |p_{f_i,j} - r_j| \tag{7.8}$$

Selecting $r_j = \pm\frac{log_e(y)}{\lambda_i \times tmax} + p_{f_i,j}$ fulfils the above constraint with randomly sampled sign.

The privacy guarantee of the combined mechanism $\mathcal{M}$ is, by the composition theorem (Theorem 1), $\sum_{i=1}^{m} \epsilon_i$.

Subsampling. In order to achieve a higher privacy guarantee, we propose to subsample the data. Given a window size w, we draw one sample from $(p_{k,i,n \cdot w}, ..., p_{k,i,(n+1) \cdot w})$ for each participant k and feature i independently where $n \in \mathbb{N}$, such that the sampling windows are non-overlapping. Notice that this subsampling approach and the corresponding window size are independent of the feature generation process. This method decreases the sensitivity further by a factor of w: $\Delta_{u,f_i,w} \leq \frac{tmax}{w} \cdot \delta_i$.

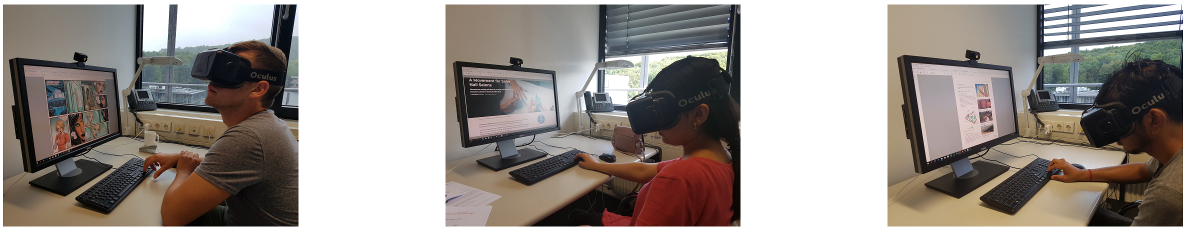
7.5 Data Collection

Given the lack of a suitable dataset for evaluating privacy-preserving eye tracking using differential privacy, we recorded our own dataset. As a utility task, we opted to detect different document types the users read, similar to a reading assistant (Kunze *et al.*, 2013c). Instead of printed documents, participants read in VR, wearing a corresponding headset. The recording of a single participant consists of three separate recording sessions, in which a participant reads one out of three different documents: a comic, online newspaper, or textbook (see Figure 7.5). All documents include a varying proportion of text and images. Each of these documents was about a 10-minute read, depending on a user's reading skill (about 30 minutes in total).

Participants. We recruited 20 participants (10 male, 10 female) aged 21 to 45 years through university mailing lists and adverts in different university buildings on campus. Most participants were BSc and MSc students from a large range of subjects (e.g. language science, psychology, business administration, computer science) and different countries (e.g. India, Pakistan, Germany, Italy). All participants had little or no experience, with eye tracking studies and had normal or corrected-to-normal vision (contact lenses).

Apparatus. The recording system consisted of a desktop computer running Windows 10, a 24" computer screen, and an Oculus DK2 virtual reality headset connected to the computer via USB. We fitted the headset with a Pupil eye tracking add-on (Kassner *et al.*, 2014) that provides state-of-the-art eye tracking capabilities. To have more flexibility in the applications used by the participants in the study, we opted for the Oculus "Virtual Desktop" that shows arbitrary application windows in the virtual environment. To record a user's eye movement data, we used the capture software provided by Pupil. We recorded a separate video from each eye and each document. Participants used the mouse to start and stop the document interaction and were free to read the documents in arbitrary order. We encouraged participants to read at their usual speed and did not tell them what exactly we were measuring.

Recording Procedure. After arriving at the lab, participants were given time to familiarise themselves with the VR system. We showed each participant how to behave



(a) Comic (b) Newspaper (c) Textbook

Figure 7.5: Each participant read three different documents: (a) comic, (b) online newspaper, and (c) textbook.

in the VR environment, given that most of them had never worn a VR headset before. We did not calibrate the eye tracker but only analysed users' eye movements from the eye videos post-hoc. This was so as not to make participants feel observed, and to be able to record natural eye movement behaviour. Before starting the actual recording, we asked participants to sign a consent form. Participants then started to interact with the VR interface, in which they were asked to read three documents floating in front of them (see Figure 7.5). After finishing reading a document, the experimental assistant stopped and saved the recording and asked participants questions on their current level of fatigue, whether they liked and understood the document, and whether they found the document difficult using a 5-point Likert scale (1: Strongly disagree to 5: Strongly agree). Participants were further asked five questions about each document to measure their text understanding. The VR headset was kept on throughout the recording.

After the recording, we asked participants to complete a questionnaire on demographics and any vision impairments. We also assessed their Big Five personality traits (John and Srivastava, 1999) using established questionnaires from psychology. In the work of this chapter we only use the given ground truth information of a user's gender from all collected (private) information, the document type, and IDs we assigned to each participant, respectively.

Eye Movement Feature Extraction. We extracted a total of 52 eye movement features, covering fixations, saccades, blinks, and pupil diameter (see Table C.1 in Appendix C). Similar to (Bulling *et al.*, 2011b), we also computed wordbook features that encode sequences of n saccades. We extracted these features using a sliding window of 30 seconds (step size of 0.5 seconds).

7.6 Evaluation

The overall goal of our evaluations was to study the effectiveness of the proposed differential privacy method and its potential as a building block for privacy-aware eye tracking. In these evaluations, gaze-based document type classification served as the utility task, while gender prediction exemplified an attacker without prior knowledge about the target, and user re-identification an attacker with prior knowledge.

7.6.1 Classifier Training

For each task, we trained a support vector machine (SVM) classifier with radial basis function (RBF) kernel and bias parameter $C = 1$ on the extracted eye movement features. We opted for an SVM due to the good performance demonstrated in a large body of work for eye-based activity recognition (Bulling *et al.*, 2011b; Steil and Bulling, 2015). As the first work of its kind, one goal was to enable readers to compare our results to the state of the art. We standardised the training data (zero mean, unit variance) before training the classifiers; the test data was standardised with the same parameters. Majority voting was used to summarise all classifications from different time points for the respective participant. We randomly sampled training and test sets with an equal distribution of samples for each of the respective classes, i.e. for the three document classes, two gender classes and 20 classes for user identification.

Document Type Classification. We trained a multi-class SVM for document type classification and used leave-one-person-out cross-validation, i.e. we trained on the data of 19 participants and tested on the remaining one – iteratively over all combinations – and averaged the performance results in the end. We envision that in the future, only differentially private data will be available; therefore, we applied our privacy-preserving mechanism to the training and test sets. However, currently there is non-noised data available as well: thus, we set up an additional experiment using clean data for training and noised data for testing.

Gender Prediction. We trained a binary SVM for gender prediction, using reported demographics as ground truth, and applied it again with a person-independent (leave-one-person-out) cross-validation. Since we are in the *without prior knowledge* threat model, we trained on differentially private and non-noised data to model both the future and current situation, as for document type classification.

User (Re-)Identification. We trained a multi-class SVM for user (re-)identification but without a leave-one-person-out evaluation scheme. Instead, we used the first half of the extracted aggregated feature vectors from each document and each participant for training. We tested on the remaining half, since here we are in the *with prior knowledge* threat model. In this scenario, we assumed a powerful attacker that was able to obtain training data from multiple people without noise and was able to map their samples to their identities. The attacker's goal was to re-identify these people when given noised samples without identity labels.

Implementing the Differential Privacy Mechanism. We applied the exponential mechanism for each of our $n = 20$ participants and for each of the $m = 52$ features, using a subsampling window size $w = 10$ to reduce sensitivity. In preliminary evaluations, we observed that subsampling alone had no negative effect on the performance of the SVM. The sensitivity for our differentially private mechanism was generated by data-driven constraints: For each feature i, we estimated δ_i by calculating the global minimum min_i

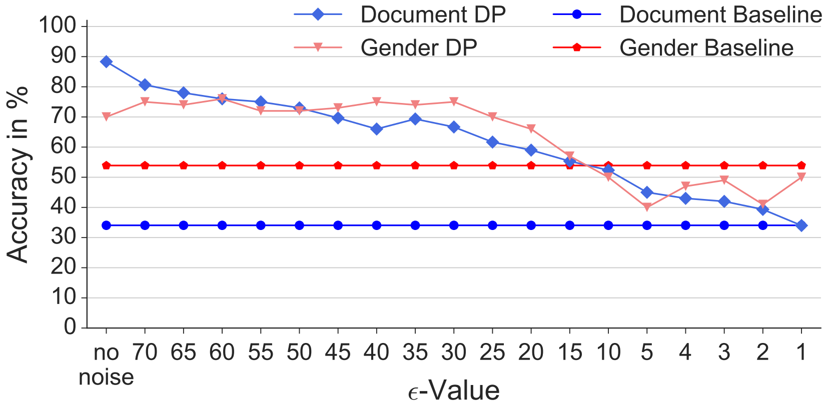


Figure 7.6: Performance for the threat model without prior knowledge trained on differentially private data.

and maximum max_i over all participants and time points and set $\delta_i = max_i - min_i$. This way, the sensitivity ensures privacy protection even of outliers. The noise we added in our study can be understood as reading-task-specific noise. For all f_i, we used the same ϵ_i so that the released data of the whole dataset is $\sum_{i=1}^{52} \epsilon_i$-private.

We repeated our experiments five times each and report averaged results to account for random subsampling and noise generation effects. As a performance metric, we report $Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$, where TP, FP, TN, and FN represent sample-based true positive, false positive, true negative, and false negative counts.

7.6.2 Without Prior Knowledge

In Figure 7.6, we first evaluated the gender prediction task, our example for the attacker *without prior knowledge*, trained on differentially private (noised) data (Gender DP) for decreasing ϵ values. As one might expect, decreasing ϵ, and thereby increasing the noise, negatively influences the testing performance when trained on differentially private data with $\epsilon < 30$. For $\epsilon = 15$, the performance almost drops to the chance level of 54% (random guessing in a slightly imbalanced case due to the leave-one-person-out cross-validation). We conclude that on our dataset, privacy of the participants' gender information is preserved for $\epsilon \leq 15$.

We then evaluated the impact of the noise level for this ϵ-value on utility (see Figure 7.6) using the SVMs trained for document type classification on noised data. As expected, noise negatively influences document type classification as well, but to a lesser extent compared to gender prediction. For privacy preservation, it is sufficient to set $\epsilon = 15$, resulting in an accuracy of about 55% for document type classification, which is still about 22% over chance level.

So far, we have assumed the SVMs were trained on noised data (Document DP). At present, to the best of our knowledge, all available eye movement datasets are not noised. To study this current situation, we trained both the gender prediction SVM and the document type classification SVM without noise and tested at various noise levels. Figure 7.7 shows the results of this evaluation. As can be seen, also in this scenario,

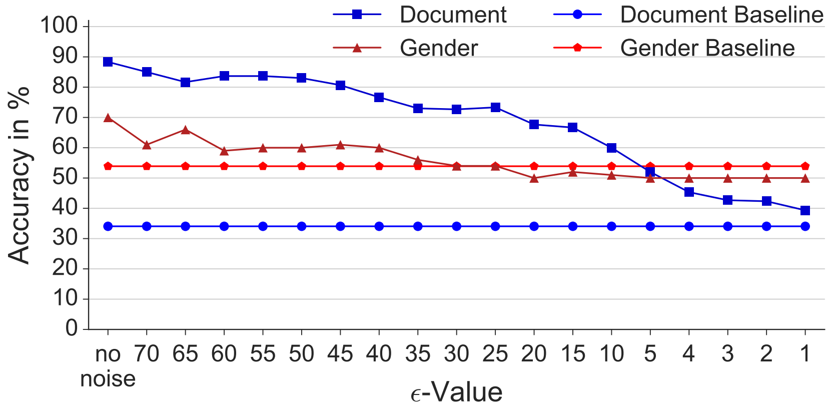


Figure 7.7: Performance for the threat model without prior knowledge trained on clean data.

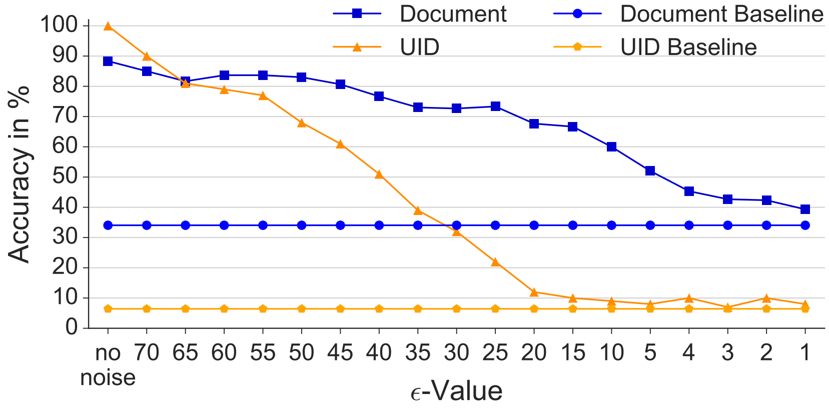


Figure 7.8: Performance for the threat model with prior knowledge trained on clean data.

privacy can be preserved: For $\epsilon = 20$, the accuracy of the gender prediction has dropped below chance level, while document type classification is still around 70%. We observed that even $\epsilon = 30$ would already preserve privacy, since training with noise seems to balance out some negative noise effects. Thus, we conclude that for both current and future situations, privacy preservation is possible while preserving most of the utility.

7.6.3 With Prior Knowledge

Finally, we evaluated in Figure 7.8 the *with prior knowledge* threat model, in which we assumed the attacker trained a SVM on the data of multiple users without noise and wanted to re-identify which person a set of noised samples belongs to. We again added the document type classification performance to be able to judge the effects on utility. As expected, the noise on the test data disturbed the attacker's classification ability: for $\epsilon = 40$, the attacker's accuracy dropped to 50%. For $\epsilon = 15$, it dropped down almost to chance level (6.4%) while the utility preserved an accuracy of about 70%. We conclude that, in this scenario as well, it is possible to preserve a user's privacy with acceptable costs on utility.

7.7 Discussion

7.7.1 Privacy Concerns in Eye Tracking

The ever-increasing availability of eye tracking to end users, e.g. in recent VR/AR headsets, in combination with the rich and sensitive information available in the eyes (e.g. on personality (Hoppe *et al.*, 2018)), creates significant challenges for protecting users' privacy. Our large-scale online survey on privacy implications of pervasive eye tracking, the first of its kind, yielded a number of interesting insights on this important, yet so far largely unexplored, topic (see Section C.3 for the full results). For example, we found that users are willing to share their eye tracking data for medical applications, such as (early) disease detection or stress level monitoring (see Figure 7.2), or for services, if these improve user experience, e.g. in VR or AR (see Figure 7.3). On the other hand, participants refused services that use eye movement data for interest identification or shopping assistance, and a majority did not like the idea of services inferring their identity, gender, sexual preference, or race. These findings are interesting, as they suggest that users are indeed willing to relinquish privacy in return for service use. They also suggest, however, that users may not be fully aware of the fact that, and to what extent, these services could also infer privacy-sensitive information from their eyes. Our proposed differential privacy approach addresses this challenge by allowing sharing of eye movement data while protecting individual privacy.

To prevent inference of users' private attributes from eye tracking data, not every data representation is suitable. Nonetheless, we identified a clear information gap on the user side, since a majority of participants agreed to share their eye tracking data in almost every data representation (see Figure C.1 in Appendix C). Participants seemed unaware of the fact that, in particular, raw eye movement data representation is inappropriate to protect their privacy. Adding noise to this data representation would not protect their private attributes either: the added noise could easily be removed by smoothing. Instead, we recommend using statistical or aggregated feature representations that summarise temporal and appearance statistics of a variety of eye movements, such as fixation, saccades, and blinks. We are the first to propose a practical solution to this challenge by using differential privacy that effectively protects private information, while at the same time maintaining data utility.

7.7.2 Privacy-Preserving Eye Tracking

Informed by our survey results, we presented a privacy-aware eye tracking method in a VR setting. This is the first of its kind to quantitatively evaluate the practicability and effectiveness of privacy-aware eye tracking. For that purpose, we study 1) two realistic threat models (*with* and *without prior knowledge* about the target user), and 2) different scenarios in training with and without clean/non-noised data. We conducted an extensive evaluation on a novel 20-participant dataset and 3) demonstrated the effectiveness of the trained threat models on two example privacy-infringing tasks, namely gender inference and user identification.

Applying differential privacy mitigates these privacy threats. The fundamental principle of differential privacy is to apply appropriate noise on the data to deteriorate the accuracy of a privacy-infringing task while maintaining that of a utility task. As such, the level of noise should be smaller than the inter-class difference in the utility task but larger than that of the privacy-infringing task.

We showed in our practical evaluations that users' privacy can be preserved with acceptable accuracy of the utility task by applying differential privacy. This conclusion was consistent across different evaluation paradigms in our example study, which aimed to perform gaze-based document type classification while preserving the privacy of users' gender and identity.

Our mechanism can be used to sanitise data not only before releasing it to the public, but also in VR/AR devices themselves, since it sanitises one user at a time. Although our example study focuses only on reading, we expect our method to generalise to any other activity involving eye tracking. Due to our data-driven approach, sensitivity can be adapted so that a similar trade-off can be found. Depending on sensitivity and data vector length, the privacy level ϵ of this trade-off may differ from the presented results. Similarly, our study was evaluated on a typical HCI dataset size, and we expect our approach to generalise to larger datasets that will be available in the future, given the rapid emergence of VR and eye tracking technology.

To conclude, the proposed method is an effective and low-cost solution to preserve users' privacy while maintaining the utility task performance.

7.8 Conclusion

In this chapter we reported the first large-scale online survey to understand users' privacy concerns about eye tracking and eye movement analysis. Motivated by the findings from this survey, we also presented the first privacy-aware gaze interface that uses differential privacy. We opted for a virtual reality gaze interface, given the significant and imminent threat potential created by upcoming eye tracking technology equipped VR headsets. Our experimental evaluations on a new 20-participant dataset demonstrated the effectiveness of the proposed approach to preserve private information while maintaining performance on a utility task – hence, implementing the principle *ensure privacy without impeding utility*.

8 Discovery of Everyday Human Activities from Long-Term Visual Behaviour Using Topic Models

Human visual behaviour has significant potential for activity recognition and computational behaviour analysis, but previous works focused on supervised methods and *recognition* of predefined activity classes based on short-term eye movement recordings. We propose a fully unsupervised method to *discover* users' everyday activities from their long-term visual behaviour. Our method combines a bag-of-words representation of visual behaviour that encodes saccades, fixations, and blinks with a latent Dirichlet allocation (LDA) topic model. We further propose different methods to encode saccades for their use in the topic model. We evaluate our method on a novel long-term gaze dataset that contains full-day recordings of natural visual behaviour of 10 participants (more than 80 hours in total). We also provide annotations for eight sample activity classes (outdoor, social interaction, focused work, travel, reading, computer work, watching media, eating) and periods with no specific activity. We show the ability of our method to discover these activities with performance competitive with that of previously published supervised methods.

8.1 Introduction

Practically everything that we do in our lives involves our eyes, and the way we move our eyes is closely linked to our goals, tasks, and intentions. These links make the eyes a particularly rich source of information about the user as demonstrated by the increasing number of works that use eye movements and closely related measures, such as pupil diameter or blink rate, for context recognition. For example, eye movement analysis has been used to recognise everyday activities, such as in the office (Bulling *et al.*, 2011b) or reading in transit (Bulling *et al.*, 2012; Ishimaru *et al.*, 2014). Moreover, the close link between eye movement and cognition promises automatic analysis of covert aspects of user state that are difficult if not impossible to assess using existing sensing modalities, such as language expertise (Kunze *et al.*, 2013a), visual memory recall (Bulling and Roggen, 2011), perceptual curiosity (Hoppe *et al.*, 2015) or cognitive load (Marshall, 2002; Tessendorf *et al.*, 2011; Chen *et al.*, 2013).

Despite these advances, previous works focused on short-term visual behaviour and supervised methods to *recognise* predefined activity classes. The availability of robust and affordable mobile head-mounted eye trackers points the way to a new class of context-aware systems that can *discover* activities from characteristic eye movement patterns, i.e. without any supervision. Unsupervised discovery of activities from eye movements has the potential to enable a range of novel applications, such as eye-based lifelogging (Ishiguro *et al.*, 2010), mental health monitoring (Vidal *et al.*, 2012b), or the quantified self (Kunze *et al.*, 2013b). The problem setting for these applications is that of post-hoc analysis of human visual behaviour. In that setting, a full-day recording of a person's visual behaviour is available at the time of analysis. The goal of the analysis is to discover characteristic visual behaviours that can then be associated to a set of desired target activity classes. These characteristic behaviours occur at arbitrary points in time and with varying durations throughout the day. Such analysis problems commonly arise in the aforementioned application domains.

So far, however, it remains unclear how much information about daily routines is contained in long-term human visual behaviour, how this information can be extracted, encoded, and modelled efficiently, and how it can be used for unsupervised discovery of human activities. The goal of this chapter is to shed some light on these questions. We collected a new long-term gaze dataset that contains natural visual behaviour of 10 participants (more than 80 hours in total). The data was collected with a state-of-the-art head-mounted eye tracker that participants wore continuously for a full day of their normal life. We annotated the dataset with eight sample activity classes (outdoor, social interaction, focused work, travel, reading, computer work, watching media, and eating) and periods with no specific activity (see Figure 8.1). The dataset and annotations are publicly available online. We further present an approach for unsupervised activity discovery that combines a bag-of-words visual behaviour representation with a latent Dirichlet allocation (LDA) topic model (see Figure 8.2). In contrast to previous works, our method is fully unsupervised, i.e. does not require manual annotation of visual behaviour. It also does not only extract information from saccade sequences but learns a more holistic model of visual behaviour from saccades, fixations, and blinks.

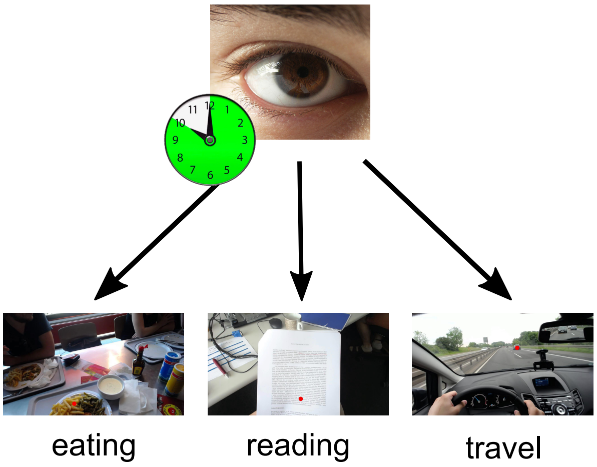


Figure 8.1: Our method takes long-term visual behaviour data (up to ten hours) as input and discovers everyday human activities, such as eating, reading, or being on travel, without supervision.

The specific contributions of this chapter are three-fold. First, we present a novel ground truth annotated long-term gaze dataset of natural human visual behaviour continuously recorded using a head-mounted video-based eye tracker in the daily life of 10 participants. Second, we propose an unsupervised method for eye-based discovery of everyday activities that combines a bag-of-words visual behaviour representation with a topic model. To this end we also propose different approaches to efficiently encode saccades, fixations, and blinks for topic modelling. Third, we present an extensive performance evaluation that shows the ability of our method to discover daily activities with performance competitive with that of previously published supervised methods for selected activities.

8.2 Related Work

Our method builds on previous works on eye movement analysis, eye-based activity and context recognition, as well as discovery of human activities using topic models.

8.2.1 Eye Movement Analysis

Eye movement analysis has a long history as a tool in experimental psychology and human vision research to better understand visual behaviour and perception. Despite its widespread use, previous works typically analysed a small set of well-known eye movement features, most notably fixation duration or fixation patterns. In an early work, Salvucci et al. described three methods based on sequence-matching and hidden Markov models for automated analysis of fixation patterns (Salvucci and Anderson, 2001). Later works used fixation analysis, for example, to identify image features that affect the perception of visual realism (Elhelw *et al.*, 2008), to train novice doctors in assessing tomography images (Dempere-Marco *et al.*, 2002), or to study differences in face recognition (Chuk *et al.*, 2014). Blink rate was shown to correlate with fatigue (Schleicher *et al.*, 2008).

The analysis of the high-frequent fluctuations in pupil diameter has emerged as a robust and well-tested measure of cognitive activity, such as high cognitive load (Marshall, 2002; Palinko *et al.*, 2010). All of these works demonstrated the significant influence of specific tasks on human visual behaviour, but they did not aim to analyse said behaviour to recognise the task at hand.

8.2.2 Eye-Based Activity Recognition

Eye-based activity recognition was first explored in a series of studies by Bulling et al. They proposed a set of eye movement features, including repetitive saccade patterns, as well as a supervised method to recognise human activities from eye movements, such as reading in transit (Bulling *et al.*, 2012), office activities (Bulling *et al.*, 2011b) or cognitive processes, such as visual memory recall processes (Bulling and Roggen, 2011). A similar approach was later used by Tessendorf et al. to recognise cognitive load for context-aware hearing instruments (Tessendorf *et al.*, 2011), as well as by Kunze et al., who showed that different document types could be recognised from visual behaviour (Kunze *et al.*, 2013c). Ishimaru et al. used eye blink frequency and head motion patterns to recognise activities, such as reading or watching TV (Ishimaru *et al.*, 2014). In human-computer interaction, recent works used specific eye movement features to recognise users' tasks, such as task transitions as well as perceptual and cognitive load (Chen *et al.*, 2013), or cognitive abilities, such as visual working memory and perceptual speed (Steichen *et al.*, 2013). More closely related to the work in this chapter, Bulling et al. described an approach to recognise four high-level contextual cues, such as interacting with somebody vs. no interaction, from long-term visual behaviour (Bulling *et al.*, 2013). However, their dataset was considerably smaller and, most importantly, their method was fully supervised.

8.2.3 Activity Discovery Using Topic Models

Topic models have been widely used to discover human activities from video (see (Niebles *et al.*, 2008) for an example) but less often from ambient and on-body sensors (see (Seiter *et al.*, 2014) for a recent analysis of different unsupervised activity discovery approaches). In an early work, Begole et al. analysed daily rhythms of computer use by clustering patterns of computer and email activity (Begole *et al.*, 2003). Barger et al. used mixture models to discover human behaviour patterns from statistics of sensor events in a smart home (Barger *et al.*, 2005). Gu et al. proposed an unsupervised approach for activity recognition based on fingerprints of object use (Gu *et al.*, 2010). They developed a wearable RFID system for object use detection and conducted a real-world data collection with seven participants in a smart home over two weeks. Farrahi et al. used topic models to infer daily routines from mobile phone data (Farrahi and Gatica-Perez, 2008) while Huynh et al. discovered daily routines from accelerometer recordings of a single user (Huynh *et al.*, 2008). We are not aware of any previous work that used topic models to discover activities from human visual behaviour.

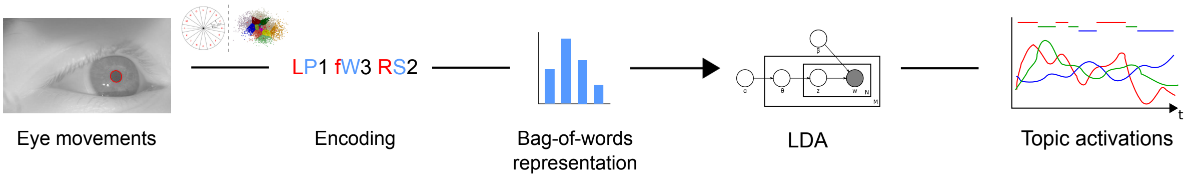


Figure 8.2: Input to our method consists of eye movements detected in the eye video. These movements are first encoded into a string sequence from which a bag-of-words representation is generated. The representation is used to learn a latent Dirichlet allocation (LDA) topic model. Output of the model is the set of topic activations that can be associated with different activities.

8.3 Activity Discovery from Visual Behaviour

We propose a method for unsupervised discovery of everyday human activities (see Figure 8.2 for an overview). Our method combines a bag-of-words visual behaviour representation with a latent Dirichlet allocation (LDA) topic model. Our model uses the full range of eye movements available in current head-mounted eye trackers, namely blinks, fixations (static states of the eyes), and saccades (fast simultaneous movements of both eyes to position gaze at a new location).

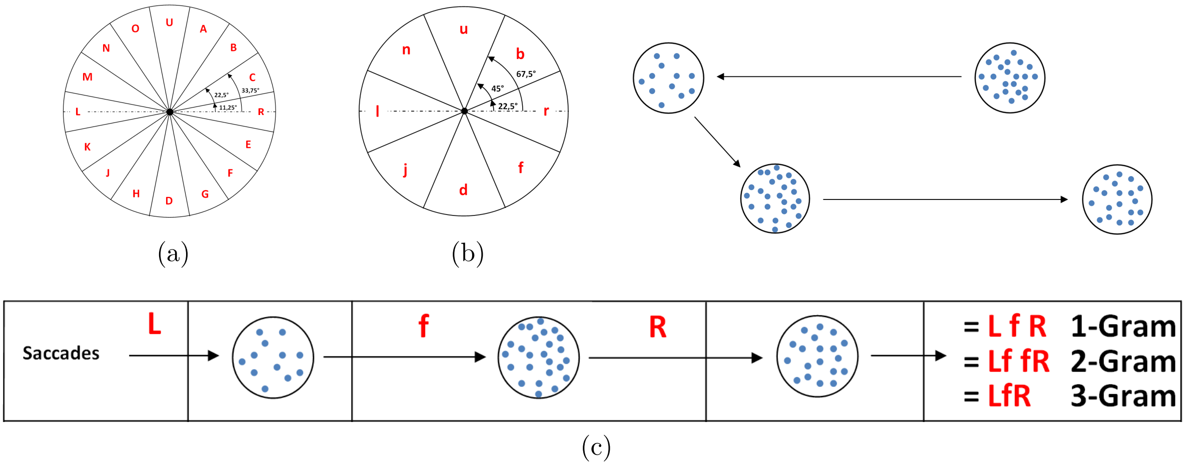
8.3.1 Eye Movement Detection

Eye movements are detected from the pupil positions provided by the eye tracker software in each eye video frame. We first identify overexposed frames and wrongly detected pupils. Specifically, we discard frames with an average grey value larger than 225, a pupil detection confidence value below 85%, or a pupil diameter smaller than 40 pixels. We found these values to work robustly in previous recordings in mobile settings with the same eye tracker.

We then detect three fundamental eye movements from the pupil positions, namely blinks, fixations, and saccades. Blinks can take place at any time and are characterised by closed eye lids. Consequently, to detect blinks, we take frames in which no pupil was detected as blink candidates. Failed pupil detections can also be caused by motion blur, e.g. during a saccade. To discriminate blinks and saccades we apply a velocity threshold of 150 pixels/sec on pupil positions. The velocity is calculated as the difference in pupil position before and after a particular blink candidate divided by the blink duration. We detect fixations using a dispersion-based algorithm (Salvucci and Goldberg, 2000). Frames are assumed to belong to a fixation if the dispersion of the corresponding pupil positions is within a maximum radius of 7.5 pixels, which we determined empirically. In addition, a fixation had to last at least for 200 ms (Holmqvist *et al.*, 2011).

8.3.2 Eye Movement Encoding

We propose four different approaches for encoding saccades into a sequence of characters. In the 1-gram approach we consider individual saccades that we encode according to



(a) (b)

(c)

Figure 8.3: Encoding of large (a) and small (b) saccades according to their direction and amplitude. Example of a resulting encoding of three consecutive saccades for the 1-gram, 2-gram, and 3-gram approach (c). Blue dots indicate individual gaze samples belonging to four fixations.

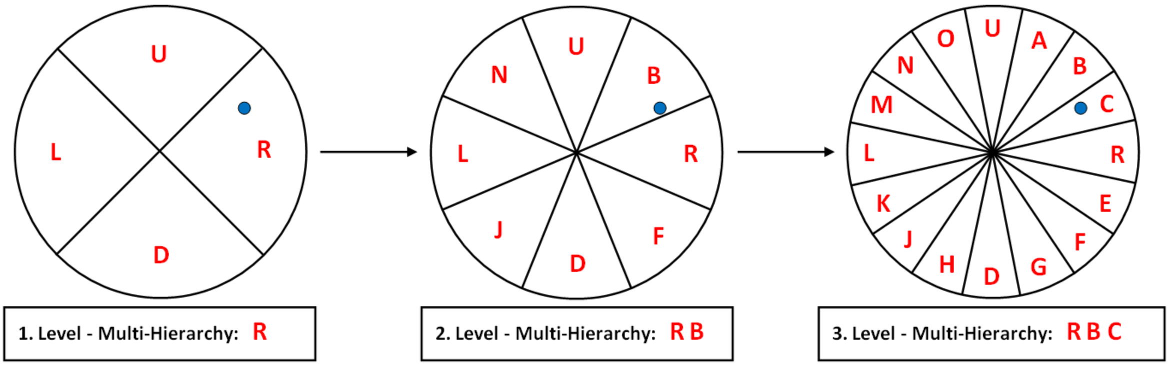

U
L
R
D
N
U
B
L
R
J
D
F
O
U
A
N
B
M
C
L
R
K
E
J
F
H
D
G
1. Level - Multi-Hierarchy: R
2. Level - Multi-Hierarchy: R B
3. Level - Multi-Hierarchy: R B C

Figure 8.4: Sample multi-hierarchy encoding of a particular saccade direction (blue dot). The saccade is encoded across three granularity levels of the discretised saccade direction space.

their direction and amplitude. Similar to (Bulling *et al.*, 2011b), the n-gram approach generalises the 1-gram approach by considering n consecutive 1-gram encodings, thereby retaining information about pre- and succeeding saccades (see Figure 8.3). In the multi-level approach we discretise the saccade direction space with three granularity levels and encode saccades across these levels (see Figure 8.4). For the fourth approach we use k-means to cluster saccades into k clusters based on their direction and amplitude and encode each cluster centroid individually (see Figure 8.5). This approach is data-driven and only requires a single parameter, the number of clusters k, instead of predefined thresholds for saccade amplitudes and directions.

We further encode fixation duration and blink rate (see Figure 8.6). Fixation duration is a well-established measure in experimental psychology and commonly used for studies on visual perception and cognition (Just and Carpenter, 1976). We encode fixation duration by first finding the person-specific minimum and maximum durations and

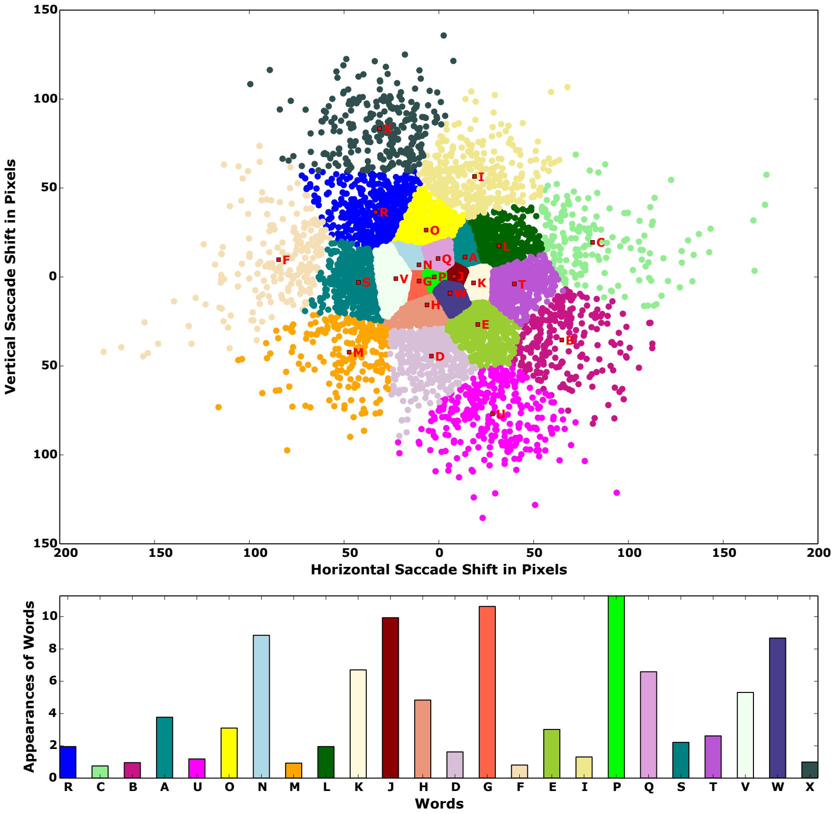


Figure 8.5: Sample k-means clustering of saccades based on their direction and amplitude for $k = 24$ of P6. Each cluster centroid is encoded with a distinct character.

then splitting this range into 10 equally-sized bins. Each bin, and consequently all fixation durations that fall into that bin, is then encoded with a distinct character. The number of blinks during a fixation we directly encode in the string sequence. Finally, we encode the combined character sequence – that still contains temporal information – into a bag-of-words representation by generating histograms of word occurrence counts (see Figure 8.5).

8.3.3 Topic Modelling

The bag-of-words visual behaviour representation serves as input to an LDA topic model (Blei *et al.*, 2003). We opted for an LDA model given that it recently proved most robust among popular topic models (Seiter *et al.*, 2014). Topic models were originally proposed in the text processing community (Hofmann, 2001) but subsequently became influential also in other domains, most notably computer vision (Csurka *et al.*, 2004) and human activity recognition (Huynh *et al.*, 2008). As introduced by Blei et al. (Blei *et al.*, 2003), topic models regard a corpus of text documents as a collection of words belonging to different topics, the so-called bag-of-words (BoW) representation. Topic models learn probability distributions of words belonging to these topics but, more importantly, also make it possible to infer the underlying topics from a corpus of documents.

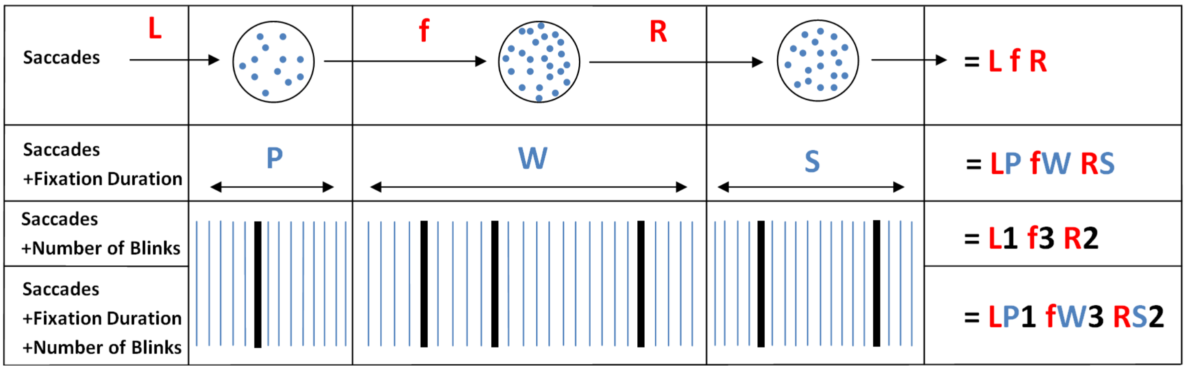


Figure 8.6: Fixation duration is binned into a person-specific histogram and each bin is encoded with a distinct character. The number of blinks is directly encoded in the character sequence.

Expressed mathematically, a document is defined as a collection of N words denoted by $\mathbf{w} = (w_1, w_2, ..., w_n)$, where w_n is the n^{th} word in the document. The document corpus C contains M documents denoted by $C = \{\mathbf{w}_1,\ \mathbf{w}_2,\ ...,\ \mathbf{w}_M\}$. In addition to the document corpus, the number of topics K and the Dirichlet prior $p(\theta_d|\alpha)$ with parameter α on the topic-document distributions $p(t|\theta_d)$ have to be determined to derive θ, which describes the topic-document distribution. By defining the number of topics, the dimensionality of the topic variable t is assumed to be known and fixed. The word probabilities are parametrised by a $K \times V$ matrix β, where $\beta_{ij} = p(w^j = 1|t^i = 1)$. To calculate the probability of a corpus $p(C|\alpha, \beta)$, the parameters α for the Dirichlet distribution and parameter β for the word distribution $p(w|t, \beta)$ have to be found to maximise the likelihood $\mathcal{L}$ over all documents $d = 1, ..., M$. The formula is given by

$$p(C|\alpha, \beta) = \mathcal{L}(\alpha, \beta) =$$

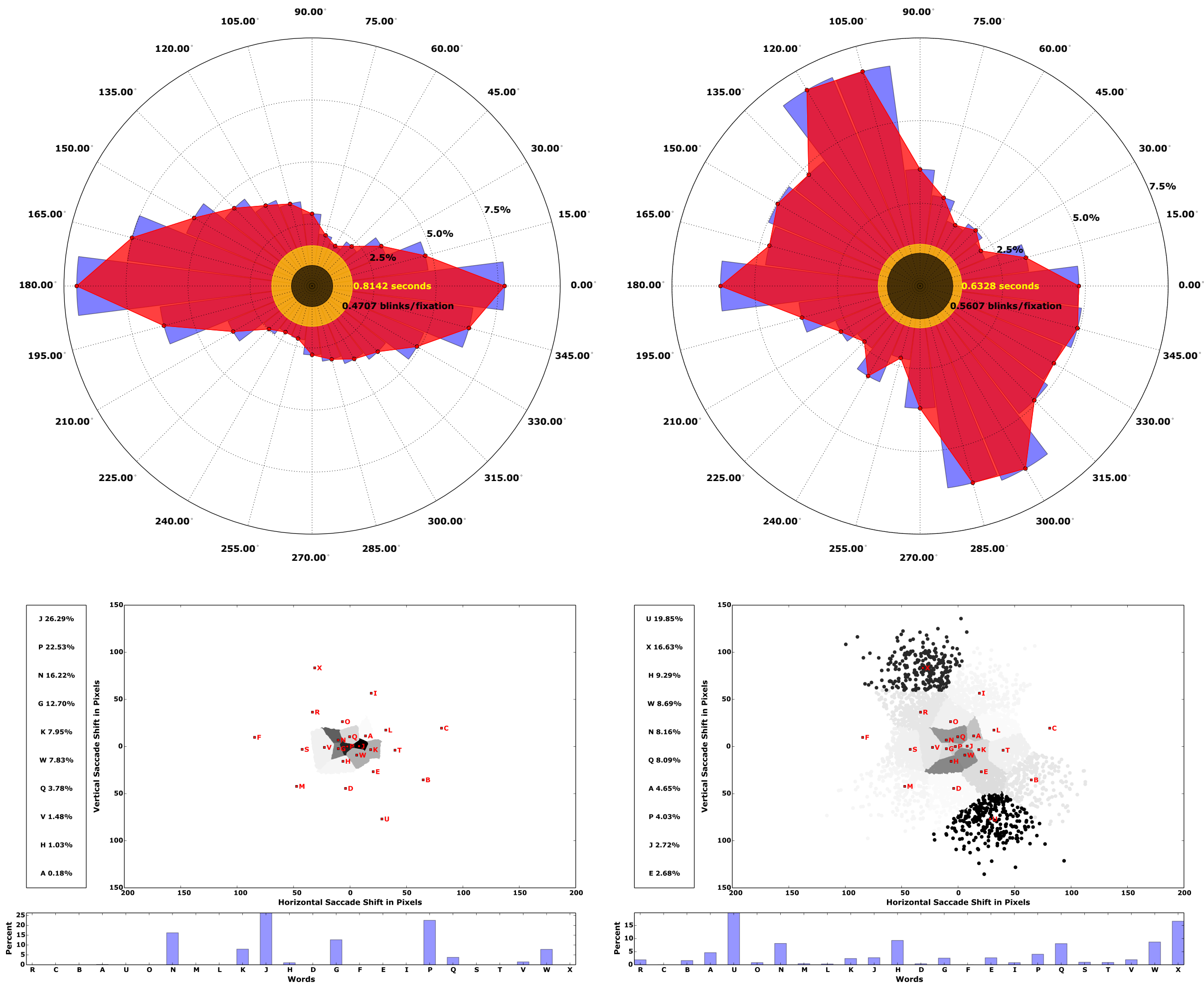
$$\prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{t=1}^{K} p(w_n^d|t_n^d, \beta) p(t_n^d|\theta_d) \right) \theta_d, \tag{8.1}$$

where each document consists of the words w_n^d with $n = 1, ..., N_d$. With α and β, θ can be derived and the corpus C can be decomposed into the following form:

$$C = \phi \cdot \theta \tag{8.2}$$

8.4 Data Collection

In this representation the word-topic distribution ϕ and the topic-document distribution θ are key to discovering activities. Following the same terminology as (Blei *et al.*, 2003), we propose to encode eye movement characteristics as words and to regard long-term visual behaviour as a corpus of text documents composed of these words, from which activities (topics) are automatically inferred. Consequently, we split the encoded visual



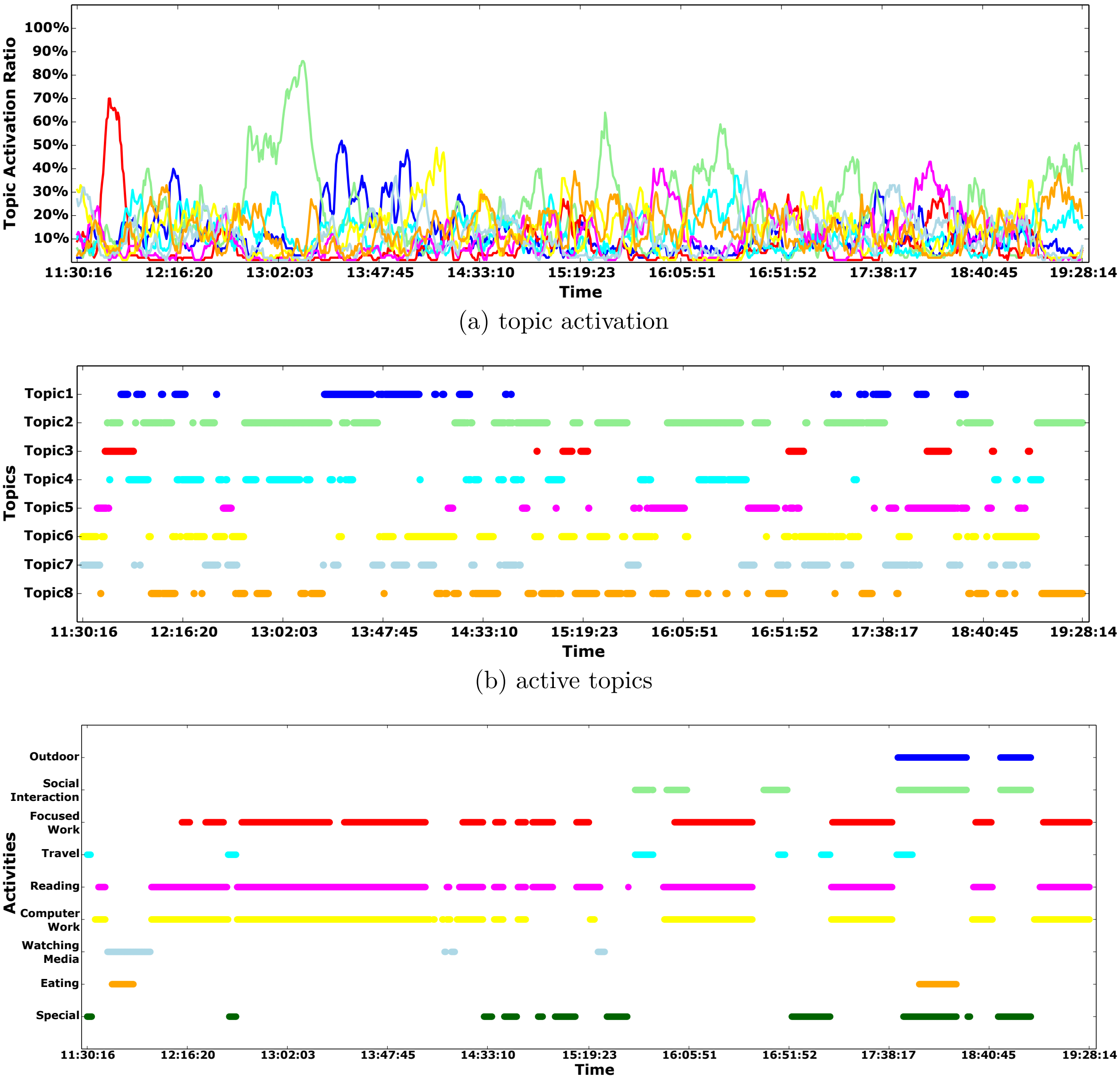
(a) reading

(b) watching media

Figure 8.7: Sample saccade direction distributions for "reading" and "watching media" (top), as well as the corresponding 24-means saccade encoding and (middle) and word-topic distributions (bottom) for P6.

behaviour sequence into a corpus of documents using a sliding window with a window size of five minutes and a step size of 30 seconds. These values were, again, determined empirically. We then run the LDA topic model with $K = 4, 6, 8, 10$ topics using a Dirichlet prior α of $50/K$, as recommended by Griffths and Steyvers (Steyvers and Griffiths, 2007). The topic model generates two outputs: 1) the word-topic distribution ϕ that describes the visual behaviour for a specific topic or, as in our our case, during a specific activity, and 2) the topic-document distribution θ that indicates if and when a topic is active in a particular document. These topic activations are then associated with the different ground truth activities.

Figure 8.7 shows sample saccade direction distributions for "reading" and "watching media" as well as the corresponding word-topic distributions. The corresponding topic-document distributions are shown in Figure 8.8b while Figure 8.8a shows the topic activations. The active topics can then be compared to the annotated ground truth



(a) topic activation

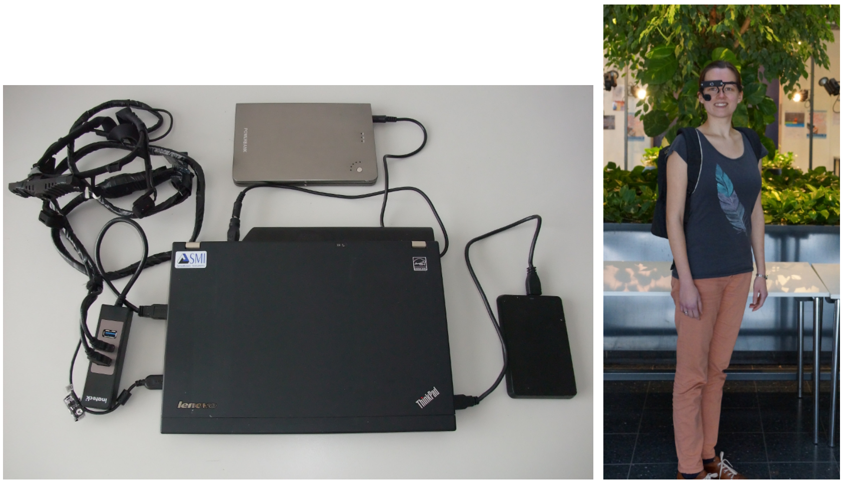
(b) active topics

(c) ground truth

Figure 8.8: Result of the topic modelling approach applied with eight topics on the 24-means encoding and the ground truth annotation of P6.

activities (see Figure 8.8c). In this example, topic 2 seems to represent “reading” while topic 3 matches best with “watching media”.

To the best of our knowledge, the only long-term dataset of human visual behaviour recorded in daily life so far is the one presented in (Bulling *et al.*, 2013). However, that dataset is not publicly available and, as mentioned before, it only contains relative eye movements of four participants recorded using a wearable electro-oculography device. We therefore collected our own long-term visual behaviour dataset using a state-of-the-art head-mounted video-based eye tracker.



(a) (b)

Figure 8.9: Recording setup consisting of a laptop with an additional external hard drive and battery pack, as well as a Pupil head-mounted eye tracker (a). Recording hardware worn by a participant (b).

8.4.1 Apparatus

The recording system consisted of a Lenovo Thinkpad X220 laptop, an additional 1TB hard drive and battery pack, as well as an external USB hub. Gaze data was collected using a Pupil head-mounted eye tracker connected to the laptop via USB (Kassner *et al.*, 2014) (see Figure 8.9). The eye tracker features two cameras: one eye camera with a resolution of 640 × 360 pixels recording a video of the right eye from close proximity, as well as an egocentric (scene) camera with a resolution of 1280 × 720 pixels. Both cameras record at 30 Hz. The battery lifetime of the system was four hours. We implemented custom recording software with a particular focus on ease of use as well as the ability to easily restart a recording if needed.

8.4.2 Procedure

We recruited 10 participants (three female) aged between 17 and 25 years through university mailing lists and adverts in university buildings. Most participants were bachelor's and master's students in computer science and chemistry. None of them had previous experience with eye tracking. After arriving in the lab, participants were first introduced to the purpose and goals of the study and could familiarise themselves with the recording system. In particular, we showed them how to start and stop the recording software, how to run the calibration procedure, and how to restart the recording. We then asked them to take the system home and wear it continuously for a full day from morning to evening. We asked participants to plug in and recharge the laptop during prolonged stationary activities, such as at their work desk. We did not impose any other restrictions on these recordings, such as which day of the week to record or which activities to perform, etc.

Activity Class	Description	P1 (m)	P2 (m)	P3 (f)	P4 (m)	P5 (m)	P6 (m)	P7 (m)	P8 (f)	P9 (m)	P10 (f)	Total
outdoor	Person is outside	134	48	6	27	6	62	0	33	0	150	466
social interaction	Person is interacting with somebody else	173	69	127	77	81	95	5	59	0	169	855
focused work	Person is doing focused work	313	34	114	170	221	221	275	214	72	243	1877
travel	Person is travelling, e.g. walking or driving	156	70	40	47	33	47	18	32	23	30	496
reading	Person is reading	347	39	182	278	282	266	350	288	83	256	2371
computer work	Person is working on the computer	189	30	135	267	277	263	327	121	81	30	1720
watching media	Person is watching media	9	280	115	114	46	37	90	36	308	62	1097
eating	Person is eating	44	43	49	34	34	32	55	47	28	56	422
special	Special events, e.g. tying shoes	49	45	97	32	95	124	52	67	79	45	685

Table 8.1: Overview of the dataset showing the amount of ground truth annotated data for each activity class and participant in minutes. Participants' gender is given in brackets (f: female, m: male). Note that annotations are non-mutually exclusive, i.e. they sum up to more than the actual dataset size.

8.4.3 Ground Truth Annotation

For evaluation purposes, the full dataset was annotated post-hoc from the scene videos by a paid human annotator with a set of nine non-mutually-exclusive ground truth activity labels (see Table 8.1 and Figure 8.8c). Specifically, we included labels for whether the participant was inside or outside (outdoor), took part in social interaction, did focused work, travelled (such as by walking or driving), read, worked on the computer, watched media (such as a movie) or ate. We further included a label for special events, such as tying shoes or packing a backpack. This selection of labels was inspired by previous works and includes a subset of activities from (Bulling *et al.*, 2011b, 2013; Shiga *et al.*, 2014).

8.4.4 Dataset

We were able to record a dataset of more than 80 hours of eye tracking data (see Table 8.1 for an overview and Figure 8.10 for sample images). The dataset comprises 7.8 hours of outdoor activities, 14.3 hours of social interaction, 31.3 hours of focused work, 8.3 hours of travel, 39.5 hours of reading, 28.7 hours of computer work, 18.3 hours of watching media, 7 hours of eating, and 11.4 hours of other (special) activities. Note that annotations are not mutually exclusive, i.e. these durations should be seen independently and sum up to more than the actual dataset size.

Most of our participants were students and wore the eye tracker through one day of their normal university life. This is reflected in the overall predominant activities, namely focused work, reading, and computer work. Otherwise, as can also be seen from the table, our dataset contains significant variability with respect to participants' daily routines and consequently the number, type, and distribution of activities that they performed. For example, while P1 wore the eye tracker during a normal working day at the university, P7 and P9 recorded at a weekend and stayed at home all day mainly reading and working on the computer (P7) or watching movies (P9) with little or no social interactions.

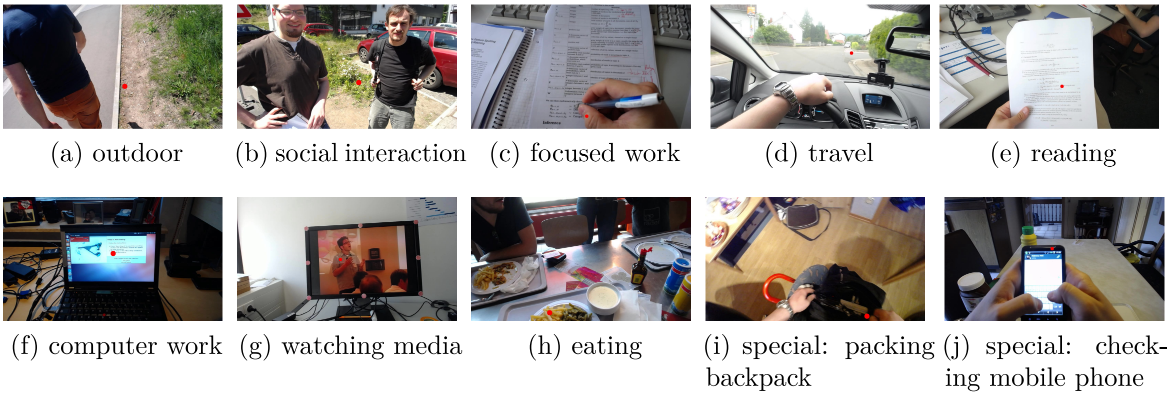

(a) outdoor
(b) social interaction
(c) focused work
(d) travel
(e) reading
(f) computer work
(g) watching media
(h) eating
(i) special: packing backpack
(j) special: checking mobile phone

Figure 8.10: Sample scene images for each activity class annotated in our dataset showing the considerable variability in terms of place and time of recording. The red dot indicates the gaze location in that particular image.

8.5 Results

Huynh et al. used topic models to discover daily routines that consisted of re-occurring activities of a single person over several days (Huynh *et al.*, 2008). Although participants' activities varied across days, their overall daily routines were still rather similar. In contrast, we deal with full-day recordings of multiple participants and a large variability with respect to the number, type, and distribution of activities that they performed, as well as their visual behaviour. In consequence, the best-performing model – specifically the best-performing saccade encoding, eye movement characteristics, as well as topic model parameters – is highly person-specific. We therefore opted to first show the best performance for each participant irrespective of the particular parameters used. In subsequent analyses we then focus on one representative participant to show the influence of different parameters on performance. In all analyses that follow, performance was calculated using the F1 score $F_1 = 2 * \frac{precision*recall}{precision+recall}$, which 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.

8.5.1 Performance for Each Participant

We first calculated the performance for each participant while optimising all free parameters of our method, i.e. saccade encoding, eye movement characteristics, as well as the number of topics in the topic model. Figure 8.11 shows the top mean F1 score for each participant with error bars visualising the range of performances for the particular subset of activities performed by the participant. As can be seen from the figure, our method achieves robust performance for discovering everyday activities across all participants independent of the particular type and distribution of activities. However, the figure also shows the considerable variability in performance for individual activities depending on the duration with which these activities were performed (cf. Table 8.1).

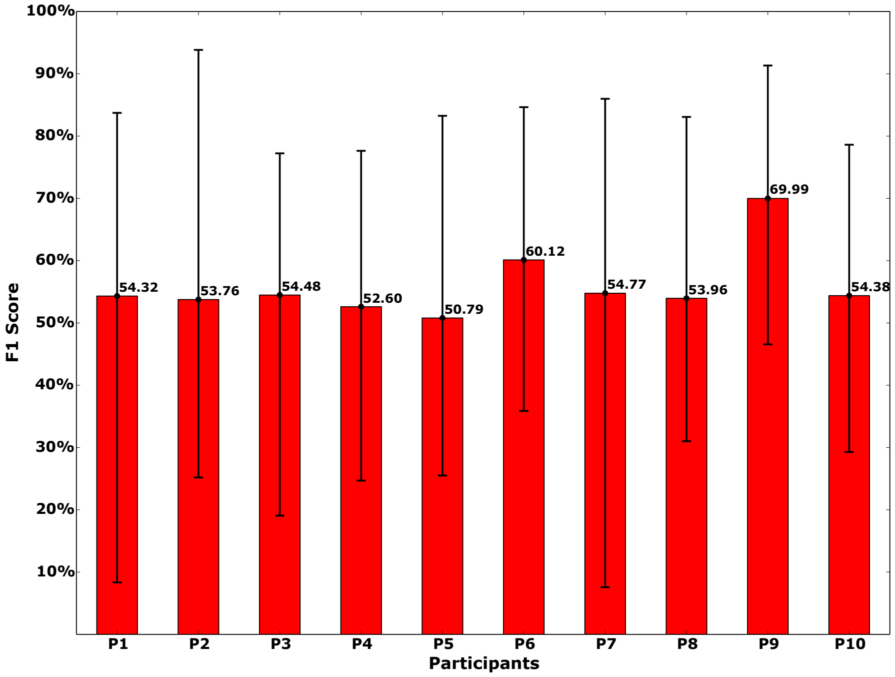


Figure 8.11: Top mean F1 scores for each participant with error bars visualising the range of performances for the particular set of activities performed by the participant irrespective of the particular saccade encoding, eye movement characteristics, or topic model parameters used.

For example, the minimum F1 score was achieved for P1 for watching media (8.34%) and P7 for social interaction (7.58%). As can be seen from Table 8.1, in both cases the respective activity was performed over considerably shorter durations than all other activities. The top F1 scores were achieved by P2 (93.83%) and P9 (91.33%) for watching media. These were also the activities performed the most among all activities.

8.5.2 Performance Across Participants

We then studied performance across all participants. As before, we optimised all free parameters of our method and calculated the mean, minimum, and maximum F1 scores for each activity. Figure 8.12 shows the top mean F1 score averaged over all participants performing the activity with error bars visualising the range of individual performances. The best performance was achieved for reading (74.75%), focused work (70.01%), and computer work (64.18%), while all other activities could be discovered with a mean F1 score of around 50%. These findings are in line with results reported in previous works that showed that reading and focused work could be recognised well using supervised learning methods (Bulling *et al.*, 2008b; Tessendorf *et al.*, 2011). Table 8.1 further shows that the good performance correlates with the duration with which these three activities were performed, i.e. the more data is available, the better the activity can be discovered by our LDA topic model.

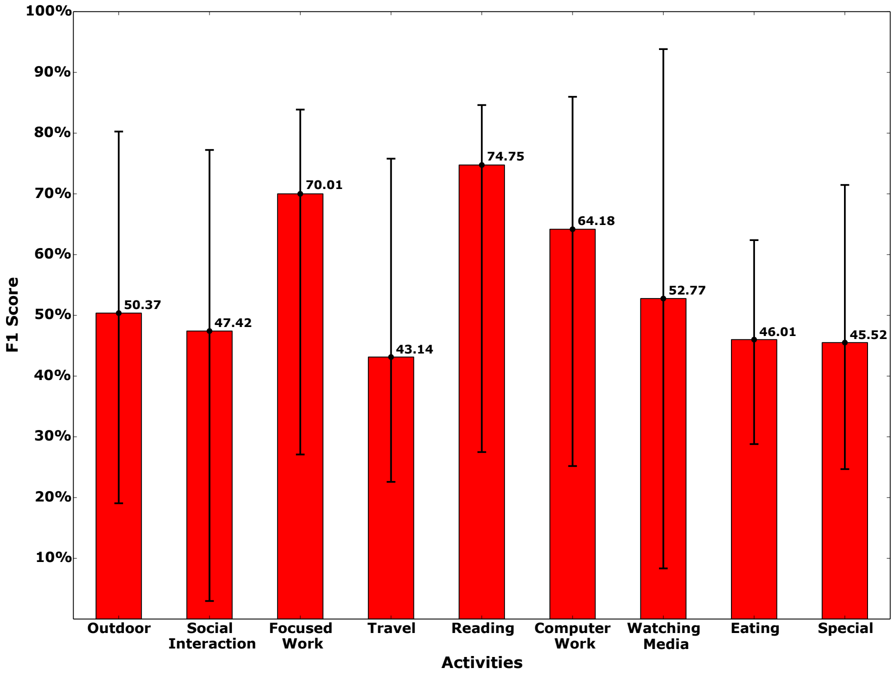


Figure 8.12: Top mean F1 score for each activity across all participants with error bars visualising the range of performances results for the participants performing the corresponding activity irrespective of the particular saccade encoding, eye movement characteristics, or topic model parameters used.

8.5.3 Impact of Different Saccade Encodings

We then evaluated the different saccade encodings because of their fundamental importance for our activity discovery method. For each encoding (1-gram, n-gram, multi-hierarchy, and k-means) we calculated the best average performance across activities and participants using all eye movement characteristics. We also swept the the number of topics $K = 4, 6, 8, 10$ in our topic model. Although not shown here, the k-means encoding with $k = 24$ and $K = 10$ topics performed best overall. Thus, we decided to use k-means encoding with $k = 24$ in all following evaluations.

As mentioned before, both the activities that participants performed and their visual behaviour was highly person-specific. Evaluating all parameters for all participants was therefore deemed infeasible. To select one representative participant, we calculated histograms over the activity durations for each participant as well as the total, and calculated the binary distances between these using the χ^2 distance metric. Based on these distance comparisons, we selected P6 for further investigation, as his activity distribution most closely resembled the distribution of the full dataset.

8.5.4 Impact of Eye Movement Characteristics

We were further interested in the impact of different eye movement characteristics on performance for individual activities. Figure 8.13 provides an overview of the

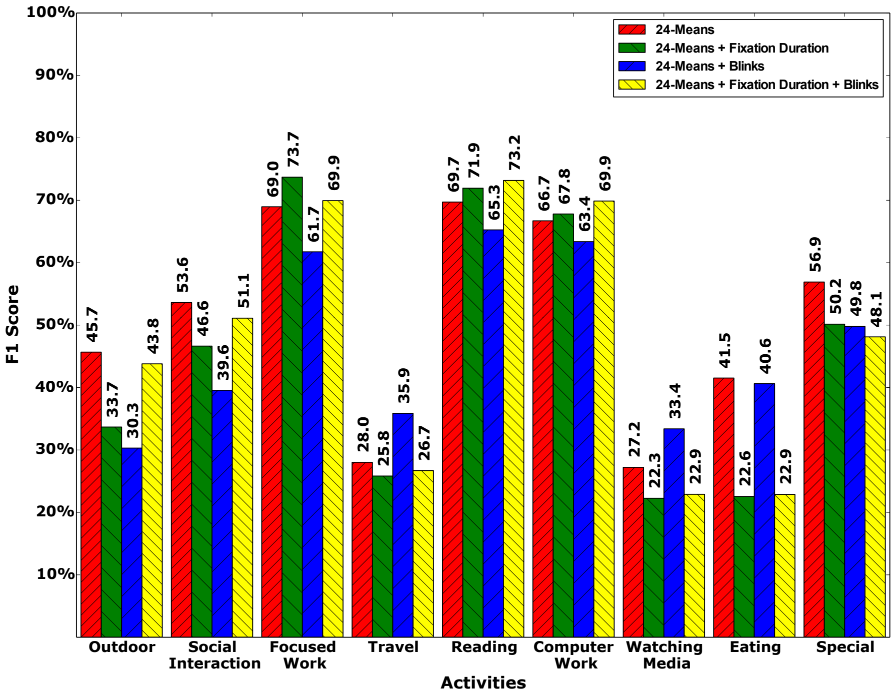


Figure 8.13: Performance comparison for different eye movement characteristics for the 24-means saccade encoding with 10 topics for P6.

performance for P6 for different eye movement characteristics using the 24-means saccade encoding for each activity. The figure shows that the best-performing eye movement characteristic is indeed activity-specific. For this specific participant, only using information about saccades achieved the best performance for four out of the nine activity classes, namely outdoor (45.7%), social interaction (53.6%), eating (41.5%), and special (56.9%). Additional information on fixation duration achieved the best performance only for focused work (73.7%) while adding information on blinks achieved best performance for travel (35.9%) and watching media (33.4%). Finally, using information about all three eye movement characteristics achieved best performance for reading (73.2%) as well as computer work (69.9%).

8.5.5 Impact of Number of Topics

The previous evaluation showed that additional eye movement characteristics can improve performance for particular activities. We further analysed the impact of different number of topics $K = 4, 6, 8, 10$ on performance. Figure 8.14 shows a performance comparison for different numbers of topics for the 24-means saccade encoding with blinks for P6. The figure shows that, similar to the different eye movement features, the number of topics affects individual activities differently. These performance differences are also linked to the duration of the activities performed by the participant (cf. Table 8.1). Generally speaking, the lower the number of topics the better the dominating activities – focused work, reading, computer work, and special – can be discovered from visual behaviour. The higher the number of topics,

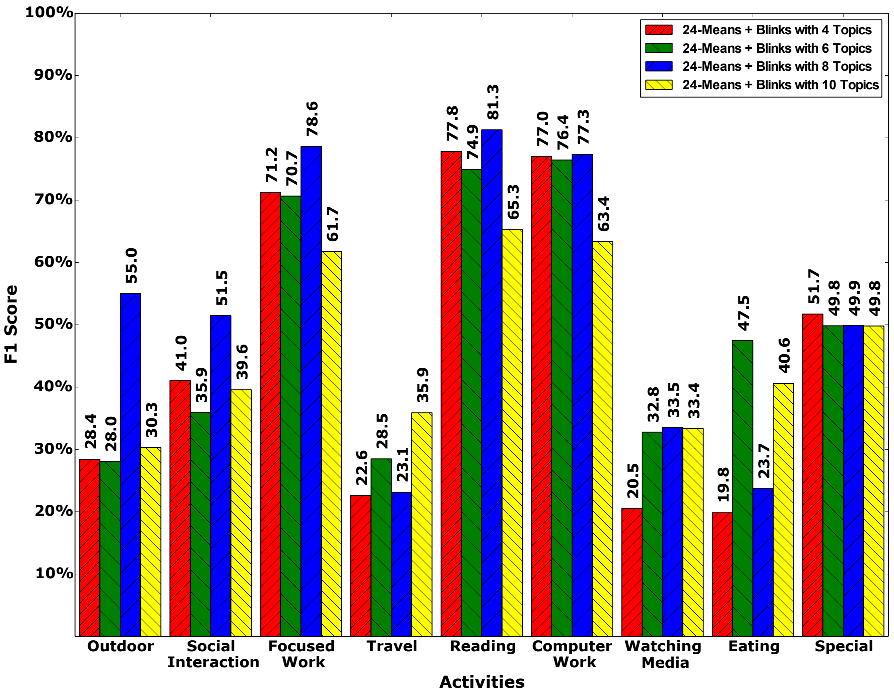


Figure 8.14: Performance comparison for different number of topics for the 24-means saccade encoding with blinks for P6.

the more activities can be discovered, but with decreased F1 scores. This can be seen in Figure 8.14 where the F1 scores are generally higher for eight topics than for ten topics. If there are many activities, the smaller the number of topics, the worse the results given that one topic will encode multiple activities.

8.5.6 Comparison with Supervised Methods

Supervised methods have previously been used to recognise reading and different office activities from eye movement (Bulling *et al.*, 2011b, 2012). We were therefore finally interested in comparing the performance achieved for discovering reading, computer work, and watching media with our unsupervised method with those used in prior work (see Figures 8.15-8.17). As shown in Figure 8.15 we were able to recognise reading with a top F1 score of 74.75% compared to the F1 score of about 70% achieved using a linear support vector machine as reported in (Bulling *et al.*, 2011b). For computer work we achieved a maximum mean F1 score of 64.18%, which is a bit lower than the 70% for browsing reported in (Bulling *et al.*, 2011b). For watching media we achieved a maximum mean F1 score of only 52.77%, while the corresponding performance for recognising watching video in (Bulling *et al.*, 2011b) was about 83%. It is important to note, however, that performance for discovering computer work and watching media is reduced because not every participant performed these activities for a sufficient amount of time. For individual participants we were able to achieve a maximum performance of over 90% F1 score for watching media.

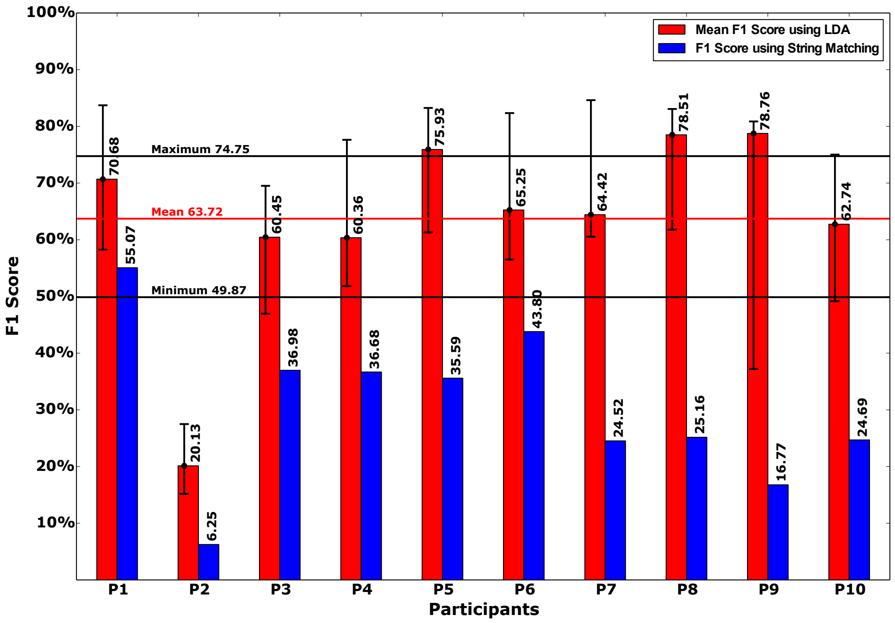


Figure 8.15: Performance comparison for "reading" for each participant using the 24-means saccade encoding with blinks and 10 topics. The blue bars show the F1 scores achieved using the string matching approach described in (Bulling *et al.*, 2008b) which moves the reading template "Rlll" over the encoded 1-gram saccade sequence and thresholds on the Levensthein distances.

To establish baseline performance results and directly compare the different methods on our new dataset, we reimplemented the string matching algorithm for reading recognition as described in (Bulling *et al.*, 2012). We also trained our own linear support vector machines and naïve Bayes classifiers for binary activity classification, the former of which was used in (Bulling *et al.*, 2011b). In a nutshell, the string matching algorithm moves a predefined reading template "Rlll" over the encoded 1-gram saccade sequence. Intuitively, the template describes the characteristic sequence of small saccades to the left while scanning a line of text, followed by the large "carriage return" saccade to the right to jump to the beginning of the next line. To detect reading, the algorithm calculates the Levensthein distance and applies a distance threshold of $T_{ed} = 3$ in each step of moving the template over the sequence and finally performs majority voting in a window of string length $W_{str} = 30$. As can be seen from Figure 8.15, our LDA topic model outperforms the string matching approach for all participants.

For the SVM algorithm we fixed the two main parameters, the cost C and the tolerance of termination criterion ϵ, to $C = 1$ and $\epsilon = 0.001$. Every feature vector consists of 56 of the 62 features described in (Bulling *et al.*, 2011b) and was computed for a time window $W_{fe} = 120s$ and a step size $S_{fe} = 1s$. Table 8.2 provides an overview of this comparison for P6. As can be seen from the table, our method shows competitive performance to the SVM in terms of F1 score, accuracy and correlation and even outperforms SVM in terms of recall. Both always outperforms the naïve Bayes classifier.

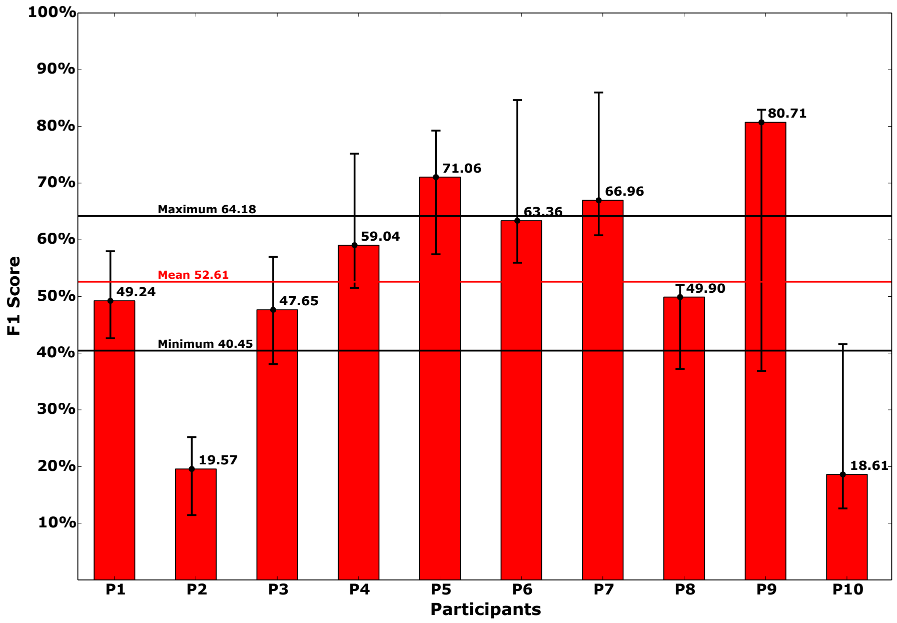


Figure 8.16: Performance comparison for "computer work" for each participant using the 24-means saccade encoding with blinks and 10 topics.

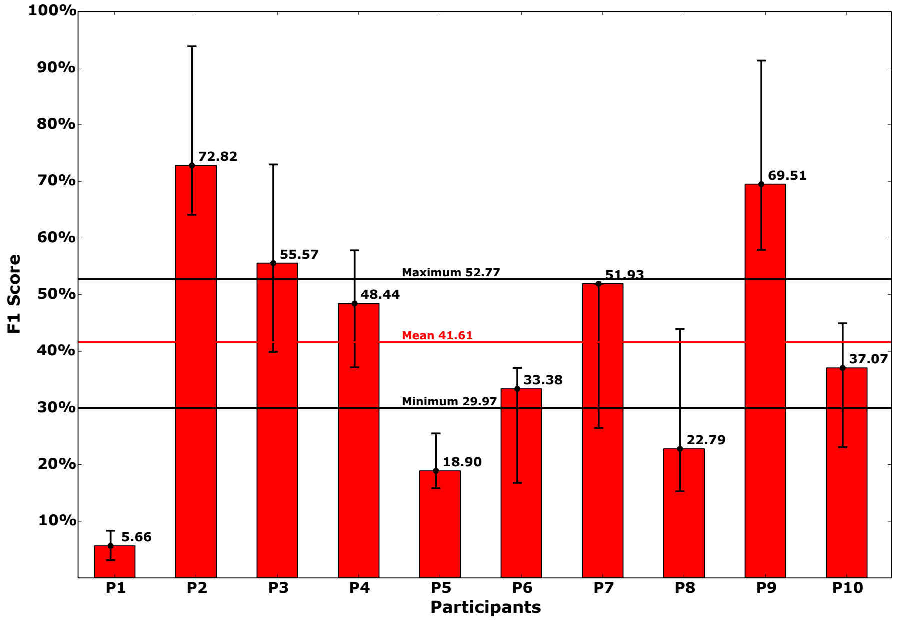


Figure 8.17: Performance comparison for "watching media" for each participant using the 24-means saccade encoding with blinks and 10 topics.

	Precision			**Recall**			**F1 Score**			**Accuracy**			**Correlation**		
Activity Class	*LDA*	*SVM*	*NB*	*LDA*	*SVM*	*NB*	*LDA*	*SVM*	*NB*	*LDA*	*SVM*	*NB*	*LDA*	*SVM*	*NB*
outdoor	38.0	68.7	29.6	100.0	85.0	69.9	55.0	76.0	41.6	81.2	94.4	89.3	0.55	0.73	0.41
social interaction	43.4	75.2	18.1	80.8	30.7	57.9	56.5	43.6	27.6	76.9	62.2	81.5	0.46	0.27	0.25
focused work	74.9	81.8	97.5	84.5	78.3	58.8	79.4	80.0	73.4	79.3	81.3	67.5	0.59	0.62	0.46
travel	23.7	69.7	93.5	73.3	23.3	11.0	35.9	35.0	19.7	78.9	78.9	38.0	0.33	0.31	0.16
reading	82.4	80.8	96.4	82.3	86.5	67.3	82.3	83.6	79.3	79.8	82.4	72.2	0.58	0.65	0.47
computer work	86.2	87.9	95.8	83.1	83.0	64.7	84.6	85.4	77.2	83.3	83.6	69.2	0.66	0.67	0.42
watching media	27.0	20.0	93.7	59.3	79.2	8.8	37.1	32.0	16.1	82.8	93.9	30.4	0.32	0.38	0.12
eating	38.3	60.8	97.2	86.1	68.7	18.0	53.0	64.5	30.4	89.4	95.6	70.6	0.53	0.62	0.34
special	41.4	57.4	96.1	92.5	79.8	28.6	57.2	66.8	44.1	66.6	86.0	40.2	0.44	0.59	0.21
Average	50.6	66.9	79.8	82.4	68.3	42.8	60.1	63.0	45.5	79.8	84.3	62.1	0.50	0.54	0.3

Table 8.2: Performance comparison for the LDA topic model, a support vector machine (SVM), and a naïve Bayes (NB) classifier in terms of precision, recall, F1 score, accuracy, and Matthews correlation coefficient for P6.

8.6 Discussion

Referring to the open questions from the introduction, results on our new 10-participant dataset demonstrate that long-term human visual behaviour does indeed contain a significant amount of information about our daily routines. We demonstrated that this information can be extracted from key eye movements that can be readily tracked with head-mounted eye trackers, namely saccades, fixations, and blinks. We further proposed and evaluated different methods to efficiently encode the extracted information into a joint bag-of-words representation. Building on this representation, we introduced LDA topic models as a versatile method to model a wide variety of human visual behaviours. We demonstrated the suitability of this whole approach for unsupervised discovery of everyday activities. Specifically, we are able to recognise reading with a top average performance of 74.75%, which is competitive with results reported in previous works using fully supervised methods (Bulling *et al.*, 2012).

Our evaluations also revealed that the best combination of methods and parameters – and consequently the performance – depend considerably on the particular user and his specific visual behaviour as well as the type, number, and distribution of activities that he performed throughout the day. Consequently, to achieve good performance, both the specific eye movement characteristics as well as the number of topics (activities) modelled in the topic model have to be optimised to the particular set of activities relevant for a particular application. While this may seem a severe limitation, supervised methods pose even stricter requirements, as the set of activity classes recognised by the system has to be defined and trained up front. In contrast, the proposed method can deal with an arbitrary number of activity classes as long as the target activities are performed sufficiently long relative to all other activities. This requirement directly stems from the fact that topic models rely on word-topic and topic-document distributions and require sufficient statistics about individual topics.

8.7 Conclusion

In this chapter we proposed a new dataset as well as a fully unsupervised approach to discover human activities from long-term visual behaviour. Our approach efficiently encodes the full range of eye movements available in current head-mounted eye trackers, namely blinks, saccades, and fixations. Our results show the significant information content available in human visual behaviour for unsupervised discovery of activities, opening up new venues for research on eye-based behavioural monitoring and lifelogging.

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

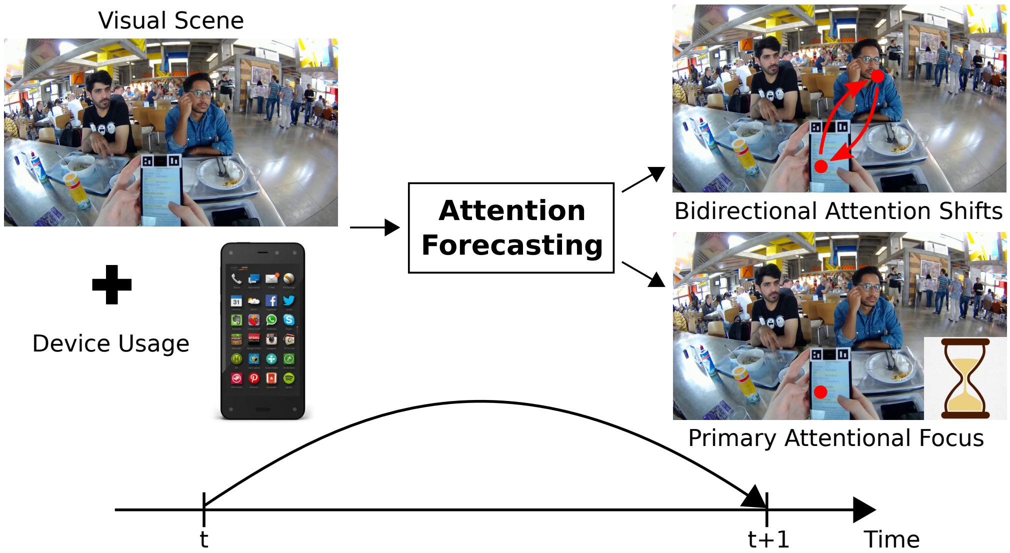


Figure 9.1: We propose a method to forecast temporal allocation of overt visual attention (gaze) during everyday interactions with a handheld mobile device. Our method uses information on users' visual scene as well as device usage to predict attention shifts between mobile device and environment and 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.*, 2010b) 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*

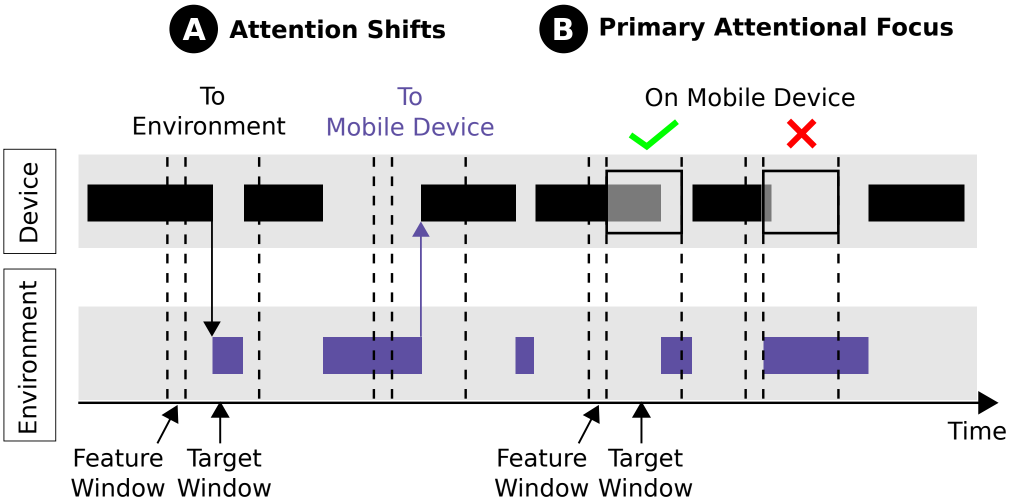


Figure 9.2: Overview of the different prediction tasks explored in this chapter: Prediction of attention shifts to the environment and (back) to the mobile device, and the primary attentional focus, i.e. whether attention is primarily on or off the device.

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

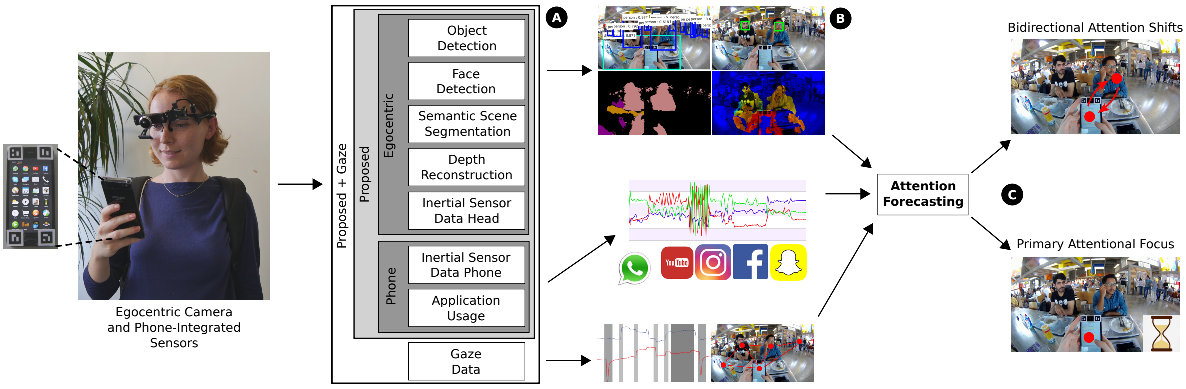


Figure 9.3: Overview of our method for attention forecasting during mobile interactions. Taking information on users' visual scene, mobile device (phone) and head inertial data, as well as on mobile app usage as input (A), our method extracts rich semantic information about the user's visual scene using 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
Proposed + Gaze	Proposed	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	**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
			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.*, 2011b, 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.*, 2015) 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

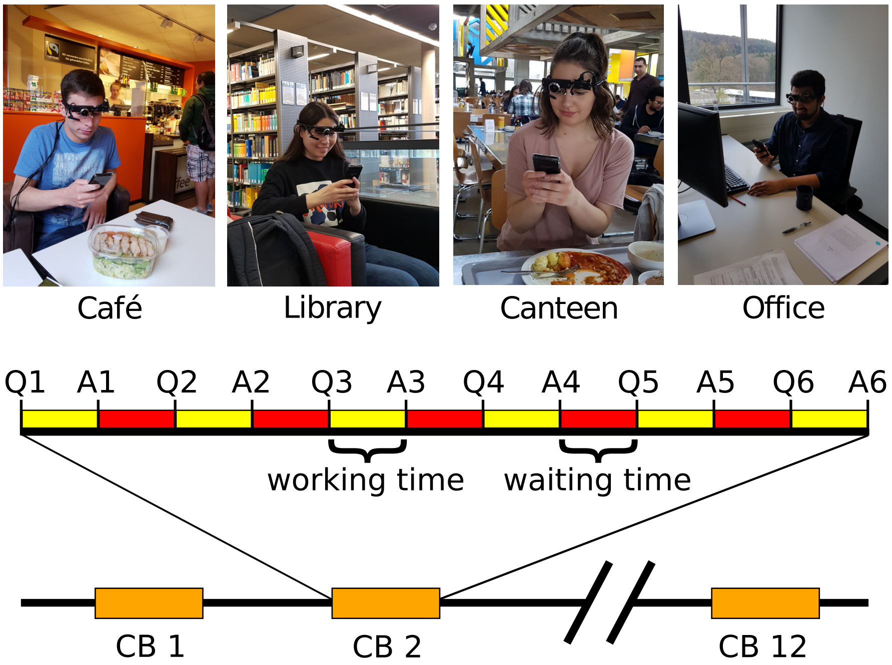


Figure 9.4: Participants were engaged in 12 chat blocks (CB) in different environments that were randomly distributed over their recording, which lasted in total about 4.5 hours. In each block, participants had to answer six questions, some of which required a short online search (Q1–Q6, working time),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

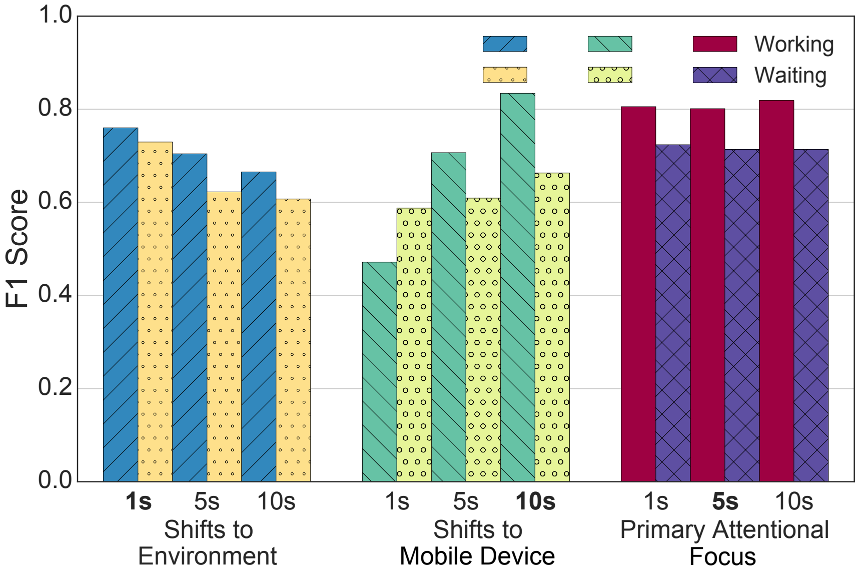


Figure 9.5: Performance analysis for shifts to environment, shifts to mobile device, and primary attentional focus for different target sizes (1s, 5s, 10s).

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

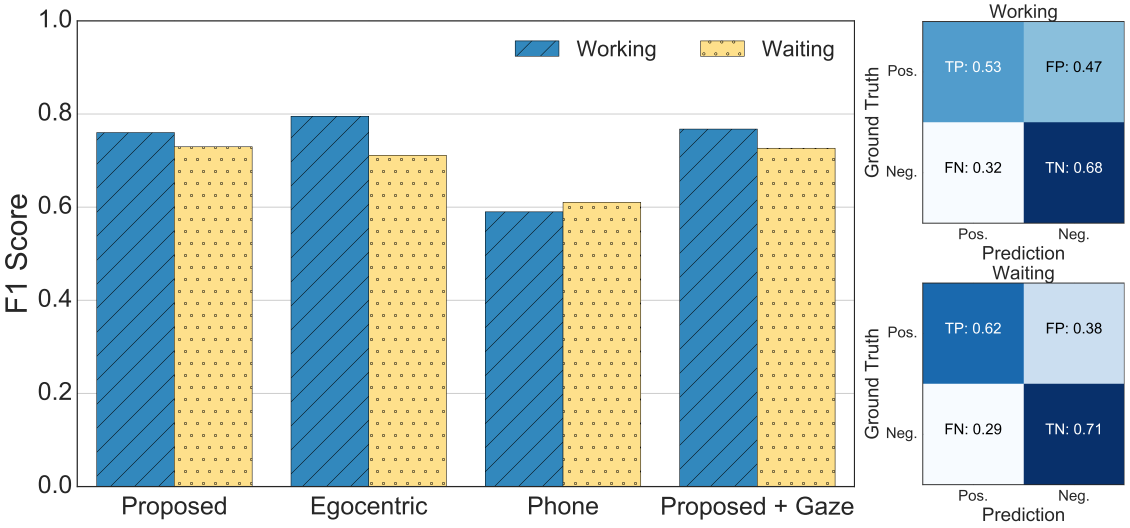


Figure 9.6: Performance for predicting *shifts to the environment* during working and waiting time segments for the different feature sets for a one-second target window, and confusion matrices for our proposed feature set.

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

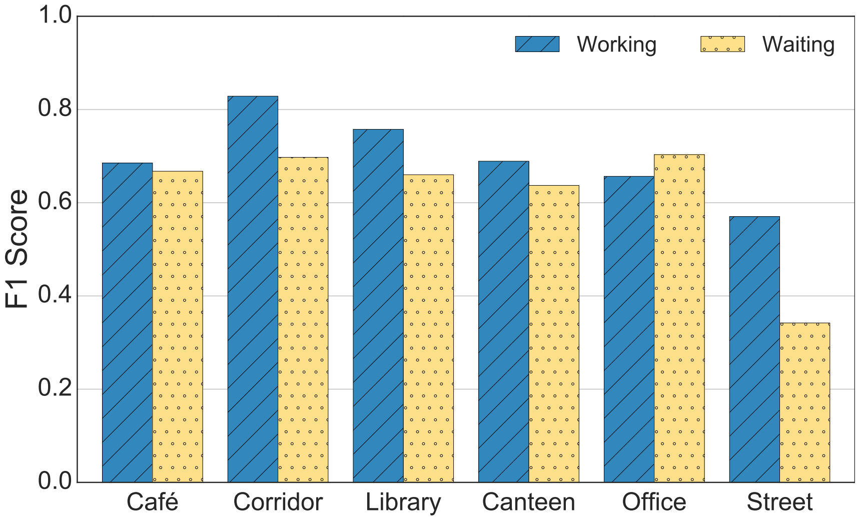
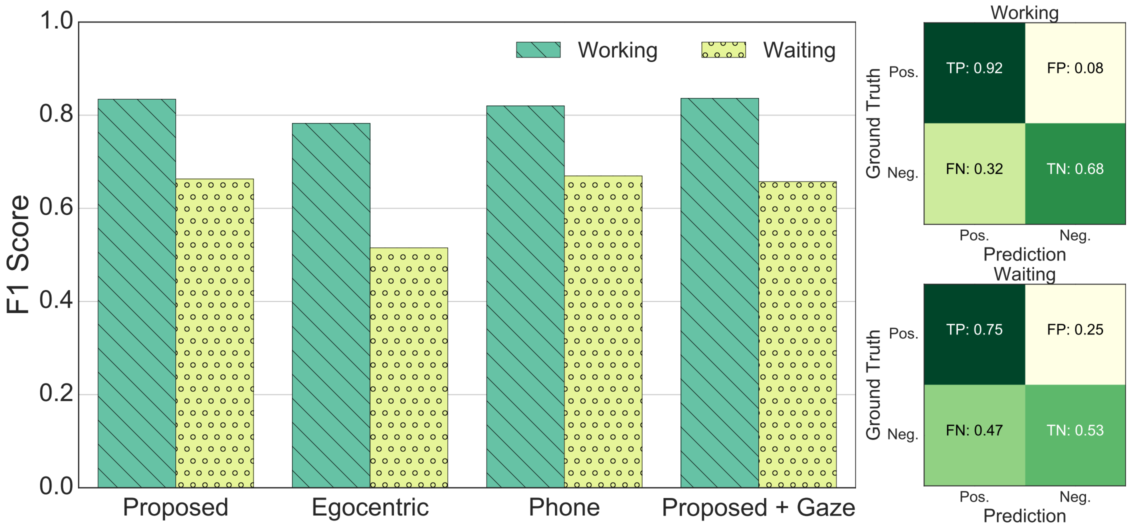


Figure 9.7: Performance for predicting shifts to the environment for different real-world environments of our proposed feature set during working and waiting time segments.



Working

	Pos.	Neg.
Pos.	TP: 0.92	FP: 0.08
Neg.	FN: 0.32	TN: 0.68

Waiting

	Pos.	Neg.
Pos.	TP: 0.75	FP: 0.25
Neg.	FN: 0.47	TN: 0.53

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.

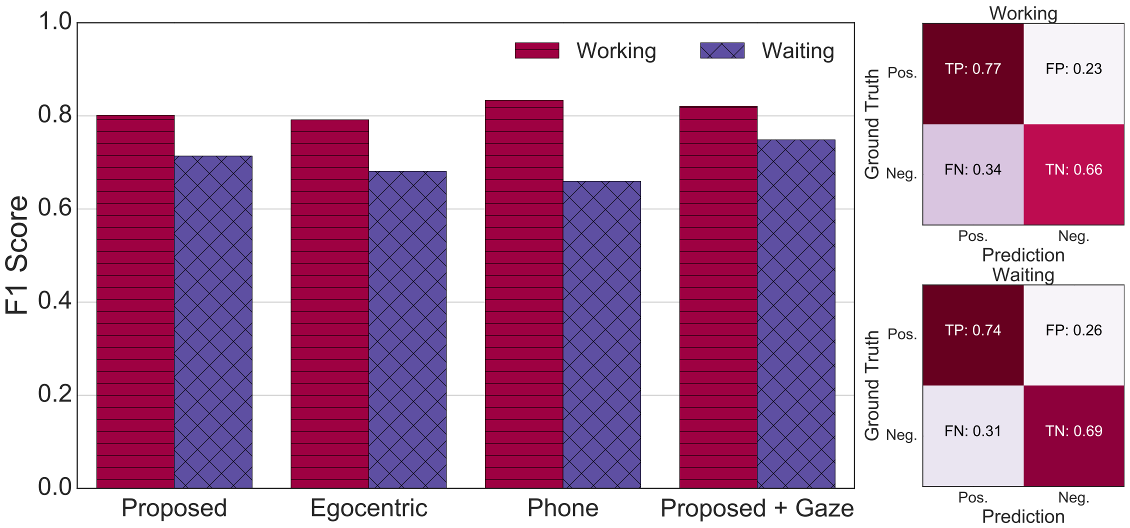


Figure 9.9: Performance for *primary attentional focus* on mobile device during working and waiting time segments for the different feature sets for a five-second target window, and confusion matrices for our proposed feature set.

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.*, 2018a) 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 of 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.

A

3D Gaze Estimation from 2D Pupil Positions on Monocular Head-Mounted Eye Trackers

A.1 3D Gaze Estimation Approaches

We first introduce detailed formulations of three approaches that are briefly presented in Chapter 3.

A.1.1 2D-to-2D Mapping Approach

As briefly described above, standard 2D gaze estimation methods assume 2D pupil positions $\boldsymbol{p}$ in the eye camera images as input, and the task is to find the polynomial mapping function from $\boldsymbol{p}$ to 2D gaze positions $\boldsymbol{s}$ in the scene camera images. 2D pupil positions are first converted into their polynomial representations $\boldsymbol{q}(\boldsymbol{p})$, and a coefficient vector $\boldsymbol{w}$ which minimises a cost function

$$E_{\mathrm{2Dto2D}}(\boldsymbol{w}) = \sum_{i=1}^{N} |\boldsymbol{s}_i - \boldsymbol{q}_i\boldsymbol{w}|^2 \tag{A.1}$$

is obtained via linear regression methods. Then any pupil positions $\boldsymbol{p}$ can be mapped to 2D gaze positions as $\boldsymbol{f} = \boldsymbol{q}\boldsymbol{w}$.

A.1.2 3D-to-3D Mapping Approach

In this case, the input to the mapping function is 3D pupil pose unit vectors $\boldsymbol{n}$. Given the calibration data $(\boldsymbol{n}_i, \boldsymbol{t}_i)_{i=1}^{N}$ with 3D calibration targets $\boldsymbol{t}$, the task is to find the rotation $\boldsymbol{R}$ and translation $\boldsymbol{T}$ between the scene and eye camera coordinate systems.

If we denote the origin of the pupil pose vectors as $\boldsymbol{e}_{cam}$, 3D gaze rays after the rotation and translation are defined as a line $\boldsymbol{e}_{cam} + \boldsymbol{T} + \lambda\boldsymbol{R}\boldsymbol{n}$, where λ parameterise the gaze line[18]. Given the calibration data, $\boldsymbol{R}$ and $\boldsymbol{T}$ are obtained by minimising distances d_i between 3D gaze targets $\boldsymbol{t}_i$ and the 3D gaze rays. In a vector form, the squared distance d_i^2 can be written as

$$\begin{aligned} d_i^2 &= \frac{|\boldsymbol{R}\boldsymbol{n}_i \times (\boldsymbol{t}_i - (\boldsymbol{e}_{cam} + \boldsymbol{T}))|^2}{|\boldsymbol{R}\boldsymbol{n}_i|^2} \\ &= |\boldsymbol{R}\boldsymbol{n}_i \times (\boldsymbol{t}_i - (\boldsymbol{e}_{cam} + \boldsymbol{T}))|^2. \end{aligned} \tag{A.2}$$

[18] Please note that λ is the parameter required to determine the 3D gaze point by intersecting the gaze ray to the scene, and does not have to be obtained during calibration stage.

Since $\boldsymbol{e}_{cam} + \boldsymbol{T}$ denotes the eyeball centre position $\boldsymbol{e}$ in the scene camera coordinate system, the cost function can be defined as

$$E_{\text{3Dto3D}}(\boldsymbol{R}, \boldsymbol{e}) = \sum_{i=1}^{N} |\boldsymbol{R}\boldsymbol{n}_i \times (\boldsymbol{t}_i - \boldsymbol{e})|^2. \tag{A.3}$$

Minimisation of Equation (A.3) can be done using nonlinear optimisation methods such as the Levenberg-Marquardt algorithm. At the initialisation step of the nonlinear optimisation, we assume $\boldsymbol{e}_0 = (0, 0, 0)$ and $\boldsymbol{R}_0 = (0, \pi, 0)$ considering the opposite direction of the scene and eye cameras in the world coordinate system.

A.1.3 2D-to-3D Mapping Approach

Another potential approach is to directly map 2D pupil positions $\boldsymbol{p}$ to 3D gaze directions $\boldsymbol{g}$. In this case, we map the polynomial feature $\boldsymbol{q}$ to unit gaze vectors $\boldsymbol{g}$ originating from the eyeball centre $\boldsymbol{e}$ in the scene camera coordinate system. $\boldsymbol{g}$ can be parameterised in a polar coordinate system as

$$\boldsymbol{g} = \begin{pmatrix} \sin\theta \\ \cos\theta \sin\phi \\ \cos\theta \cos\phi \end{pmatrix}, \tag{A.4}$$

and we assume a linear mapping from the polynomial feature $\boldsymbol{q}$ to the angle vector as

$$\boldsymbol{\alpha} = (\theta, \phi) = \boldsymbol{q}\boldsymbol{w}. \tag{A.5}$$

Given the 3D calibration data $(\boldsymbol{p}_i, \boldsymbol{t}_i)_{i=1}^{N}$, $\boldsymbol{w}$ can be obtained by minimising distances d_i between 3D gaze targets $\boldsymbol{t}_i$ and the gaze rays. Therefore, similarly to the 3D-to-3D mapping case, the target cost function to be minimised is

$$E_{\text{2Dto3D}}(\boldsymbol{w}, \boldsymbol{e}) = \sum_{i=1}^{N} |\boldsymbol{g}(\boldsymbol{q}_i\boldsymbol{w}) \times (\boldsymbol{t}_i - \boldsymbol{e})|^2. \tag{A.6}$$

In order to initialise the parameters for nonlinear optimisation, we first set $\boldsymbol{e}_0 = (0, 0, 0)$. Then using the polar coordinates of gaze targets $\boldsymbol{t}_i = (\theta_i, \phi_i)$, the initial $\boldsymbol{w_0}$ can be obtained by solving the linear regression problem

$$E(\boldsymbol{w}) = \sum_{i=1}^{N} |(\boldsymbol{\theta_i}, \boldsymbol{\phi_i}) - \boldsymbol{q}_i\boldsymbol{w}|^2. \tag{A.7}$$

A.2 Extended Analysis

In this section, we provide extended analysis on the different performance taking single and multiple calibration depth combinations into account.

A.2.1 Simulation Study

Figure A.1 shows the error for all three mapping approaches on the simulation data by fixing the calibration depth in a similar manner as in Mardanbegi and Hansen's work (Mardanbegi and Hansen, 2012). Figure A.1a and Figure A.1b are corresponding to performances using one and three calibration depth, respectively. Each plot shows the mean angular error distribution over test depths, and each colour corresponds to a certain calibration depth. The error bars describe the corresponding standard deviations. Dashed lines correspond to the 2D-to-3D mapping, dotted lines correspond to the 3D-to-3D mapping, and solid lines correspond to the 2D-to-2D mapping.

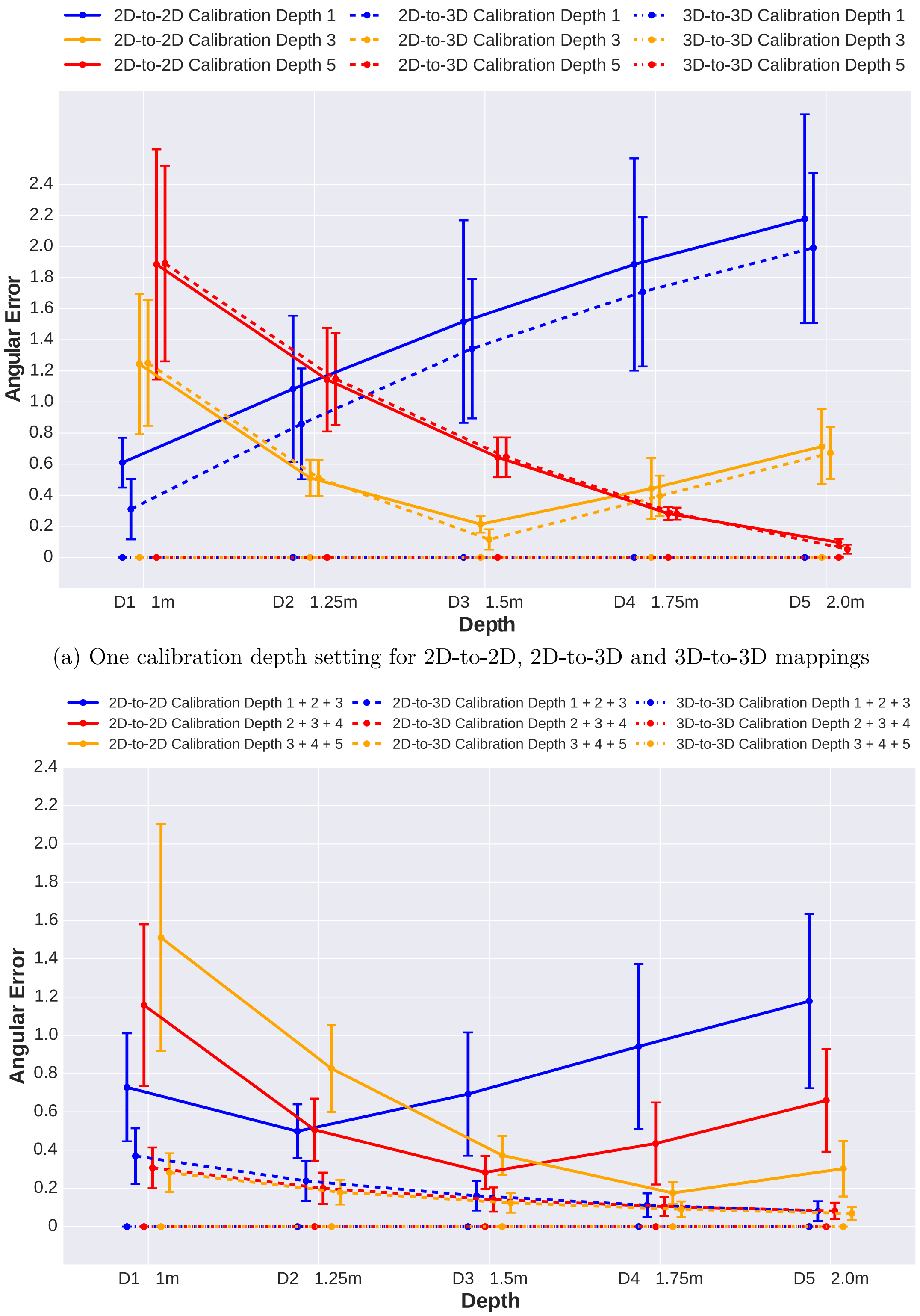
With one calibration depth (Figure A.1a), the performance of the 2D-to-3D mapping is always better than the 2D-to-2D case. However, we can observe that the parallax error is still present in the 2D-to-3D case, which indicates the fundamental limitations of the approximated mapping approach. With three calibration depth (Figure A.1b), the 2D-to-3D mapping approach performs significantly better than in Figure A.1a and the parallax error reaches a near zero level. However, there is a tendency for the error to become larger as the test depth becomes closer to the camera, which indicates the limitations of the proposed mapping function. The performance of the 2D-to-2D mapping is also improved, but we can see that the increased number of calibration depths cannot be a fundamental solution to the parallax error issue. For the 3D-to-3D mapping, the angular error is close to zero even for only one calibration depth. Taking more calibration depths into account does not lead to a further improvement.

A.2.2 Real-World Study

Similarly, we show a detailed comparison of the 2D-to-2D and 2D-to-3D mapping approaches using the real-world data. Figure A.2a displays the mean angular error for both approaches taking only one calibration depth over all 14 participants in the same manner as in Figure A.1a. For both mapping approaches, each calibration depth setting performed best for the corresponding test depth, and the error increased with an increased test distance from the calibration depth. However, for the 2D-to-2D approach the angular error values over all distances are smaller than for the 2D-to-3D case, except for the case where the calibration depth and test depth are the same.

This behaviour changes for an increasing number of calibration depths, as can be seen in Figure A.2b, where we used three different calibration depths as in Figure A.1b. The 2D-to-3D mapping approach performs better than the 2D-to-2D mapping for nearly all combinations, except for the test depth D1, exploiting the additional 3D information collected during calibration to improve the gaze direction estimation.

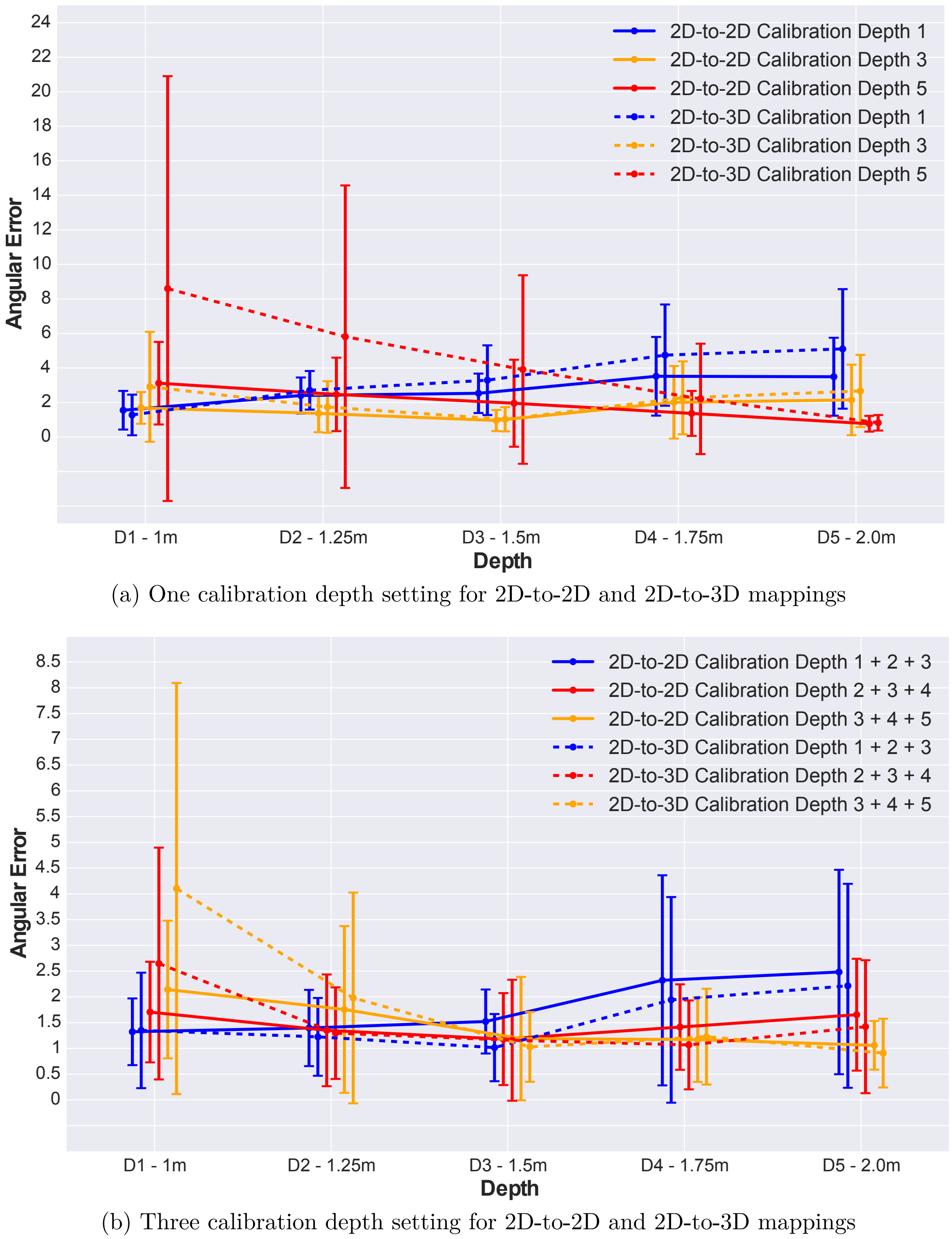
Figure A.3 shows the mean angular errors with respect to the offset between the calibration and test depths for the one calibration depth setting. The negative distance values on the horizontal axis indicate cases where the test depth is closer than the calibration depth, and vice versa for the positive distance values. As can be seen, the 2D-to-3D mapping approach tends to produce higher error if the test depth distance from the calibration depth increases.



(a) One calibration depth setting for 2D-to-2D, 2D-to-3D and 3D-to-3D mappings

(b) Three calibration depth setting for 2D-to-2D, 2D-to-3D and 3D-to-3D mappings

Figure A.1: Comparison of parallax error. Vertical axis shows the mean angular error values over all test depths (D1-D5). Dashed lines correspond to the 2D-to-3D mapping, Dotted lines correspond to 3D-to-3D and solid lines correspond to the 2D-to-2D mapping. Each colour represents one of the different calibration depth settings.



(a) One calibration depth setting for 2D-to-2D and 2D-to-3D mappings

(b) Three calibration depth setting for 2D-to-2D and 2D-to-3D mappings

Figure A.2: Comparison of one and three calibration depths considering the mean angular error values over all participants and for every test depth (D1-D5). The error bars show the corresponding standard deviation for every test depth.

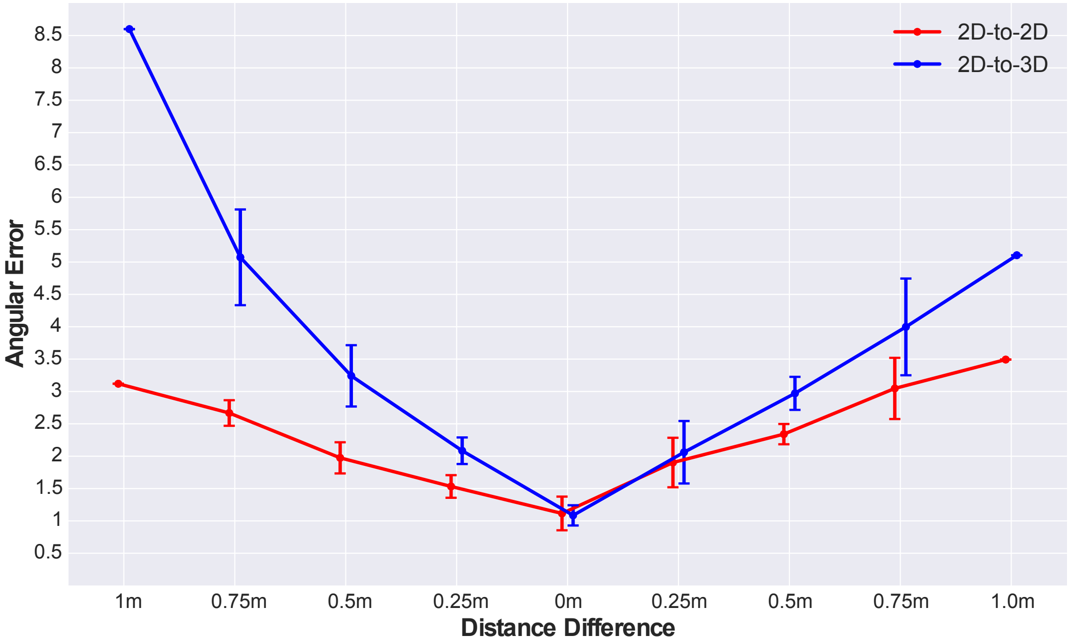


Figure A.3: The effect of distance from the calibration depth for 2D-to-2D and 2D-to-3D, taking one calibration depth into account. Every point describes the mean angular error with respect to the offset between the calibration and test depth. The error bars provide the corresponding standard deviation.

B

PrivacEye: Privacy-Preserving Head-Mounted Eye Tracking Using Egocentric Scene Image and Eye Movement Features

B.1 Data Annotation Scheme

Annotations were performed using Advene (Aubert *et al.*, 2012). Participants were asked to annotate continuous video segments showing the same situation, environment, or activity. They could also introduce new segments in case a privacy-relevant feature in the scene changed, e.g., when a participant switched to a sensitive app on the mobile phone. Participants were asked to annotate each of these segments according to the annotation scheme shown in Table B.1, specifically scene content (Q1-7) and privacy sensitivity ratings (Q8-11). Privacy sensitivity was rated on a 7-point Likert scale (see Figure B.1) ranging from 1 (fully inappropriate) to 7 (fully appropriate). As we expected our participants to have difficulties understanding the concept of "privacy sensitivity", Q8 was rephrased for the annotation to "How appropriate is it that a camera is in the scene?".

#	Question	Example Annotation
1.	What is the current environment you are in?	office, library, street, canteen
2.	Is this an indoor or outdoor environment?	indoor, outdoor
3.	What is your current activity in the video segment?	talking, texting, walking
4.	Are private objects present in the scene?	schedule, notes, wallet
5.	Are devices with potentially sensitive content present in the scene?	laptop, mobile phone
6.	Is a person present that you personally know?	yes, no
7.	Is the scene a public or a private place?	private, public, mixed
8.	How appropriate is it that a camera is in the scene?	Likert scale (1: fully inappropriate – 7: fully appropriate)
9.	How appropriate is it that a camera is continuously recording the scene?	
10.	How confident are you in a confined sharing (e.g. with friends and relatives) of the recorded imagery?	
11.	How confident are you in a public sharing of the recorded imagery?	

Table B.1: Annotation scheme used by the participants to annotate their recordings.

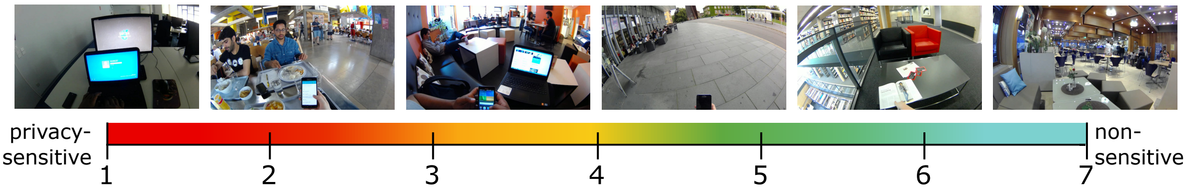


Figure B.1: Sample images showing daily situations ranging from “privacy-sensitive”, such as password entry or social interactions, to “non-sensitive”, such as walking down a road or sitting in a café.

B.2 Eye Movement Features

Table B.2 summarises the features that we extracted from fixations, saccades, blinks, pupil diameter, and a user’s scan paths. Similar to (Bulling *et al.*, 2011b), each saccade is encoded as a character forming words of length n (wordbook). We extracted these features on a sliding window of 30 seconds (step size of 1 second).

Fixation (8)	rate, mean, max, var of durations, mean/var of var pupil position within one fixation
Saccades (12)	rate/ratio of (small/large/right/left) saccades, mean, max, variance of amplitudes
Combined (1)	ratio saccades to fixations
Wordbooks (24)	number of non-zero entries, max and min entries, and their difference for n-grams with n <= 4
Blinks (3)	rate, mean/var blink duration
Pupil Diameter (4)	mean/var of mean/var during fixations

Table B.2: We extracted 52 eye movement features to describe a user’s eye movement behaviour. The number of features per category is given in parentheses.

B.3 CNN Network Architecture

Inspired by prior work on predicting privacy-sensitive pictures posted in social networks (Orekondy *et al.*, 2017), we used a pre-trained GoogleNet, a 22-layer deep convolutional neural network (Szegedy *et al.*, 2015). We adapted the original GoogleNet model for our specific prediction task by adding two additional fully connected (FC) layers (see Figure B.2). The first layer was used to reduce the feature dimensionality from 1024 to 68 and the second one, a Softmax layer, to calculate the prediction scores.

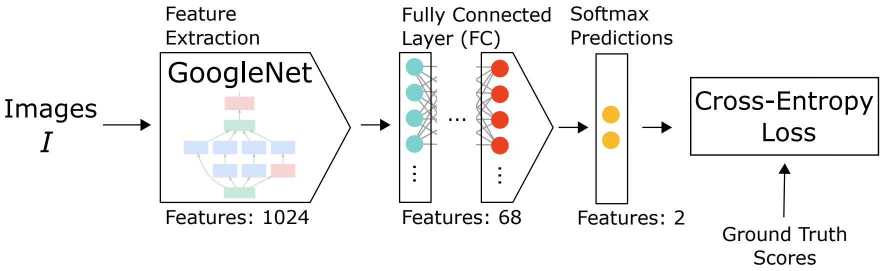


Figure B.2: Our method for detecting privacy-sensitive situations is based on a pre-trained GoogleNet model that we adapted with a fully connected (FC) and a Softmax layer. Cross-entropy loss is used for training the model.

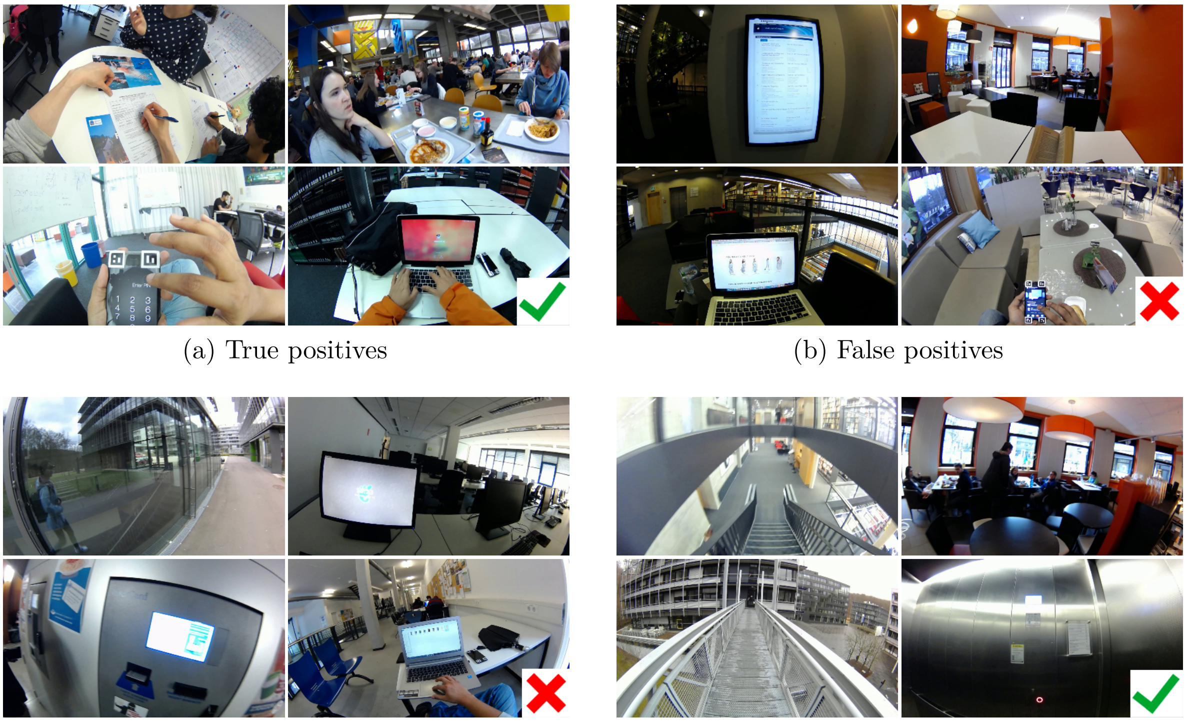
Output of our model was a score for each first-person image indicating whether the situation visible in that image was privacy-sensitive or not. The cross-entropy loss was used to train the model.

B.4 Error Case Analysis

For *PrivacEye*, it is not only important to detect the privacy-sensitive situations (TP), but equally important to detect non-sensitive situations (TN), which are relevant to grant a good user experience.

Our results suggest that the combination *SVM/SVM* performs best for the person-specific case. In the following, we detail its performance on data recorded in different environments and during different activities. We detail on the occurrence of false positives, i.e., cases where the camera is de-activated in a non-sensitive situation, as well as false negatives, i.e., cases where the camera remains active although the scene is privacy-sensitive. Examples such as in Figure B.3 show that, while false positives would be rather unproblematic in a realistic usage scenario, false negatives are critical and might lead to misclosures. Thus, our argumentation focuses on eliminating false negatives. While *PrivacEye* correctly identifies signing a document, social interactions, and screen interactions as privacy-sensitive, false positives contain reading a book or standing in front of a public display. In the latter cases *PrivacEye* would act too restrictively and the de-activation of the scene camera would lead to a loss of functionality (e.g. tracking). False negative cases include, e.g., reflections (when standing in front of a window), self-luminous screens, or cases that are under-represented in our dataset (e.g. entering a pin at the ATM).

Figure B.4 provides a detailed overview of true positives and false negatives with respect to the labelled environments (Figure B.4, left) and activities (Figure B.4, right). For each label two stacked bars are shown: *PrivacEye*'s prediction (top row) and the ground truth annotation (GT, bottom row). The prediction's result defines the "cut-off" between closed shutter (left, privacy-sensitive) and open shutter (right, non-sensitive), which is displayed as vertical bar. Segments that were predicted to be privacy-sensitive, include both true positives (TP, red) and false positives (FP, yellow-green) are shown left of the "cut-off". Similarly, those segments that were predicted to be non-sensitive,



(a) True positives

(b) False positives

(c) False negatives

(d) True negatives

Figure B.3: Examples for (a) correct detection of “privacy-sensitive” situations, (b) incorrect detection of “non-sensitive” situations, (c) incorrect detection of “privacy-sensitive” situations, and (d) correct detection of “non-sensitive” situations.

including true negatives (TN, yellow-green) and false negatives (FN, red), are displayed right of the “cut-off”. While false positives (FP) (i.e. non-sensitive situations classified as sensitive) are not problematic, as they do not create the risk of misclosures, false negatives (FN) are critical. Thus, we focus our discussion on the false negatives (red, top, right). A comparison of true positives (TP) and false negatives (FN) shows that *PrivacEye* performs well within most environments, e.g., offices or corridors. In these environments true positives outweigh false negatives. However, in the computer room environment, where a lot of screens with potentially problematic content (which the wearer might not even be aware of at recording time) are present, performance drops. Misclassifications between personal displays, e.g., laptops and public displays (e.g. room occupancy plans) are a likely reason for the larger amount of false negatives (FN). Future work might aim to combine *PrivacEye* with an image-based classifier trained for screen contents (cf. (Korayem *et al.*, 2016)), which, however, would come at the cost of excluding also non-sensitive screens from the footage. Future work might specifically target these situations to increase accuracy. For the activities outlined in Figure B.4 (right), *PrivacEye* works best while eating/drinking and in media interactions. Also, the results are promising for detecting social interactions. The performance for password entry, however, is still limited. Although the results show that it is possible to detect password entry, the amount of false negatives (FN) is comparatively high. This is likely caused by the under-representation of this activity, which typically lasts only a few

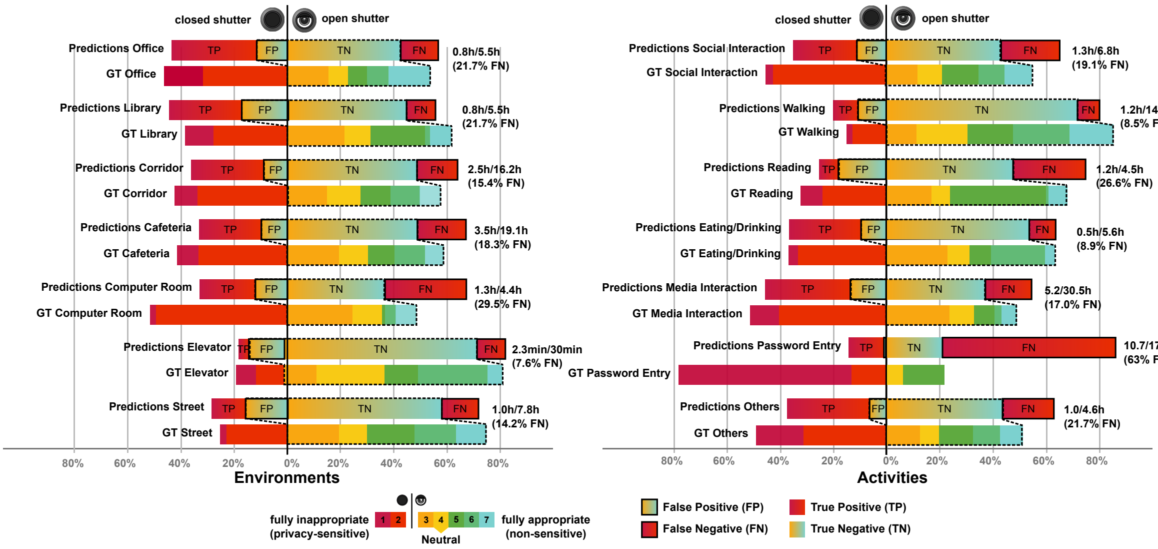


Figure B.4: Error case analysis for different environments (left) and activities (right) showing the "cut-off" between closed shutter (left, *privacy-sensitive*) and open shutter (right, *non-sensitive*) with *PrivacEye* prediction and the corresponding ground truth (GT). False positives (FP) are *non-sensitive* but protected (closed shutter), false negatives (FN) are *privacy-sensitive* but unprotected (open shutter).

seconds in our dataset. Future work might be able to eliminate this by specifically training for password and PIN entry, possibly enabling the classifier to better distinguish between PIN entry and, e.g., reading.

B.5 Interview Protocol

During the interviews, participants were encouraged to interact with state-of-the-art head-mounted displays (Vuzix M300 and Sony SmartEyeglass) and our prototype. Participants were presented with the fully functional *PrivacEye* prototype, which was used to illustrate three scenarios: 1) interpersonal conversations, 2) sensitive objects (a credit card and a passport), and 3) sensitive contents on a device screen. Due to the time required to gather person-specific training data for each interviewee as well as runtime restrictions, the scenarios were presented using the Wizard-of-Oz method. This is also advantageous, as the laboratory-style study environment – with white walls, an interviewer and no distractors present – might have induced different eye movement patterns than a natural environment. Also, potential errors of the system, caused by its prototypical implementation, might have caused participant bias toward the concept. To prevent these issues, the shutter was controlled remotely by an experimental assistant. This way, the interviewees commented on the concept and vision of *PrivacEye* and not on the actual proof-of-concept implementation, which – complementing the afore-described evaluation – provides a more comprehensive and universal set of results altogether. The semi-structured interview was based on the following questions:

Q1 Would you be willing to wear something that would block someone from being able to record you?

Q2 If technically feasible, would you expect the devices themselves, instead of their user, to protect your privacy automatically?

Q3 Would you feel different about being around someone who is wearing those kinds of intelligent glasses than about those commercially available today? Why?

Q4 If you were using AR glasses, would you be concerned about accidentally recording any sensitive information belonging to you?

Q5 How would you feel about (such) a system automatically taking care that you do not capture any sensitive information?

Q6 How do you think the eye tracking works? What can the system infer from your eye data?

Q7 How would you feel about having your eye movements tracked by augmented reality glasses?

The questions were designed following a "funnel principle", with increasing specificity towards the end of the interview. We started with four more general questions (not listed above), such as "Do you think recording with those glasses is similar or different to recording with a cell phone? Why?", based on (Denning *et al.*, 2014). This provided the participant with some time to familiarise herself with the topic before being presented with the proof-of-concept prototype (use case "bystander privacy") after Q1 and the use cases "sensitive objects" (e.g., credit card, passport) and "sensitive data" (e.g. login data) after Q4. Eye tracking functionality was demonstrated after Q5. While acquiescence and other forms of interviewer effects cannot be ruled out completely, this step-by-step presentation of the prototype and its scenarios ensured that the participants voiced their own ideas first, before being directed towards discussing the actual concept of the *PrivacEye* prototype. Each participant was asked for his/her perspectives on the *PrivacEye*'s concept (Q2-Q5) and eye tracking (Q6 and Q7). The interviews were audio recorded and transcribed for later analysis. Subsequently, qualitative analysis was performed following inductive category development (Mayring, 2014).

Privacy-Aware Eye Tracking Using Differential Privacy

C.1 Survey Results

We conducted a large-scale online survey to shed light on users' privacy concerns related to eye tracking technology and the information that can be inferred from eye movement data. We advertised our survey on social platforms (Facebook, WeChat) and local mailing lists for study announcements. The survey opened with general questions about eye tracking and VR technologies; continued with questions about future use and applications, data sharing and privacy (especially regarding with whom users are willing to share their data); and concluded with questions about the participants' willingness to share different eye movement representations. Participants answered each question on a 7-point Likert scale (1: Strongly disagree to 7: Strongly agree). To simplify the analysis, we merged scores 1 to 3 to "Disagree" and 5 to 7 to "Agree". At the end we asked for demographic information and offered a raffle.

C.1.1 Data Representation

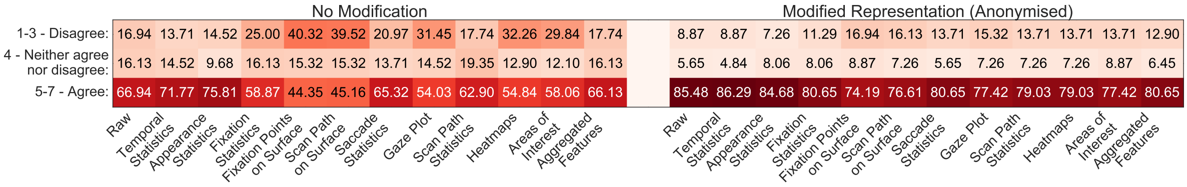


Figure C.1: Survey results (Data Representations): What kind of data representation would you agree to share (No Modification), and does this behaviour change if the data is anonymised prior to sharing (Anonymised)?

C.2 Statistical Tests

The following tables correspond to the Figures 7.2, 7.3, and C.1. We found nearly all answers for the provided questions to be significantly different from equal distribution tested with Pearson's chi-squared test ($p < 0.001$, $dof = 6$). Additionally, we calculated the skewness and observed that the majority of questions show a significant difference to the corresponding normal distribution ($p < 0.1$).

C.2.1 Statistical Tests Corresponding to Figure 7.2

	Services												Private Attributes					
	Diseases Detection	Natural VR Interaction	Visual Search Target Detection	User Interface Interaction	Understandable Website Content	Reading Skill Improvement	Learning Skill Improvement	Stress Level Monitoring	Interest Identification	Activity Recognition	Shopping Assistance		Sexual Preference	Gender	Age	Mood and Emotions	Race	Identity
chi-squared	91.19	54.84	25.48	43.1	24.92	55.63	68.84	40.73	73.58	19.5	114.9		138.39	51.0	32.48	28.87	74.15	189.31
p-value	1.7e-17	5.0e-10	2.8e-4	1.1e-7	3.5e-4	3.5e-10	7.1e-13	3.3e-7	7.5e-14	3.4e-3	1.9e-22		2.2e-27	3.0e-9	1.3e-5	6.4e-5	5.8e-14	3.6e-38
z-score skew	-4.64	-3.07	-0.94	-3.15	0.54	-3.39	-4.05	-3.03	3.45	0.79	4.28		4.31	0.71	-1.04	-0.35	2.99	4.79
p-value skew	3.5e-6	2.1e-3	0.35	1.6e-3	0.59	7.0e-4	5.0e-5	2.5e-3	5.5e-4	0.43	1.9e-5		1.6e-5	0.48	0.30	0.72	2.8e-3	1.7e-6

C.2.2 Statistical Tests Corresponding to Figure 7.3

	Sharing		Owner										Environment				Application	
	Eye Tracking Data		Governmental Agency (non-health)	Governmental Health Authority	Local Company	International Company	Private Company (user's country)	Private Company (foreign country)	User Himself (home cloud)	Company Internal Use (intranet)	Research Institute		Public	Private	Constrained		In Exchange for Benefits	VR/AR
chi-squared	26.73		44.79	22.89	43.1	55.97	61.39	77.42	72.0	26.16	88.94		39.03	30.0	31.81		40.5	32.26
p-value	1.6e-4		5.2e-8	8.4e-4	1.1e-7	3.0e-10	2.4e-11	1.2e-14	1.6e-13	2.1e-4	5.0e-17		7.1e-7	4.0e-5	1.8e-5		3.6e-7	1.5e-5
z-score skew	-0.97		1.59	-1.53	1.5	1.8	1.76	2.06	-4.58	1.55	-4.95		2.1	1.7	-1.97		2.55	-1.99
p-value skew	0.33		0.11	0.13	0.13	0.07	0.08	0.04	4.7e-6	0.12	7.5e-7		0.04	0.09	0.05		0.01	0.05

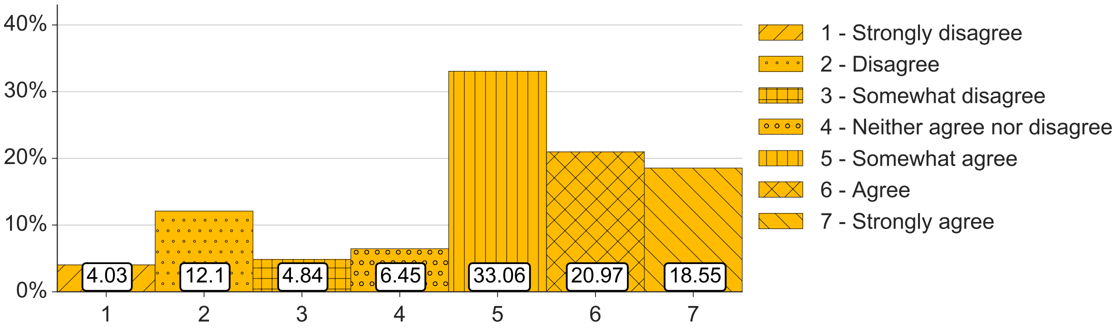
C.2.3 Statistical Tests Corresponding to Figure C.1

	No Modification													Modified Representation (Anonymised)											
	Raw	Temporal Statistics	Appearance Statistics	Fixation Statistics	Fixation Points on Surface	Scan Path on Surface	Saccade Statistics	Gaze Plot	Scan Path Statistics	Heatmaps	Areas of Interest	Aggregated Features		Raw	Temporal Statistics	Appearance Statistics	Fixation Statistics	Fixation Points on Surface	Scan Path on Surface	Saccade Statistics	Gaze Plot	Scan Path Statistics	Heatmaps	Areas of Interest	Aggregated Features
chi-squared	41.74	57.1	63.76	29.77	13.85	9.79	40.05	21.76	48.52	19.61	26.16	49.31		95.26	96.61	91.08	74.71	51.45	58.68	75.84	61.39	68.5	68.16	62.74	73.24
p-value	2.1e-7	1.8e-10	7.7e-12	4.3e-5	3.1e-2	1.3e-1	4.5e-7	1.3e-3	9.3e-9	3.2e-3	2.1e-4	6.5e-9		2.4e-18	1.3e-18	1.8e-17	4.4e-14	2.4e-9	8.4e-11	2.6e-14	2.4e-11	8.3e-13	9.7e-13	1.3e-11	8.8e-14
z-score skew	-3.25	-3.78	-3.98	-2.13	-0.78	-0.9	-2.98	-1.87	-2.9	-1.84	-2.12	-3.5		-4.8	-4.45	-4.28	-4.44	-3.82	-3.91	-3.77	-4.0	-4.24	-4.14	-3.98	-4.5
p-value skew	1.1e-3	1.6e-4	7.0e-5	0.03	0.44	0.37	2.9e-3	0.06	3.7e-3	0.07	0.03	4.7e-4		1.6e-6	8.6e-6	1.9e-5	9.1e-6	1.3e-4	9.4e-5	1.6e-4	6.4e-5	2.2e-5	3.4e-5	6.9e-5	6.6e-6

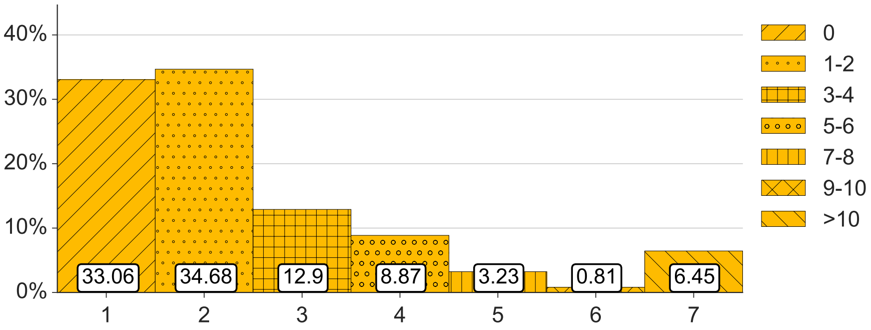
C.3 Survey Evaluation

C.3.1 Eye Tracking and Virtual Reality (VR) Technologies

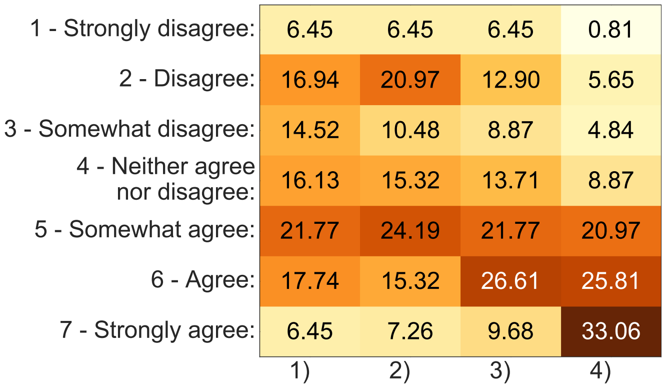
1. I am familiar with eye tracking technology.



2. How many eye tracking applications or experiments have you used or participated?



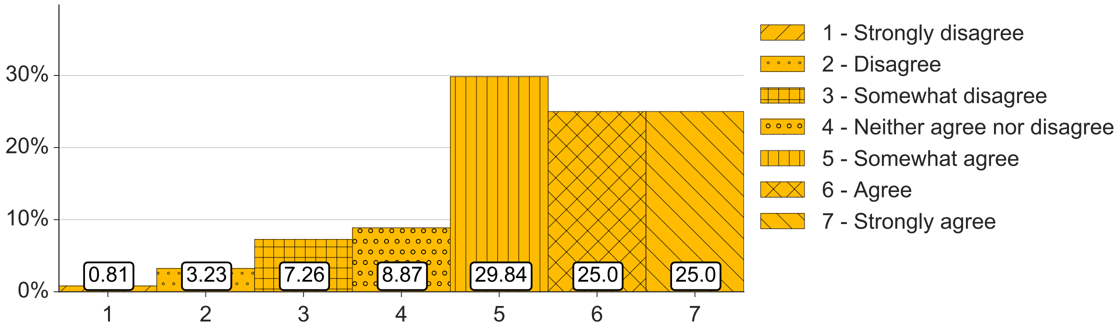
3. I am concerned about eye tracking technology in terms of ...



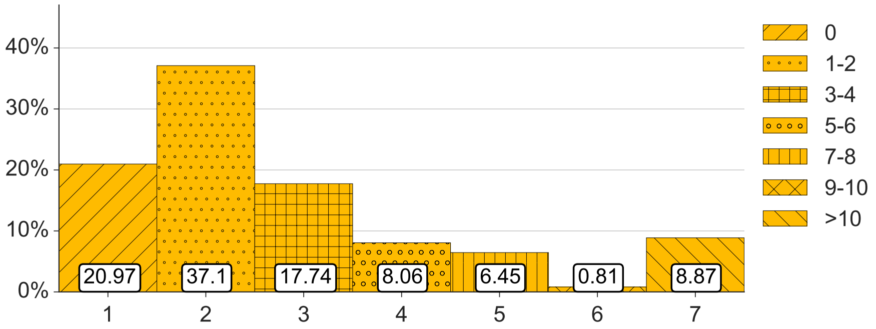
	1)	2)	3)	4)
1 - Strongly disagree:	6.45	6.45	6.45	0.81
2 - Disagree:	16.94	20.97	12.90	5.65
3 - Somewhat disagree:	14.52	10.48	8.87	4.84
4 - Neither agree nor disagree:	16.13	15.32	13.71	8.87
5 - Somewhat agree:	21.77	24.19	21.77	20.97
6 - Agree:	17.74	15.32	26.61	25.81
7 - Strongly agree:	6.45	7.26	9.68	33.06

1) social acceptability (e.g.: How I am perceived by other people?)
2) mental comfortability (e.g.: increase/decrease mental workload)
3) physical comfortability (e.g.: increase/decrease physical workload)
4) privacy

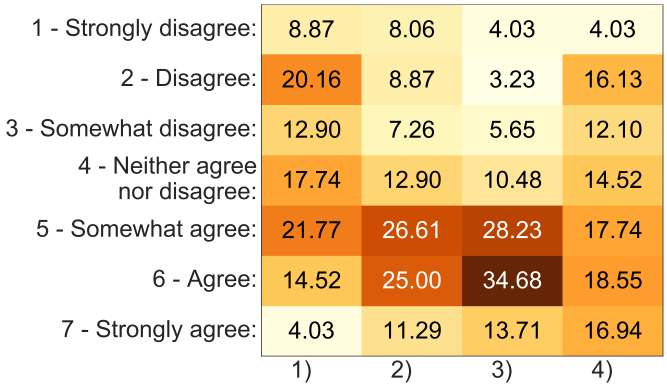
4. I am familiar with virtual reality (VR) and augmented reality (AR) technology.



5. How many VR applications or experiments have you used or participated?



6. I am concerned about VR technology in terms of ...

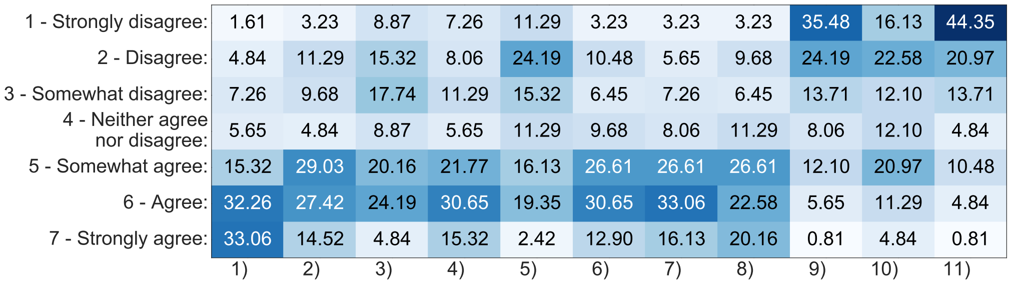


	1)	2)	3)	4)
1 - Strongly disagree:	8.87	8.06	4.03	4.03
2 - Disagree:	20.16	8.87	3.23	16.13
3 - Somewhat disagree:	12.90	7.26	5.65	12.10
4 - Neither agree nor disagree:	17.74	12.90	10.48	14.52
5 - Somewhat agree:	21.77	26.61	28.23	17.74
6 - Agree:	14.52	25.00	34.68	18.55
7 - Strongly agree:	4.03	11.29	13.71	16.94

1) social acceptability (e.g.: How I am perceived by other people?)
2) mental comfortability (e.g.: increase/decrease mental workload)
3) physical comfortability (e.g.: increase/decrease physical workload)
4) privacy

C.3.2 Future Use of Eye Tracking Data

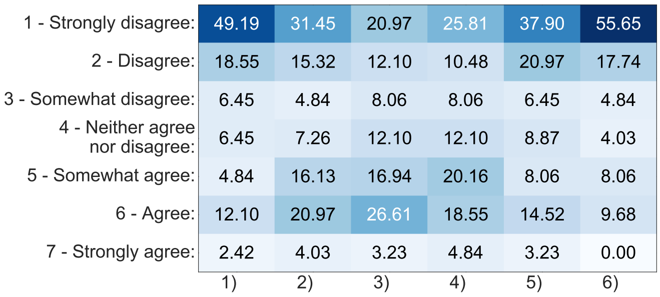
1. Would you agree to share eye tracking data ...



1 - Strongly disagree:	1.61	3.23	8.87	7.26	11.29	3.23	3.23	3.23	35.48	16.13	44.35
2 - Disagree:	4.84	11.29	15.32	8.06	24.19	10.48	5.65	9.68	24.19	22.58	20.97
3 - Somewhat disagree:	7.26	9.68	17.74	11.29	15.32	6.45	7.26	6.45	13.71	12.10	13.71
4 - Neither agree nor disagree:	5.65	4.84	8.87	5.65	11.29	9.68	8.06	11.29	8.06	12.10	4.84
5 - Somewhat agree:	15.32	29.03	20.16	21.77	16.13	26.61	26.61	26.61	12.10	20.97	10.48
6 - Agree:	32.26	27.42	24.19	30.65	19.35	30.65	33.06	22.58	5.65	11.29	4.84
7 - Strongly agree:	33.06	14.52	4.84	15.32	2.42	12.90	16.13	20.16	0.81	4.84	0.81
	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)

1) for early detection of mental and psychological disease like dementia or Parkinson?
2) to enable hands-free interaction with displays and in virtual reality? You could type letters or select a button by gaze interaction or interact with items or persons in VR more naturally.
3) to identify the target of your visual search to provide information about the target you are looking at (e.g. the name of a person, information about a product, etc.)?
4) to improve interactions with user interfaces and devices, e.g. to make them more intuitive or faster?
5) to allow apps and websites to provide content easy to understand?
6) to analyse reading ability and propose methods to improve your reading skills or to change reading material in terms of appearance (enlarging text, highlighting current line)?
7) to analyse and improve your learning skills?
8) monitor your stress level and to provide early-stage healthcare intervention?
9) to identify your interests, e.g. what you like or dislike, and guide you like a shopping assistance, or to steer advertisement?
10) to identify activity specific patterns which could be used for activity tracking, lifelogging or self-quantifying, (e.g. reading, watching TV, playing a video game, computer work, etc.)?
11) to analyse your shopping behaviour on websites or within shopping malls to improve product placement?

2. Would you agree to share eye tracking data to identify your ... for a better service (e.g. entertainment, news, business, education, etc.)?

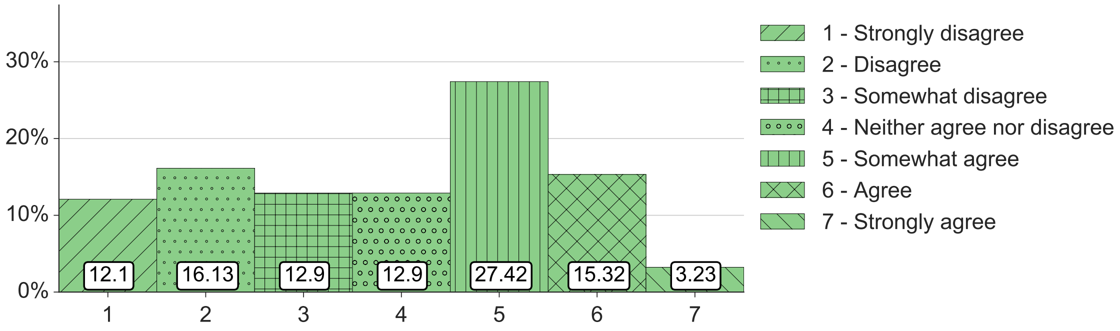


1 - Strongly disagree:	49.19	31.45	20.97	25.81	37.90	55.65
2 - Disagree:	18.55	15.32	12.10	10.48	20.97	17.74
3 - Somewhat disagree:	6.45	4.84	8.06	8.06	6.45	4.84
4 - Neither agree nor disagree:	6.45	7.26	12.10	12.10	8.87	4.03
5 - Somewhat agree:	4.84	16.13	16.94	20.16	8.06	8.06
6 - Agree:	12.10	20.97	26.61	18.55	14.52	9.68
7 - Strongly agree:	2.42	4.03	3.23	4.84	3.23	0.00
	1)	2)	3)	4)	5)	6)

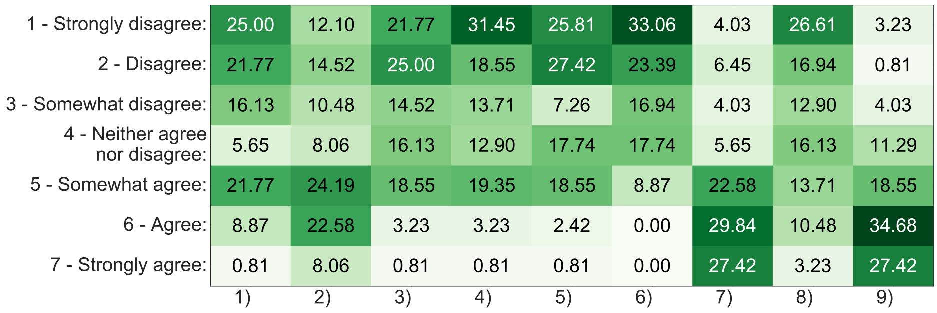
1) sexual preferences
2) gender
3) age
4) mood and emotions
5) race
6) identity

C.3.3 Sharing Eye Tracking Data

1. Would you share your eye tracking data?



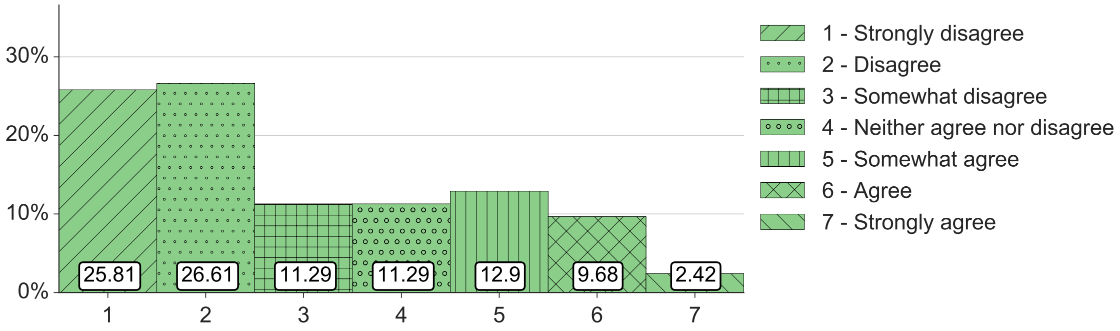
2. In general, I would trust a manufacturer and I am willing to share my eye tracking data if it is operated/owned by ...



	1)	2)	3)	4)	5)	6)	7)	8)	9)
1 - Strongly disagree:	25.00	12.10	21.77	31.45	25.81	33.06	4.03	26.61	3.23
2 - Disagree:	21.77	14.52	25.00	18.55	27.42	23.39	6.45	16.94	0.81
3 - Somewhat disagree:	16.13	10.48	14.52	13.71	7.26	16.94	4.03	12.90	4.03
4 - Neither agree nor disagree:	5.65	8.06	16.13	12.90	17.74	17.74	5.65	16.13	11.29
5 - Somewhat agree:	21.77	24.19	18.55	19.35	18.55	8.87	22.58	13.71	18.55
6 - Agree:	8.87	22.58	3.23	3.23	2.42	0.00	29.84	10.48	34.68
7 - Strongly agree:	0.81	8.06	0.81	0.81	0.81	0.00	27.42	3.23	27.42

1) a governmental agency (non-health related).
2) a governmental health authority (e.g., city, state/province, federal/national).
3) a recognised local company.
4) a recognised international company.
5) a recognised private company in user's country.
6) a recognised private company in foreign country.
7) the user himself (home cloud).
8) company internal use (intranet).
9) research institute.

3. Would you share your eye tracking data in exchange for benefits like shopping assistance, activity logging, etc.?



4. Would you share your eye tracking data if the data was collected in …

	1)	2)	3)
1 - Strongly disagree:	25.81	22.58	7.26
2 - Disagree:	24.19	23.39	12.10
3 - Somewhat disagree:	13.71	12.10	12.90
4 - Neither agree nor disagree:	9.68	16.94	13.71
5 - Somewhat agree:	15.32	13.71	26.61
6 - Agree:	10.48	9.68	22.58
7 - Strongly agree:	0.81	1.61	4.84

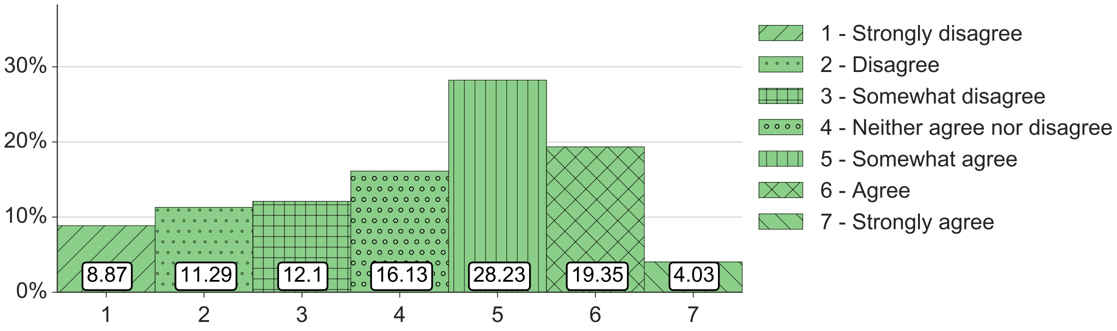
1) public, e.g. train station or park?
2) a private environment, e.g. office or home?
3) constrained environments, e.g. a specific room or place?

5. Would you share eye tracking data if the data was collected in one of the following places?

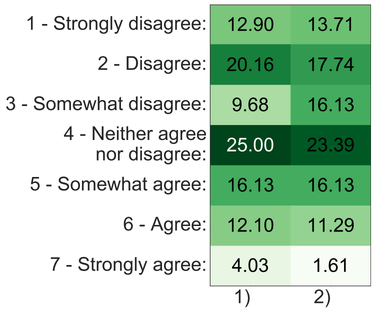
	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)
1 - Strongly disagree:	29.03	25.81	25.00	23.39	16.13	60.48	29.84	12.10	32.26	33.06	34.68
2 - Disagree:	25.00	25.00	22.58	20.97	20.97	19.35	26.61	15.32	19.35	20.97	20.97
3 - Somewhat disagree:	9.68	20.97	14.52	8.06	7.26	9.68	10.48	12.10	10.48	10.48	8.87
4 - Neither agree nor disagree:	5.65	8.06	9.68	14.52	13.71	4.84	8.06	11.29	10.48	9.68	6.45
5 - Somewhat agree:	20.16	11.29	14.52	17.74	20.16	1.61	12.90	18.55	14.52	12.90	12.10
6 - Agree:	8.06	8.06	12.10	10.48	15.32	3.23	10.48	19.35	11.29	10.48	14.52
7 - Strongly agree:	2.42	0.81	1.61	4.84	6.45	0.81	1.61	11.29	1.61	2.42	2.42

1) department store
2) friend's home
3) public transport
4) home
5) library
6) restroom
7) workplace
8) car
9) lobby (e.g. hotel)
10) cafe
11) street

6. Would you share your eye tracking data if the data was collected in VR or AR?



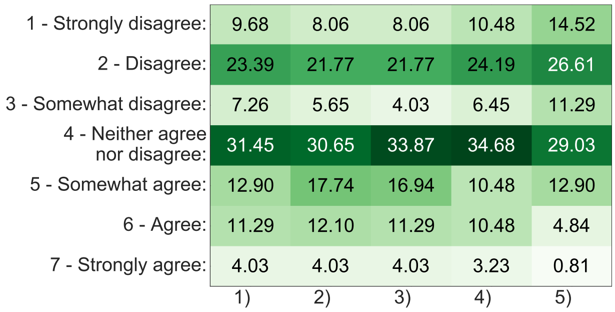
7. Would you share your eye tracking data if the data was collected ...



	1)	2)
1 - Strongly disagree:	12.90	13.71
2 - Disagree:	20.16	17.74
3 - Somewhat disagree:	9.68	16.13
4 - Neither agree nor disagree:	25.00	23.39
5 - Somewhat agree:	16.13	16.13
6 - Agree:	12.10	11.29
7 - Strongly agree:	4.03	1.61

1) indoor?
2) outdoor?

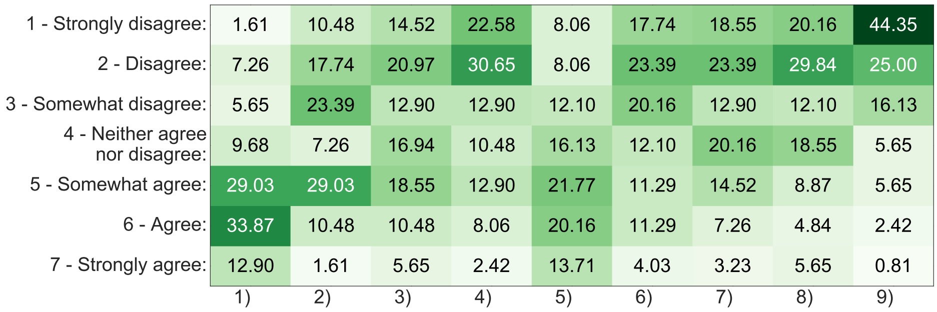
8. Would you share eye tracking data if the collected data was recorded in the ...



	1)	2)	3)	4)	5)
1 - Strongly disagree:	9.68	8.06	8.06	10.48	14.52
2 - Disagree:	23.39	21.77	21.77	24.19	26.61
3 - Somewhat disagree:	7.26	5.65	4.03	6.45	11.29
4 - Neither agree nor disagree:	31.45	30.65	33.87	34.68	29.03
5 - Somewhat agree:	12.90	17.74	16.94	10.48	12.90
6 - Agree:	11.29	12.10	11.29	10.48	4.84
7 - Strongly agree:	4.03	4.03	4.03	3.23	0.81

1) morning?
2) noon?
3) afternoon?
4) evening?
5) night?

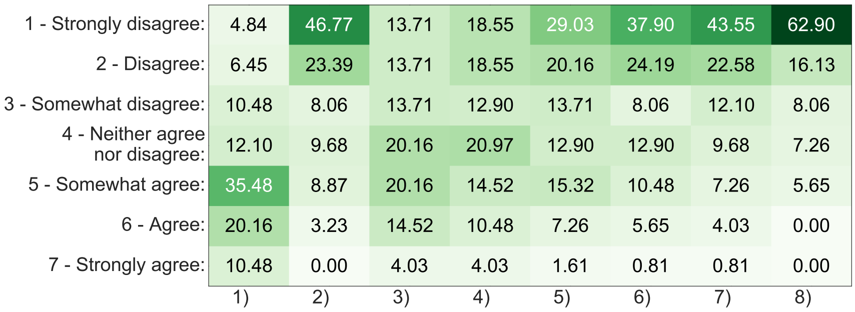
9. Would you share eye tracking data if the recording duration was restricted to one of the following options?



	1)	2)	3)	4)	5)	6)	7)	8)	9)
1 - Strongly disagree:	1.61	10.48	14.52	22.58	8.06	17.74	18.55	20.16	44.35
2 - Disagree:	7.26	17.74	20.97	30.65	8.06	23.39	23.39	29.84	25.00
3 - Somewhat disagree:	5.65	23.39	12.90	12.90	12.10	20.16	12.90	12.10	16.13
4 - Neither agree nor disagree:	9.68	7.26	16.94	10.48	16.13	12.10	20.16	18.55	5.65
5 - Somewhat agree:	29.03	29.03	18.55	12.90	21.77	11.29	14.52	8.87	5.65
6 - Agree:	33.87	10.48	10.48	8.06	20.16	11.29	7.26	4.84	2.42
7 - Strongly agree:	12.90	1.61	5.65	2.42	13.71	4.03	3.23	5.65	0.81

1) for a specific application with user allowance
2) automatic data recording if eye tracking data is necessary for usage
3) selected hours per day
4) during work time (at work)
5) for personal use
6) during free time
7) during work days
8) during weekend
9) whole day recording

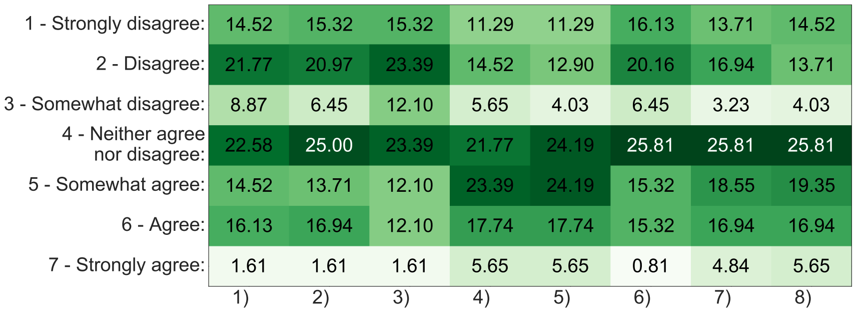
10. Would you share eye tracking data if the collected data was saved for ...



	1)	2)	3)	4)	5)	6)	7)	8)
1 - Strongly disagree:	4.84	46.77	13.71	18.55	29.03	37.90	43.55	62.90
2 - Disagree:	6.45	23.39	13.71	18.55	20.16	24.19	22.58	16.13
3 - Somewhat disagree:	10.48	8.06	13.71	12.90	13.71	8.06	12.10	8.06
4 - Neither agree nor disagree:	12.10	9.68	20.16	20.97	12.90	12.90	9.68	7.26
5 - Somewhat agree:	35.48	8.87	20.16	14.52	15.32	10.48	7.26	5.65
6 - Agree:	20.16	3.23	14.52	10.48	7.26	5.65	4.03	0.00
7 - Strongly agree:	10.48	0.00	4.03	4.03	1.61	0.81	0.81	0.00

1) application specific purpose (e.g. direct application feedback, VR interaction and gaming, etc.)?
2) an unspecified amount of time?
3) a day?
4) a week?
5) a month?
6) a half year?
7) a year?
8) forever?

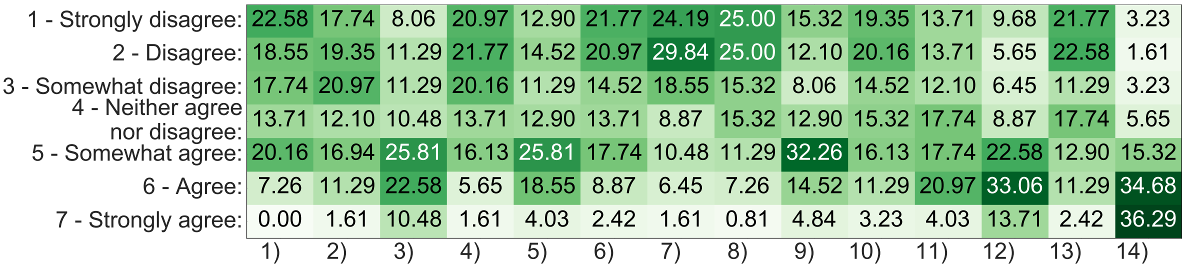
11. Would you share eye tracking data if the data was collected during one of the following emotions?



	1)	2)	3)	4)	5)	6)	7)	8)
1 - Strongly disagree:	14.52	15.32	15.32	11.29	11.29	16.13	13.71	14.52
2 - Disagree:	21.77	20.97	23.39	14.52	12.90	20.16	16.94	13.71
3 - Somewhat disagree:	8.87	6.45	12.10	5.65	4.03	6.45	3.23	4.03
4 - Neither agree nor disagree:	22.58	25.00	23.39	21.77	24.19	25.81	25.81	25.81
5 - Somewhat agree:	14.52	13.71	12.10	23.39	24.19	15.32	18.55	19.35
6 - Agree:	16.13	16.94	12.10	17.74	17.74	15.32	16.94	16.94
7 - Strongly agree:	1.61	1.61	1.61	5.65	5.65	0.81	4.84	5.65

1) fear
2) anger
3) sadness
4) joy
5) surprise
6) disgust
7) trust
8) anticipation

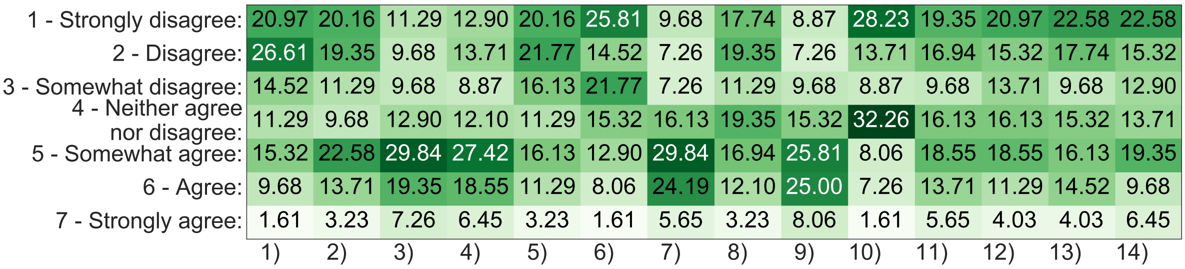
12. Would you share eye tracking data if it was collected to run an application of the following categories?



	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)	12)	13)	14)
1 - Strongly disagree:	22.58	17.74	8.06	20.97	12.90	21.77	24.19	25.00	15.32	19.35	13.71	9.68	21.77	3.23
2 - Disagree:	18.55	19.35	11.29	21.77	14.52	20.97	29.84	25.00	12.10	20.16	13.71	5.65	22.58	1.61
3 - Somewhat disagree:	17.74	20.97	11.29	20.16	11.29	14.52	18.55	15.32	8.06	14.52	12.10	6.45	11.29	3.23
4 - Neither agree nor disagree:	13.71	12.10	10.48	13.71	12.90	13.71	8.87	15.32	12.90	15.32	17.74	8.87	17.74	5.65
5 - Somewhat agree:	20.16	16.94	25.81	16.13	25.81	17.74	10.48	11.29	32.26	16.13	17.74	22.58	12.90	15.32
6 - Agree:	7.26	11.29	22.58	5.65	18.55	8.87	6.45	7.26	14.52	11.29	20.97	33.06	11.29	34.68
7 - Strongly agree:	0.00	1.61	10.48	1.61	4.03	2.42	1.61	0.81	4.84	3.23	4.03	13.71	2.42	36.29

1) utilities (e.g. taxi app, bank app, etc.)
2) entertainment (e.g. streaming, chatting, watching videos, etc.)
3) games (e.g. VR)
4) news
5) productivity
6) lifestyle
7) social networking
8) business
9) education/parenting
10) travel
11) book
12) health/medical and fitness
13) food and drink
14) research (anonymised data storage)

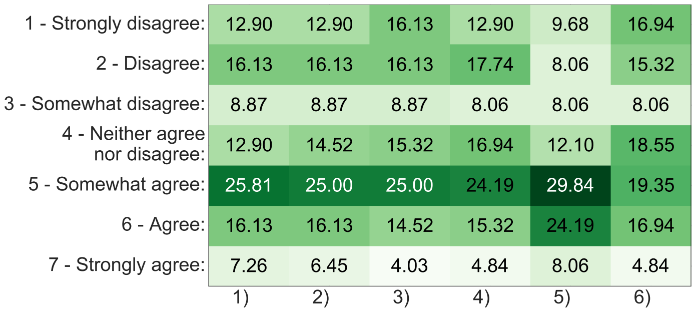
13. Would you share eye tracking data if it was collected during one of the following activities?



	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)	12)	13)	14)
1 - Strongly disagree:	20.97	20.16	11.29	12.90	20.16	25.81	9.68	17.74	8.87	28.23	19.35	20.97	22.58	22.58
2 - Disagree:	26.61	19.35	9.68	13.71	21.77	14.52	7.26	19.35	7.26	13.71	16.94	15.32	17.74	15.32
3 - Somewhat disagree:	14.52	11.29	9.68	8.87	16.13	21.77	7.26	11.29	9.68	8.87	9.68	13.71	9.68	12.90
4 - Neither agree nor disagree:	11.29	9.68	12.90	12.10	11.29	15.32	16.13	19.35	15.32	32.26	16.13	16.13	15.32	13.71
5 - Somewhat agree:	15.32	22.58	29.84	27.42	16.13	12.90	29.84	16.94	25.81	8.06	18.55	18.55	16.13	19.35
6 - Agree:	9.68	13.71	19.35	18.55	11.29	8.06	24.19	12.10	25.00	7.26	13.71	11.29	14.52	9.68
7 - Strongly agree:	1.61	3.23	7.26	6.45	3.23	1.61	5.65	3.23	8.06	1.61	5.65	4.03	4.03	6.45

1) office work
2) computer work
3) reading
4) writing
5) browsing
6) social interaction
7) gaming
8) eating, drinking
9) driving
10) smoking
11) walking
12) mobile phone interaction
13) watching TV
14) concentrated work

14. Would you share eye tracking data if it was collected while you are interacting with one of the following devices?



	1)	2)	3)	4)	5)	6)
1 - Strongly disagree:	12.90	12.90	16.13	12.90	9.68	16.94
2 - Disagree:	16.13	16.13	16.13	17.74	8.06	15.32
3 - Somewhat disagree:	8.87	8.87	8.87	8.06	8.06	8.06
4 - Neither agree nor disagree:	12.90	14.52	15.32	16.94	12.10	18.55
5 - Somewhat agree:	25.81	25.00	25.00	24.19	29.84	19.35
6 - Agree:	16.13	16.13	14.52	15.32	24.19	16.94
7 - Strongly agree:	7.26	6.45	4.03	4.84	8.06	4.84

1) desktop computer

2) laptop

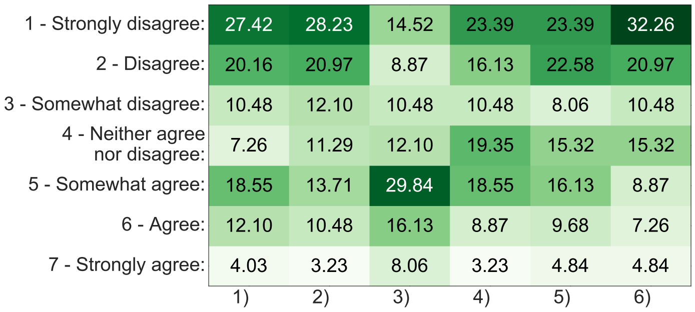
3) mobile phone

4) tablet

5) book

6) TV

15. Would you share eye tracking data if the data is collected while you are interacting with one of the following persons?



	1)	2)	3)	4)	5)	6)
1 - Strongly disagree:	27.42	28.23	14.52	23.39	23.39	32.26
2 - Disagree:	20.16	20.97	8.87	16.13	22.58	20.97
3 - Somewhat disagree:	10.48	12.10	10.48	10.48	8.06	10.48
4 - Neither agree nor disagree:	7.26	11.29	12.10	19.35	15.32	15.32
5 - Somewhat agree:	18.55	13.71	29.84	18.55	16.13	8.87
6 - Agree:	12.10	10.48	16.13	8.87	9.68	7.26
7 - Strongly agree:	4.03	3.23	8.06	3.23	4.84	4.84

1) friends

2) relatives

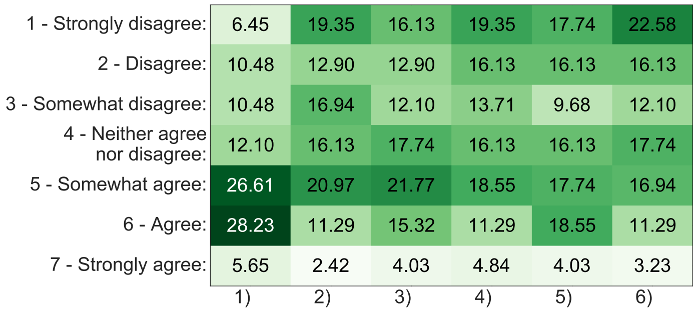
3) pets

4) foreigners/strangers

5) working colleagues

6) boss

16. Would you share eye tracking data if the kind of recorded was restricted to ...



	1)	2)	3)	4)	5)	6)
1 - Strongly disagree:	6.45	19.35	16.13	19.35	17.74	22.58
2 - Disagree:	10.48	12.90	12.90	16.13	16.13	16.13
3 - Somewhat disagree:	10.48	16.94	12.10	13.71	9.68	12.10
4 - Neither agree nor disagree:	12.10	16.13	17.74	16.13	16.13	17.74
5 - Somewhat agree:	26.61	20.97	21.77	18.55	17.74	16.94
6 - Agree:	28.23	11.29	15.32	11.29	18.55	11.29
7 - Strongly agree:	5.65	2.42	4.03	4.84	4.03	3.23

1) gaze or pupil behaviour?

2) scene video content?

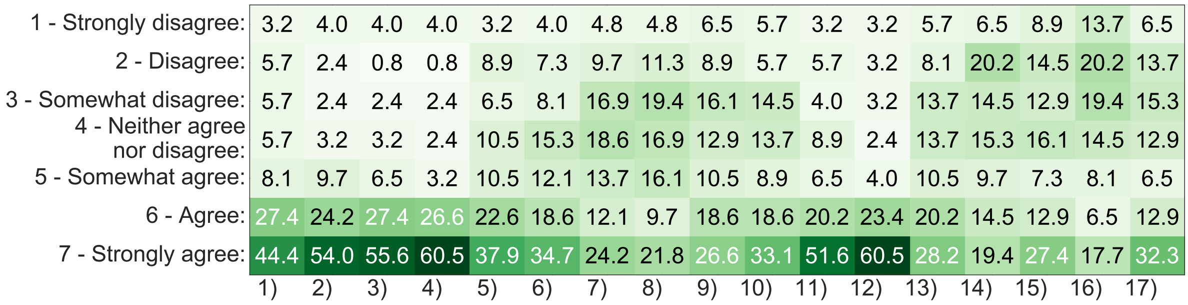
3) eye video content?

4) gaze or pupil behaviour + scene video content?

5) gaze or pupil behaviour + eye video content?

6) gaze or pupil behaviour + scene video content + eye video content?

17. Suppose you want to create an anonymous online identity in order to share your eye tracking data. Would you "hide" the following ADDITIONAL personal information?



	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)	12)	13)	14)	15)	16)	17)
1 - Strongly disagree:	3.2	4.0	4.0	4.0	3.2	4.0	4.8	4.8	6.5	5.7	3.2	3.2	5.7	6.5	8.9	13.7	6.5
2 - Disagree:	5.7	2.4	0.8	0.8	8.9	7.3	9.7	11.3	8.9	5.7	5.7	3.2	8.1	20.2	14.5	20.2	13.7
3 - Somewhat disagree:	5.7	2.4	2.4	2.4	6.5	8.1	16.9	19.4	16.1	14.5	4.0	3.2	13.7	14.5	12.9	19.4	15.3
4 - Neither agree nor disagree:	5.7	3.2	3.2	2.4	10.5	15.3	18.6	16.9	12.9	13.7	8.9	2.4	13.7	15.3	16.1	14.5	12.9
5 - Somewhat agree:	8.1	9.7	6.5	3.2	10.5	12.1	13.7	16.1	10.5	8.9	6.5	4.0	10.5	9.7	7.3	8.1	6.5
6 - Agree:	27.4	24.2	27.4	26.6	22.6	18.6	12.1	9.7	18.6	18.6	20.2	23.4	20.2	14.5	12.9	6.5	12.9
7 - Strongly agree:	44.4	54.0	55.6	60.5	37.9	34.7	24.2	21.8	26.6	33.1	51.6	60.5	28.2	19.4	27.4	17.7	32.3

1) first name

2) last name

3) identifiable profile picture

4) residential address

5) city where I live

6) occupation and employment information

7) hobbies

8) interests

9) current location information (e.g., kitchen, public transport, office)

10) my health condition(s)

11) email address

12) phone number

13) age and date of birth

14) gender

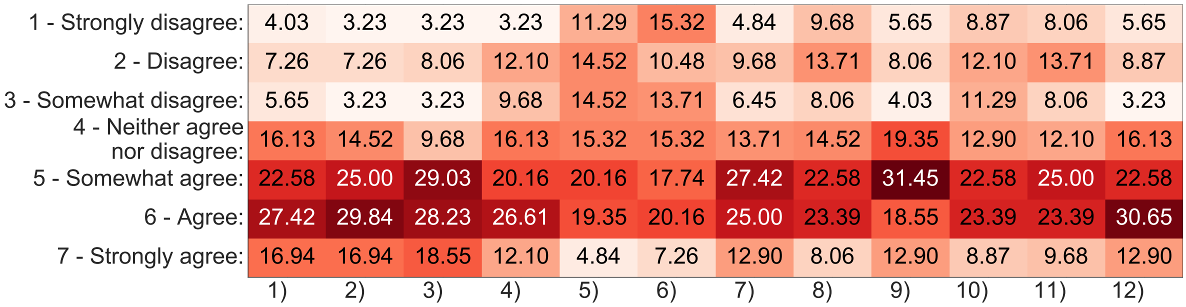
15) race

16) eye colour

17) iris image

C.3.4 Eye Tracking Data Representations

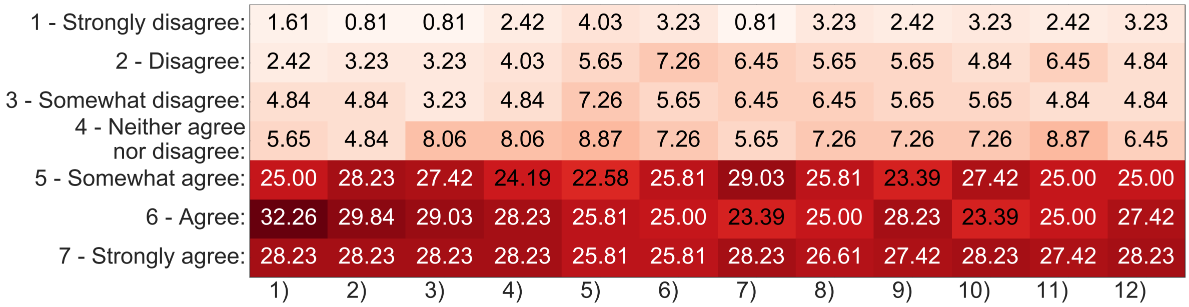
1. Would you agree to share eye tracking data which consists of ...



	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)	12)
1 - Strongly disagree:	4.03	3.23	3.23	3.23	11.29	15.32	4.84	9.68	5.65	8.87	8.06	5.65
2 - Disagree:	7.26	7.26	8.06	12.10	14.52	10.48	9.68	13.71	8.06	12.10	13.71	8.87
3 - Somewhat disagree:	5.65	3.23	3.23	9.68	14.52	13.71	6.45	8.06	4.03	11.29	8.06	3.23
4 - Neither agree nor disagree:	16.13	14.52	9.68	16.13	15.32	15.32	13.71	14.52	19.35	12.90	12.10	16.13
5 - Somewhat agree:	22.58	25.00	29.03	20.16	20.16	17.74	27.42	22.58	31.45	22.58	25.00	22.58
6 - Agree:	27.42	29.84	28.23	26.61	19.35	20.16	25.00	23.39	18.55	23.39	23.39	30.65
7 - Strongly agree:	16.94	16.94	18.55	12.10	4.84	7.26	12.90	8.06	12.90	8.87	9.68	12.90

1) the raw x and y gaze or pupil position over time?
2) statistics of steady (fixations) and dynamic (saccades) state of the eyes (when fixations and saccades take place)?
3) statistics of steady (fixations) and dynamic (saccades) state of the eyes (how often fixations and saccades appear in a given time range)?
4) statistics of eye tracking data which describe the number of fixations, fixation duration as well as their spatial distribution on public displays, computer monitors, or in VR environment?
5) fixation points on public displays, computer monitors, or in VR environment?
6) scan path information, the concatenation of gaze movements on public displays, computer monitors, or in VR environment?
7) scan path statistics, e.g. whether after a gaze movement to left is followed by a movement upwards, on public displays, computer monitors, or in VR environment?
8) fixations with duration and scan path information (Gaze Plot), the concatenation of gaze movements and fixational behaviour, on public displays, computer monitors, or in VR environment?
9) fixations with duration and scan path information statistics, e.g. whether after a gaze movement to left is followed by a movement upwards and how long the following fixation lasts, on public displays, computer monitors, or in VR environment?
10) heatmaps, user's gaze distribution on public displays, computer monitors, or in VR environment?
11) statistics of gaze distribution on areas of interests (AOIs) on public displays, computer monitors, or in VR environment?
12) aggregated features, given as so-called feature vectors, where each entry of such a vector describe a feature of user's behaviour like average blinking rate, fixation duration, ratio of saccadic movements, etc. within a given time window?

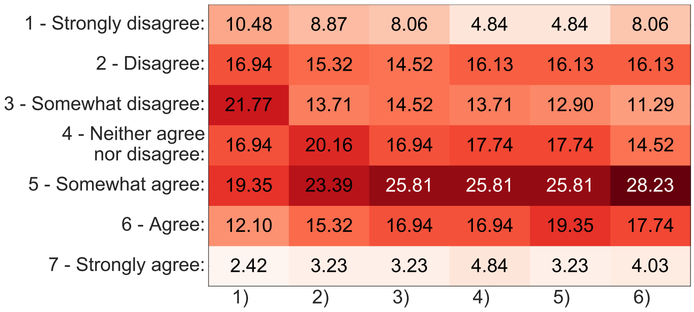
2. Imagine your eye tracking data could be modified so that it is anonymous, i.e. indistinguishable from that of another user. Would you agree to share ...



	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)	12)
1 - Strongly disagree:	1.61	0.81	0.81	2.42	4.03	3.23	0.81	3.23	2.42	3.23	2.42	3.23
2 - Disagree:	2.42	3.23	3.23	4.03	5.65	7.26	6.45	5.65	5.65	4.84	6.45	4.84
3 - Somewhat disagree:	4.84	4.84	3.23	4.84	7.26	5.65	6.45	6.45	5.65	5.65	4.84	4.84
4 - Neither agree nor disagree:	5.65	4.84	8.06	8.06	8.87	7.26	5.65	7.26	7.26	7.26	8.87	6.45
5 - Somewhat agree:	25.00	28.23	27.42	24.19	22.58	25.81	29.03	25.81	23.39	27.42	25.00	25.00
6 - Agree:	32.26	29.84	29.03	28.23	25.81	25.00	23.39	25.00	28.23	23.39	25.00	27.42
7 - Strongly agree:	28.23	28.23	28.23	28.23	25.81	25.81	28.23	26.61	27.42	28.23	27.42	28.23

1) the raw x and y gaze or pupil position over time?
2) statistics of steady (fixations) and dynamic (saccades) state of the eyes (when fixations and saccades take place)?
3) statistics of steady (fixations) and dynamic (saccades) state of the eyes (how often fixations and saccades appear in a given time range)?
4) statistics of eye tracking data which describe the number of fixations, fixation duration as well as their spatial distribution on public displays, computer monitors, or in VR environment?
5) fixation points on public displays, computer monitors, or in VR environment?
6) scan path information, the concatenation of gaze movements on public displays, computer monitors, or in VR environment?
7) scan path statistics, e.g. whether after a gaze movement to left is followed by a movement upwards, on public displays, computer monitors, or in VR environment?
8) fixations with duration and scan path information (Gaze Plot), the concatenation of gaze movements and fixational behaviour, on public displays, computer monitors, or in VR environment?
9) fixations with duration and scan path information statistics, e.g. whether after a gaze movement to left is followed by a movement upwards and how long the following fixation lasts, on public displays, computer monitors, or in VR environment?
10) heatmaps, user's gaze distribution on public displays, computer monitors, or in VR environment?
11) statistics of gaze distribution on areas of interests (AOIs) on public displays, computer monitors, or in VR environment?
12) aggregated features, given as so-called feature vectors, where each entry of such a vector describe a feature of user's behaviour like average blinking rate, fixation duration, ratio of saccadic movements, etc. within a given time window?

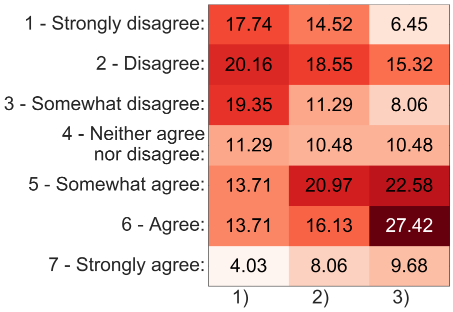
3. Would you agree to share scene video information of public displays, computer monitors, or in VR environment which consists of ...



	1)	2)	3)	4)	5)	6)
1 - Strongly disagree:	10.48	8.87	8.06	4.84	4.84	8.06
2 - Disagree:	16.94	15.32	14.52	16.13	16.13	16.13
3 - Somewhat disagree:	21.77	13.71	14.52	13.71	12.90	11.29
4 - Neither agree nor disagree:	16.94	20.16	16.94	17.74	17.74	14.52
5 - Somewhat agree:	19.35	23.39	25.81	25.81	25.81	28.23
6 - Agree:	12.10	15.32	16.94	16.94	19.35	17.74
7 - Strongly agree:	2.42	3.23	3.23	4.84	3.23	4.03

1) the scene content during a whole eye tracking recording?
2) single frame from each fixation?
3) gaze stripes, sequence of image frame from each fixation?
4) tiny image patches around the gaze position during a whole eye tracking recording?
5) tiny image patches around the gaze position from each fixation?
6) the whole scene video but with blurred out surrounding and clear object of interest?

4. Would you agree to share eye video information recorded from eye tracking camera which consists of ...



	1)	2)	3)
1 - Strongly disagree:	17.74	14.52	6.45
2 - Disagree:	20.16	18.55	15.32
3 - Somewhat disagree:	19.35	11.29	8.06
4 - Neither agree nor disagree:	11.29	10.48	10.48
5 - Somewhat agree:	13.71	20.97	22.58
6 - Agree:	13.71	16.13	27.42
7 - Strongly agree:	4.03	8.06	9.68

1) the whole eye video with visible iris and surrounding facial expressions?
2) share the whole eye video with visible iris but blurred surrounding facial expressions?
3) the whole eye video but only showing the pupil centre without iris or facial expressions?

C.4 Eye Movement Feature Extraction

Table C.1 summarises the features that we extracted from fixations, saccades, blinks, pupil diameter, and a user's scan paths. Similar to Bulling *et al.* (2011b), each saccade is encoded as a character forming words of length n (wordbook). We extracted these features on a sliding window of 30 seconds (step size of 0.5 seconds).

Fixation (8)	rate, mean, max, var of durations, mean/var of var pupil position within one fixation
Saccades (12)	rate/ratio of (small/large/right/left) saccades, mean, max, variance of amplitudes
Combined (1)	ratio saccades to fixations
Wordbooks (24)	number of non-zero entries, max and min entries, and their difference for n-grams with n <= 4
Blinks (3)	rate, mean/var blink duration
Pupil Diameter (4)	mean/var of mean/var during fixations

Table C.1: We extracted 52 eye movement features to describe a user's eye movement behaviour. The number of features per category is given in parentheses.

List of Figures

List of Tables

Bibliography

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.* (2016). TensorFlow: Large-scale Machine Learning on Heterogeneous Distributed Systems, *arXiv preprint arXiv:1603.04467*. Cited on page 84.

E. Abdulin, I. Rigas, and O. Komogortsev (2016). Eye Movement Biometrics on Wearable Devices: What Are the Limits?, in *Proc. of the ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI) 2016*. Cited on page 81.

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 (UbiComp) 2016*. Cited on page 160.

D. Adams, A. Bah, C. Barwulor, N. Musaby, K. Pitkin, and E. M. Redmiles (2018). Ethics Emerging: The Story of Privacy and Security Perceptions in Virtual Reality, in *Proc. of the Symposium on Usable Privacy and Security (SOUPS) 2018*. Cited on page 119.

A. I. Adiba, N. Tanaka, and J. Miyake (2016). An Adjustable Gaze Tracking System and Its Application for Automatic Discrimination of Interest Objects, *IEEE/ASME Transactions on Mechatronics*, vol. 21(2), pp. 973–979. Cited on page 20.

P. Aditya, R. Sen, P. Druschel, S. Joon Oh, R. Benenson, M. Fritz, B. Schiele, B. Bhattacharjee, and T. T. Wu (2016). I-Pic: A Platform for Privacy-Compliant Image Capture, in *Proc. of the International Conference on Mobile Systems, Applications, and Services (MobiSys) 2016*. Cited on page 98.

D. Akkil, J. Kangas, J. Rantala, P. Isokoski, O. Spakov, and R. Raisamo (2015). Glance Awareness and Gaze Interaction in Smartwatches, in *Proc. of the ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI) 2015*. Cited on pages 20 and 78.

H. S. Al-Khalifa and R. P. George (2010). Eye Tracking and e-Learning: Seeing Through Your Students' Eyes, *eLearn*, vol. 2010(6), p. 8. Cited on page 3.

W. Albert and T. Tullis (2013). *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*, Newnes. Cited on page 14.

N. Ali, M. Salim, H. Mohadis, and W. Ahmad (2016). User Perception Towards the Use of Wearable Cameras, *New Zealand Journal of CHI*, vol. 1(1), pp. 1–13. Cited on page 16.

F. Alt, A. S. Shirazi, A. Schmidt, and J. Mennenöh (2012). Increasing the User's Attention on the Web: Using Implicit Interaction Based on Gaze Behavior to Tailor Content, in *Proc. of the Nordic Conference on Human-Computer Interaction: Making Sense Through Design 2012*. Cited on pages 3 and 13.

N. Anantrasirichai, I. D. Gilchrist, and D. R. Bull (2016). Fixation Identification for Low-Sample-Rate Mobile Eye Trackers, in *Proc. of the IEEE International Conference on Image Processing (ICIP) 2016*. Cited on page 63.

R. Andersson, L. Larsson, K. Holmqvist, M. Stridh, and M. Nyström (2017). One Algorithm to Rule Them All? An Evaluation and Discussion of Ten Eye Movement Event-Detection Algorithms, *Behavior Research Methods*, vol. 49(2), pp. 616–637. Cited on pages 12 and 62.

T. Appel, T. Santini, and E. Kasneci (2016). Brightness- and Motion-Based Blink Detection for Head-Mounted Eye Trackers, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2016*. Cited on page 12.

N. M. Arar, H. Gao, and J.-P. Thiran (2015). Robust Gaze Estimation Based on Adaptive Fusion of Multiple Cameras, in *Proc. of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG) 2015*. Cited on page 81.

O. Aubert, Y. Prié, and D. Schmitt (2012). Advene As a Tailorable Hypervideo Authoring Tool: A Case Study, in *Proc. of the ACM Symposium on Document Engineering (DocEng) 2012*. Cited on pages 65, 104, and 185.

S. Baluja and D. Pomerleau (1994). Non-Intrusive Gaze Tracking Using Artificial Neural Networks, in *Proc. of the Advances in Neural Information Processing Systems 1994*. Cited on pages 9 and 84.

T. S. Barger, D. E. Brown, and M. Alwan (2005). Health-Status Monitoring Through Analysis of Behavioral Patterns, *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 35(1), pp. 22–27. Cited on page 138.

L. Baruh and Z. Cemalcılar (2014). It Is More Than Personal: Development and Validation of a Multidimensional Privacy Orientation Scale, *Personality and Individual Differences*, vol. 70, pp. 165–170. Cited on page 104.

M. Barz, F. Daiber, and A. Bulling (2016). Prediction of Gaze Estimation Error for Error-Aware Gaze-Based Interfaces, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2016*. Cited on pages 5 and 92.

R. Bednarik, T. Kinnunen, A. Mihaila, and P. Fränti (2005). Eye-Movements as a Biometric, in *Proc. of the Scandinavian Conference on Image Analysis (SCIA) 2005*. Cited on pages 17 and 120.

R. Bednarik, H. Vrzakova, and M. Hradis (2012). What Do You Want to Do Next: A Novel Approach for Intent Prediction in Gaze-Based Interaction, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2012*. Cited on pages 19 and 78.

J. B. Begole, J. C. Tang, and R. Hill (2003). Rhythm Modeling, Visualizations and Applications, in *Proc. of ACM Symposium on User Interface Software and Technology (UIST) 2003*. Cited on page 138.

D. J. Berg, S. E. Boehnke, R. A. Marino, D. P. Munoz, and L. Itti (2009). Free Viewing of Dynamic Stimuli by Humans and Monkeys, *Journal of Vision (JOV)*, vol. 9(5), pp. 19–19. Cited on pages 13 and 62.

J. R. Bergstrom and A. Schall (2014). *Eye Tracking in User Experience Design*, Elsevier. Cited on pages 16 and 19.

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 163.

D. Beymer and M. Flickner (2003). Eye Gaze Tracking Using an Active Stereo Head, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2003*. Cited on pages 10, 31, and 50.

T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl (2014). State-of-the-Art of Visualization for Eye Tracking Data, in *Proc. of the Eurographics Conference on Visualization (EuroVis) 2014*. Cited on page 13.

T. Blascheck, K. Kurzhals, M. Raschke, S. Strohmaier, D. Weiskopf, and T. Ertl (2016). AOI Hierarchies for Visual Exploration of Fixation Sequences, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2016*. Cited on page 62.

D. M. Blei, A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet Allocation, *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022. Cited on pages 141 and 142.

P. Blignaut (2009). Fixation Identification: The Optimum Threshold for a Dispersion Algorithm, *Attention, Perception, & Psychophysics*, vol. 71(4), pp. 881–895. Cited on pages 12 and 62.

P. Blignaut, K. Holmqvist, M. Nyström, and R. Dewhurst (2014). Improving the Accuracy of Video-Based Eye Tracking in Real Time Through Post-Calibration Regression, in *Current Trends in Eye Tracking Research 2014*, pp. 77–100, Springer. Cited on page 3.

J. Bohn, V. Coroamă, M. Langheinrich, F. Mattern, and M. Rohs (2005). Social, Economic, and Ethical Implications of Ambient Intelligence and Ubiquitous Computing, in *Ambient Intelligence 2005*, pp. 5–29, Springer. Cited on pages 34 and 101.

F. H. Borsato and C. H. Morimoto (2016). Episcleral Surface Tracking: Challenges and Possibilities for Using Mice Sensors for Wearable Eye Tracking, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2016*. Cited on page 81.

C. Braunagel, E. Kasneci, W. Stolzmann, and W. Rosenstiel (2015). Driver-Activity Recognition in the Context of Conditionally Autonomous Driving, in *Proc. of the IEEE International Conference on Intelligent Transportation Systems (ITSC) 2015*. Cited on page 12.

A. Bulling (2016). Pervasive Attentive User Interfaces, *IEEE Computer*, vol. 49(1), pp. 94–98. Cited on pages 24, 50, and 158.

A. Bulling and H. Gellersen (2010). Toward Mobile Eye-Based Human-Computer Interaction, *IEEE Pervasive Computing*, (4), pp. 8–12. Cited on pages 17, 78, and 118.

A. Bulling and K. Kunze (2016). Eyewear Computers for Human-Computer Interaction, *ACM Interactions*, vol. 23(3), pp. 70–73. Cited on pages 17 and 98.

A. Bulling and D. Roggen (2011). Recognition of Visual Memory Recall Processes Using Eye Movement Analysis, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2011*. Cited on pages 19, 78, 136, and 138.

A. Bulling, D. Roggen, and G. Tröster (2008a). It's in Your Eyes: Towards Context-awareness and Mobile HCI Using Wearable EOG Goggles, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2008*. Cited on page 81.

A. Bulling, D. Roggen, and G. Tröster (2009a). Wearable EOG Goggles: Seamless Sensing and Context-Awareness in Everyday Environments, *Journal of Ambient Intelligence and Smart Environments*, vol. 1(2), pp. 157–171. Cited on page 81.

A. Bulling, D. Roggen, and G. Tröster (2011a). What's in the Eyes for Context-Awareness?, *IEEE Pervasive Computing*, vol. 10(2), pp. 48 – 57. Cited on pages 17, 35, 78, and 118.

A. Bulling, J. A. Ward, and H. Gellersen (2012). Multimodal Recognition of Reading Activity in Transit Using Body-Worn Sensors, *ACM Transactions on Applied Perception*, vol. 9(1), pp. 2:1–2:21. Cited on pages 19, 36, 68, 104, 136, 138, 151, 152, and 154.

A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster (2009b). Eye Movement Analysis for Activity Recognition, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2009*. Cited on page 78.

A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster (2011b). 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 pages 14, 19, 22, 60, 61, 90, 99, 103, 104, 128, 129, 136, 138, 140, 146, 151, 152, 164, 186, and 206.

A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster (2008b). Robust Recognition of Reading Activity in Transit Using Wearable Electrooculography, in *Proc. of the International Conference on Pervasive Computing (Pervasive) 2008*. Cited on pages 78, 148, 152, and 213.

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 (CHI) 2013*. Cited on pages 18, 19, 24, 78, 99, 119, 138, 144, 146, and 164.

A. Bulling and T. O. Zander (2014). Cognition-Aware Computing, *IEEE Pervasive Computing*, vol. 13(3), pp. 80–83. Cited on pages 18, 60, 78, and 119.

K. Caine (2009). *Exploring Everyday Privacy Behaviors and Misclosures*, Georgia Institute of Technology. Cited on pages 101 and 104.

V. Cantoni, C. Galdi, M. Nappi, M. Porta, and D. Riccio (2015). GANT: Gaze Analysis Technique for Human Identification, *Pattern Recognition*, vol. 48(4), pp. 1027–1038. Cited on pages 18 and 119.

S. Castagnos, N. Jones, and P. Pu (2009). Recommenders' Influence on Buyers' Decision Process, in *Proc. of the ACM Recommender Systems Conference (RecSys) 2009*. Cited on page 14.

S. Castagnos, N. Jones, and P. Pu (2010). Eye-Tracking Product Recommenders' Usage, in *Proc. of the ACM Recommender Systems Conference (RecSys) 2010*. Cited on page 14.

S. Castagnos and P. Pu (2010). Consumer Decision Patterns Through Eye Gaze Analysis, in *Proc. of the Workshop on Eye Gaze on Intelligent Human Machine Interaction 2010*. Cited on page 14.

J. J. Cerrolaza, A. Villanueva, and R. Cabeza (2012). Study of Polynomial Mapping Functions in Video-Oculography Eye Trackers, *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 19(2), p. 10. Cited on page 9.

L. Chan and K. Minamizawa (2017). FrontFace: Facilitating Communication Between HMD Users and Outsiders Using Front-Facing-Screen HMDs, in *Proc. of the ACM International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI) 2017*. Cited on page 114.

J. Chen and Q. Ji (2011). Probabilistic Gaze Estimation Without Active Personal Calibration, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 10.

L. Chen and P. Pu (2010). Eye-Tracking Study of User Behavior in Recommender Interfaces, in *Proc. of the International Conference on User Modeling, Adaptation, and Personalization 2010*. Cited on page 14.

L. Chen and F. Wang (2016). An Eye-Tracking Study: Implication to Implicit Critiquing Feedback Elicitation in Recommender Systems, in *Proc. of the ACM Conference on User Modeling Adaptation and Personalization (UMAP) 2016*. Cited on page 14.

S. Chen, J. Epps, and F. Chen (2013). Automatic and Continuous User Task Analysis via Eye Activity, in *Proc. of the ACM International Conference on Intelligent User Interfaces (IUI) 2013*. Cited on pages 19, 136, and 138.

S. Cheng, X. Liu, P. Yan, J. Zhou, and S. Sun (2010a). Adaptive User Interface of Product Recommendation Based on Eye-Tracking, in *Proc. of the Workshop on Eye Gaze in Intelligent Human Machine Interaction 2010*. Cited on page 42.

Z. Cheng, B. Gao, and T.-Y. Liu (2010b). Actively Predicting Diverse Search Intent from User Browsing Behaviors, in *Proc. of the World Wide Web Conference (WWW) 2010*. Cited on page 161.

F. Chollet *et al.* (2015). *Keras*, `https://github.com/fchollet/keras`. Cited on page 84.

S. Chowdhury, M. S. Ferdous, and J. M. Jose (2016). Exploring Lifelog Sharing and Privacy, in *Adj. Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2016*. Cited on page 100.

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 (UbiComp) 2016*. Cited on pages 158 and 160.

L. Chuang, A. Duchowski, P. Qvarfordt, and D. Weiskopf (2019). Ubiquitous Gaze Sensing and Interaction (Dagstuhl Seminar 18252), *Dagstuhl Reports*, vol. 8(6), pp. 77–148. Cited on page 19.

T. Chuk, A. B. Chan, and J. H. Hsiao (2014). Understanding Eye Movements in Face Recognition Using Hidden Markov Models, *Journal of Vision (JOV)*, vol. 14(11). Cited on page 137.

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 167.

L. Cowen, L. J. Ball, and J. Delin (2002). An Eye Movement Analysis of Web Page Usability, in *People and Computers XVI-Memorable yet Invisible 2002*, pp. 317–335, Springer. Cited on page 13.

E. C. Crowe and N. H. Narayanan (2000). Comparing Interfaces Based on What Users Watch and Do, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2000*. Cited on page 13.

G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray (2004). Visual Categorization with Bags of Keypoints, in *Proc. of the European Conference on Computer Vision Workshops (ECCVW) 2004*. Cited on page 141.

K. M. Dalton, B. M. Nacewicz, T. Johnstone, H. S. Schaefer, M. A. Gernsbacher, H. H. Goldsmith, A. L. Alexander, and R. J. Davidson (2005). Gaze Fixation and the Neural Circuitry of Face Processing in Autism, *Nature Neuroscience*, vol. 8(4), pp. 519–526. Cited on page 60.

P. M. Daye, D. C. Roberts, D. S. Zee, and L. M. Optican (2015). Vestibulo-Ocular Reflex Suppression During Head-Fixed Saccades Reveals Gaze Feedback Control, *Journal of Neuroscience*, vol. 35(3), pp. 1192–1198. Cited on page 21.

J. DeBlasio and B. N. Walker (2009). Documentation in a Medical Setting: Effects of Technology on Perceived Quality of Care, in *Proc. of the Human Factors and Ergonomics Society Annual Meeting 2009*. Cited on page 16.

L. Dempere-Marco, X.-P. Hu, S. L. MacDonald, S. M. Ellis, D. M. Hansell, and G.-Z. Yang (2002). The Use of Visual Search for Knowledge Gathering in Image Decision Support, *IEEE Transactions on Medical Imaging*, vol. 21(7), pp. 741–754. Cited on page 137.

T. Denning, Z. Dehlawi, and T. Kohno (2014). In Situ with Bystanders of Augmented Reality Glasses: Perspectives on Recording and Privacy-mediating Technologies, in *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI) 2014*. Cited on pages 18, 34, 100, 101, 114, and 190.

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 (ISWC) 2016*. Cited on page 160.

S. D'Mello, A. Olney, C. Williams, and P. Hays (2012). Gaze Tutor: A Gaze-Reactive Intelligent Tutoring System, *International Journal of Human-Computer Studies*, vol. 70(5), pp. 377–398. Cited on page 60.

T. D'Orazio, M. Leo, G. Cicirelli, and A. Distante (2004). An Algorithm for Real Time Eye Detection in Face Images, in *Proc. of the International Conference on Pattern Recognition (ICPR) 2004*. Cited on page 7.

A. Doshi and M. M. Trivedi (2012). Head and Eye Gaze Dynamics During Visual Attention Shifts in Complex Environments, *Journal of Vision (JOV)*, vol. 12(2), pp. 1–16. Cited on page 21.

J. Duchi, E. Hazan, and Y. Singer (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, *Journal of Machine Learning Research (JMLR)*, vol. 12(Jul), pp. 2121–2159. Cited on page 84.

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 pages 3 and 19.

A. T. Duchowski (2007). Eye Tracking Methodology, *Theory and Practice*, vol. 328, p. 614. Cited on pages 5 and 8.

A. T. Duchowski, D. H. House, J. Gestring, R. Congdon, L. Świrski, N. A. Dodgson, K. Krejtz, and I. Krejtz (2014). Comparing Estimated Gaze Depth in Virtual and Physical Environments, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2014*. Cited on pages 31 and 50.

C. Dwork, F. McSherry, K. Nissim, and A. Smith (2006). Calibrating Noise to Sensitivity in Private Data Analysis, in *Proc. of the Theory of Cryptography Conference (TCC) 2006*. Cited on page 125.

C. Dwork, A. Roth, *et al.* (2014). The Algorithmic Foundations of Differential Privacy, *Foundations and Trends® in Theoretical Computer Science*, vol. 9(3–4), pp. 211–407. Cited on pages 23, 124, and 125.

M. L. Dybdal, J. S. Agustin, and J. P. Hansen (2012). Gaze Input for Mobile Devices by Dwell and Gestures, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2012*. Cited on page 19.

S. Eberz, K. B. Rasmussen, V. Lenders, and I. Martinovic (2016). Looks Like Eve: Exposing Insider Threats Using Eye Movement Biometrics, *ACM Transactions on Privacy and Security (TOPS)*, vol. 19(1), p. 1. Cited on pages 17 and 120.

C. Ehmke and S. Wilson (2007). Identifying Web Usability Problems from Eye-Tracking Data, in *Proc. of the British HCI Group Annual Conference on People and Computers: HCI... But Not as We Know It – Volume 1, 2007*. Cited on pages 14, 16, and 19.

M. Elhelw, M. Nicolaou, A. Chung, G.-Z. Yang, and M. S. Atkins (2008). A Gaze-Based Study for Investigating the Perception of Visual Realism in Simulated Scenes, *ACM Transactions on Applied Perception*, vol. 5(1), p. 3. Cited on page 137.

J. Engel, T. Schöps, and D. Cremers (2014). LSD-SLAM: Large-Scale Direct Monocular SLAM, in *Proc. of the European Conference on Computer Vision (ECCV) 2014*. Cited on page 13.

B. Ens, T. Grossman, F. Anderson, J. Matejka, and G. Fitzmaurice (2015). Candid Interaction: Revealing Hidden Mobile and Wearable Computing Activities, in *Proc. of the ACM Symposium on User Interface Software and Technology (UIST) 2015*. Cited on pages 34, 101, and 113.

Z. Erickson, J. Compiano, and R. Shin (2014). Neural Networks for Improving Wearable Device Security. Cited on pages 18 and 100.

C. W. Eriksen and J. E. Hoffman (1972). Temporal and Spatial Characteristics of Selective Encoding from Visual Displays, *Attention, Perception, & Psychophysics*, vol. 12(2), pp. 201–204. Cited on pages 64, 68, 69, and 208.

C. W. Eriksen and Y.-y. Yeh (1985). Allocation of Attention in the Visual Field, *Journal of Experimental Psychology: Human Perception and Performance*, vol. 11(5), p. 583. Cited on pages 17 and 98.

M. Eriksson and N. P. Papanikolopoulos (2001). Driver Fatigue: A Vision-Based Approach to Automatic Diagnosis, *Transportation Research Part C: Emerging Technologies*, vol. 9(6), pp. 399–413. Cited on page 3.

A. Esteves, E. Velloso, A. Bulling, and H. Gellersen (2015). Orbits: Enabling Gaze Interaction in Smart Watches Using Moving Targets, in *Proc. of the ACM Symposium on User Interface Software and Technology (UIST) 2015*. Cited on pages 20 and 78.

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 (UbiComp) 2016*. Cited on pages 158 and 160.

M. Faber, R. Bixler, and S. K. D'Mello (2017). An Automated Behavioral Measure of Mind Wandering During Computerized Reading, *Behavior Research Methods*, pp. 1–17. Cited on pages 18, 60, and 119.

L. Fan and L. Xiong (2012). Adaptively Sharing Time-Series with Differential Privacy, *arXiv preprint arXiv:1202.3461*. Cited on page 120.

A. Faro, D. Giordano, C. Pino, and C. Spampinato (2010). Visual Attention for Implicit Relevance Feedback in a Content Based Image Retrieval, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2010*. Cited on page 42.

K. Farrahi and D. Gatica-Perez (2008). What Did You Do Today? Discovering Daily Routines from Large-Scale Mobile Data, in *Proc. of the ACM International Conference on Multimedia (MM) 2008*. Cited on page 138.

G. C. Feng and P. C. Yuen (2001). Multi-Cues Eye Detection on Gray Intensity Image, *Pattern Recognition*, vol. 34(5), pp. 1033–1046. Cited on page 7.

M. S. Ferdous, S. Chowdhury, and J. M. Jose (2017). Analysing Privacy in Visual Lifelogging, *Pervasive and Mobile Computing*. Cited on page 100.

V. P. Ferrera (2000). Task-Dependent Modulation of the Sensorimotor Transformation for Smooth Pursuit Eye Movements, *Journal of Neurophysiology*, vol. 84(6), pp. 2725–2738. Cited on pages 12 and 62.

M. Fetter (2007). Vestibulo-Ocular Reflex, in *Neuro-Ophthalmology 2007*, vol. 40, pp. 35–51, Karger Publishers. Cited on page 21.

J. M. Findlay and I. D. Gilchrist (2003). *Active Vision: The Psychology of Looking and Seeing*, no. 37, Oxford University Press. Cited on page 169.

G. Fischer (2001). User Modeling in Human–Computer Interaction, *User Modeling and User-Adapted Interaction*, vol. 11(1-2), pp. 65–86. Cited on pages 17 and 98.

D. R. Flatla, C. Gutwin, L. E. Nacke, S. Bateman, and R. L. Mandryk (2011). Calibration Games: Making Calibration Tasks Enjoyable by Adding Motivating Game Elements, in *Proc. of the ACM Symposium on User Interface Software and Technology (UIST) 2011*. Cited on page 10.

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 160.

A. Fogelton and W. Benesova (2016). Eye Blink Detection Based on Motion Vectors Analysis, *Computer Vision and Image Understanding*, vol. 148, pp. 23–33. Cited on page 12.

M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart (2014). Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing, in *Proc. of the USENIX Security Symposium 2014*. Cited on page 120.

W. Fuhl, T. Kübler, K. Sippel, W. Rosenstiel, and E. Kasneci (2015). ExCuSe: Robust Pupil Detection in Real-World Scenarios, in *Proc. of the International Conference on Computer Analysis of Images and Patterns (CAIP) 2015*. Cited on pages 8, 80, and 82.

W. Fuhl, T. Santini, D. Geisler, T. Kübler, W. Rosenstiel, and E. Kasneci (2016a). Eyes Wide Open? Eyelid Location and Eye Aperture Estimation for Pervasive Eye Tracking in Real-World Scenarios, in *Adj. Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2016*. Cited on page 7.

W. Fuhl, T. Santini, G. Kasneci, and E. Kasneci (2016b). PupilNet: Convolutional Neural Networks for Robust Pupil Detection, *arXiv preprint arXiv:1601.04902*. Cited on pages 8 and 80.

W. Fuhl, T. C. Santini, T. Kübler, and E. Kasneci (2016c). ElSe: Ellipse Selection for Robust Pupil Detection in Real-World Environments, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2016*. Cited on pages 8 and 80.

W. Fuhl, M. Tonsen, A. Bulling, and E. Kasneci (2016d). Pupil Detection for Head-Mounted Eye Tracking in the Wild: An Evaluation of the State of the Art, *Machine Vision and Applications*, vol. 27(8), pp. 1275–1288. Cited on page 8.

K. A. Funes Mora and J.-M. Odobez (2013). Person Independent 3D Gaze Estimation from Remote RGB-D Cameras, in *Proc. of the IEEE International Conference on Image Processing (ICIP) 2013*. Cited on page 10.

S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez (2014). Automatic Generation and Detection of Highly Reliable Fiducial Markers Under Occlusion, *Pattern Recognition*, vol. 47(6), pp. 2280–2292. Cited on page 53.

K. R. Gegenfurtner (2016). The Interaction Between Vision and Eye Movements, *Perception*, vol. 45(12), pp. 1333–1357. Cited on page 21.

E. Goffman (2006). The Presentation of Self, *Life as Theater: A Dramaturgical Sourcebook*, pp. 129–139. Cited on page 16.

J. H. Goldberg and X. P. Kotval (1999). Computer Interface Evaluation Using Eye Movements: Methods and Constructs, *International Journal of Industrial Ergonomics*, vol. 24(6), pp. 631–645. Cited on page 19.

L. A. Granka, T. Joachims, and G. Gay (2004). Eye-Tracking Analysis of User Behavior in WWW Search, in *Proc. of the International ACM SIGIR Conference on Research and Development in Information Retrieval 2004*. Cited on page 14.

K. Grauman, M. Betke, J. Gips, and G. R. Bradski (2001). Communication via Eye Blinks–Detection and Duration Analysis in Real Time, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2001*. Cited on page 12.

K. Grauman, M. Betke, J. Lombardi, J. Gips, and G. R. Bradski (2003). Communication via Eye Blinks and Eyebrow Raises: Video-Based Human-Computer Interfaces, *Universal Access in the Information Society*, vol. 2(4), pp. 359–373. Cited on page 12.

T. Gu, S. Chen, X. Tao, and J. Lu (2010). An Unsupervised Approach to Activity Recognition and Segmentation Based on Object-Use Fingerprints, *Data & Knowledge Engineering*, vol. 69(6), pp. 533–544. Cited on page 138.

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 (CSCW) 2017*. Cited on pages 25, 37, and 158.

J. Häkkilä, F. Vahabpour, A. Colley, J. Väyrynen, and T. Koskela (2015). Design Probes Study on User Perceptions of a Smart Glasses Concept, in *Proc. of the ACM International Conference on Mobile and Ubiquitous Multimedia (MUM) 2015*. Cited on page 101.

P. W. Hallinan (1991). Recognizing Human Eyes, in *Geometric Methods in Computer Vision 1991*. Cited on page 8.

R. I. Hammoud (2008). *Passive Eye Monitoring: Algorithms, Applications and Experiments*, Springer Science & Business Media. Cited on page 8.

D. W. Hansen, J. P. Hansen, M. Nielsen, A. S. Johansen, and M. B. Stegmann (2002). Eye Typing Using Markov and Active Appearance Models, in *Proc. of the IEEE Workshop on Applications of Computer Vision (WACV) 2002*. Cited on page 8.

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 (TPAMI)*, vol. 32(3), pp. 478–500. Cited on pages 6 and 8.

D. W. Hansen and A. E. Pece (2005). Eye Tracking in the Wild, *Computer Vision and Image Understanding*, vol. 98(1), pp. 155–181. Cited on page 7.

J. P. Hansen, A. S. Johansen, D. W. Hansen, K. Itoh, and S. Mashino (2003). Command Without a Click: Dwell Time Typing by Mouse and Gaze Selections, in *Proc. of Human-Computer Interaction–INTERACT 2003*. Cited on pages 17, 35, 98, and 118.

A. Harvey (2010). *Camoflash-Anti-Paparazzi Clutch*, `http://ahprojects.com/projects/camoflash/`. Cited on pages 18 and 100.

A. Harvey (2012). CVDazzle: Camouflage from Computer Vision, *Technical Report*. Cited on pages 18, 98, and 100.

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 165.

R. Herpers, M. Michaelis, K.-H. Lichtenauer, and G. Sommer (1996). Edge and Keypoint Detection in Facial Regions, in *Proc. of the IEEE International Conference on Automatic Face and Gesture Recognition (FG) 1996*. Cited on page 7.

E. H. Hess and J. M. Polt (1960). Pupil Size as Related to Interest Value of Visual Stimuli, *Science*, vol. 132(3423), pp. 349–350. Cited on pages 17 and 119.

R. S. Hessels, D. C. Niehorster, C. Kemner, and I. T. Hooge (2017). Noise-Robust Fixation Detection in Eye Movement Data: Identification by Two-Means Clustering (I2MC), *Behavior Research Methods*, vol. 49(5), pp. 1802–1823. Cited on page 60.

K. Higuch, R. Yonetani, and Y. Sato (2016). Can Eye Help You?: Effects of Visualizing Eye Fixations on Remote Collaboration Scenarios for Physical Tasks, in *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI) 2016*. Cited on pages 60 and 62.

S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood (2006). SenseCam: A Retrospective Memory Aid, in *Proc. of the International Conference on Ubiquitous Computing (UbiComp) 2006*. Cited on page 24.

T. Hofmann (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, vol. 42(1–2), pp. 177–196. Cited on page 141.

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 (ETRA) 2012*. Cited on pages 158 and 160.

K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer (2011). *Eye Tracking: A Comprehensive Guide to Methods and Measures*, Oxford University Press. Cited on pages 3, 12, 32, 60, 62, 68, 69, 139, and 208.

P. S. Holzman, L. R. Proctor, D. L. Levy, N. J. Yasillo, H. Y. Meltzer, and S. W. Hurt (1974). Eye-Tracking Dysfunctions in Schizophrenic Patients and Their Relatives, *Archives of General Psychiatry*, vol. 31(2), pp. 143–151. Cited on pages 17, 19, 41, and 119.

S. Hoppe and A. Bulling (2016). End-to-End Eye Movement Detection Using Convolutional Neural Networks, *arXiv preprint arXiv:1609.02452*. Cited on pages 13, 40, and 62.

S. Hoppe, T. Loetscher, S. Morey, and A. Bulling (2015). Recognition of Curiosity Using Eye Movement Analysis, in *Adj. Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2015*. Cited on pages 20, 60, 78, and 136.

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 pages 3, 18, 20, 60, 99, 104, 119, and 132.

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 ACM Conference on Human Factors in Computing Systems (CHI) 2015*. Cited on pages 99 and 112.

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 (UbiComp) 2014*. Cited on pages 100, 104, and 112.

C.-F. Hsu, A. Chen, C.-H. Hsu, C.-Y. Huang, C.-L. Lei, and K.-T. Chen (2017). Is Foveated Rendering Perceivable in Virtual Reality? Exploring the Efficiency and Consistency of Quality Assessment Methods, in *Proc. of the ACM International Conference on Multimedia (MM) 2017*. Cited on pages 17 and 41.

J. Huang, D. Ii, X. Shao, and H. Wechsler (1998). Pose Discriminiation and Eye Detection Using Support Vector Machines (SVM), in *Face Recognition 1998*, pp. 528–536, Springer. Cited on page 8.

J. Huang and H. Wechsler (1999). Eye Detection Using Optimal Wavelet Packets and Radial Basis Functions (RBFs), *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 13, pp. 1009–1026. Cited on page 8.

M. X. Huang, T. C. Kwok, G. Ngai, H. V. Leong, and S. C. Chan (2014). Building a Self-Learning Eye Gaze Model from User Interaction Data, in *Proc. of the ACM International Conference on Multimedia (MM) 2014*. Cited on page 10.

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 160.

Y. Huang, M. Cai, H. Kera, R. Yonetani, K. Higuchi, and Y. Sato (2017). Temporal Localization and Spatial Segmentation of Joint Attention in Multiple First-Person Videos, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*. Cited on page 62.

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 (ECCV) 2018*. Cited on page 176.

J. T. Hutton, J. Nagel, and R. B. Loewenson (1984). Eye Tracking Dysfunction in Alzheimer-Type Dementia, *Neurology*, vol. 34(1), pp. 99–99. Cited on pages 17, 19, 41, and 119.

T. Huynh, M. Fritz, and B. Schiele (2008). Discovery of Activity Patterns Using Topic Models, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2008*. Cited on pages 138, 141, and 147.

D. E. Irwin (1992). Visual Memory Within and Across Fixations, in *Eye Movements and Visual Cognition 1992*, pp. 146–165, Springer. Cited on pages 64, 65, 67, and 208.

Y. Ishiguro, A. Mujibiya, T. Miyaki, and J. Rekimoto (2010). Aided Eyes: Eye Activity Sensing for Daily Life, in *Proc. of the Augmented Human International Conference (AH) 2010*. Cited on pages 24, 41, 81, 83, and 136.

S. Ishimaru, J. Weppner, K. Kunze, K. Kise, A. Dengel, P. Lukowicz, and A. Bulling (2014). In the Blink of an Eye – Combining Head Motion and Eye Blink Frequency for Activity Recognition with Google Glass, in *Proc. of the Augmented Human International Conference (AH) 2014*. Cited on pages 19, 136, and 138.

L. Itti and P. Baldi (2005). A Principled Approach to Detecting Surprising Events in Video, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2005*. Cited on page 24.

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 161.

L. Itti and C. Koch (2001). Computational Modelling of Visual Attention, *Nature Reviews Neuroscience*, vol. 2(3), p. 194. Cited on pages 17 and 98.

R. J. Jacob and K. S. Karn (2003). Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises, in *The Mind's Eye 2003*, pp. 573–605, Elsevier. Cited on pages 3, 12, 16, and 19.

A.-H. Javadi, Z. Hakimi, M. Barati, V. Walsh, and L. Tcheang (2015). SET: A Pupil Detection Method Using Sinusoidal Approximation, *Frontiers in Neuroengineering*, vol. 8. Cited on pages 8 and 80.

Q. Ji and X. Yang (2002). Real-Time Eye, Gaze, and Face Pose Tracking for Monitoring Driver Vigilance, *Real-Time Imaging*, vol. 8(5), pp. 357–377. Cited on page 9.

X. Jiang, G. Tien, D. Huang, B. Zheng, and M. S. Atkins (2013). Capturing and Evaluating Blinks from Video-Based Eyetrackers, *Behavior Research Methods*, vol. 45(3), pp. 656–663. Cited on page 12.

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 (INFOCOM) 2016*. Cited on page 160.

O. P. John and S. Srivastava (1999). The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives, *Handbook of Personality: Theory and Research*, vol. 2(1999), pp. 102–138. Cited on page 128.

J. A. Jones, J. E. Swan II, G. Singh, E. Kolstad, and S. R. Ellis (2008). The Effects of Virtual Reality, Augmented Reality, and Motion Parallax on Egocentric Depth Perception, in *Proc. of the ACM Symposium on Applied Perception in Graphics and Visualization 2008*. Cited on page 17.

J. Jung, Y. Matsuba, R. Mallipeddi, H. Funaya, K. Ikeda, and M. Lee (2013). Evolutionary Programming Based Recommendation System for Online Shopping, in *Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA) 2013*. Cited on page 42.

M. A. Just and P. A. Carpenter (1976). Eye Fixations and Cognitive Processes, *Cognitive Psychology*, vol. 8(4), pp. 441–480. Cited on page 140.

M. Kandemir and S. Kaski (2012). Learning Relevance from Natural Eye Movements in Pervasive Interfaces, in *Proc. of the ACM International Conference on Multimodal Interaction (ICMI) 2012*. Cited on pages 13 and 20.

E. Kasneci, G. Kasneci, T. C. Kübler, and W. Rosenstiel (2015). Online Recognition of Fixations, Saccades, and Smooth Pursuits for Automated Analysis of Traffic Hazard Perception, in *Artificial Neural Networks 2015*, pp. 411–434, Springer. Cited on pages 13 and 62.

P. Kasprowski (2004). *Human Identification Using Eye Movements*, Ph.D. thesis. Cited on pages 17 and 120.

P. Kasprowski and J. Ober (2003). Eye Movement Tracking for Human Identification, in *Proc. of the BIOMETRICS World Conference 2003*. Cited on pages 17 and 120.

P. Kasprowski and J. Ober (2005). Enhancing Eye-Movement-Based Biometric Identification Method by Using Voting Classifiers, in *Proc. of the Biometric Technology for Human Identification II 2005*. Cited on pages 17 and 120.

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 (UbiComp) 2014*. Cited on pages 8, 11, 21, 31, 32, 36, 38, 51, 53, 60, 65, 67, 80, 84, 92, 93, 98, 102, 127, 145, 166, 167, 168, and 210.

A. E. Kaufman, A. Bandopadhay, and B. D. Shaviv (1993). An Eye Tracking Computer User Interface, in *Proc. of the IEEE Research Properties in Virtual Reality Symposium 1993*. Cited on page 2.

H. Kera, R. Yonetani, K. Higuchi, and Y. Sato (2016). Discovering Objects of Joint Attention via First-Person Sensing, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2016*. Cited on page 62.

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 (CHI) 2010*. Cited on pages 25, 37, and 158.

M. Khamis, O. Saltuk, A. Hang, K. Stolz, A. Bulling, and F. Alt (2016). TextPursuits: Using Text for Pursuits-Based Interaction and Calibration on Public Displays, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2016*. Cited on page 10.

D. Kim, S. Choi, J. Choi, H. Shin, and K. Sohn (2011). Visual Fatigue Monitoring System Based on Eye-Movement and Eye-Blink Detection, in *Proc. of the Stereoscopic Displays and Applications XXII 2011*. Cited on page 12.

E. S. Kim, A. Naples, G. V. Gearty, Q. Wang, S. Wallace, C. Wall, M. Perlmutter, J. Kowitt, L. Friedlaender, B. Reichow, F. Volkmar, and F. Shic (2014). Development of an Untethered, Mobile, Low-Cost Head-Mounted Eye Tracker, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2014*. Cited on page 80.

K.-N. Kim and R. Ramakrishna (1999). Vision-Based Eye-Gaze Tracking for Human Computer Interface, in *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics (SMC) 1999*. Cited on page 7.

D. E. King (2009). Dlib-ml: A Machine Learning Toolkit, *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758. Cited on page 165.

T. Kinnunen, F. Sedlak, and R. Bednarik (2010). Towards Task-Independent Person Authentication Using Eye Movement Signals, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2010*. Cited on page 120.

T. Kinsman, P. Bajorski, and J. B. Pelz (2010). Hierarchical Image Clustering for Analyzing Eye Tracking Videos, in *Proc. of the Western New York Image Processing Workshop (WNYIPW) 2010*. Cited on page 63.

T. Kinsman, K. Evans, G. Sweeney, T. Keane, and J. Pelz (2012). Ego-Motion Compensation Improves Fixation Detection in Wearable Eye Tracking, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2012*. Cited on pages 22 and 62.

K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto (2011). Fast Unsupervised Ego-Action Learning for First-Person Sports Videos, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 20.

C. Koch and S. Ullman (1987). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry, in *Matters of Intelligence 1987*, pp. 115–141, Springer. Cited on page 10.

M. Koelle, S. Boll, T. Olsson, J. Williamson, H. Profita, S. Kane, and R. Mitchell (2018a). (Un) Acceptable!?!: Re-Thinking the Social Acceptability of Emerging Technologies, in *Proc. of the ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI) 2018*. Cited on pages 16 and 34.

M. Koelle, W. Heuten, and S. Boll (2017). Are You Hiding It? Usage Habits of Lifelogging Camera Wearers, in *Proc. of the ACM International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI) 2017*. Cited on page 98.

M. Koelle, M. Kranz, and A. Möller (2015). Don't Look at Me That Way!: Understanding User Attitudes Towards Data Glasses Usage, in *Proc. of the ACM International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI) 2015*. Cited on pages 16, 18, 23, 34, 100, and 101.

M. Koelle, K. Wolf, and S. Boll (2018b). Beyond LED Status Lights – Design Requirements of Privacy Notices for Body-Worn Cameras, in *Proc. of the ACM International Conference on Tangible, Embedded, and Embodied Interaction (TEI) 2018*. Cited on pages 23, 98, 101, and 106.

S. Kolski, K. Macek, L. Spinello, and R. Siegwart (2007). Secure Autonomous Driving in Dynamic Environments: From Object Detection to Safe Driving, in *Proc. of the Workshop on Safe Navigation in Open and Dynamic Environments: Applications to Autonomous Vehicles at the IEEE International Conference on Intelligent Robots and Systems (IROS) 2007*. Cited on page 24.

O. V. Komogortsev and C. D. Holland (2013). Biometric Authentication via Complex Oculomotor Behavior, in *Proc. of the IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS) 2013*. Cited on pages 17 and 120.

O. V. Komogortsev, S. Jayarathna, C. R. Aragon, and M. Mahmoud (2010). Biometric Identification via an Oculomotor Plant Mathematical Model, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2010*. Cited on pages 17 and 120.

O. V. Komogortsev and A. Karpov (2013). Automated Classification and Scoring of Smooth Pursuit Eye Movements in the Presence of Fixations and Saccades, *Behavior Research Methods*, vol. 45(1), pp. 203–215. Cited on pages 12 and 62.

M. Korayem, R. Templeman, D. Chen, D. Crandall, and A. Kapadia (2014). ScreenAvoider: Protecting Computer Screens from Ubiquitous Cameras, *arXiv preprint arXiv:1412.0008*. Cited on page 113.

M. Korayem, R. Templeman, D. Chen, D. Crandall, and A. Kapadia (2016). Enhancing Lifelogging Privacy by Detecting Screens, in *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI) 2016*. Cited on pages 18, 100, 101, and 188.

R. Kothari and J. L. Mitchell (1996). Detection of Eye Locations in Unconstrained Visual Images, in *Proc. of the IEEE International Conference on Image Processing (ICIP) 1996*. Cited on page 7.

K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba (2016). Eye Tracking for Everyone, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016*. Cited on pages 8, 80, and 90.

K. Krombholz, A. Dabrowski, M. Smith, and E. Weippl (2015). Ok Glass, Leave Me Alone: Towards a Systematization of Privacy Enhancing Technologies for Wearable Computing, in *Proc. of the International Conference on Financial Cryptography and Data Security 2015*. Cited on pages 18 and 100.

K. Krombholz, A. Dabrowski, M. Smith, and E. Weippl (2017). Exploring design directions for wearable privacy, in *Proc. of the Workshop on Usable Security (USEC) 2017*. Cited on page 18.

C. A. Kuechenmeister, P. H. Linton, T. V. Mueller, and H. B. White (1977). Eye Tracking in Relation to Age, Sex, and Illness, *Archives of General Psychiatry*, vol. 34(5), pp. 578–579. Cited on pages 17, 19, 41, and 119.

K. Kunze, H. Kawaichi, K. Yoshimura, and K. Kise (2013a). Towards Inferring Language Expertise Using Eye Tracking, in *Proc. of the ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI) 2013*. Cited on pages 20 and 136.

K. Kunze, H. Kawaichi, K. Yoshimura, and K. Kise (2013b). The Wordometer – Estimating the Number of Words Read Using Document Image Retrieval and Mobile Eye Tracking, in *Proc. of the International Conference on Document Analysis and Recognition (ICDAR) 2013*. Cited on pages 20, 123, and 136.

K. Kunze, K. Masai, M. Inami, Ö. Sacakli, M. Liwicki, A. Dengel, S. Ishimaru, and K. Kise (2015). Quantifying Reading Habits: Counting How Many Words You Read, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2015*. Cited on page 123.

K. Kunze, Y. Utsumi, Y. Shiga, K. Kise, and A. Bulling (2013c). I Know What You Are Reading: Recognition of Document Types Using Mobile Eye Tracking, in *Proc. of the ACM International Symposium on Wearable Computers (ISWC) 2013*. Cited on pages 20, 123, 127, and 138.

K. Kurzhals, M. Hlawatsch, M. Burch, and D. Weiskopf (2016a). Fixation-Image Charts, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2016*. Cited on page 63.

K. Kurzhals, M. Hlawatsch, F. Heimerl, M. Burch, T. Ertl, and D. Weiskopf (2016b). Gaze Stripes: Image-Based Visualization of Eye Tracking Data, *IEEE Transactions on Visualization and Computer Graphics*, vol. 22(1), pp. 1005–1014. Cited on pages 15 and 63.

K. Kurzhals, M. Hlawatsch, C. Seeger, and D. Weiskopf (2017). Visual Analytics for Mobile Eye Tracking, *IEEE Transactions on Visualization and Computer Graphics*, vol. 23(1), pp. 301–310. Cited on pages 15, 32, 60, 62, 63, and 64.

K.-M. Lam and H. Yan (1996). Locating and Extracting the Eye in Human Face Images, *Pattern Recognition*, vol. 29(5), pp. 771–779. Cited on page 7.

M. Land and B. Tatler (2009). *Looking and Acting: Vision and Eye Movements in Natural Behaviour*, Oxford University Press. Cited on page 20.

C. Lander, S. Gehring, A. Krüger, S. Boring, and A. Bulling (2015). GazeProjector: Accurate Gaze Estimation and Seamless Gaze Interaction Across Multiple Displays, in *Proc. of the ACM Symposium on User Interface Software and Technology (UIST) 2015*. Cited on pages 20 and 78.

C. Lander, S. Gehring, M. Löchtefeld, A. Bulling, and A. Krüger (2017a). Eyemirror: Mobile Calibration-Free Gaze Approximation Using Corneal Imaging, in *Proc. of the ACM International Conference on Mobile and Ubiquitous Multimedia (MUM) 2017*. Cited on page 11.

C. Lander, F. Kosmalla, F. Wiehr, and S. Gehring (2017b). Using Corneal Imaging for Measuring a Human's Visual Attention, in *Proc. of the ACM International Symposium on Wearable Computers (ISWC) 2017*. Cited on page 11.

C. Lander and A. Krüger (2018). EyeSense: Towards Information Extraction on Corneal Images, in *Proc. of the ACM International Joint Conference and International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp/ISWC) 2018*. Cited on page 11.

C. Lander, M. Löchtefeld, and A. Krüger (2018a). hEYEbrid: A Hybrid Approach for Mobile Calibration-Free Gaze Estimation, *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 1(4), p. 149. Cited on page 11.

C. Lander, M. Speicher, F. Kerber, and A. Krüger (2018b). Towards Fixation Extraction in Corneal Imaging Based Eye Tracking Data, in *Proc. of the ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI) 2018*. Cited on page 11.

V. Laurutis and D. Robinson (1986). The Vestibulo-Ocular Reflex During Human Saccadic Eye Movements, *The Journal of Physiology*, vol. 373(1), pp. 209–233. Cited on page 21.

H. Le, T. Dang, and F. Liu (2013). Eye Blink Detection for Smart Glasses, in *Proc. of the IEEE International Symposium on Multimedia (ISM) 2013*. Cited on page 12.

Y. J. Lee, J. Ghosh, and K. Grauman (2012). Discovering Important People and Objects for Egocentric Video Summarization, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 24.

A. Lefohn, B. Budge, P. Shirley, R. Caruso, and E. Reinhard (2003). An Ocularist's Approach to Human Iris Synthesis, *IEEE Transactions on Computer Graphics and Applications*, vol. 23(6), pp. 70–75. Cited on pages 31 and 52.

C.-s. R. Li, W.-h. Lin, Y.-y. Yang, C.-c. Huang, T.-w. Chen, and Y.-c. Chen (2002). Impairment of Temporal Attention in Patients with Schizophrenia, *Neuroreport*, vol. 13(11), pp. 1427–1430. Cited on page 12.

D. Li and D. Parkhurst (2006). Open-Source Software for Real-Time Visible-Spectrum Eye Tracking, in *Proc. of the Conference on Communication by Gaze Interaction (COGAIN) 2006*. Cited on page 80.

D. Li, D. Winfield, and D. J. Parkhurst (2005). Starburst: A Hybrid Algorithm for Video-Based Eye Tracking Combining Feature-Based and Model-Based Approaches, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2005*. Cited on pages 8 and 80.

J. Li, G. Ngai, H. V. Leong, and S. C. Chan (2016). Your Eye Tells How Well You Comprehend, in *Proc. of the Computer Software and Applications Conference (COMPSAC) 2016*. Cited on page 60.

D. J. Liebling and S. Preibusch (2014). Privacy Considerations for a Pervasive Eye Tracking World, in *Adj. Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2014*. Cited on pages 113 and 118.

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 (CVPR) 2015*. Cited on page 165.

S. S. Liu, A. Rawicz, T. Ma, C. Zhang, K. Lin, S. Rezaei, and E. Wu (2018). An Eye-Gaze Tracking and Human Computer Interface System for People with ALS and Other Locked-In Diseases, *Canadian Medical and Biological Engineering Society (CMBES) Proceedings*, vol. 33(1). Cited on page 17.

X. Long, O. K. Tonguz, and A. Kiderman (2007). A High Speed Eye Tracking System with Robust Pupil Center Estimation Algorithm, in *Proc. of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2007*. Cited on page 80.

F. Lu, Y. Sugano, T. Okabe, and Y. Sato (2014). Adaptive Linear Regression for Appearance-Based Gaze Estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 36(10), pp. 2033–2046. Cited on pages 80 and 83.

G. Luo, N. M. Rensing, E. Weststrate, and E. Peli (2005). Registration of an On-Axis See-Through Head-Mounted Display and Camera System, *Optical Engineering*, vol. 44(2). Cited on page 17.

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 (CVPR) 2016*. Cited on pages 20 and 176.

J. J. MacInnes, S. Iqbal, J. Pearson, and E. N. Johnson (2018). Wearable Eye-Tracking for Research: Automated Dynamic Gaze Mapping and Accuracy/Precision Comparisons Across Devices, *bioRxiv*. Cited on page 4.

A. J. Maeder and C. B. Fookes (2003). A Visual Attention Approach to Personal Identification. Cited on pages 17 and 120.

P. Majaranta and A. Bulling (2014). *Eye Tracking and Eye-Based Human-Computer Interaction*, pp. 39–65, Springer Publishing London. Cited on page 78.

P. Majaranta and K.-J. Räihä (2002). Twenty Years of Eye Typing: Systems and Design Issues, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2002*. Cited on page 20.

H. Manabe and M. Fukumoto (2006). Full-Time Wearable Headphone-Type Gaze Detector, in *Proc. of the ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI) 2006*. Cited on page 81.

M. Mansouryar, J. Steil, Y. Sugano, and A. Bulling (2016). 3D Gaze Estimation from 2D Pupil Positions on Monocular Head-Mounted Eye Trackers, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2016*. Cited on pages 28, 31, and 80.

D. Mardanbegi and D. W. Hansen (2012). Parallax Error in the Monocular Head-Mounted Eye Trackers, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2012*. Cited on pages 5, 31, 50, and 181.

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 (CHI) 2015*. Cited on pages 25, 37, and 158.

S. P. Marshall (2002). The Index of Cognitive Activity: Measuring Cognitive Workload, in *Proc. of the IEEE Conference on Human Factors and Power Plants 2002*. Cited on pages 14, 136, and 138.

B. Massé, S. Ba, and R. Horaud (2017). Tracking Gaze and Visual Focus of Attention of People Involved in Social Interaction, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on page 60.

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 (UbiComp) 2016*. Cited on pages 158 and 160.

G. Matthews, W. Middleton, B. Gilmartin, and M. Bullimore (1991). Pupillary Diameter and Cognitive Load, *Journal of Psychophysiology*. Cited on pages 17 and 119.

A. Mayberry, P. Hu, B. Marlin, C. Salthouse, and D. Ganesan (2014). iShadow: Design of a Wearable, Real-Time Mobile Gaze Tracker, in *Proc. of the International Conference on Mobile Systems, Applications, and Services (MobiSys) 2014*. Cited on page 80.

P. Mayring (2014). *Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution*. Cited on pages 110 and 190.

L. K. McIntire, R. A. McKinley, C. Goodyear, and J. P. McIntire (2014). Detection of Vigilance Performance Using Eye Blinks, *Applied Ergonomics*, vol. 45(2), pp. 354–362. Cited on page 12.

M. L. Mele and S. Federici (2012). Gaze and Eye-Tracking Solutions for Psychological Research, *Cognitive Processing*, vol. 13(1), pp. 261–265. Cited on page 3.

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 158 and 160.

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 (IJCAI) 2016*. Cited on page 161.

C. S. Montero, J. Alexander, M. T. Marshall, and S. Subramanian (2010). Would You Do That? – Understanding Social Acceptance of Gestural Interfaces, in *Proc. of the ACM International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI) 2010*. Cited on pages 16 and 22.

T. Moriyama, T. Kanade, J. F. Cohn, J. Xiao, Z. Ambadar, J. Gao, and H. Imamura (2002). Automatic Recognition of Eye Blinking in Spontaneously Occurring Behavior, in *Proc. of Object Recognition Supported by User Interaction for Service Robots 2002*. Cited on page 12.

V. G. Motti and K. Caine (2016). Towards a Visual Vocabulary for Privacy Concepts, in *Proc. of the Human Factors and Ergonomics Society Annual Meeting 2016*. Cited on page 114.

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 (ETRA) 2019*. Cited on page 10.

S. M. Munn and J. B. Pelz (2008). 3D Point-of-Regard, Position and Head Orientation from a Portable Monocular Video-Based Eye Tracker, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2008*. Cited on pages 4, 31, and 50.

T. Nagamatsu, R. Sugano, Y. Iwamoto, J. Kamahara, and N. Tanaka (2010). User-Calibration-Free Gaze Tracking with Estimation of the Horizontal Angles Between the Visual and the Optical Axes of Both Eyes, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2010*. Cited on pages 10, 31, and 50.

Y. I. Nakano and R. Ishii (2010). Estimating User's Engagement from Eye-Gaze Behaviors in Human-Agent Conversations, in *Proc. of the ACM International Conference on Intelligent User Interfaces (IUI) 2010*. Cited on page 14.

A. Nakazawa and C. Nitschke (2012). Point of Gaze Estimation Through Corneal Surface Reflection in an Active Illumination Environment, *Proc. of the European Conference on Computer Vision (ECCV)*, pp. 159–172. Cited on page 80.

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 pages 16, 33, 78, and 176.

C. Nguyen and F. Liu (2016). Gaze-Based Notetaking for Learning from Lecture Videos, in *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI) 2016*. Cited on page 60.

J. C. Niebles, H. Wang, and L. Fei-Fei (2008). Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words, *International Journal of Computer Vision*, vol. 79(3), pp. 299–318. Cited on page 138.

J. Nielsen (1994). *Usability Engineering*, Elsevier. Cited on page 16.

M. Nixon (1985). Eye Spacing Measurement for Facial Recognition, in *Proc. of the Applications of Digital Image Processing VIII 1985*. Cited on page 7.

B. Noris, J.-B. Keller, and A. Billard (2011). A Wearable Gaze Tracking System for Children in Unconstrained Environments, *Computer Vision and Image Understanding*, vol. 115(4), pp. 476–486. Cited on page 80.

M. Nyström, R. Andersson, K. Holmqvist, and J. Van De Weijer (2013). The Influence of Calibration Method and Eye Physiology on Eyetracking Data Quality, *Behavior Research Methods*, vol. 45(1), pp. 272–288. Cited on page 3.

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 (UbiComp) 2016*. Cited on pages 160 and 162.

K. Ogaki, K. M. Kitani, Y. Sugano, and Y. Sato (2012). Coupling Eye-Motion and Ego-Motion Features for First-Person Activity Recognition, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2012*. Cited on page 20.

K. Ooms, L. Dupont, L. Lapon, and S. Popelka (2015). Accuracy and Precision of Fixation Locations Recorded with the Low-Cost Eye Tribe Tracker in Different Experimental Setups, *Journal of Eye Movement Research*, vol. 8(1). Cited on page 3.

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 176.

T. Orekondy, B. Schiele, and M. Fritz (2017). Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images, in *Proc. of the IEEE International Conference on Computer Vision (ICCV) 2017*. Cited on pages 103 and 186.

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 (CHI) 2005*. Cited on pages 24, 37, 158, 160, 166, 167, and 168.

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 (ETRA) 2014*. Cited on page 160.

L. Paletta, K. Santner, and G. Fritz (2013a). An Integrated System for 3D Gaze Recovery and Semantic Analysis of Human Attention, *arXiv preprint arXiv:1307.7848*. Cited on page 13.

L. Paletta, K. Santner, G. Fritz, A. Hofmann, G. Lodron, G. Thallinger, and H. Mayer (2013b). A Computer Vision System for Attention Mapping in SLAM Based 3D Models, *arXiv preprint arXiv:1305.1163*. Cited on page 13.

L. Paletta, K. Santner, G. Fritz, H. Mayer, and J. Schrammel (2013c). 3D Attention: Measurement of Visual Saliency Using Eye Tracking Glasses, in *Proc. of the ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI) 2013*. Cited on page 13.

O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman (2010). Estimating Cognitive Load Using Remote Eye Tracking in a Driving Simulator, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2010*. Cited on pages 3, 14, and 138.

A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn (2016). Towards Foveated Rendering for Gaze-Tracked Virtual Reality, *ACM Transactions on Graphics (TOG)*, vol. 35(6), p. 179. Cited on pages 17 and 41.

A. Peréz, M. L. Córdoba, A. Garcia, R. Méndez, M. Munoz, J. L. Pedraza, and F. Sanchez (2003). A Precise Eye-Gaze Detection and Tracking System, in *Proc. of the International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG) 2003*. Cited on page 7.

A. J. Perez, S. Zeadally, and S. Griffith (2017). Bystanders' Privacy, *IT Professional*, vol. 19(3), pp. 61–65. Cited on pages 18 and 98.

T. Pfeiffer and P. Renner (2014). EyeSee3D: A Low-Cost Approach for Analyzing Mobile 3D Eye Tracking Data Using Computer Vision and Augmented Reality Technology, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2014*. Cited on pages 4, 31, and 50.

K. Pfeuffer, J. Alexander, M. K. Chong, and H. Gellersen (2014). Gaze-touch: Combining Gaze with Multi-Touch for Interaction on the Same Surface, in *Proc. of the ACM Symposium on User Interface Software and Technology (UIST) 2014*. Cited on page 19.

K. Pfeuffer, M. Vidal, J. Turner, A. Bulling, and H. Gellersen (2013). Pursuit Calibration: Making Gaze Calibration Less Tedious and More Flexible, in *Proc. of the ACM Symposium on User Interface Software and Technology (UIST) 2013*. Cited on page 10.

P. Piasek, K. Irving, and A. F. Smeaton (2014). Using Lifelogging to Help Construct the Identity of People with Dementia, in *Proc. of the Conference on Interfaces and Human Computer Interaction (IHCI) 2014*. Cited on page 24.

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 (IMWUT)*, vol. 1(3), p. 91. Cited on page 160.

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 (UbiComp) 2015*. Cited on pages 158 and 160.

B. R. Pires, M. Devyver, A. Tsukada, and T. Kanade (2013). Unwrapping the Eye for Visible-spectrum Gaze Tracking on Wearable Devices, in *Proc. of IEEE Workshop on Applications of Computer Vision (WACV) 2013*. Cited on page 80.

A. Plopski, C. Nitschke, K. Kiyokawa, D. Schmalstieg, and H. Takemura (2015). Hybrid Eye Tracking: Combining Iris Contour and Corneal Imaging, in *Proc. of the International Conference on Artificial Reality and Telexistence & Eurographics Symposium on Virtual Environments (ICAT-EGVE) 2015*. Cited on page 80.

D. F. Pontillo, T. B. Kinsman, and J. B. Pelz (2010). SemantiCode: Using Content Similarity and Database-Driven Matching to Code Wearable Eyetracker Gaze Data, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2010*. Cited on pages 62 and 63.

A. Poole and L. J. Ball (2005). Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects, in *C. Ghaoui (Ed.): Encyclopedia of Human-Computer Interaction. Pennsylvania: Idea Group, Inc. 2005*. Cited on page 14.

A. Poole and L. J. Ball (2006). Eye Tracking in HCI and Usability Research, in *Encyclopedia of Human Computer Interaction 2006*, pp. 211–219, IGI Global. Cited on page 16.

R. S. Portnoff, L. N. Lee, S. Egelman, P. Mishra, D. Leung, and D. Wagner (2015). Somebody's Watching Me? Assessing the Effectiveness of Webcam Indicator Lights, in *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI) 2015*. Cited on page 114.

Z. Pousman, G. Iachello, R. Fithian, J. Moghazy, and J. Stasko (2004). Design Iterations for a Location-Aware Event Planner, *Personal and Ubiquitous Computing*, vol. 8(2), pp. 117–125. Cited on page 114.

S. Preibusch (2014). Eye-Tracking. Privacy Interfaces for the Next Ubiquitous Modality, in *Proc. of the W3C Workshop on Privacy and User-Centric Controls 2014*. Cited on page 113.

B. A. Price, A. Stuart, G. Calikli, C. McCormick, V. Mehta, L. Hutton, A. K. Bandara, M. Levine, and B. Nuseibeh (2017). Logging You, Logging Me: A Replicable Study of Privacy and Sharing Behaviour in Groups of Visual Lifeloggers, *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 1(2), pp. 22:1–22:18. Cited on pages 100, 104, 112, and 113.

H. Profita, R. Albaghli, L. Findlater, P. Jaeger, and S. K. Kane (2016). The AT Effect: How Disability Affects the Perceived Social Acceptability of Head-Mounted Display Use, in *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI) 2016*. Cited on pages 16, 18, and 100.

A. Pyrgelis, C. Troncoso, and E. De Cristofaro (2017). Knock Knock, Who's There? Membership Inference on Aggregate Location Data, *arXiv preprint arXiv:1708.06145*. Cited on page 120.

P. Qvarfordt and S. Zhai (2005). Conversing with the User Based on Eye-Gaze Patterns, in *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI) 2005*. Cited on pages 3 and 19.

N. Raval, A. Srivastava, K. Lebeck, L. Cox, and A. Machanavajjhala (2014). MarkIt: Privacy Markers for Protecting Visual Secrets, in *Adj. Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2014*. Cited on pages 18, 100, and 101.

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 161.

K. Rayner (1998). Eye Movements in Reading and Information Processing: 20 Years of Research, *Psychological Bulletin*, vol. 124(3), p. 372. Cited on pages 3 and 13.

K. Rayner (2012). *Eye Movements and Visual Cognition: Scene Perception and Reading*, Springer Science & Business Media. Cited on page 20.

K. Rayner and A. Pollatsek (1992). Eye Movements and Scene Perception, *Canadian Journal of Psychology*, vol. 46(3), p. 342. Cited on page 3.

K. Rayner, C. M. Rotello, A. J. Stewart, J. Keir, and S. A. Duffy (2001). Integrating Text and Pictorial Information: Eye Movements When Looking at Print Advertisements, *Journal of Experimental Psychology: Applied*, vol. 7(3), p. 219. Cited on page 3.

M. Reani, N. Peek, and C. Jay (2018). An Investigation of the Effects of n-Gram Length in Scanpath Analysis for Eye-Tracking Research, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2018*. Cited on page 14.

M. J. Reinders, R. Koch, and J. J. Gerbrands (1996). Locating Facial Features in Image Sequences Using Neural Networks, in *Proc. of the IEEE International Conference on Automatic Face and Gesture Recognition (FG) 1996*. Cited on page 7.

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 165.

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 16, 33, 78, and 176.

D. A. Robinson (1963). A Method of Measuring Eye Movement Using a Scleral Search Coil in a Magnetic Field, *IEEE Transactions on Bio-medical Electronics*, vol. 10(4), pp. 137–145. Cited on page 2.

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 pages 24 and 158.

R. Ruddarraju, A. Haro, and I. Essa (2003a). Fast Multiple Camera Head Pose Tracking, in *Proc. of the Vision Interface Conference 2003*. Cited on page 81.

R. Ruddarraju, A. Haro, K. Nagel, Q. T. Tran, I. A. Essa, G. Abowd, and E. D. Mynatt (2003b). Perceptual User Interfaces Using Vision-Based Eye Tracking, in *Proc. of the International Conference on Multimodal Interfaces (ICMI) 2003*. Cited on page 81.

S. Sah, S. Kulhare, A. Gray, S. Venugopalan, E. Prud'hommeaux, and R. Ptucha (2017). Semantic Text Summarization of Long Videos, in *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV) 2017*. Cited on page 24.

N. Saleheen, S. Chakraborty, N. Ali, M. M. Rahman, S. M. Hossain, R. Bari, E. Buder, M. Srivastava, and S. Kumar (2016). mSieve: Differential Behavioral Privacy in Time Series of Mobile Sensor Data, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2016*. Cited on page 120.

D. D. Salvucci and J. R. Anderson (1998). Tracing Eye Movement Protocols with Cognitive Process Models, in *Proc. of the Conference of the Cognitive Science Society 1998*. Cited on page 19.

D. D. Salvucci and J. R. Anderson (2001). Automated Eye-Movement Protocol Analysis, *Human-Computer Interaction*, vol. 16(1), pp. 39–86. Cited on pages 24 and 137.

D. D. Salvucci and J. H. Goldberg (2000). Identifying Fixations and Saccades in Eye-Tracking Protocols, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2000*. Cited on pages 12, 60, 62, 67, and 139.

F. Samaria and S. Young (1994). HMM-Based Architecture for Face Identification, *Image and Vision Computing*, vol. 12(8), pp. 537–543. Cited on page 8.

N. Sammaknejad, H. Pouretemad, C. Eslahchi, A. Salahirad, and A. Alinejad (2017). Gender Classification Based on Eye Movements: A Processing Effect During Passive Face Viewing, *Advances in Cognitive Psychology*, vol. 13(3), p. 232. Cited on pages 18 and 119.

J. San Agustin, J. P. Hansen, and M. Tall (2010a). Gaze-Based Interaction with Public Displays Using Off-the-Shelf Components, in *Adj. Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2010*. Cited on page 3.

J. San Agustin, H. Skovsgaard, E. Mollenbach, M. Barret, M. Tall, D. W. Hansen, and J. P. Hansen (2010b). Evaluation of a Low-Cost Open-Source Gaze Tracker, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2010*. Cited on page 80.

T. Santini, W. Fuhl, T. Kübler, and E. Kasneci (2016). Bayesian Identification of Fixations, Saccades, and Smooth Pursuits, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2016*. Cited on pages 13 and 62.

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 165.

H. Sattar, A. Bulling, and M. Fritz (2017a). Predicting the Category and Attributes of Visual Search Targets Using Deep Gaze Pooling, in *Proc. of the IEEE International Conference on Computer Vision Workshops (ICCVW) 2017*. Cited on page 63.

H. Sattar, M. Fritz, and A. Bulling (2017b). Visual Decoding of Targets During Visual Search from Human Eye Fixations, Technical report. Cited on pages 19 and 78.

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 (CVPR) 2015*. Cited on pages 3, 19, 63, and 78.

J. Schiff, M. Meingast, D. K. Mulligan, S. Sastry, and K. Goldberg (2007). Respectful Cameras: Detecting Visual Markers in Real-Time to Address Privacy Concerns, in *Proc. of the IEEE International Conference on Intelligent Robots and Systems (IROS) 2007*. Cited on page 98.

R. Schlegel, A. Kapadia, and A. J. Lee (2011). Eyeing Your Exposure: Quantifying and Controlling Information Sharing for Improved Privacy, in *Proc. of the Symposium on Usable Privacy and Security (SOUPS) 2011*. Cited on page 114.

R. Schleicher, N. Galley, S. Briest, and L. Galley (2008). Blinks and Saccades as Indicators of Fatigue in Sleepiness Warnings: Looking Tired?, *Ergonomics*, vol. 51(7), pp. 982–1010. Cited on pages 12 and 137.

T. Schneider, B. Schauerte, and R. Stiefelhagen (2014). Manifold Alignment for Person Independent Appearance-Based Gaze Estimation, in *Proc. of the IEEE International Conference on Pattern Recognition (ICPR) 2014*. Cited on page 10.

J. Schrammel, G. Regal, and M. Tscheligi (2014). Attention Approximation of Mobile Users Towards Their Environment, in *Proc. of the ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI) 2014*. Cited on page 13.

J. Seiter, O. Amft, M. Rossi, and G. Tröster (2014). Discovery of Activity Composites Using Topic Models: An Analysis of Unsupervised Methods, *Pervasive and Mobile Computing*, vol. 15, pp. 215 – 227. Cited on pages 138 and 141.

Y. Shiga, T. Toyama, Y. Utsumi, K. Kise, and A. Dengel (2014). Daily Activity Recognition Combining Gaze Motion and Visual Features, in *Adj. Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2014*. Cited on pages 63 and 146.

K. Shinohara (2017). *Design for Social Accessibility: Incorporating Social Factors in the Design of Accessible Technologies*, Ph.D. thesis. Cited on page 16.

J. Shu, R. Zheng, and P. Hui (2016). Cardea: Context-Aware Visual Privacy Protection from Pervasive Cameras, *arXiv preprint arXiv:1610.00889*. Cited on pages 18 and 101.

A. R. Silva, S. Pinho, L. M. Macedo, and C. J. Moulin (2013). Benefits of SenseCam Review on Neuropsychological Test Performance, *American Journal of Preventive Medicine*, vol. 44(3), pp. 302–307. Cited on page 24.

S. A. Sirohey and A. Rosenfeld (2001). Eye Detection in a Face Image Using Linear and Nonlinear Filters, *Pattern Recognition*, vol. 34(7), pp. 1367–1391. Cited on page 7.

R. Sousa, M. Wäny, P. Santos, and F. Morgado-Dias (2017). NanEye – An Endoscopy Sensor with 3D Image Synchronization, *IEEE Sensors Journal*, vol. 17, pp. 623–631. Cited on page 83.

O. Špakov and D. Miniotas (2007). Visualization of Eye Gaze Data Using Heat Maps, *Elektronika ir elektrotechnika*, pp. 55–58. Cited on page 13.

B. Steichen, G. Carenini, and C. Conati (2013). User-Adaptive Information Visualization: Using Eye Gaze Data to Infer Visualization Tasks and User Cognitive Abilities, in *Proc. of the ACM International Conference on Intelligent User Interfaces (IUI) 2013*. Cited on pages 19 and 138.

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 (UbiComp) 2015*. Cited on pages 18, 23, 28, 36, 60, 61, 78, 99, 104, 119, 129, and 164.

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 (ETRA) 2019*. Cited on pages 28 and 35.

J. Steil, M. X. Huang, and A. Bulling (2018a). Fixation Detection for Head-Mounted Eye Tracking Based on Visual Similarity of Gaze Targets, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2018*. Cited on pages 28 and 32.

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 (ETRA) 2019*. Cited on pages 28 and 34.

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 ACM International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI) 2018*. Cited on pages 28, 34, 37, and 104.

J. Steil, M. Tonsen, Y. Sugano, and A. Bulling (2019c). InvisibleEye: Fully Embedded Mobile Eye Tracking Using Appearance-Based Gaze Estimation, *GetMobile: Mobile Computing and Communications*, vol. 23(2), pp. 13–20. Cited on pages 28 and 33.

S. Stellmach and R. Dachselt (2012). Look & Touch: Gaze-Supported Target Acquisition, in *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI) 2012*. Cited on pages 17, 20, 35, and 118.

S. Stellmach and R. Dachselt (2013). Still Looking: Investigating Seamless Gaze-Supported Selection, Positioning, and Manipulation of Distant Targets, in *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI) 2013*. Cited on pages 3, 20, and 78.

M. Steyvers and T. Griffiths (2007). Probabilistic Topic Models, *Handbook of Latent Semantic Analysis*, vol. 427(7), pp. 424–440. Cited on page 143.

R. Stiefelhagen, J. Yang, and A. Waibel (1997a). A Model-Based Gaze Tracking System, *International Journal on Artificial Intelligence Tools*, vol. 6(02), pp. 193–209. Cited on page 7.

R. Stiefelhagen, J. Yang, and A. Waibel (1997b). Tracking Eyes and Monitoring Eye Gaze, in *Proc. of the Workshop on Perceptual User Interfaces 1997*. Cited on page 7.

Y.-C. Su and K. Grauman (2016). Detecting Engagement in Egocentric Video, in *Proc. of the European Conference on Computer Vision 2016*. Cited on page 20.

Y. Sugano and A. Bulling (2015). Self-Calibrating Head-Mounted Eye Trackers Using Egocentric Visual Saliency, in *Proc. of the ACM Symposium on User Interface Software and Technology (UIST) 2015*. Cited on pages 10, 32, 61, 65, 93, 95, and 104.

Y. Sugano, Y. Matsushita, and Y. Sato (2014). Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014*. Cited on pages 10 and 80.

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 (ECCV) 2008*. Cited on page 10.

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 (UIST) 2016*. Cited on pages 13, 17, 90, 98, and 174.

M. Suzuki, N. Yamamoto, O. Yamamoto, T. Nakano, and S. Yamamoto (2006). Measurement of Driver's Consciousness by Image Processing-A Method for Presuming Driver's Drowsiness by Eye-Blinks Coping with Individual Differences, in *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics (SMC) 2006*. Cited on page 12.

L. Świrski, A. Bulling, and N. Dodgson (2012). Robust Real-Time Pupil Tracking in Highly Off-Axis Images, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2012*. Cited on pages 8, 80, and 82.

L. Świrski and N. Dodgson (2014). Rendering Synthetic Ground Truth Images for Eye Tracker Evaluation, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2014*. Cited on page 82.

L. Świrski and N. A. Dodgson (2013). A Fully-Automatic, Temporal Approach to Single Camera, Glint-Free 3D Eye Model Fitting, in *Proc. of the International Workshop on Pervasive Eye Tracking and Mobile Gaze-Based Interaction (PETMEI) 2013*. Cited on pages 10, 31, 43, 50, 51, and 80.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich (2015). Going Deeper with Convolutions, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on pages 103 and 186.

K. Takemura, K. Takahashi, J. Takamatsu, and T. Ogasawara (2014). Estimating 3-D Point-of-Regard in a Real Environment Using a Head-Mounted Eye-Tracking System, *IEEE Transactions on Human-Machine Systems*, vol. 44(4), pp. 531–536. Cited on pages 4, 13, 31, and 50.

R. Templeman, M. Korayem, D. J. Crandall, and A. Kapadia (2014). PlaceAvoider: Steering First-Person Cameras Away from Sensitive Spaces, in *Proc. of the Network and Distributed System Security Symposium (NDSS) 2014*. Cited on pages 18, 100, and 101.

B. Tessendorf, A. Bulling, D. Roggen, T. Stiefmeier, M. Feilner, P. Derleth, and G. Tröster (2011). Recognition of Hearing Needs from Body and Eye Movements to Improve Hearing Instruments, in *Proc. of the International Conference on Pervasive Computing (Pervasive) 2011*. Cited on pages 136, 138, and 148.

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 (IMWUT)*, vol. 1(3), pp. 106:1–106:21. Cited on pages 17, 18, 28, 32, 33, 60, 98, 118, and 176.

M. Tonsen, X. Zhang, Y. Sugano, and A. Bulling (2016). Labelled Pupils in the Wild: A Dataset for Studying Pupil Detection in Unconstrained Environments, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2016*. Cited on pages 82 and 95.

C. Topal, S. Gunal, O. Koçdeviren, A. Doğan, and Ö. N. Gerek (2014). A Low-Computational Approach on Gaze Estimation with Eye Touch System, *IEEE Transactions on Cybernetics*, vol. 44(2), pp. 228–239. Cited on pages 81 and 83.

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 (ICCV) 2015*. Cited on page 176.

K. Truong, S. Patel, J. Summet, and G. Abowd (2005). Preventing Camera Recording by Designing a Capture-Resistant Environment, *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pp. 903–903. Cited on page 98.

H. Y. Tsang, M. Tory, and C. Swindells (2010). eSeeTrack – Visualizing Sequential Fixation Patterns, *IEEE Transactions on Visualization and Computer Graphics*, vol. 16(6), pp. 953–962. Cited on page 63.

A. Tsukada, M. Shino, M. Devyver, and T. Kanade (2011). Illumination-Free Gaze Estimation Method for First-Person Vision Wearable Device, in *Proc. of the IEEE International Conference on Computer Vision Workshops (ICCVW) 2011*. Cited on page 80.

J. Turner, A. Bulling, J. Alexander, and H. Gellersen (2014). Cross-Device Gaze-Supported Point-to-Point Content Transfer, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2014*. Cited on pages 20 and 78.

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 (UbiComp) 2015*. Cited on page 160.

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 160.

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 (UbiComp) 2016*. Cited on pages 158 and 160.

T. Urruty, S. Lew, N. Ihadaddene, and D. A. Simovici (2007). Detecting Eye Fixations by Projection Clustering, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 3(4), p. 5. Cited on pages 13, 32, 60, and 62.

A. Utsumi, K. Okamoto, N. Hagita, and K. Takahashi (2012). Gaze Tracking in Wide Area Using Multiple Camera Observations, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2012*. Cited on page 81.

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 (PETMEI) 2012*. Cited on page 160.

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 page 165.

J. D. Velásquez (2013). Combining Eye-Tracking Technologies with Web Usage Mining for Identifying Website Keyobjects, *Engineering Applications of Artificial Intelligence*, vol. 26(5-6), pp. 1469–1478. Cited on pages 13 and 42.

R. Vertegaal *et al.* (2003). Attentive User Interfaces, *Communications of the ACM*, vol. 46(3), pp. 30–33. Cited on pages 17, 35, 98, and 118.

M. Vidal, A. Bulling, and H. Gellersen (2012a). Detection of Smooth Pursuits Using Eye Movement Shape Features, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2012*. Cited on pages 13 and 62.

M. Vidal, K. Pfeuffer, A. Bulling, and H. Gellersen (2013). Pursuits: Eye-Based Interaction with Moving Targets, in *Proc. of the ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI) 2013*. Cited on pages 19 and 78.

M. Vidal, J. Turner, A. Bulling, and H. Gellersen (2012b). Wearable Eye Tracking for Mental Health Monitoring, *Computer Communications*, vol. 35(11), pp. 1306–1311. Cited on page 136.

A. Villanueva and R. Cabeza (2008). A Novel Gaze Estimation System with One Calibration Point, *IEEE Transactions on Systems, Man, and Cybernetics (SMC), Part B (Cybernetics)*, vol. 38(4), pp. 1123–1138. Cited on page 9.

M. Voit and R. Stiefelhagen (2006). Tracking Head Pose and Focus of Attention with Multiple Far-Field Cameras, in *Proc. of the International Conference on Multimodal Interfaces (ICMI) 2006*. Cited on page 81.

L. Wang, S. Guo, W. Huang, and Y. Qiao (2015). Places205-VGGNet Models for Scene Recognition, *arXiv preprint arXiv:1508.01667*. Cited on page 165.

J. A. Ward, P. Lukowicz, and G. Tröster (2006). Evaluating Performance in Continuous Context Recognition Using Event-Driven Error Characterisation, in *Proc. of the International Symposium on Location-and Context-Awareness (LoCA) 2006*. Cited on page 67.

D. Weber and S. Mayer (2014). *LogEverything*, `https://github.com/hcilab-org/LogEverything/`. Cited on page 167.

A. F. Westin (2003). Social and Political Dimensions of Privacy, *Journal of Social Issues*, vol. 59(2), pp. 431–453. Cited on page 104.

S. A. Winder and M. Brown (2007). Learning Local Image Descriptors, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2007*. Cited on page 65.

E. Wood, T. Baltrusaitis, L.-P. Morency, P. Robinson, and A. Bulling (2016a). A 3D Morphable Eye Region Model for Gaze Estimation, in *Proc. of the European Conference on Computer Vision (ECCV) 2016*. Cited on pages 82, 85, 86, and 209.

E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling (2016b). Learning an Appearance-Based Gaze Estimator from One Million Synthesised Images, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2016*. Cited on page 33.

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 (ICCV) 2015*. Cited on pages 79 and 82.

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 (ETRA) 2014*. Cited on pages 7, 158, 160, 162, and 174.

X. Xie, R. Sudhakar, and H. Zhuang (1994). On Improving Eye Feature Extraction Using Deformable Templates, *Pattern Recognition*, vol. 27(6), pp. 791–799. Cited on page 8.

J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh (2015). Gaze-Enabled Egocentric Video Summarization via Constrained Submodular Maximization, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on page 62.

S. Xu, H. Jiang, and F. Lau (2008). Personalized Online Document, Image and Video Recommendation via Commodity Eye-Tracking, in *Proc. of the ACM Recommender Systems Conference (RecSys) 2008*. Cited on pages 14 and 42.

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 (PSIVT) 2011*. Cited on page 161.

T. Yamada, S. Gohshi, and I. Echizen (2013). Privacy Visor: Method Based on Light Absorbing and Reflecting Properties for Preventing Face Image Detection, in *Proc. of the IEEE International Conference on Systems, Man, and Cybernetics (SMC) 2013*. Cited on pages 18, 100, and 101.

F. Yang, X. Yu, J. Huang, P. Yang, and D. Metaxas (2012). Robust Eyelid Tracking for Fatigue Detection, in *Proc. of the IEEE International Conference on Image Processing (ICIP) 2012*. Cited on page 12.

J. Yang, R. Stiefelhagen, U. Meier, and A. Waibel (1998). Real-Time Face and Facial Feature Tracking and Applications, in *Proc. of the International Conference on Auditory-Visual Speech Processing (AVSP) 1998*. Cited on page 7.

D. Young, H. Tunley, and R. Samuels (1995). *Specialised Hough Transform and Active Contour Methods for Real-Time Eye Tracking*, University of Sussex, Cognitive & Computing Science. Cited on page 7.

J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan (2017). iPrivacy: Image Privacy Protection by Identifying Sensitive Objects via Deep Multi-Task Learning, *IEEE Transactions on Information Forensics and Security*, vol. 12(5), pp. 1005–1016. Cited on pages 100 and 113.

L. H. Yu and M. Eizenman (2004). A New Methodology for Determining Point-of-Gaze in Head-Mounted Eye Tracking Systems, *IEEE Transactions on Biomedical Engineering*, vol. 51(10), pp. 1765–1773. Cited on page 9.

A. L. Yuille, P. W. Hallinan, and D. S. Cohen (1992). Feature Extraction from Faces Using Deformable Templates, *International Journal of Computer Vision*, vol. 8(2), pp. 99–111. Cited on page 7.

S. Zagoruyko and N. Komodakis (2015). Learning to Compare Image Patches via Convolutional Neural Networks, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*. Cited on pages 32, 61, 64, 65, and 208.

G. J. Zelinsky, H. Adeli, Y. Peng, and D. Samaras (2013). Modelling Eye Movements in a Categorical Search Task, *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368(1628). Cited on pages 19 and 78.

R. Zemblys, D. C. Niehorster, and K. Holmqvist (2018). gazeNet: End-to-End Eye-Movement Event Detection with Deep Neural Networks, *Behavior Research Methods*, pp. 1–25. Cited on pages 13 and 40.

R. Zemblys, D. C. Niehorster, O. Komogortsev, and K. Holmqvist (2017). Using Machine Learning to Detect Events in Eye-Tracking Data, *Behavior Research Methods*, pp. 1–22. Cited on pages 13 and 62.

S. Zhai, C. Morimoto, and S. Ihde (1999). Manual and Gaze Input Cascaded (MAGIC) Pointing, in *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI) 1999*. Cited on pages 3 and 19.

L. Zhang (1996). Estimation of Eye and Mouth Corner Point Positions in a Knowledge-Based Coding System, in *Proc. of the Digital Compression Technologies and Systems for Video Communications 1996*. Cited on page 7.

M. Zhang, K. T. Ma, J. H. 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 (CVPR) 2017*. Cited on page 176.

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 (CVPR) 2015*. Cited on pages 8, 79, 80, and 174.

X. Zhang, Y. Sugano, M. Fritz, and A. Bulling (2017b). It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2017*. Cited on pages 8 and 79.

X. Zhang, Y. Sugano, M. Fritz, and A. Bulling (2018a). MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on page 174.

Y. Zhang, W. Hu, W. Xu, C. T. Chou, and J. Hu (2018b). Continuous Authentication Using Eye Movement Response of Implicit Visual Stimuli, *Proc. of the ACM on*

Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT), vol. 1(4), p. 177. Cited on pages 17 and 120.

Y. Zhang, H. J. Müller, M. K. Chong, A. Bulling, and H. Gellersen (2014). GazeHorizon: Enabling Passers-by to Interact with Public Displays by Gaze, in *Proc. of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) 2014*. Cited on pages 19 and 78.

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 (ICCV) 2015*. Cited on page 165.

H. Zhong, J. Shi, and M. Visontai (2004). Detecting Unusual Activity in Video, in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2004*. Cited on page 24.

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 161.

T. Zhu, G. Li, W. Zhou, and S. Y. Philip (2017). Differentially Private Data Publishing and Analysis: A Survey, *IEEE Transactions on Knowledge and Data Engineering*, vol. 29(8), pp. 1619–1638. Cited on page 120.

Z. Zhu, K. Fujimura, and Q. Ji (2002a). Real-Time Eye Detection and Tracking Under Various Light Conditions, in *Proc. of the ACM International Symposium on Eye Tracking Research and Applications (ETRA) 2002*. Cited on page 8.

Z. Zhu, Q. Ji, K. Fujimura, and K. Lee (2002b). Combining Kalman Filtering and Mean Shift for Real Time Eye Tracking Under Active IR Illumination, in *Proc. of Object Recognition Supported by User Interaction for Service Robots 2002*. Cited on page 8.

M. Ziefle and C. Röcker (2010). Acceptance of Pervasive Healthcare Systems: A Comparison of Different Implementation Concepts, in *Proc. of the International Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health) 2010*. Cited on page 16.