Robust Eye Tracking Based on Adaptive Fusion of Multiple Cameras

THÈSE N^{O} 7933 (2017)

PRÉSENTÉE LE 29 SEPTEMBRE 2017
À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE TRAITEMENT DES SIGNAUX 5
PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Nuri Murat ARAR

acceptée sur proposition du jury:

Dr J.-M. Sallese, président du jury
Prof. J.-Ph. Thiran, directeur de thèse
Dr D. W. Hansen, rapporteur
Dr H. K. Ekenel, rapporteur
Dr J.-M. Odobez, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2017

Canım ailem ve arkadaşlarıma . . .

To my family and friends . . .

If you cannot do great things, do small things in a great way.
— Napoleon Hill

Acknowledgements

My PhD journey, the past five years I have spent at LTS5, has been one of the most fabulous chapters in my life. During this long and exciting journey filled with great experiences, I have learned and grown up a lot, met great colleagues and friends, and built countless memories together. Here, I would like to take the opportunity to express my gratitude to these great people. This thesis would not have been possible without your sincere efforts of guidance, assistance, support, and encouragement.

First and foremost, I would like to express my deepest gratitude to my thesis supervisor Prof. Jean-Philippe Thiran. None of this would have been possible if you had not given me the opportunity to join your lab. Thank you, Jean-Philippe, for your guidance and endless support during all these years. Also, I would like to acknowledge the Swiss Commission for Technology and Innovation (CTI) and Logitech Europe SA, who funded and supported this work. I would particularly like to thank Olivier Theytaz, Regis Croisonnier, Yves Moser, and Patrick Salamin from Logitech, for their valuable contributions and discussions that pushed this work forward.

I would also like to express my gratitude to my mentors, Dr. Hua Gao and Prof. Hazım Kemal Ekenel. Thank you very much, Hua and Hazım, for your guidance, inspiring ideas, and endless help during both my master and doctoral studies.

I am also grateful to the members of my thesis jury: Dr Jean-Michel Sallese, the president of the jury, Prof. Dan Witzner Hansen, Dr Jean-Marc Odobez, and Prof. Hazım Kemal Ekenel, for having accepted to be part of that jury, and for their insightful comments and valuable feedback that have shaped the final version of this thesis manuscript.

A very special thanks goes to my dear office mates, who have also become three of my best friends. Anıl, Christina, and Gabriel, I am so grateful to have you around me at all times. You made my PhD journey much easier and absolutely more fun. Thank you, Anıl, for welcoming and taking care of me when I first arrived in Switzerland. Thanks also to you that I met many other great people in Lausanne with whom we have shared many great moments. Thank you for being such a great friend and for everything you have done. Gabriel, *my twin PhD brother*, we started our PhDs on the same day, and now successfully completing it only one day apart! Not to mention that our PhD footprints (courses, candidacy exam, summer schools, internship, trips, sport activities, parties, etc.) are almost identical. We have literally lived the PhD life together. I believe that this is more than just a coincidence, and I am really grateful to share so many moments together. *Thanks bro'!* Christina, my first close friend in Lausanne, you were always such a great and caring friend. Thank you very much for constantly being so nice. I will always

remember our not only insightful but also entertaining discussions, and our amazing trips.

I would also like to thank my colleagues and friends from EPFL with whom I shared a lot of amazing moments. Firstly, I want to thank to the past and present members of the *Face Group*: Anıl, Hua, Gabriel, Marina, Christophe, Damien and Saeed for your collaboration and everything we have shared. Also, I want to thank the rest of LTS5 family (Mario, Tom, Didrik, Vijay, Christophe Paccolat, Ricardo, Rosie, Frank, Alia, Alessandra, Eleni, Elda, Tobi, Anna, Rafael, Dimitris, Elena, Mina, and many others), and my LTS corridor buddies (Sasan, Kirell, Sophia, Dorina, Francesca, Renata, Ana, Hamed, and many others), and my non-LTS friends (Sami, Gülcan, Onur, Deniz, Emre, Eray, and many others).

Furthermore, I would very specially like to express my gratitude to my friends in Lausanne who made every moment of my life enjoyable and with whom I have shared numerous fabulous moments. Thank you very much, Mario, Tom, Didrik, Irene, Sasan, Vijay, Marina, Ricardo, Chris, Caroline, Antigoni, Gerard, Sophia, Emre, and *The Turkish gang*: Emrah, Damla, Anıl, Ezgi, Deniz, Şeniz, Mustafa, Cansaran, Tuğba, Ebru, Gökhan, and my basketball team, *Phoenix Mediterraneo*, Eray, Agras, Skylos, Kotsanis, Giannopoulos, Pantelis, Costas, and all others who I fail to list here. Here, I would like to make a special mention of my adventure buddies and my best friends, Tom and Mario, for always being such great guys. Our adrenaline-charged trips and adventures tremendously helped me to deal with the stress of this intense PhD journey. Our legendary backpacking in India will probably remain at the top of my craziest trips for a very long time. *TomTom*, thank you very much for all the amazing moments and experiences that we shared together, for always being so enthusiastic and energetic, and for motivating us to do various crazy activities. Mario, *my brother*, thank you so much for always being there for me, you have certainly been the best one can ever imagine. Although we come from very different cultures, we are so much alike in many ways and I have never felt this close to a friend before. Thanks a lot for your help and support especially at difficult times, for always pushing our limits towards being better and stronger, and for building many great memories together.

I would also like to thank my friends who have been physically far away but were always present: Sinem, Çağrı, Kaan, Belma, Burak, Sinan, Turgut, Emre, Açıkel, Gülbey, Tunahan, and many others; I thank all of you for our enduring friendship that I always value so much.

I would like to especially express my deepest gratitude to my parents and my beloved brother Şamil for their endless love, support, and encouragement throughout this journey. I am so grateful for all the values you have instilled in me and for all the trust, encouragement, and freedom you have given me.

Finally, and most importantly, I would like to thank my girlfriend Marili, who has not only helped and encouraged me at every stage of this thesis but has also been an important part of my life. You have been so wonderful, caring, supporting, and understanding during this period, especially the past stressful year. I honestly cannot even imagine how it would have been without you. Thank you so much for always being there by my side and everything you have done.

Lausanne, August 5th 2017 Nuri Murat Arar

Abstract

Eye and gaze movements play an essential role in identifying individuals' emotional states, cognitive activities, interests, and attention among other behavioral traits. Besides, they are natural, fast, and implicitly reflect the targets of interest, which makes them a highly valuable input modality in human-computer interfaces. Therefore, tracking gaze movements, in other words, eye tracking is of great interest to a large number of disciplines, including human behaviour research, neuroscience, medicine, and human-computer interaction.

Tracking gaze movements accurately is a challenging task, especially under unconstrained conditions. Over the last two decades, significant advances have been made in improving the gaze estimation accuracy. However, these improvements have been achieved mostly under controlled settings. Meanwhile, several concerns have arisen, such as the complexity, inflexibility and cost of the setups, increased user effort, and high sensitivity to varying real-world conditions. Despite various attempts and promising enhancements, existing eye tracking systems are still inadequate to overcome most of these concerns, which prevent them from being widely used.

In this thesis, we revisit these concerns and introduce a novel multi-camera eye tracking framework. The proposed framework achieves a high estimation accuracy while requiring a minimal user effort and a non-intrusive flexible setup. In addition, it provides improved robustness to large head movements, illumination changes, use of eye wear, and eye type variations across users. We develop a novel real-time gaze estimation framework based on adaptive fusion of multiple single-camera systems, in which the gaze estimation relies on projective geometry. Besides, to ease the user calibration procedure, we investigate several methods to model the subject-specific estimation bias, and consequently, propose a novel approach based on weighted regularized least squares regression. The proposed method provides a better calibration modeling than state-of-the-art methods, particularly when using low-resolution and limited calibration data. Being able to operate with low-resolution data also enables to utilize a large field-of-view setup, so that large head movements are allowed.

To address aforementioned robustness concerns, we propose to leverage multiple eye appearances simultaneously acquired from various views. In comparison with conventional single view approach, the main benefit of our approach is to more reliably detect gaze features under challenging conditions, especially when they are obstructed due to large head pose or movements, or eye glasses effects. We further propose an adaptive fusion mechanism to effectively combine the gaze outputs obtained from multi-view appearances. To this effect, our mechanism firstly determines the estimation reliability of each gaze output and then performs a reliability-based

weighted fusion to compute the overall point of regard. In addition, to address illumination and eye type robustness, the setup is built upon active illumination and robust feature detection methods are developed. The proposed framework and methods are validated through extensive simulations and user experiments featuring 20 subjects. The results demonstrate that our framework provides not only a significant improvement in gaze estimation accuracy but also a notable robustness to real-world conditions, making it suitable for a large spectrum of applications.

Key words: eye tracking, gaze estimation, multi-camera eye tracking, adaptive fusion, multi-camera fusion, robust eye tracking, convenient user calibration, real-time eye tracking, computer vision, human-computer interaction.

Résumé

Les mouvements du regard et des yeux jouent un rôle essentiel dans l'identification des états émotionnels des individus, ainsi que de leurs activités cognitives, leurs intérêts et leur attention parmi d'autres traits de comportement. De plus, ils sont naturels, rapides et reflètent implicitement les centres d'intérêts, ce qui fait d'eux une entrée utile pour les interfaces homme-machine. De fait, suivre les mouvements du regard – autrement dit, le suivi oculaire ou « eye tracking » - présente un grand intérêt pour un grand nombre de disciplines, y compris la recherche sur le comportement humain, les neurosciences, la médecine et l'interaction homme-machine.

Suivre les mouvements du regard de manière précise relève du défi, en particulier avec des conditions sans contraintes. Au cours des deux dernières décennies, des avancées significatives ont été réalisées dans l'amélioration de la précision de l'estimation du regard. Cependant, ces améliorations ont principalement concerné des expériences réalisées avec des réglages contrôlés. Entre-temps, plusieurs préoccupations ont émergé, concernant notamment la complexité, l'inflexibilité et le coût des installations, l'effort accru de l'utilisateur, et la forte sensibilité aux conditions variables du monde réel. Malgré de nombreuses tentatives et des améliorations encourageantes, les systèmes actuels de suivi oculaire n'arrivent toujours pas à répondre à ces préoccupations, ce qui les empêche d'être utilisés à grande échelle.

Dans cette thèse, nous nous intéressons à ces préoccupations et nous proposons un nouveau système à caméras multiples pour le suivi oculaire. Le système proposé permet une grande précision de l'estimation tout en nécessitant un effort minimal d'utilisation et une installation flexible et non-intrusive. De plus, il présente une meilleure robustesse face aux larges mouvements de tête, aux changements d'éclairage, à l'utilisation de lunettes et à la variabilité du type d'œil pour différents utilisateurs. Tout d'abord, nous avons développé un nouveau système d'estimation du regard en temps réel fondé sur une fusion adaptative de plusieurs systèmes à caméra unique, dans lesquels l'estimation du regard repose sur une géométrie projective. Ensuite, pour faciliter la procédure de calibration de l'utilisateur, nous avons exploré plusieurs méthodes pour modéliser de manière efficace le biais de l'estimation en fonction du sujet, et proposer par conséquent une nouvelle approche qui repose sur une méthode de régression utilisant une pondération de moindres carrés régularisés. Cette modélisation fournie de meilleures modélisations de calibrations que l'état de l'art, en particulier pour des données à faible résolution ainsi qu'avec peu de données de calibration. Fonctionner avec des données de faible résolution nous permet également d'exploiter des installations avec un large champ de vue, et a pour conséquence d'autoriser les larges mouvements de tête.

Résumé

Pour répondre aux préoccupations susmentionnées concernant la robustesse, nous proposons d'exploiter de multiples apparences de yeux acquises simultanément à partir de diverses vues. En comparaison avec l'approche conventionnelle utilisant une seule vue, le principal bénéfice de notre approche concerne la détection plus fiable des caractéristiques du regard dans des conditions difficiles, en particulier lors de larges mouvements de tête et des effets dus aux lunettes. De plus, nous proposons un mécanisme de fusion adaptative pour combiner efficacement les résultats fournis par les différentes vues. A cet effet, notre mécanisme détermine d'abord la fiabilité de l'estimation de chaque résultat du regard et réalise ensuite une fusion pondérée reposant sur la fiabilité pour calculer le point du regard global. En outre, pour répondre à la question de l'éclairage et de la robustesse face au type d'œil, le système utilise un éclairage actif et des méthodes robustes de détection de caractéristiques ont été développées. Le système et les méthodes que nous proposons ont été validés par un grand nombre de simulations et d'expériences sur 20 sujets. Nos résultats ont démontré que notre système ne permet pas seulement une amélioration significative de la précision de l'estimation du regard mais aussi une robustesse notable pour des conditions réelles, ce qui le rend approprié pour un large spectre d'applications.

Mots clefs : suivi oculaire, estimation du regard, suivi oculaire à caméras multiples, fusion adaptative, fusion à caméras multiples, suivi oculaire robuste, calibration facile d'utilisation, estimation du regard en temps réel, vision par ordinateur, interaction homme-machine.

Contents

List of Abbreviations

2D	2-dimensional
3D	3-dimensional
3DMM	3D morphable model
CR	cross ratio
CNN	convolutional neural network
EPFL	École Polytechnique Fédérale de Lausanne
fps	frames per second
FoV	field-of-view
HCI	human-computer interaction
LED	light emitting diode
LoS	line of sight
LoG	line of gaze
NIR	near-infrared
PoR	point of regard

List of Figures

List of Tables

1 Introduction

Human visual system provides us with the ability to observe our surroundings through the communication between the eye and the core of the central nervous system, the brain. The eye is the major component of our visual system, and is essentially a sensory organ, which receives stimuli from the physical environment in the form of light waves. It sends those stimuli as electrical signals to the brain to be interpreted as images. Our eyes can distinguish between 8 to 10 million colors and they are capable of spotting a single photon [Tinsley et al., 2016]. Therefore, they allow us to perceive colors and depth in intricate detail. Although the eyes are mainly acknowledged for providing the observer with vision, *"the eyes are the mirror of the soul and reflect everything that seems to be hidden"* as stated by famous author Paulo Coelho. They as well provide significant cues indicating observer's emotional state, cognitive processes, visual attention, interest, and inter-personal interactions [Underwood, 2005, Duchowski, 2007].

Eye and gaze movements provide explicit inputs to point out the observed surrounding. Since they are natural and fast, they are considered as an essential modality for visually mediated human-computer interfaces. They hold a significant potential to enhance user experience in human-computer interaction, such as for controlling and navigation. Besides, they provide non verbal communication signals to perceive the human-human and human-computer interactions. In this regard, they play an important role in human behaviour research, such as to identify emotional states and expressions [Alghowinem et al., 2014, Filik et al., 2017], cognitive activity [Eckstein et al., 2016], attention and interest [Valenti et al., 2012, Borji and Itti, 2013]. Therefore, tracking eye and gaze movements, also known as eye tracking, is relevant to a wide range of disciplines, including sociology, psychology, psycholinguistics, cognitive science, neuroscience, education, medicine, marketing research, usability testing, gaming research, human-computer interaction, and computer vision.

The study of eye movements goes as early as 1878 when Louis Émile Javal, a French ophthalmologist, noticed that people do not read smoothly across a page, but rather pause on some words while moving quickly through others [Javal, 1878]. Ever since, there has been growing interest in eye tracking technology from a wide variety of domains. Over more than a hundred years,

significant achievements have been attained by both industry and scientific community to advance the eye tracking technology. Especially over the last three decades, promising contributions have been made. Recently, with the advances in computer technologies, the popularity of eye tracking is on the rise.

On the other hand, despite valuable efforts and significant improvements in eye tracking research, there still remains several concerns, such as the setup complexity and flexibility, cost, user calibration procedures, tracking accuracy and robustness to varying real-world conditions. These concerns hinder the eye tracking from becoming a pervasive technology. Hence, there is still room for further research efforts to address such limitations.

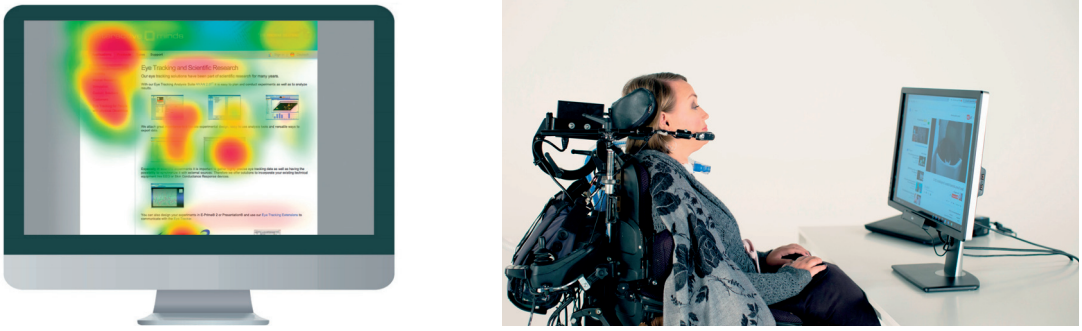
In this thesis, we aim to address some of the major concerns in eye tracking and present a novel eye tracking system and methodology, which has a promising potential to be used in a large spectrum of applications.

In the following sections, we firstly present various eye tracking applications from a wide range of domains to further explain our motivation and the impact of eye tracking research. We further describe the main challenges in eye tracking and explain the objectives of this thesis. We then present our approach to meet these objectives and list the major contributions to the literature. Lastly, we give the organization of the thesis.

1.1 Motivation & Applications

Eye trackers have traditionally proven themselves valuable tools for diagnostic studies, and nowadays eye tracking research is in its new era, distinguished by the emergence of interactive applications, as described in [Duchowski, 2002]. Therefore, they have numerous potential applications and are of high value for a wide variety of domains, including among many others, human behaviour research, cognitive science, neuroscience, marketing research, usability testing, artificial intelligence, and human-computer interactions.

Regarding its diagnostics use, eye trackers provide objective and quantitative evidence of individuals' visual and attentional processes. For instance, diagnostic studies in psychology and sociology fields can greatly benefit from eye tracking since gaze movements reveal important signals for individual's personality traits and social behaviours. In this context, eye tracking can facilitate detection of one's emotional states [Hortensius et al., 2014, Alghowinem et al., 2014, Wells et al., 2016, Filik et al., 2017], attention [Borji and Itti, 2013, Kuo et al., 2014, Valuch et al., 2015] and arousal [Bradley et al., 2008]. In addition, it enables to examine inter-personal behaviors in social settings [Gatica-Perez et al., 2005, Terburg et al., 2011]. Cognitive science researchers also frequently use eye trackers to understand individual's cognitive development [Eckstein et al., 2016], cognitive load [Palinko et al., 2010, Bartels and Marshall, 2012], and decision making process [Pärnamets et al., 2015]. An interesting use of eye tracking can be found in in-car driving scenarios, in which driver's fatigue [Eriksson and Papanikotopoulos, 1997] and



(a) (b)

Figure 1.1 – Example applications of eye tracking: (a) monitoring customer/user behaviours in marketing research and usability testing studies, (b) assisting individuals with different physical and cognitive limitations (image courtesy of ©Tobii Dynavox).

attention [Fletcher and Zelinsky, 2009] are detected, so that certain security measures can be taken. Moreover, eye tracking can be used as a highly valuable decision support tool in neuroscience since certain neurological disorders involve ocular control and attention dysfunctions. For example, it can be utilized to diagnose certain disorders, such as autism in children, attention deficit hyperactivity disorder (ADHD), fetal alcohol spectrum disorder (FASD), Parkinson's disease [Tseng et al., 2013], schizophrenia [Levy et al., 2010], and consciousness disorders [Ting et al., 2014]. Furthermore, marketing research and usability testing studies take advantage of eye trackers to assess consumers' attention and users' response to different products and designs by evaluating gaze patterns and identifying attentions, e.g., advertisements, web or software interfaces, as shown in Figure 1.1a.

In addition to their roles in diagnostics, eye trackers serve as an essential input modality for human-computer interfaces with vision-based applications. For instance, the point of regard can be utilized as an alternative pointing device or assistive input to a traditional computer mouse in order to control gaze-based interfaces (e.g., activate, select, zoom, scroll) [Salvucci and Anderson, 2000, Zhu and Ji, 2004, Hansen and Hansen, 2006, Reale et al., 2011, Topal et al., 2014]. Eye tracking has also a number of potential uses in virtual reality, augmented reality, and gaming research. For instance, in virtual reality, where a realistic and immersive simulation of a 3-dimensional (3D) 360-degree interactive environment is created, the virtual environment is experienced or controlled by the movements of a user. Similarly in augmented reality and gaming, user movements play significant interactive roles. Therefore, eye tracking can be benefited so as to enhance the user experience with natural realistic controlling and navigation. Furthermore, since gaze interaction does not require the movement of any muscles, eye tracking can be considered as an ideal assistive technology for people with rehabilitative disabilities (e.g., paralysis, spinal cord injury, repetitive strain injury, severe carpal tunnel) and motor disabilities (e.g., amyotrophic lateral sclerosis (ALS), cerebral palsy) by enabling gaze-based typing and controlling the interfaces [Betke et al., 2002, Majaranta, 2011], as can be seen in Figure 1.1b.

1.2 Objectives and Approach

The primary objective of eye trackers is to determine gaze. Depending on the context and methodology, gaze estimation can refer to determining either the *Line of Sight (LoS)* in 3D or the *Point of Regard (PoR)* in 2D. The *LoS* describes where a user is looking in 3D world coordinate system, whereas the *PoR* denotes where the *LoS* intersects with the scene, typically a screen or an object. In this thesis, the term *"gaze"* is used to indicate the *PoR* on a screen unless stated otherwise.

Eye tracking development and gaze estimation require to address several challenges and to consider various trade-offs depending on the use case and application scenario. In this respect, we identify the main challenges and desired attributes in eye tracking as follows:

- **Intrusiveness:** Considering the user experience, an ideal eye tracker should include minimal intrusiveness and obstruction while maintaining high tracking performance. In this regard, early systems were highly intrusive since they required attaching a number of electrodes around the eye, or placing a reflective contact lens onto the eye [Young and Sheena, 1975]. In spite of the high accuracy and robustness, such techniques were later avoided due to their intrusiveness. More recently introduced head-mounted trackers, which comprise of a camera and light sources placed on a helmet or eye glasses, also accommodate high accuracy and significant head movement tolerance. Nevertheless, they are as well not frequently preferred by users due to their intrusive nature. As opposed to early systems and head-mounted trackers, remote sensors based eye trackers enable a non-intrusive user experience. On the other hand, their accuracy and robustness are notably lower than the intrusive ones. Still, remote eye trackers are mostly preferred, and their use will perhaps be compulsory for future eye trackers.

- **Setup complexity & flexibility:** Most of the existing eye trackers have a complex hardware setup. These setups usually require camera and geometric scene calibrations, which further increase the system complexity. In addition, such systems are highly inflexible, such that any modification in the setup requires a re-calibration. Therefore, a highly desirable attribute for eye trackers would be to have a simple and flexible setup, in which it can automatically adapt to the modifications without requiring explicit calibration of geometry and cameras.

- **Real-time tracking:** In eye tracking, it is of great importance to capture even very small details in gaze patterns as subtle changes in eye and gaze movements convey vital information for many disciplines. In this respect, detecting saccades, one of the fastest movements produced by the human body, and precise duration of fixations is only possible with computationally light-weight real-time eye trackers. Besides, for interactive applications, a rapid gaze processing is crucial to obtain natural human-computer interaction. Therefore, the development of highly complex models and tracking setups that generate gaze outputs at a low frequency would not be valuable in practical terms, even if

they enable a high accuracy and robustness.

- **Cost:** Most of the current eye trackers use complex and expensive hardware setups (e.g., high-resolution cameras and sensors, high-quality lenses). In addition, the market is relatively small. Consequently, the prices of the current eye tracking systems remain too high for general public use. Thus, one of the main challenges is to enable accurate eye tracking with less complicated hardware setups and possibly lower quality data.

- **User calibration:** In order to reach high tracking performance, user calibration is inevitable for existing eye trackers and gaze estimation models. As the user calibration procedure is tedious for the users, it can significantly harm the user experience. Hence, one of the main challenges in eye tracking research is to develop models that achieve high performance while requiring a minimal effort from the users.

- **Accuracy & precision:** In order for eye tracking to be effectively profited in most of the aforementioned disciplines, an important requirement is to determine the user gaze with high precision and accuracy, e.g., lower than 1° of visual angular error. In fact, very high estimation performances can already be achieved by some of the existing eye trackers. However, such systems have to sacrifice from one or more of the other desirable attributes. They are often not affordable and require intrusive or complex setups, high data resolutions, or extensive user calibration procedures. In this regard, one of the main challenges is to reach high tracking performance while satisfying as many of the mentioned criteria as possible.

- **Robustness against varying real-world conditions:** In current eye tracking systems, one of the most critical limitations is the high intolerance to varying real-world conditions, including head pose changes, head movements in 3D, ambient illumination variations, use of eye wear, and between-subject variations in eye phenotype (color and shape). In this regard, significant efforts have been devoted to improve the tracking robustness against head movements. Detailed eye modeling and use of multiple light sources enable promising improvements. In addition, having sensors with large field-of-view (FoV)s plays a significant role in practice in order to accommodate large head movements. Nevertheless, this usually brings another challenge, that is to deal with low resolution eye data. Moreover, high sensitivity to uncontrolled illumination conditions is a critical limitation, particularly for outdoor applications. Besides, robustness to eye wear and between-subject eye type variations constitutes an important practical problem for certain users. However, these concerns have been only partially investigated in the literature. In overall, the desired level of robustness in eye tracking has unfortunately not been achieved, therefore, further improvements are essential in eye tracking research.

In eye tracking literature, there exist numerous studies towards overcoming the aforementioned challenges, as will be explained in detail in the next chapters. Despite significant efforts and promising advances, existing eye tracking systems are still inadequate in dealing with most of

these challenges. This is, in fact, the major reason that prevents eye tracking technology from becoming a pervasive technology, or even being widely used. Hence, there is still room for further research efforts to advance eye tracking systems. In this respect, we set the main objectives of this thesis as follows:

- to develop a real-time eye tracking system and methodology that provides high gaze estimation accuracy and high tolerance to unconstrained conditions by requiring minimal user effort and by using a non-intrusive and flexible hardware setup.
- to validate its efficacy with extensive simulations and user experiments under challenging real-world scenarios, such as head pose changes and large head/body movements, varying illumination conditions, use of eye glasses and contact lenses, and between-subject eye type variations.

In order to meet our objectives, we revisit the aforementioned challenges and desirable attributes, and therefore, introduce a multi-camera eye tracking system and methodology. First of all, we design a multi-camera eye tracking setup that is non-intrusive, flexible and adaptable to different application scenarios. Our design enables simultaneously acquiring multiple eye appearances from various views. The setup aims to obtain a large working volume to allow for large head movements. It consists of multiple camera sensors equipped with large FoV lenses, therefore, the methodology is designed to operate with low-resolution eye data. The main benefit of our design is to enable leveraging multiple eye appearances in order to reliably detect gaze features under challenging tracking conditions, especially when they are obstructed in conventional single view appearance due to large head pose and movements, disturbances or occlusions caused by eye glasses. In addition, the setup is based on active near-infrared (NIR) illumination, so the system is more robust to varying ambient illumination conditions.

Once the local gaze features are extracted on the acquired eye appearances, they are used to estimate multiple gaze outputs. In our methodology, the gaze estimation relies on a cross ratio-based method due to its particular advantages, such as no camera or geometric scene calibrations being needed. The estimated gaze outputs are then combined by an adaptive fusion mechanism to compute user's overall PoR. The proposed adaptive fusion mechanism first determines the estimation reliability of each gaze output by considering various reliability criteria. Then, a reliability-based weighted fusion of the available gaze outputs is performed to compute the overall PoR. Besides, we propose a novel subject-specific bias compensation method based on weighted least-squares regression to ease the burden of user calibration. In comparison with widely used homography-based calibration methods, our method enables a more generalizable estimation bias modeling, particularly when the calibration data is limited in amount and quality. Hence, the proposed method aims to minimize the user effort, which leads to more convenient and user-friendly calibration procedure. Moreover, the computational complexity of our complete eye tracking framework is notably low, therefore, a real-time tracking is easily achieved without requiring significant computational optimization efforts.

1.3 Main Contributions

The main contributions of this thesis can be summarized as follows:

- A novel real-time multi-camera eye tracking framework is designed and implemented in order to address some of the major concerns in existing eye tracking systems. It consists of a non-intrusive, flexible, and adaptable hardware setup. Differently from the conventional single view tracking setups, the proposed setup leverages multiple eye appearances simultaneously acquired from various views. It accurately operates in real-time with low-resolution data and allows for a large working volume, owing to multiple cameras' combined large FoV.

- A computationally simple multi-camera gaze estimation methodology is proposed. The estimation of the gaze relies on a simple cross-ratio geometry in projective space. In addition, the setup and cameras do not require any hardware or scene calibration. Hence, the suggested methodology not only provides a computationally efficient gaze estimation that can operate in real-time, but also enables a flexible uncalibrated setup that can effortlessly be adapted for various applications.

  The framework and methodology have been published in the proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG) [Arar et al., 2015a] and extended in [Arar and Thiran, 2017] (under review). A patent, US #9,411,417, has also been granted [Arar et al., 2016b].

- In order to effectively benefit from a multi-camera system, we propose an adaptive fusion mechanism to combine the gaze outputs obtained from individual camera systems. This mechanism firstly determines the estimation reliability of each gaze output, and then performs a reliability-based weighted fusion. To determine the gaze reliabilities, we suggest to exploit various reliability indicators, such as subject-specific gazing behaviors, momentary head poses with respect to each camera, distances to the cameras, statistics calculated from the calibration data, etc. In comparison with more complex systems, the proposed methodology achieves highly competitive estimation accuracies under challenging experimental scenarios, including large head movements and pose changes, varying illumination, use of eye wear, and between-subject variations.

  The proposed methods have been published in the proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA) [Arar and Thiran, 2016] and submitted for a journal publication [Arar and Thiran, 2017] (under review).

- A comprehensive investigation on regression-based calibration methods is conducted in order to ease the user calibration procedure. In this regard, we propose a novel subject-specific estimation bias modeling based on a weighted regularized least-squares regression method. Our method enables an effective estimation bias modeling and achieves a better generalization than the state-of-the-art calibration methods, particularly when the calibration data is limited in size and quality.

  This work has initially been published in the proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV) [Arar et al., 2015b] and extended in IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) [Arar et al., 2016a].

- In order to examine the efficacy of the proposed framework and methods, extensive evaluations on both simulated data and user experiments were conducted. In simulations, we analyze in detail the influence of increasing the number of cameras (up to 36) in various configurations. The trade-off between the tracking performance and setup complexity is presented. In user experiments, natural and realistic human-computer interaction (HCI) scenarios were targeted and the users were asked to follow some conventional and newly introduced experimental scenarios. Firstly, a database consisting of 10 users performing natural gazing scenarios was collected in order to validate the efficacy of the proposed user calibration method as well as the multi-camera concept. Later, a larger database featuring 20 users, which includes subjects with diverse eye types and eye wear (eye glasses, contact lenses), was collected. The users performed eight experiments under varying illumination conditions and head movements to demonstrate the system's robustness to real-world conditions. In all experiments, the users were asked to interact with the system as natural as possible and no chin rest was required to keep their head stable. In addition, realistic and reliable evaluation schemes were introduced.

This work was supported by Logitech Europe SA and by the Swiss Commission for Technology and Innovation (CTI) under grant number 13594.1 PFFLR-ES.

1.4 Thesis Organization

The rest of the thesis is organized as follows:

Chapter 2: Introduction to Eye Tracking This chapter presents an overview of existing eye tracking techniques, followed by a comprehensive literature review.

Chapter 3: Robust Real-Time Multi-Camera Gaze Estimation Framework This chapter describes the proposed gaze estimation framework, from data capturing and gaze features detection to gaze estimation and real-time implementation. It explains the main processes of our gaze estimation framework.

Chapter 4: Regression Based User Calibration This chapter addresses the subject-specific calibration challenge in eye tracking. Firstly, it introduces the previous efforts in the literature and then describes the investigated regression based user calibration methods. It also describes the evaluations on simulated data and user experiments, followed by a comparison with the state-of-the-art and discussions.

Chapter 5: Robust Eye Tracking Based on Adaptive Multi-Camera Fusion This chapter mainly addresses robustness concerns in eye tracking. Firstly, it presents the existing studies in the literature in detail. Then, it describes the details of the proposed adaptive fusion approaches. It further explains the evaluations on simulated data and user experiments to examine our framework's tracking performance under challenging real-world conditions using various setup configurations. Lastly, a comprehensive comparison with the previous work is given together with discussions and acquired insights.

Chapter 6: Conclusion This chapter reviews the key contributions presented in this thesis and discuss the benefits they bring about. In addition, it explains the current limitations and describes the future perspectives to address them.

Contributions that are not presented in this manuscript A certain number of other contributions to the computer vision literature have been made throughout this thesis. These works are the outcomes of either in-lab and international collaborations or my research internship at IBM Zurich Research Lab. In this manuscript, these works are not presented due to their irrelevance to the main thesis topic. Yet, we simply list them hereafter and let the interested reader check the corresponding publications.

(i) Robust face recognition using local appearance models based on curvature Gabor wavelets [Arar et al., 2012].

(ii) Multi-view facial expression recognition using partial least squares [Güney et al., 2013].

(iii) Improved facial action unit detection by combining multiple local curvature Gabor binary patterns (LCGBP) [Yüce et al., 2013].

(iv) Automatic immunostaining quality assessment and sensitivity analysis of the process parameters towards the standardization of immunostaining [Arar et al., 2017a, Arar et al., 2017b]. Two patents on an automated method for process parameter optimization for tissue section immunostaining are in the filing process.

2 Introduction to Eye Tracking

This chapter describes briefly the origins of eye tracking in Section 2.1. Section 2.2 then explains the techniques used for the gaze estimation. Section 2.3 presents a literature review, particularly on the works that are relevant to this thesis. Lastly, the conclusions are given in Section 2.4.

2.1 Eye Tracking Origins

Eye movement studies has been around since the late 1800s. Initially, eye tracking research focused mainly on studying how people read. In 1878, Louis Émile Javal, a French ophthalmologist, made an observation that readers pause on some words while moving quickly through others [Javal, 1878]. These pauses are referred to as eye fixations. Later on, researchers continued to conduct eye tracking studies using naked-eye observations to better understand and evaluate these

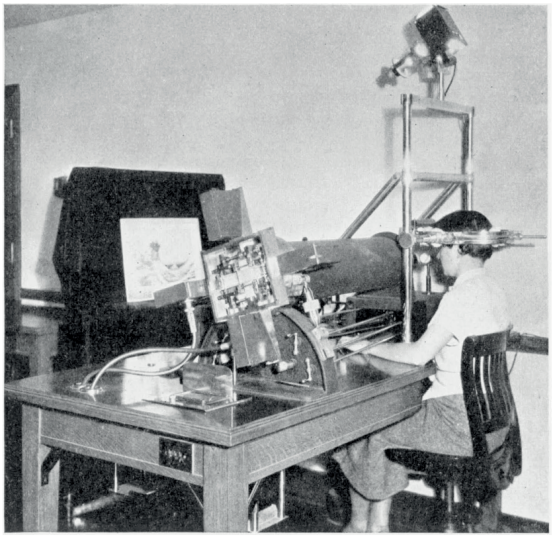


Figure 2.1 – One of the earliest eye trackers, developed by Thomas Buswell in 1935 (image courtesy of © EyeSee).

eye fixations. In fact, this is still an important research question trying to be answered even today.

The first known eye tracker device was built by Edmund Huey later in 1908 to track eye movements during the reading process. The device was very intrusive as readers had to wear a special contact lens with a pointer attached to it. The pointer changed its position following the movements of the eye, so that Huey could observe where a reader was looking and which words he or she pauses on [Huey, 1908].

In 1930s, educational psychologists Guy Thomas Buswell and Charles H. Judd developed the first non-intrusive eye-tracking device that used light beams which were reflected on reader's eyes and recorded them on film. Their research led to many leaps in the field of education and literacy.

Later, towards 1970s, eye tracking studies and research continued to rapidly grow. The main focus was still to study how people read. In the 1980s, Just and Carpenter came up with the *Strong eye-mind hypothesis*, *"there is no appreciable lag between what is fixated and what is processed"* [Just and Carpenter, 1980]. This hypothesis states the existence of a direct correlation between the gaze fixations and the cognitive process. Despite this thesis was questioned because of the idea of covert attention, it was taken as granted by most of the researchers. The 1980s also ushered in the start of real-time eye tracking for human-computer interaction. Early works mostly focused on assisting individuals with disabilities [Levine, 1981, Hutchinson et al., 1989]. Besides, scientists analyzed how users navigated through and interacted with computer command windows. In addition, marketing research started to utilize eye tracking to measure the effectiveness of advertisements in magazines.

Today, the interest in eye tracking is continuing to grow and the eye tracking technology is continuing to advance. Therefore, it is highly likely that eye tracking technology will be integrated in many more aspects of our lives in the future.

2.2 Eye Tracking Techniques

Various techniques have been developed since the earliest attempts in order to track eye and gaze movements. These techniques can mainly be divided into three categories depending on the employed hardware and methodology, namely, *Electro-oculography*, *Contact lens based*, and *Video-oculography*, as described in [Duchowski, 2002].

2.2.1 Electro-oculography

Electro-oculography technique relies on the existence of an electrical field that changes its potential as the eye moves in its orbit (Figure 2.2a). It is considered as highly intrusive as it requires the electrodes to be placed on the skin around the eyes in order to detect the changes in the electric potential [Young and Sheena, 1975]. On the other hand, the main advantage of this

(a) (b)

Figure 2.2 – Intrusive eye tracking techniques based on: (a) *electro-oculography*, (b) *contact lens (search coil)* method [Duchowski, 2000].

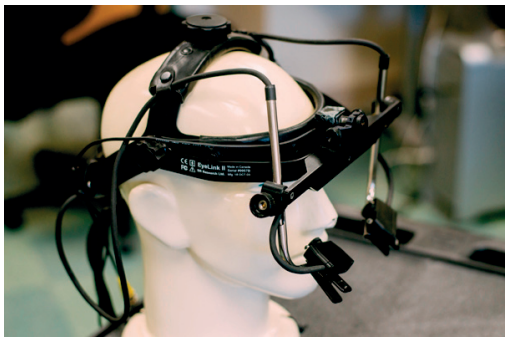
technique is that it can be used with contact lenses and eye glasses without sacrificing from the accuracy. In addition, the tracking performance is highly insensitive to the changes in the head movements.

2.2.2 Contact lens-Based

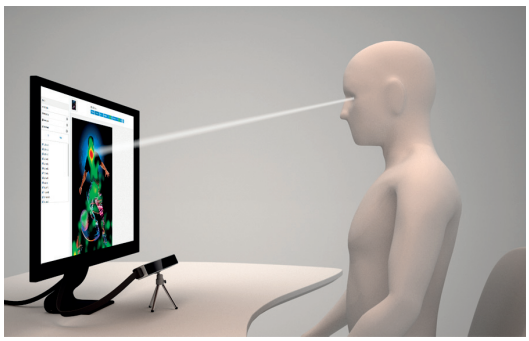
This technique is perhaps the most accurate technique amongst the three categories, however, it requires the user to wear a special contact lens that are connected to wires (Figure 2.2b). Despite being very accurate, its highly intrusive nature and the unresolved health concerns due to the use of high frequency electro-magnetic fields make the technique impractical for non-laboratory conditions [Young and Sheena, 1975].

2.2.3 Video-oculography

Video-oculography technique uses one or more cameras to observe the eye movements and determine the line of gaze (LoG) or the point of regard (PoR). This technique can be implemented as a head-mounted or as a remote sensors-based system. *Head-mounted eye trackers* comprise of a camera and light sources placed on a helmet or a pair of glasses (e.g., [Babcock and Pelz, 2004, Noris et al., 2011]), as can be seen in Figure 2.3a. They enable mobile gaze interaction with head-mounted displays, and may be preferred for applications that require large and fast head movements. Nevertheless, they are impractical to be used in applications which require continuous gaze monitoring over long periods of time, e.g., monitoring driver behavior, aids for motor-disabled persons, due to their intrusive nature. *Remote sensors-based eye trackers*, on the other hand, capture the eyes of a user by one or several remote sensors or video cameras (Figure 2.3b). In this technique, the gaze is estimated through employing image processing and computer vision methods on the captured eye images. In fact, remote eye trackers can be



(a)



(b)

Figure 2.3 – Video-oculography, (a) *head-mounted eye tracker* (image courtesy of © SR Research), (b) *remote sensors-based eye tracker* (image courtesy of © The Eye Tribe).

considered as the most popular eye trackers due to their practical advantages. The most prominent advantage lies in their non-intrusiveness. Therefore, especially for interactive applications, they are highly preferred although they provide lower accuracy and head movement tolerance in comparison with *head-mounted eye trackers*. Since this thesis targets a natural and user-friendly eye tracking, our focus will mostly be on remote sensors-based gaze estimation methods.

2.3 Remote Sensors-Based Eye Tracking

As described in recent surveys on remote gaze estimation and eye tracking by [Morimoto and Mimica, 2005] and [Hansen and Ji, 2010], remote sensors-based eye tracking methods can mainly be categorized into two groups, namely, *feature-based methods* and *appearance-based methods*. As the name implies, *feature-based methods* utilize local features extracted on eye images, such as pupil center, contours, eye corners, and reflections on the cornea (i.e., glints), to determine the gaze. On the other hand, *appearance-based methods* do not explicitly extract features, but rather use the image content as the input. They map the image features directly to the gaze points. The system and hardware requirements of *appearance-based methods* tend to be simpler than those of *feature-based methods*. They simply require an ordinary camera such as a webcam. Also, they require neither camera nor geometric scene calibration. Nevertheless, they are restricted to particular applications due to their limitations in the estimation accuracy and head movement robustness. *Feature-based methods* enable higher estimation accuracies than *appearance-based methods* since the extracted gaze features are formally related to the gaze points through the geometry of the system and eye physiology. In addition, the detection of gaze features is straight-forward. Due to such benefits, they have become the most popular method for gaze estimation. A comparison of gaze estimation methods regarding the setup complexity, hardware requirements and calibration, user calibration, estimation accuracy, and implicit tracking robustness to real-world conditions, can be found in Table 2.1.

Table 2.1 – Comparison of gaze estimation methods regarding the setup complexity, hardware requirements and calibration, user calibration, estimation accuracy, and implicit tracking robustness to head movements, illumination variations and eye wear.

	Feature-based			Appearance-based
	3D Model	**Regression**	**Cross Ratio**	
Setup Complexity	High	Medium	Medium	Low
System Calibration	Camera & Scene	-	-	-
Hardware Requirements:				
- Cameras	2+ Infrared (stereo)	1+ Infrared	1+ Infrared	1+
- Lights	2+ Infrared	2+ Infrared	4+ Infrared	-
User Calibration	Required	Critical	Required	Optional
Gaze Accuracy Error	$< 1^\circ$	$\sim 1-2^\circ$	$\sim 1-2^\circ$	$> 2^\circ$
Implicit Robustness:				
- Head Movements	Medium-High	Low-Medium	Low-Medium	Low
- Illumination Variations	High	High	High	Low
- Eye Wear	Low	Low	Low	Medium

In the following sections, we firstly describe the working principles and dynamics of each approach, and then present a review of the existing methods. Section 2.3.1 and Section 2.3.2 explain *appearance-based* and *feature-based* methods, respectively. Note that this chapter aims to present a high level understanding of the aforementioned methods. A more detailed reviews are provided in Chapter 4 and Chapter 5 with emphases on the major focuses of this thesis, i.e., user calibration and robustness to real-world conditions, respectively.

2.3.1 Appearance-Based Gaze Estimation

Appearance-based methods avoid local gaze features detection, but rather use the image content as the input map. Instead of explicitly modeling the eye in 2-dimensional (2D) or 3-dimensional (3D), they learn a direct mapping between the image features and the gaze points. The input dimensionality is much higher than *feature-based methods*. Therefore, the success of these methods relies on how well the training data covers the variation in the test data. In this respect, unless large amounts of training data is provided to handle variations due to user identity, head pose, eye ball pose, illumination, scale, etc., they suffer from the generalization problem, particularly for subject-independent mapping. Early efforts used artificial neural networks to directly map the eye image pixels to the gaze points on a screen. Their systems required thousands of training samples and a fixed head pose to obtain acceptable gaze estimation accuracies.

As a pioneering work, [Baluja and Pomerleau, 1994] use 2000 cropped eye images as input to a multi-layer neural network and their system achieved an accuracy of 1.5° while allowing for certain head movements. Similarly, [Xu et al., 1998] used 3000 training images to achieve a comparable accuracy. Later, alternative methods were proposed under similar conditions such as limited head pose variations and in-session calibrations. For instance, [Kar-Han Tan et al., 2002] proposed to use linear interpolation to reconstruct a test sample from a local appearance manifold

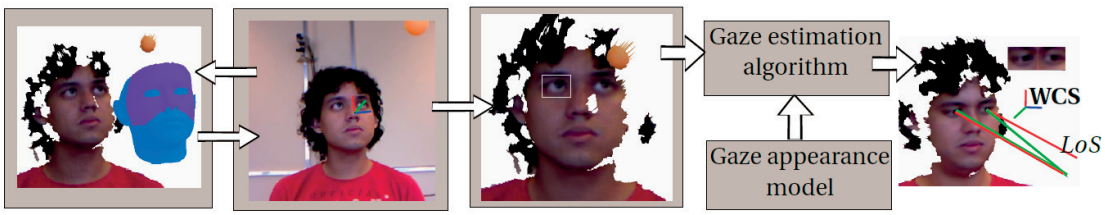


Figure 2.4 – An example appearance-based gaze estimation pipeline [Funes Mora, 2015].

within the training data. They leveraged the topology information, encoded as 2D space of gaze parameters, to constrain the samples selection. They managed to effectively reduce the number of training samples while obtaining an acceptable accuracy. In addition, [Hansen and Pece, 2005] proposed a tracking method based on particle filtering and the expectation-maximization contour algorithm for robust iris tracking. To perform gaze estimation with this method, they required users to gaze at four calibration points as a lower bound. Their system achieved an accuracy of ~4° under limited head movements.

Later on, alternative approaches have been proposed mainly to reduce the number of labeled training samples. For example, [Sugano et al., 2010] proposed a novel method that automatically collects labeled samples by utilizing saliency prior from a video. [Lu et al., 2011] introduced an adaptive linear regression method that automatically selects training samples for mapping. These methods worked well under controlled conditions, such as fixed head pose using a chin-rest, fixed illumination settings, and well aligned eye images. However, their performance degraded greatly when user head was not stationary. Moreover, [Funes-Mora and Odobez, 2012] leveraged RGB-D cameras to directly handle eye appearance variation by generating frontal view eye images used as input to adaptive linear regression. They later proposed a framework for 3D gaze estimation, as shown in Figure 2.4. Thanks to the depth measurements and the fitted 3D facial mesh, they improved the framework's robustness to head pose and between-user appearance variations [Mora and Odobez, 2016].

As the performance of *appearance-based methods* heavily relies on the data variability for the training of the model, several recent efforts have been devoted to capture larger data variability. In this context, large-scale datasets were collected, such as MPIIGaze [Zhang et al., 2015] and GazeCapture [Krafka et al., 2016]. The authors were then trained convolutional neural network (CNN)s on this large datasets to learn robust mappings. They achieved significant accuracy improvements over the state-of-the-art *appearance-based methods* with an error of ~4°. Despite such models trained on large datasets provided head pose and illumination change tolerance to a certain extent, collecting such datasets to acquire sufficient data variation is still cumbersome and impractical. Instead, *learning-by-synthesis* approaches [Lu et al., 2012, Sugano et al., 2014, Wood et al., 2016b, Wood et al., 2016a] were introduced to increase the data variability using the synthesized eye images. For example, [Lu et al., 2012] synthesized additional eye images of various head poses using pixel displacements applied on the real images captured at particular head poses. Although the method allowed certain head pose tolerance, the estimation

accuracies were poor. Besides, this technique could not improve the robustness to subject or environmental variations. [Sugano et al., 2014] collected a fully calibrated multi-view gaze dataset (UT Multi-view Gaze dataset) from eight synchronized webcams, and performed a 3D eye region reconstruction in order to generate dense training data of eye images. [Wood et al., 2016b] presented a method to rapidly synthesize large amounts of variable eye region images as training data. Their eye region model was derived from high-resolution 3D face scans, and enabled image-based lighting to cover a range of illumination conditions. To demonstrate the efficacy of the method, they synthesized over a million eye images and learned a gaze estimator using k-nearest-neighbors. Despite the simplicity of the classifier employed, they achieved $\sim 10^{\circ}$ accuracy error on a cross-dataset evaluation on MPIIGaze dataset, and outperformed the CNN-based method described in [Zhang et al., 2015].

As a conclusion, *appearance-based methods* have an important advantage over the other methods, that is to not require a particular hardware setup and user calibration. The recent advancements in the synthesizing and rendering technology together with learning successful models from large-scale datasets using deep learning techniques have brought back a considerable attention to *appearance-based methods* since they remarkably improve the estimation accuracy and the head pose and illumination variations tolerance. There is no doubt that these methods have a great potential to make eye tracking a pervasive technology. However, the current estimation accuracy and robustness performances are still insufficient to be utilized for the applications that require precise gaze estimation ($<1^{\circ}$).

2.3.2 Feature-Based Gaze Estimation

Local eye features, such as pupil center, cornea center, corneal reflections (i.e., glints), pupil or iris contours, eye corners, etc., and their cross-relations convey significant information regarding the gaze. *Feature-based methods* rely on extracting and thereby mapping some of these features to the gaze points. Since the eyeball has a complex structure, as can be seen from Figure 2.5, and most eye parameters differ for every subject, the majority of the methods require a user calibration, so that subject-specific eye parameters can be estimated for a more accurate gaze computation.

In this thesis, we further categorize *feature-based methods* into three groups, namely, *3D model-based methods*, *regression-based methods*, and *cross ratio-based methods*. *3D model-based methods* compute the gaze from the eye features obtained from a 3D geometric model of the eye, whereas *regression-based methods* assume a direct mapping from the eye features to the gaze points. On the other hand, *cross ratio-based methods* compute the gaze point by leveraging the cross ratio property of the projective space. The following subsections describe these three approaches and the related work from the literature.

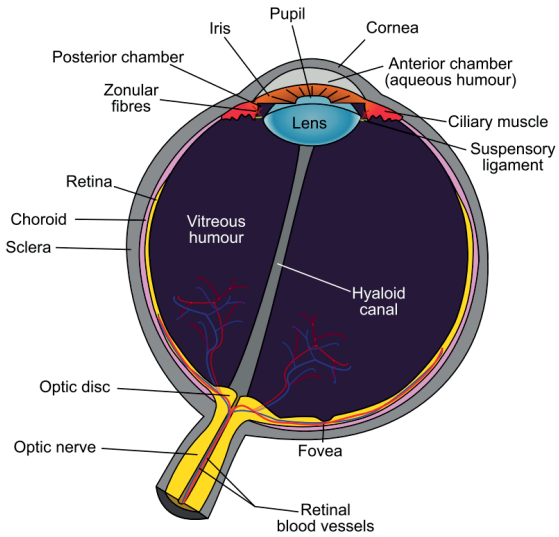


Figure 2.5 – Overview of the eyeball.

3D Model-Based Methods

3D model-based methods estimate the gaze by modeling the eyeball and the setup in 3D. The schematic illustrations of the human eye and the hardware setup configuration are illustrated in Figure 2.6. The general approach is to localize the eye features, such as cornea, pupil center, glints, and to calculate their relations through fully calibrated sensors. The gaze directions can then be modeled as the optical axis, also known as line of sight (LoS), which is the line connecting the pupil center, cornea center, and the eyeball center. Nevertheless, the true direction of the gaze is assumed to be the visual axis, also known as LoG, which is the line connecting the fovea and the center of the cornea. Unfortunately, it is not possible to directly estimate the visual axis since the fovea can not be explicitly detected. Instead, the LoG can be estimated by taking the subject-specific angular offset between the visual and optical axes into account. The visual and optical axes intersect at the cornea center, the nodal point of the eye. A user calibration, which is a procedure to collect ground-truth data from a subject fixating at target gaze points, is required to compute the angle between the axes. In this context, the general theory and the details of gaze estimation using pupil center and corneal reflections are described in [Guestrin and Eizenman, 2006].

3D model-based methods can be considered as the most accurate and robust gaze estimation approach, owing to the sophisticated 3D modeling of the eye and the environment. Knowledge of 3D location of the eyeball center or the corneal center is a direct indicator for the head location in 3D space and may obviate explicit head location models. The estimation of these points is therefore the cornerstone of most head pose invariant models. Although they offer large freedom of movement and high estimation accuracy ($<1°$), they have a significant disadvantage, that is a fully-calibrated complex setup is required. More specifically, they utilize either stereo vision setups or depth sensors in order to accurately obtain the 3D eyeball model. Therefore, camera

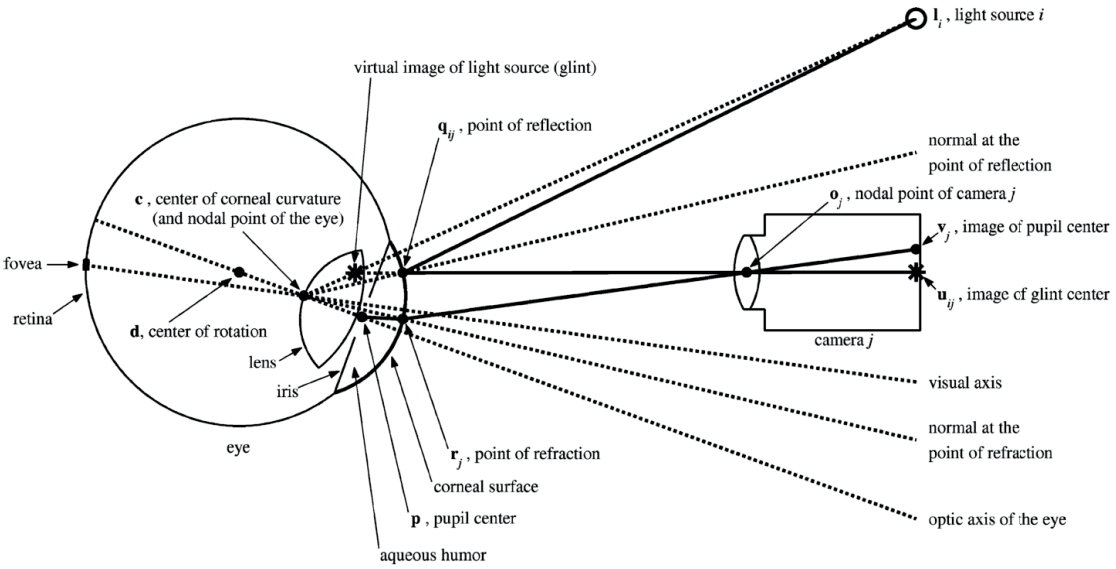


Figure 2.6 – Schematic representations of the human eye, light source, camera and projections. The visual and optical axis of the eye correspond to LoG and LoS, respectively. Image taken from [Guestrin and Eizenman, 2006].

and geometric scene calibrations are mandatory to model the eye and setup, i.e., light sources, cameras, and monitor.

Early efforts were in favor of using multiple stereo systems and pan-tilt units to allow for head movements. For instance, [Beymer and Flickner, 2003] proposed a setup comprised of two stereo systems, which were utilized using a pan-tilt unit. A wide field-of-view (FoV) stereo system was employed to detect the face and a mechanically steered narrow FoV stereo system was used to track the eye at a high resolution. In addition to the camera and geometric calibration, the system required a user calibration by fixating at a number of calibration points to determine subject-specific parameters. The system achieved a high accuracy ($<1^\circ$) under natural head movements. Similarly, [Ohno and Mukawa, 2004] suggested a slightly less complex system consisting of a stereo eye positioning unit to detect the user eye position in 3D, and a gaze tracking unit to estimate the gaze direction from the eye image taken by a near-infrared (NIR)-sensitive camera. The gaze tracking unit was placed on a pan-tilt stand so that it can change its direction to detect the user eye. In addition, [Shih and Liu, 2004] proposed a novel 3D gaze estimation method, which required a single stereo setup with two light sources. Under limited head pose scenarios, they also achieved an accuracy of $<1^\circ$.

[Guestrin and Eizenman, 2006] highlighted that the hardware setup configuration is crucial for *3D model-based methods* to achieve high accuracy and robustness against head movements. When a single camera and single light source are used, the gaze can be estimated only for a single head pose, whereas adding multiple light sources to a single camera setup improves the head pose tolerance. Consequently, they proposed a method, which required at least two synchronized

cameras and four light sources, to achieve a high estimation accuracy using a single point user calibration procedure [Guestrin and Eizenman, 2007].

Although *3D model-based methods* are implicitly more tolerant to head pose variations and head movements, they still suffer from inaccuracy under large head movements. One of the main reasons is that most systems are faced with the trade-off between the head movement range and eye data resolution. In early efforts, e.g., [Beymer and Flickner, 2003, Ohno and Mukawa, 2004], multi-camera systems were mostly utilized, such that a wide FoV stereo system was required to allow free head movements in addition to a narrow FoV one to capture eye images with high resolution. These systems were mostly interconnected through a pan-tilt unit which mechanically reoriented the narrow FoV camera to the users' eye according to the feedback of the wide FoV camera system. Despite enabling a high accuracy and robustness, the use of pan-tilt unit increased the setup complexity and the cost. Later on, such mechanical units were avoided and focused more on introducing more robust models. [Hennessey et al., 2006] presented a single camera non-stereo system based on ray tracing. They achieved an accurate gaze estimation ($< 1°$) while allowing for natural head movements. In addition, [Guestrin and Eizenman, 2007] introduced a method that used the centers of the pupil and at least two glints, which were estimated from the eye images captured by at least two cameras. Their system achieved $<1°$ accuracy error by tolerating head movements in a volume of almost 1 dm^3. Recently, [Sun et al., 2015] proposed a Kinect sensor-based technique that could handle low resolution eye data. Their system used a parametrized iris model to locate the iris center for gaze feature extraction. Thereby, the gaze direction was determined based on a 3D geometric eye model by computing the 3D position of the eyeball center and iris center.

Regression-Based Methods

Regression-based methods detect local eye features, e.g., pupil center, cornea center, glints, pupil or iris contours, eye corners, and compute certain gaze features as shown in Figure 2.7. These features are then mapped directly to the gaze points on a monitor. To do so, a user calibration process is required, in which ground-truth gaze data is collected from the user and a regression is computed between the gaze features and displayed gaze points. Once the calibration is performed, the mapping function is ready to be used during the tracking session.

Despite the estimation accuracies obtained by *regression-based methods* are mostly lower in comparison with *3D model-based methods*, they are much simpler to construct since they often do not require fully-calibrated setups. The history of *regression-based methods* goes as early as 1974 when [Merchant et al., 1974] introduced a real-time eye tracker using a single camera and a NIR light source. They used the pupil-glint vectors as gaze features and performed a linear interpolation to estimate the gaze points on the monitor. Later, many approaches have been proposed, and the most popular approach is perhaps based on polynomial regression. [White et al., 1993] and [Morimoto et al., 2000] proposed to use polynomial mapping of the pupil-glint vector to the PoRs. Later, more sophisticated alternatives, such as Gaussian processes [Hansen et al.,

Figure 2.7 – Pupil-glint vectors that are fed into the mapping function to estimate the gaze. Image taken from [Sesma-sanchez et al., 2012].

2002], generalized regression neural networks [Zhu and Ji, 2004], support vector regressions [Zhu et al., 2006], were proposed for the mapping of gaze features to the monitor coordinates.

Contrary to *3D model-based methods*, *regression-based methods* are considered as approximation models since they indirectly model the eye physiology, geometry, and optical properties. In this regard, their head movement tolerance is significantly lower than *3D model-based methods*. The estimation accuracy significantly degrades under head movements. The main reason relates to the non-linear changes in the gaze features. particularly when the user moves away from the calibration position. One of the main challenges in *regression-based* gaze estimation is to learn a head movement invariant mapping. In order to address this challenge, multiple glints based approaches have been suggested. [White et al., 1993] proposed to use a second light source, which permitted differentiation of head movement from eye rotation in the camera image. Using two glints as points of reference and exploiting spatial symmetries, they proposed a spatially dynamic calibration method to compensate for lateral head translation. A thorough review of polynomial-based regression methods using two glints was later presented in [Cerrolaza et al., 2008]. They evaluated various models using different pupil-glint vectors and polynomial functions.

In addition, [Sesma-sanchez et al., 2012] studied how binocular information can improve the accuracy and robustness against head movements for the polynomial based systems using one or two glints. They proposed alternative mapping features that relies on on the commonly used pupil-glint vector using different distances as the normalization factor. [Cerrolaza et al., 2012] suggested two calibration strategies to reduce the errors caused by head movements. Despite achieving promising results, most of the above efforts required to fix the users' head using a chin rest. Therefore, it is difficult to determine the efficacy of the proposed methods under free-head conditions. Differently from the majority of the *regression-based methods*, [Zhu and Ji, 2007] proposed a stereo-based system, which achieved an acceptable accuracy (~2°) while allowing for larger head movements without requiring the use of a chin rest. They estimated the optical axis of the user's eye in 3D by directly applying triangulation techniques on the glints and pupil center. They also suggested that 3D head pose information can be used to compensate for the bias caused by head movements. However, the main drawback of this system is that a fully-calibrated stereo setup was utilized to obtain 3D information, such that camera and geometric scene calibrations were required.

Cross Ratio-Based Methods

Cross ratio-based methods take advantage of the cross ratio property, a fundamental invariant of the 2D projective space, in order to estimate the PoR. Under 2D projective geometry transformations, neither the distances nor the ratios of distances are preserved. However, the cross ratio (also known as double ratio and anharmonic ratio), in other words, a ratio of ratios of distances, is preserved [Birchfield, 1998]. In the original work of [Yoo et al., 2002], the authors placed four active light sources to the corners of a monitor, which created four glints on the cornea surface. In other words, the monitor was projected on the cornea as a polygon, whose vertices were the glints. In this setup, the gazed point on the monitor was assumed to correspond to the pupil center. Thereon, the cross-ratio property between the screen plane, the camera plane, and a tangential plane to the cornea was used to estimate the PoR on the monitor. The geometric setup configuration and projective relations between the monitor, camera image, and corneal plane are shown in Figure 2.8.

Cross ratio-based methods are simple and fast, and besides share advantages of both *appearance-based methods* and *feature-based methods*. First of all, similar to *appearance-based methods*, they do not require any camera or geometry calibration. Also, they achieve acceptable accuracies while allowing for certain head movement tolerance similar to *3D model-based methods*. Unfortunately, the performance of *cross ratio-based methods* might be limited in accuracy and robustness due to the simplifications assumed. There are two major sources of estimation bias in *cross ratio-based methods* as described in [Kang et al., 2008]. First, the model assumes that the pupil center and glints lie on the same plane. They are, in fact, not coplanar because the cornea has a spherical surface. Second, the model computes the PoR on the basis of eye ball's optical axis (LoS) rather

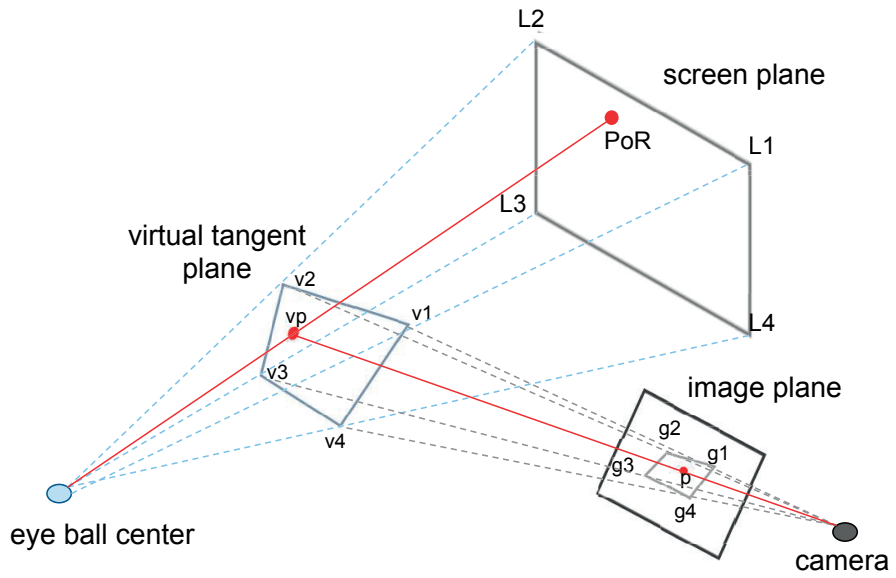


Figure 2.8 – Geometric setup of *cross ratio-based methods* and the projective relations between the monitor plane, camera image plane, and virtual tangent plane.

than the visual axis (LoG). Consequently, a user calibration is essential for *cross ratio-based methods* in order to compensate for the subject-specific estimation bias.

In the literature, several efforts have been made in order to enhance the tracking accuracy and robustness of *cross ratio-based methods* through the user calibration methods. In the original system introduced by [Yoo et al., 2002], there was not any subject-specific bias correction. Later, they refined their method by several enhancements in feature detection and they introduced a technique to compensate for cornea's non-coplanarity using an additional light emitting diode (LED) illuminator in their hardware setup [Yoo and Chung, 2005]. Even though the calibration did not consider the correction for the axes difference, it significantly improved the estimation accuracy. In a similar approach, [Coutinho and Morimoto, 2006] proposed a method to compensate for the axes difference for the first time. Yet, their system required a fifth LED in the hardware setup similar to [Yoo and Chung, 2005]. Later, homography-based bias correction modeling was introduced by [Kang et al., 2007]. They simplified the error correction using a similar calibration procedure but eliminated the need for the fifth LED. It outperformed all previous methods despite having a simpler hardware setup. [Hansen et al., 2010] then proposed a normalized homography mapping to further improve the tracking robustness against perspective distortions.

In *cross ratio-based* gaze estimation, homography-based user calibration methods are widely accepted by the eye tracking community as the state-of-the-art bias correction technique. They have proven to successfully work when there is no large head movements during the tracking. Nonetheless, their accuracies to compensate for the estimation bias are significantly affected by the large head movements. Consequently, alternative techniques have been developed to address the robustness against large head movements. The majority of these efforts suggested solutions by adapting the bias correction to the changes in head movements, e.g., [Coutinho and Morimoto, 2013, Huang et al., 2014]. For instance, [Coutinho and Morimoto, 2013] proposed methods to adaptively correct the bias displacement vector with respect to head movements. In addition, [Huang et al., 2014] proposed an adaptive homography calibration, which is an offline-trained model on simulated data. The model successfully adapted the homography calibration with respect to the head movements. On the other hand, an important limitation in [Coutinho and Morimoto, 2013] and [Huang et al., 2014] is that they utilize a chin rest to keep the head pose fixed during their evaluation. Therefore, the evaluations could not take the head pose variations and continuous head movements into account. Besides, using a chin rest significantly harms the user experience and is impractical for real-world human-computer interaction (HCI) applications.

2.4 Conclusion

Eye tracking has a long research history of over a hundred years and its popularity is now on the rise due to the growing interest from diverse disciplines. *Video-oculography* technique is undoubtedly more preferable due to its non-intrusive nature in comparison with *electro-oculography* and *contact lens* based techniques.

Video-oculography technique can be implemented as *head mounted* trackers and *remote sensors-based* trackers depending on the application scenario. *Remote sensors-based* trackers are mostly preferred over *head mounted* ones since they provide more natural user experience. *Remote sensors-based* gaze estimation methods can be categorized into two groups, namely, *appearance-based methods* and *feature-based methods*.

Appearance-based methods have an important advantage over *feature-based methods*, that is, their hardware and system requirements are considerably lower. In other words, they simply require an uncalibrated ordinary camera. The recent advances in the synthesizing and rendering technology as well as those in machine learning, such as convolutional neural networks, have attracted a significant amount of attention to *appearance-based methods* in the community. Recently, notable improvements in accuracy and robustness have been made, and a great potential has been shown. However, despite their promise, the current performances are still inadequate for them to be utilized for the applications which require precise gaze estimation.

Feature-based methods are still the most widely preferred methods since they significantly outperform *appearance-based methods* in terms of the accuracy. These methods can be analyzed under three categories, each of which has its own advantages and disadvantages. Among all categories, *3D model-based* methods enable the highest accuracy and head movement tolerance, owing to sophisticated 3D eye and environment modeling. However, their setup complexity is considerable higher. They mostly require fully calibrated setups, e.g., stereo vision systems or Kinect-like depth sensors. Unlike *3D model-based* methods, *regression-based* and *cross ratio-based* methods do not require fully calibrated setups as they rely on 2D gaze features and approximations. Despite their simplicity, their accuracies are competitive with those of *3D model-based* systems.

The tracking robustness is also a crucial concern in eye tracking research. There is no doubt that the performances of the existing systems are highly affected by certain factors, including the data resolution, the quality of user calibration, head pose changes, head movements, illumination variations, and subject-specific factors, such as eye wear and eye type. Although numerous efforts, as reviewed in detail in Section 5.1, have been devoted to address such factors and promising results have been achieved, the current performances are still far from ideal. Hence, alternative eye tracking systems that achieve high accuracies using simple, flexible, and user-friendly setups are needed. In addition to the setup complexity and high accuracy, the robustness to real-world conditions and user calibration convenience must be considered as important evaluation criteria. In this regard, new models or frameworks are necessary to improve eye trackers' sensitivity to low resolution eye data, user calibration, head movements, changes in illumination, eye glasses, and eye type variations.

In this thesis, considering the efforts in the literature and the state of the eye tracking research, we design a novel eye tracking framework, which aims to address the aforementioned challenges in eye tracking. First of all, we develop a multi-camera eye tracking framework, which uses a non-intrusive, flexible, and adaptable setup. Within this framework, we suggest to employ a

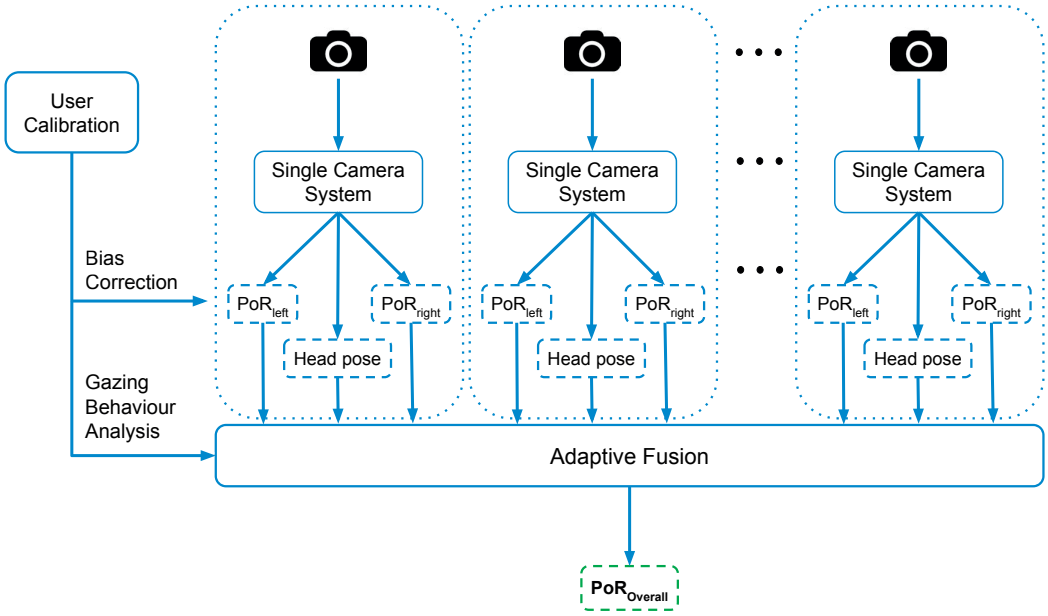
cross ratio-based gaze estimation method, which avoids camera and geometric scene calibrations, enables fast gaze processing for real-time tracking, and operates fairly with low resolution data. The details of the framework are explained in detail in Chapter 3. Secondly, we investigate regression-based techniques to address the user calibration convenience. We present novel methods, which enable to model the subject-specific estimation bias when the calibration data is limited in size and quality, as explained in detail with extensive evaluations in Chapter 4. Lastly, we optimize the proposed multi-camera framework in a way to improve the overall accuracy, tracking availability, and more importantly, tracking robustness to challenging conditions. The details of the adaptive multi-camera fusion mechanism and our comprehensive evaluations are described in Chapter 5. Besides, in our experiments and evaluations, we strictly target natural HCI. We avoid protocols which contradict with it, such as using a chin rest, and also introduce new, more realistic experiments and evaluation schemes.

3 Robust Real-Time Multi-Camera Gaze Estimation Framework

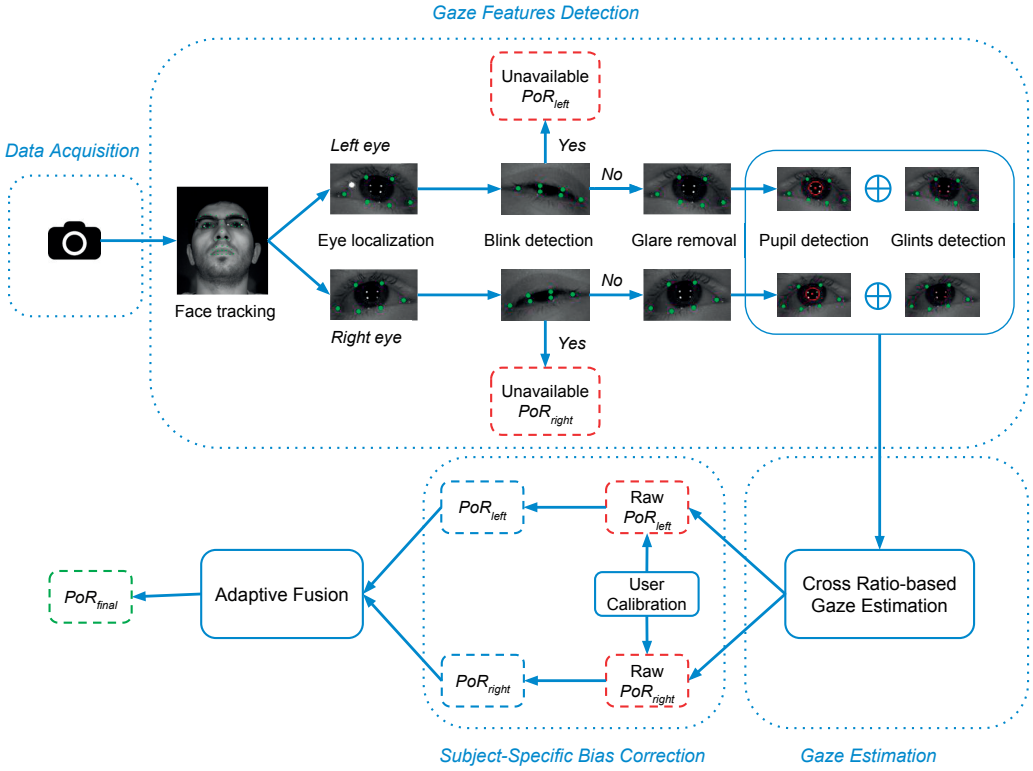
In this chapter, we present the details of the proposed gaze estimation framework. Considering the objectives of this thesis (Section 1.2), together with the acquired insights from the previous efforts (Section 2.4), we design a novel multi-camera gaze estimation framework, which enables real-time accurate eye tracking using low-resolution eye data. Meanwhile, we aim our methodology to operate with a simple and flexible hardware setup, which does not require any camera or geometric scene calibration, as required by the majority of the existing systems. In this respect, we propose a multi-camera gaze estimation framework, which comprises of independently operating single-camera systems. As the estimation of the gaze relies on a simple cross-ratio geometry in each single-camera system, the multi-camera framework is capable of real-time tracking. Besides, it only requires an uncalibrated setup that can effortlessly be adapted for various applications.

An overview of the proposed framework is illustrated in Figure 3.1. It comprises of simultaneously operating multiple single-camera systems, each of which consists of four consecutive processes: i) data acquisition, ii) gaze features detection, iii) gaze estimation, and iv) subject-specific bias correction. Gaze outputs obtained from all single-camera systems are then fed into an adaptive fusion mechanism to output an overall point of regard (PoR) per frame.

In the rest of this chapter, we firstly describe the details of the framework, especially with an emphasis from the data acquisition until cross-ratio based gaze estimation. Then, the remaining two main processes, i.e., subject-specific calibration and adaptive fusion, are briefly explained. Since these two processes constitute the two main contributions of this thesis, they are explained separately in detail in the next chapters. This chapter also presents the real-time implementation of the framework in Section 3.6. Finally, the discussions and conclusions are given in Section 3.7.



(a) Overview of the multi-camera gaze estimation framework.



(b) Overview of the single-camera system.

Figure 3.1 – Overview of the proposed framework.

3.1 Data Acquisition

In the proposed framework, the data is acquired using an uncalibrated multi-camera setup. The setup is completely remote and flexible, such that depending on the application scenario, the number of the cameras and their positioning can be alternated without requiring any camera or geometric system calibration. Since the gaze estimation relies on cross ratio technique, the setup operates under active near-infrared (NIR) illumination. Using active lighting rather than natural one, in fact, brings an important advantage. The tracking performance becomes less sensitive to the variations in ambient illumination, i.e., the performance does not drastically change under indoor or outdoor lighting, or under total darkness, as analyzed in detail in Chapter 5.

In this thesis, we primarily target screen-based (desktop) tracking scenarios due to a high number of potential use cases. In this regard, considering the trade-off between the accuracy and the total cost, we developed a prototypical three-camera setup as can be seen in Figure 3.2.

The prototypical setup consists of three PointGrey Flea3 monochrome cameras, four groups of NIR light emitting diode (LED)s for the active illumination, and a controller unit for the synchronization. Each camera has a medium image resolution (1280×1024 pixels), and is equipped with a large field-of-view (FoV) manual focus lens, i.e., focal length is 8 mm and diagonal FoV is 58°. The cameras are placed around a 24-inch monitor: one of the cameras is located slightly below the monitor, whereas the other two are placed on the left and right sides of the monitor. In order to create the corneal reflections (glints), NIR LEDs with 850 nm wavelength are placed on the corners of the monitor. In addition, band-pass filters around 850 nm are mounted to the lenses to filter the ambient light out. Besides, a micro-controller is programmed to synchronize all the cameras, so that the images can be captured simultaneously by all cameras

Figure 3.2 – Example three-camera setup.

at 30 frames per second (fps). In fact, one of the responsibilities of the micro-controller is to optimize the light emissions regarding the eye safety. Therefore, we synchronize the cameras' shutters with LEDs' emission duration. In addition, a comprehensive quantitative analysis, which discusses impacts of employing different number of cameras in various configurations, is given Chapter 5.

The current setup provides us multiple eye appearances simultaneously acquired from various views. In order to obtain a large working volume to allow for large head movements, we employed large FoV lenses. Therefore, in comparison with the majority of the existing eye trackers, the acquired data resolution is rather low in our system, e.g., image and eye resolutions are 1280×1024 and ~90×50 pixels, respectively. Besides, it is important to note that we put a great emphasis on acquiring the data in a natural manner during our user experiments. The users were explicitly asked to interact with our system the way they feel the most natural and comfortable. For instance, although a chin rest has widely been used by the previous work, it was strictly avoided in our evaluations, so that the users could perform head pose changes and continuous head movements. Sample frames acquired using the current three-camera setup are shown in Figure 3.3.

Figure 3.3 – Sample frames acquired from three subjects using the current hardware setup: (left column) right side camera, (middle column) bottom camera, (right column) left side camera.

3.2 Gaze Features Detection

Once the data is acquired, the system proceeds to detect gaze features to feed into the gaze estimation module. In this section, we describe the details of the gaze features detection, which mainly consists of five consecutive processes, namely, eye localization, blink detection, glare removal, glint detection, and pupil detection.

3.2.1 Eye Localization

Our system starts with the eye localization where the existence of eyes is determined. In order to localize and track the eyes we utilize state-of-the-art robust non-rigid face trackers. In this regard, we used active appearance models (AAM)-based ([Cootes et al., 2001]) and constrained local models-based ([Saragih et al., 2011]) face trackers in the early phases. Currently, we use a supervised decent method (SDM)-based ([Xiong and De la Torre, 2013]) face tracker due to its advantages over the previous trackers. The SDM method assumes that an accurate final face shape with 66 landmarks can be estimated with a cascade of regression models given an initial shape. Viola & Jones face detector [Viola and Jones, 2004] is used to initialize the shape. The face tracker then fits the mean shape in the initial frame and continues the fitting in the succeeding frames. Once the shape fitting reaches convergence, we extract the eye regions by considering the landmarks around the eyes. It is important to note that the extracted eye regions are used as raw, in other words, neither registration nor scaling is performed in order to ensure any particular eye resolution. On the extracted eye regions, we check whether there is any eye blink or not. If there is no blink, we firstly remove the glares on the eye glasses, if they exist, and then continue with the gaze features detection. The gaze features include four glints and the pupil center.

3.2.2 Blink Detection

In order to determine whether there is any eye blink, we analyze the positioning of the landmarks around the eyes. More specifically, we measure the vertical opening (height) of both eyes relative to the eye width. As illustrated in Figure 3.4, if the average of the ratio of eye height to eye width for both eyes is significantly lower (< 0.15) than the open eye form (~ 0.5), we determine that a natural eye blink occurs. Once an eye blink is detected on a frame, the system skips the following processes as no gaze features are available or reliable. Therefore, no gaze output is generated for the frame. Since an eye blink is on average completed within 100 to 200 milliseconds after the peak closure of eyelids, the system does not also output any PoR for the corresponding number of frames upon detection of an eye blink. On the other hand, if the system misses an eye blink, the system proceeds with the feature detection, and naturally no features are detected as the pupil area is not visible due to the blink. Hence, the performance of the system does not depend on the blink detection process. The blink detection process is rather used for computational efficiency as well as providing additional information.

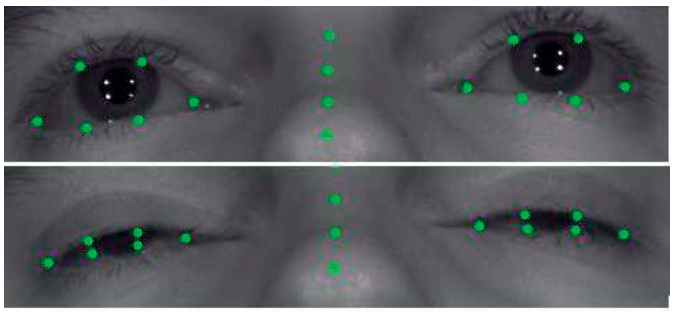


Figure 3.4 – The positioning of facial landmarks in case of (top) no eye blink, (bottom) an eye closure during a blink.

3.2.3 Glare Removal

Glare removal can be considered as a preprocessing of the input eye image to ease the actual feature detection. It mainly aims to clear out the noisy blobs, particularly the specular reflections caused by the eye glasses reflections, which might confuse the glints and pupil detectors. Since the glares on the eye glasses have considerably higher intensities than the rest, we employed well-known image processing techniques for the removal of the glare(s). More specifically, we firstly perform a global thresholding operation, followed by a few morphological erosion and dilation operations to obtain the binary mask of the detected glare. Then, we clean the detected glares by filling them up with the approximated average intensity calculated around the glares. A sample impact of the glare removal process is shown in Figure 3.5. In the resulting eye image, the gaze features can more reliably be detected when the patches do not overlap with them, as illustrated in Figure 3.5c.

(a) (b) (c)

Figure 3.5 – A sample glare removal process: (a) input eye image with a glare, (b) obtained binary mask of the glare, (c) output eye image after the glare removal.

3.2.4 Glint Detection

Well-known image processing algorithms are employed to precisely localize the glints. Firstly, histogram equalization is performed on the input image. It is then followed by a thresholding operation to obtain an initial binary mask indicating the candidate glints. This time, instead of a global thresholding, we use spatial adaptive thresholding in order to take into account spatial

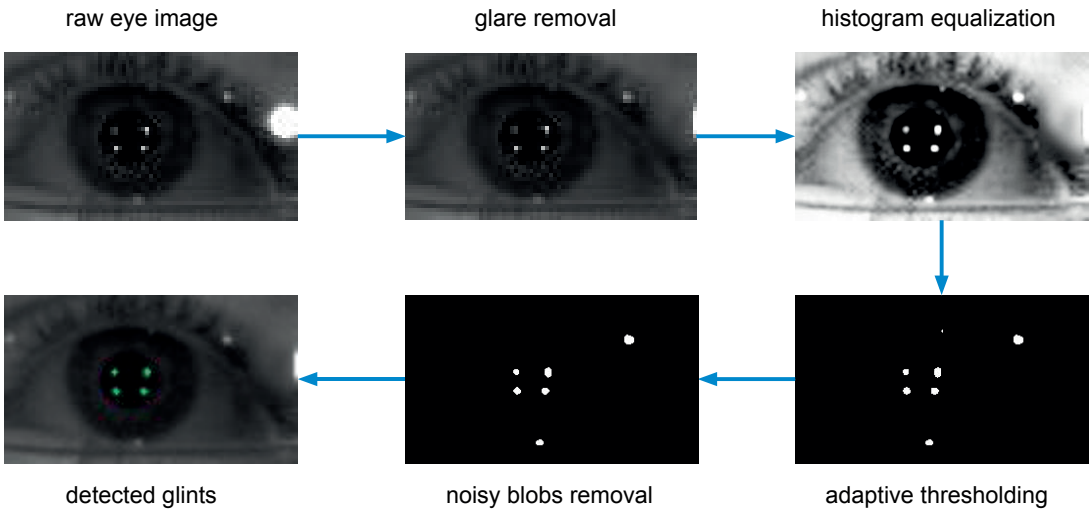


Figure 3.6 – Overview of the glint detection process.

variations in illumination. Adaptive thresholding tunes thresholds for small regions of the image rather than a global threshold value for the whole image. Therefore, various thresholds are applied for different regions of the same image, resulting in more stable thresholding under varying illumination. We use OpenCV's adaptive thresholding function. The parameters *block size* and *C* (i.e., a constant subtracted from the mean) are set to 10% of the original image width and -100, respectively. The actual threshold value, *T(x,y)*, is a mean of the *block size* $\times$ *block size* neighborhood of (x, y) minus *C*. Following the adaptive thresholding, the resulting binary image is processed by morphological operations to get rid of the small noisy blobs. We then perform a connected component analysis to obtain the candidate glints. In the resulting binary image, we expect to find four blobs that form a trapezium since they emerge by the reflections of four light sources located on the corners of the computer monitor. If there exist four or more candidate glints in the binary image, we consider the shapes formed by any four-glints combination. Finally, the set of candidates whose convex hull has the highest match with a template shape representing the screen are considered as the detected glints. Figure 3.6 illustrates the overview of the glint detection process.

3.2.5 Pupil Detection

In comparison with glint detection process, pupil detection is a more troublesome process since the intensity of the pupil is more similar to its surrounding pixels. In this regard, two different approaches are performed throughout the thesis, namely, bright-pupil based detection and dark-pupil based detection. Bright-pupil based approach leverages an optical phenomenon, which is similar to the red-eye effect in colored photography. This phenomenon is generated by placing an additional light source in the optical axis of the camera. It enables a high-contrast pupil region, and so, the pupil can more easily be detected. For this reason, it is widely preferred over the dark-

pupil based detection in the literature. Thus, we as well utilized a bright-pupil based approach in the early phases of our development. However, we later switched to a dark-pupil based one due to the following limitations of the bright-pupil based method. Firstly, the bright-pupil response is related to the size of the pupil. In this respect, it is highly affected by the ambient illumination conditions. In addition, user's age and ethnicity play an important role in the pupil response, such that it works well for Caucasians and Hispanics, whereas the response is much poorer for Asians, as clearly shown in [Nguyen et al., 2002]. Besides, we observed that placing an additional light source per camera, especially in a multi-camera setting, significantly influences the eye glasses robustness due to the additional reflections caused by the increased number of light sources. In fact, using additional light sources may also harm the eyes of the user and increases the system's total power consumption. Please note that a more complete discussion on this issue is given in Chapter 5. The following subsections describe the details of both methods employed in this thesis.

Bright-pupil based approach

This approach is originally suggested by [Ebisawa, 1998] in order to robustly detect the pupil by leveraging the bright-pupil effect, which is generated when a light source is located in the optical axis of the camera. The main advantage of this approach is that the difference of dark and bright pupil images results in a high contrast pupil region. Inspired by Ebisawa's technique, we generated dark and bright-pupil effects by switching between off-axis and on-axis light sources in consecutive frames. When these images are obtained from a high frame rate camera, the difference image provides a high contrast pupil region as can be seen in Figure 3.7. Once such a high contrast pupil region is obtained, we then proceed with the segmentation of the pupil and its center by performing a very similar image processing as in glint detection.

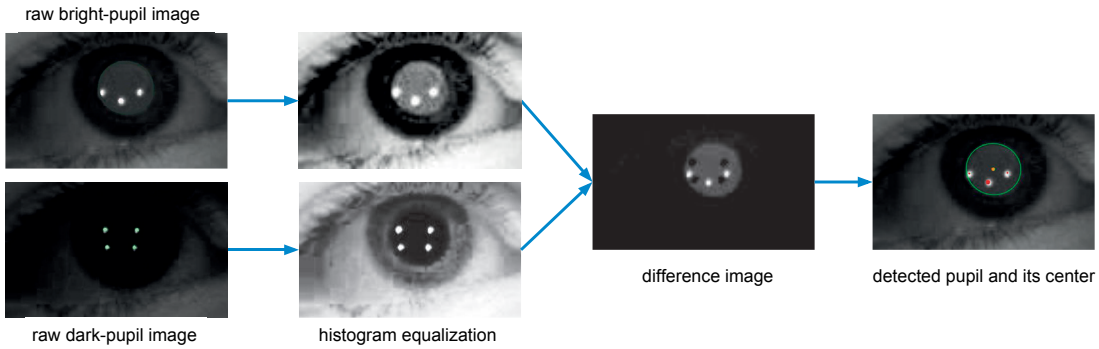


Figure 3.7 – Pupil detection using the bright-pupil approach.

Dark-pupil based approach

Dark-pupil based detection naturally requires a more sophisticated image processing in comparison to bright-pupil effect based approach since it does not exploit any special optical effect.

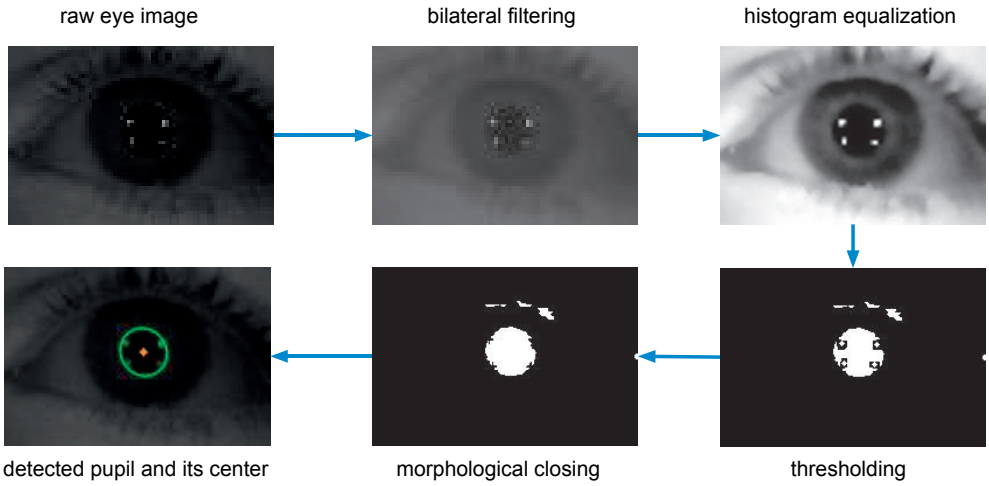


Figure 3.8 – Pupil detection using the dark-pupil approach.

Similar to glint detection process, we utilize well-known image processing algorithms on dark-pupil images. First of all, we perform bilateral filtering on the raw dark-pupil eye image to smooth the pupil region while still keeping the pupil-to-iris edges sharp. We then equalize the histogram to further enhance the contrast. Next, we approximate the average intensity within the pupil's dark regions. As some of the glints are within the pupil region, we discard the glints while calculating the average pupil intensity. We then remove the glints by filling them with the approximated average intensity. On the resulting image, we apply global thresholding by considering the average intensity within the pupil. We then invert the image to highlight the pupil blob. Nevertheless, few other blobs, which are at least as dark as the pupil region, such as eye lashes, eye lids, shades, etc., also remain in the binary image. In order to separate the actual pupil region from the noisy blobs, we iterate over all of them and apply morphological operators to determine the candidate pupil blobs. Among the candidates, we determine the final pupil by considering the shape, size, and the location of the blobs. Lastly, we determine the pupil center by calculating the center of gravity of the found pupil blob. The dark-pupil based detection process is shown with intermediate steps in Figure 3.8.

3.3 Gaze Estimation Based on Cross Ratios

Cross ratio-based gaze estimation methods leverage the cross ratio property, a fundamental invariant of the 2D projective space, in order to estimate the gaze. Although neither the distances nor the ratios of distances are preserved under 2D projective geometry transformations, the cross ratio, a ratio of ratios of distances, is preserved [Birchfield, 1998]. In eye tracking, the cross ratios were exploited to compute the PoR for the first time by [Yoo et al., 2002]. The authors placed four active light sources to the corners of a monitor, which created four glints on the cornea surface. In other words, the monitor was projected on the cornea as a polygon, whose vertices were the glints. In their setup, the gazed point on the monitor was assumed to correspond to the pupil center.

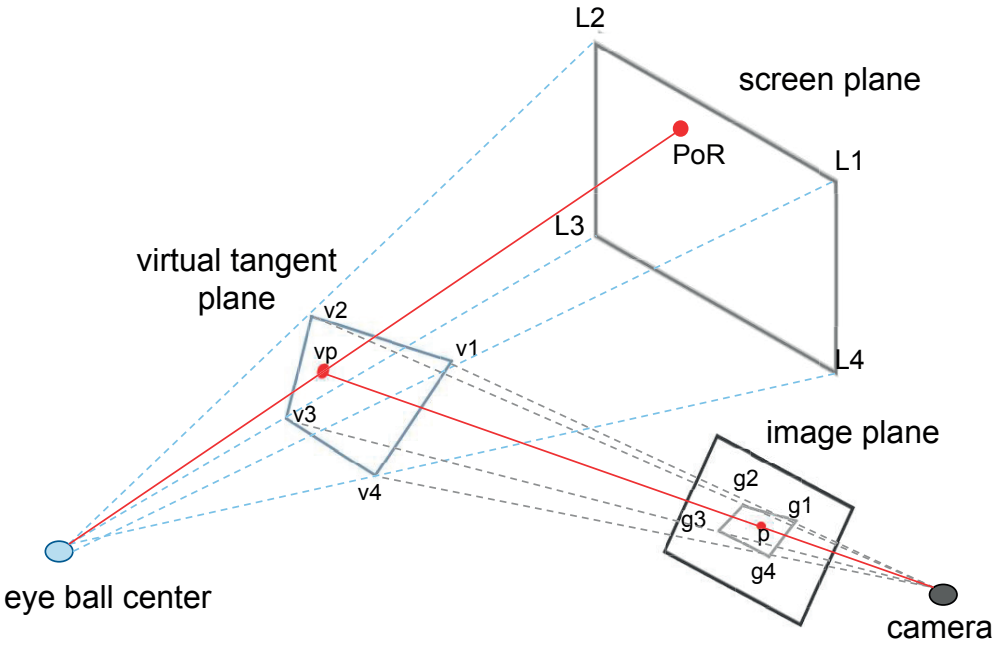


Figure 3.9 – Geometric setup in *cross ratio-based* gaze estimation.

Thereon, the cross-ratio property between the screen plane, the camera plane, and a tangential plane to the cornea was used to estimate the PoR on the monitor. In *cross ratio-based* gaze estimation, the main advantage is that it enables a competitive accuracy while allowing for certain head movement tolerance using an uncalibrated setup. On the negative side, its performance is limited in accuracy and robustness due to the simplifications assumed [Kang et al., 2008].

In this thesis, we employ the original cross ratio-based technique [Yoo et al., 2002] for the estimation of the PoR. Fig. 3.9 illustrates the projective relationships between the camera plane, tangential cornea plane, and the monitor place.

In *cross ratio-based* gaze estimation, a virtual tangent plane on the cornea surface, where the four glints (v_1, v_2, v_3, v_4) lie on, is assumed to exist. Hence, the polygon formed by the glints is the projection of the monitor. Another projection takes place from the corneal plane to the image plane, obtaining the glints (g_1, g_2, g_3, g_4) and the projection of the pupil center, p. As the virtual tangent plane on the cornea has the same planar projective transformation of the monitor and image planes, the pupil center on image plane corresponds to the PoR on the monitor.

The PoR on the monitor can be computed by the equality of the cross-ratios on the monitor plane, $CR_{monitor}$ and the camera image plane, CR_{image} (Fig. 3.10). The cross ratio is defined for four collinear points as:

$$CR(p_1, p_2, p_3, p_4) = \frac{|p_1p_2||p_3p_4|}{|p_1p_3||p_2p_4|}, \tag{3.1}$$

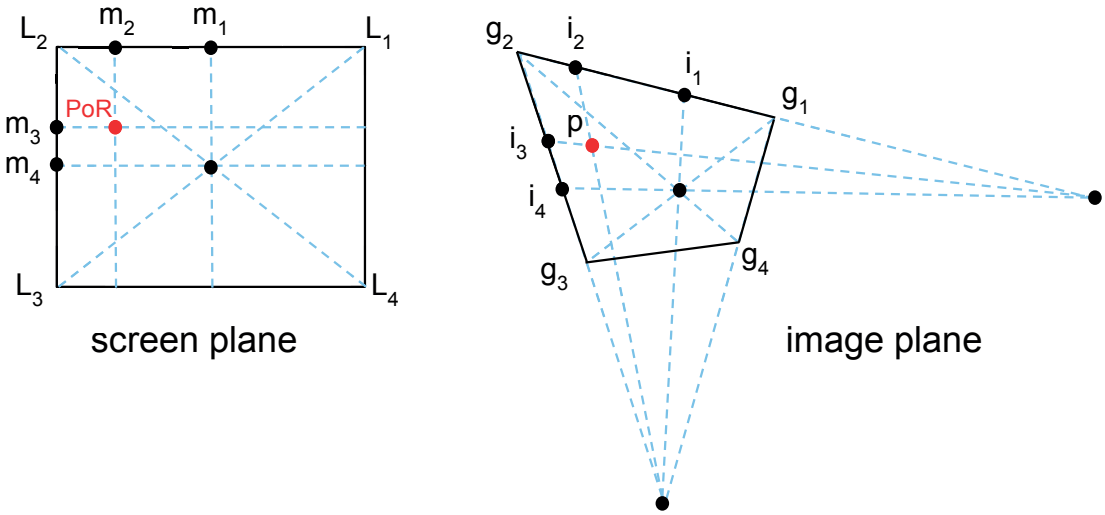


Figure 3.10 – Cross-ratio of image and screen points.

where

$$|p_i p_j| = det \begin{bmatrix} p_i^x & p_j^x \\ p_i^y & p_j^y \end{bmatrix}. \tag{3.2}$$

The cross-ratio on the x axis of the monitor plane can be computed as follows:

$$CR^x_{monitor}(L_1, m_1, m_2, L_2) = \frac{(w - \frac{w}{2})\ \ \hat{p}_x}{(w - \hat{p}_x),\ \ \frac{w}{2}} = \frac{\hat{p}_x}{w - \hat{p}_x}, \tag{3.3}$$

where w is the width of the monitor and $\hat{p}_x$ is the x coordinate of the estimated gaze point p.

The corresponding cross-ratio of the image plane is:

$$CR^x_{image}(g_1, i_1, i_2, g_2) = \frac{|g_1 i_1||i_2 g_2|}{|g_1 i_2||i_1 g_2|}. \tag{3.4}$$

Since the cross-ratios of both configurations are equal, the estimated x coordinate of the PoR, $\hat{p}_x$, can be calculated as follows:

$$\hat{p}_x = \frac{w}{1 + CR^x_{image}}. \tag{3.5}$$

A similar derivation on the y axis gives the estimated y coordinate of the PoR, $\hat{p}_y$, as follows:

$$\hat{p}_y = \frac{h \cdot CR^y_{image}}{1 + CR^y_{image}}, \tag{3.6}$$

where h is the height of the monitor.

3.4 Subject-Specific Calibration

As explained in the previous chapters, a subject-specific estimation bias correction, in other words, a user calibration is crucial for remote gaze estimation. It is employed either to learn a direct mapping between the image or gaze features to the monitor coordinates for *appearance-based* and *regression-based* methods, or to compansate for the estimation bias caused by subject-specific eye parameters for *cross ratio-based* and *3D model-based* methods. In *cross ratio based gaze estimation*, the estimation bias is largely introduced by the simplifying assumptions, which are not valid in practice. [Kang et al., 2008] identified two major sources of estimation bias. First, the model assumes that the pupil center and glints lie on the same plane. In fact, they are not coplanar since the cornea has a spherical surface. Second, the model computes the PoR on the basis of eye ball's optical axis rather than the visual axis, which is the real line of gaze. As these relates to person-specific eye parameters, a calibration is required in order to model the subject-specific estimation bias correction. The calibration procedure is performed once, prior to the use of the system. The users are asked to look at N calibration points on the monitor for K frames long. Subject-specific bias correction, $\mathcal{F}$, can be learned by minimizing the distances between the estimated gaze positions and the corresponding calibration points on the monitor as follows:

$$min \sum_{i}^{N} \sum_{j}^{K} \|\mathbf{P}_i - \mathcal{F}(\mathbf{z}_{i,j})\|, \tag{3.7}$$

where $\mathbf{P}_i$ and $\mathbf{z}_{i,j}$ are i^{th} calibration point and estimated PoRs for this point, respectively.

As mentioned in Section 2.3, many techniques have been devoted to compensate for the estimation bias in *cross ratio-based* gaze estimation. In this thesis, however, we suggest a novel regression-based user calibration methodology since our main objectives differ from most of the previous efforts. As one of the main objectives, we put an emphasis on developing a calibration method that can sufficiently model the estimation bias when there is minimal user effort and when the data is noisy due to low resolution tracking. In this respect, we developed a novel regression-based subject-specific calibration methodology. Instead of employing a classical regression method, we propose to utilize a weighted regression scheme, in which the calibration point clusters and/or

individual calibration samples have varying impacts in the overall regression according to the measured data quality. The details of this method together with comprehensive evaluations are explained in more detail in the next chapter (Chapter 4).

3.5 Adaptive Fusion Scheme

The proposed multi-camera gaze estimation framework, which consists of individual single-camera trackers, is mainly designed to permit natural head movements. To this effect, each single-camera system has the ability to simultaneously track both eyes. Two PoRs can be computed in each frame, in other words, two gaze sensors exist per camera. Consequently, in a multi-camera setup with C cameras, the system is able to generate a total of 2C PoRs per frame. The overall PoR can then be computed by combining the available PoRs obtained from all sensors. For instance, one can perform a simple averaging of available PoRs as the most straightforward solution. However, it is important to notice that the estimation reliability of each sensor may change with respect to various factors, such as the viewing angles of the cameras, targeted gaze location, subject-specific gaze behaviours, eye glasses effects, etc. Thus, an effective fusion ought to take the estimation reliability of individual PoRs into account. This can, in fact, enable a significant improvement in the overall estimation accuracy. In this thesis, we propose to combine the available gaze outputs in a weighted manner, in which the weights correspond to the reliability of individual gaze outputs, as follows:

$$\mathbf{z}^* = \sum_c \sum_e \mathbf{z}_\mathbf{c}^\mathbf{e} w_c^e \tag{3.8}$$

$$\sum_c \sum_e w_c^e = 1, \quad e \in \{Left, Right\}, \quad c \in \{1, 2, .., C\},$$

where $\mathbf{z}^*$ is the overall PoR and, w_c^R and w_c^L are the weights for the right and left eye's PoRs from the c^{th} camera, respectively. In case one of the PoRs can not be calculated for a given frame, then the weight of the missing PoR is set to zero. We do not report an overall PoR in case both PoRs of all the cameras are unavailable for a given frame. In order to determine the reliability of individual gaze outputs, in other words, the weights of the PoRs, we investigated various efforts. The details of the proposed adaptive fusion methods are described in detail in Chapter 5.

3.6 Real-time Implementation

One of the high-priority objectives of this thesis is to achieve real-time eye tracking performance, so that the eye tracker can be utilized in practice. In this regard, we developed a complete multi-camera gaze estimation library in C++. In order to implement image processing and computer vision algorithms, we mostly utilized Open Computer Vision (OpenCV) library[1]. Localization of facial landmarks were performed using an SDM-based face tracker[2]. Furthermore, to achieve real-time tracking performance, Open Multi-Processing (OpenMP[3]) application programming interface was utilized for the parallellization of our library implementation.

The computational complexity of the system is lower than *3D model-based* methods as the gaze estimation is based on simple 2D cross ratio geometry. This enables to achieve a real-time implementation without requiring any performance optimization. In our implementation, the most computationally expensive process is face detection/tracking. Gaze estimation for both eyes using cross ratio algorithm, user calibration, and adaptive fusion processes require much lower computational effort. For instance, these three processes take only ~8 ms on a PC with Intel i7 3.2 GHz processor, whereas face tracking itself takes ~24 ms. Our current three-camera system can simultaneously output PoRs for both eyes as well as an overall PoR at ~30 fps with a mean estimation accuracy error of ~1° of visual angle. We also note that there is still much room for computationally improving our implementation to reach higher frame rates. For instance, the computationally expensive face tracking process can be replaced with a simpler or faster face or eye region tracker, such as local binary features (LBF)-based face tracking [Ren et al., 2014], or one millisecond face alignment with an ensemble of regression trees [Kazemi and Sullivan, 2014]. As the feature extraction process does not require precisely located facial landmarks from a face tracker, but rather needs a rough estimate of the eye region, simpler trackers can be employed to reach higher frame rates while achieving similar estimation accuracies. In addition, the implementation of remaining processes can further be optimized. We leave such computational improvements as future work.

3.7 Conclusion

In this chapter, we propose a novel gaze estimation framework based on multiple cameras. Despite utilizing multiple cameras, the proposed framework relies on a computationally light eye tracking methodology, such that it enables an accurate real-time eye tracking. Our framework consists of independently operating single-camera systems, each of which has a large FoV and operates with low-resolution eye data. In each single-camera system, the gaze estimation relies on simple cross-ratio geometry, which only requires an uncalibrated hardware setup. Therefore, the overall framework uses a simple and flexible setup that can effortlessly be adapted for various applications. In addition, our framework puts a particular emphasis on robust feature detection.

[1] http://opencv.org/
[2] www.humansensing.cs.cmu.edu/intraface/download_functions_cpp.html
[3] http://www.openmp.org/

In this respect, to address the eye glasses tolerance, a glare removal process is applied prior to the glint detection as a preprocessing. Also, a dark pupil-based approach is suggested for the pupil detection in order to improve the robustness to illumination and eye type variations. Overall, we believe the proposed framework is highly valuable for many types of scenarios in human-computer interaction applications. The efficacy of the framework is demonstrated using extensive quantitative evaluations on simulated and real data in Chapter 4 and 5.

4 Regression-Based User Calibration

In this chapter, we present the details of the subject-specific bias correction process, which is one of the main processes of our framework. As mentioned in the previous chapters, subject-specific bias correction, in other words, user calibration, is inevitable for the great majority of the gaze estimation techniques in order to reach high accuracies through compensating for the estimation bias caused by person-specific eye parameters, e.g., the angular offset between the visual and optical axis of the eye ball, the cornea radius and curvature, distance between the pupil center and corneal center, refraction of the aqueous humor and cornea. Therefore, some user effort is required by the eye trackers prior to the actual tracking so that a number of gaze samples with ground truth can be collected. The data can be collected either by explicitly asking the users to gaze at a certain number of target points, or by implicitly acquiring it during the initial interaction with the system. The collected data is then used for modeling the person-specific parameters explicitly or implicitly depending on the employed gaze estimation technique. In any case, such a calibration procedure is tedious for the users, and may significantly harm the user experience. To address this issue, we firstly investigate the potential drawbacks of the existing user calibration method, particularly in relation with using limited and low-resolution data. We then carry out an extensive study of regression techniques together with widely accepted homography-based methods, and consequently, develop a novel weighted regression-based calibration technique that enables a high estimation accuracy with minimal user effort, leading to a convenient user calibration in eye tracking.

In the rest of this chapter, we firstly review the existing methods from the literature in Section 4.1. In Section 4.2, we present in detail the investigated methods and propose a novel weighted regression-based method. We then describe the experimental evaluations on simulated data and user experiments together with discussions and acquired insights in Section 4.3. Lastly, we give our conclusions in Section 4.5.

Note that the majority of the work included in this chapter has been published in the proceedings of *IEEE Winter Conference on Applications of Computer Vision (WACV)* [Arar et al., 2015b] and in *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)* [Arar et al., 2016a].

4.1 Related Work

Gaze-based interfaces aim to accurately map user gaze to the screen coordinates. As previously discussed in Chapter 2, for interactive applications, remote sensors-based gaze estimation methods are preferred due to mainly their non-intrusive nature for the users. These methods can be categorized into two groups, namely, *feature-based methods* and *appearance-based methods* [Hansen and Ji, 2010]. *Feature-based methods* can further be examined under three groups such as *3D model-based*, *regression-based*, and *cross ratio-based* methods. Despite the term *"user calibration"* is used in all groups, it practically refers to different responsibilities in each group. For *appearance-based* and *regression-based* methods, it corresponds to learning a direct mapping to the gaze points from the image content and from the gaze features, respectively. On the other hand, it is required to compensate for the estimation bias due to the variations in subject-specific eye parameters in *3D model-based* and *cross ratio-based* methods. In the former case, the person-specific parameters are explicitly calculated through user calibration, whereas in the latter case, an estimation bias model, which computes an offset from the point of regard (PoR), is usually learned. The following subsections explains the related work from the literature in each category.

4.1.1 Appearance-Based Methods

Appearance based methods learn a direct mapping between the image features and the gaze points rather explicitly modeling the eye in 2D or 3D. Consequently, the input dimensionality is much higher than *feature based methods*. In this regard, one of the major challenge is to handle variations due to user identity, head pose, eye ball pose, illumination, scale, etc., particularly for person-independent mapping. Therefore, the success of these methods rely on how well the training data, i.e., calibration data, covers the variation in the test data.

Early efforts used few thousands of eye images to learn a mapping using multi-layer neural networks [Baluja and Pomerleau, 1994, Xu et al., 1998]. Besides, alternative approaches that were based on Gaussian Process interpolation [Hansen et al., 2002] and appearance manifolds [Kar-Han Tan et al., 2002] have been proposed. These methods significantly reduced the number of labeled training samples required. Later, [Sugano et al., 2010] proposed a novel method to automatically collects samples by utilizing saliency priors from a video. [Lu et al., 2011] then introduced an adaptive linear regression method that automatically selects training samples for mapping. They adaptively find the subset of training samples where the test sample is most linearly representable by solving the problem via ℓ^1-optimization. The method enabled to infer the gaze from variant resolution eye images using much fewer training samples than the previous methods. While high tracking performances were achieved by such methods under controlled conditions, e.g., fixed head pose using a chin rest, stable illumination, the performances degraded greatly when user head was not stationary. The reason was that the head motion deformed the input eye appearance, and so, it significantly differed from original training images even if they all corresponded to the same gaze direction.

[Funes-Mora and Odobez, 2012] used RGB-D sensors to directly handle eye appearance variation by generating frontal view eye images used as input to an adaptive linear regression. They later proposed a framework for 3D gaze estimation which is less sensitive against head pose and inter-user appearance variations, owing to the depth measurements and the fitted 3D facial mesh [Mora and Odobez, 2016]. Their system achieved ∼4° and ∼6° estimation accuracies in subject-specific and subject-independent settings, respectively. From a different perspective, [Alnajar et al., 2013] proposed a method to auto-calibrate gaze estimators based on the observation that humans produce similar gaze patterns when looking at a stimulus. They used the gaze patterns of individuals to estimate the gaze points for new subjects without explicit calibration. They achieved 4.3° estimation accuracy error without a chin rest.

Recent efforts emphasized on subject-independent gaze estimation through capturing larger data variability together with training effective models through convolutional neural network (CNN)s. For instance, [Zhang et al., 2015] collected a large dataset, MPIIGaze dataset, which contains around 214 thousand images from 15 participants during a period over three months. On this large dataset, they trained a multimodal CNN to learn the mapping from the 3-dimensional (3D) head poses and eye images to the gaze directions in the camera coordinate system. They achieved promising results, about ∼ 10°, under uncontrolled, also known as wild, conditions. In a similar effort, [Krafka et al., 2016] introduced another large-scale dataset (GazeCapture) for eye tracking, which contains almost 2.5 million frames from over 1450 people collected using mobile devices (iPhones and iPads). They recruit users from the Amazon Mechanical Turk crowdsourcing platform. Thereupon, they trained a CNN on the collected dataset, and the learned model achieved a significant accuracy improvement over the state-of-the-art *appearance based methods* with an accuracy error of < 4°.

Despite the fact that training models on large datasets provided head pose and illumination tolerance to a certain extent, collecting such datasets is still troublesome and impractical. Alternatively, *learning-by-synthesis* approaches were introduced to increase the data variability using the synthetic eye images. For example, [Sugano et al., 2014] collected a fully calibrated multi-view gaze dataset, UT Multi-view Gaze dataset, from eight synchronized webcams, and performed a 3D eye region reconstruction in order to generate dense training data of eye images. They learned a random regression forest on the synthesized images, and showed improved results. However, their rigid and low-resolution 3D eye models failed to accurately reconstruct the eyeball due to its complex material. [Lu et al., 2014, Lu et al., 2015] synthesized additional eye images of various head poses from the captured real eye images of certain head poses by warping them with pixel displacements rather than using 3D graphics techniques. Although they allowed certain head pose tolerance, the sensitivity to subject and environment variations remain as major concerns.

[Wood et al., 2016b] presented a method to rapidly synthesize large amounts of variable eye region images as training data. Their eye region model was derived from high-resolution 3D face scans, and enabled image-based lighting to cover a range of illumination conditions. In addition, the system enabled modelling up to ±30° deviations in pitch and yaw for head pose. To demonstrate the efficacy of the method, they synthesized over a million eye images and

learned a gaze estimator using k-nearest-neighbors. Despite the simplicity of the classifier employed, they achieved $\sim 10^\circ$ accuracy error on the cross-dataset evaluation on MPIIGaze dataset, and outperformed the CNN-based method described in [Zhang et al., 2015]. Later, the authors alternatively suggested to leverage the benefits of both *appearance based methods* and *3D model-based methods* by fitting a 3D morphable model of the facial eye region to an input eye image using *analysis-by-synthesis*, and the fitted model parameters enabled to obtain 3D gaze information [Wood et al., 2016a]. The use of 3D morphable model (3DMM) brings head pose and illumination tolerance, and the system outperforms [Zhang et al., 2015] with an accuracy error of $< 10^\circ$ on Columbia [Smith et al., 2013] and EyeIdiap [Funes-Mora et al., 2014] datasets. Despite the promising results achieved, limitations still remain such that the system takes several seconds per image to compute the gaze, therefore, it does not enable real-time processing. Also, the method can be trapped in a local minima and further robustness improvements are necessary to be utilized in real-world scenarios and applications.

To sum up, the current accuracy and robustness performances of *appearance-based methods* are not yet comparable to those of *feature-based methods*. Thus, they are still not highly suitable to be utilized for human-computer interaction (HCI) applications that require high accuracy gaze estimation, e.g., $<1^\circ$. Nevertheless, the recent advances in synthesizing and rendering technology together with learning successful models from large scale datasets using deep learning techniques hold a great promise for *appearance based methods*. Once this approach enables to sufficiently capture the data variability, particularly for subject-independent settings, it can lead to significant improvements in the estimation accuracy and robustness.

4.1.2 Feature-Based Methods

3D Model-Based Methods

3D model based methods estimate the 3D gaze direction by modeling the eye in 3D. The intersection between scene geometry and gaze direction is computed as PoR. Compared to other gaze estimation techniques, they offer greater freedom of movement and higher estimation accuracy, owing to detailed modeling of the eye in 3D using fully-calibrated complex hardware setups. As they are based on accurate 3D modeling of user eye, estimation of the subject-specific eye parameters such as corneal ball radii, pupil radial offset, refractive index of the aqueous humor and cornea, foveal angle deviations, etc., is very crucial to achieve accurate gaze estimation. In order to precisely determine these parameters, the geometric relationship of the eye and scene in 3D needs to be examined during the user calibration. In other words, a set of linear system of equations must be derived, and then solved by leveraging the acquired calibration data with ground truth labels.

User calibration is mostly performed by asking the users to gaze at several points displayed, e.g., 9 uniformly distributed points, on a monitor, as in [Beymer and Flickner, 2003, Guestrin and Eizenman, 2006, Park, 2007, Lai et al., 2015]. Besides, various other approaches have been

devoted to improve the user calibration convenience. In this respect, the use of multiple cameras enabled to infer the gaze accurately by requiring a simple calibration procedure. For instance, [Guestrin and Eizenman, 2007] proposed a methodology that achieves <1° accuracy by requiring a simple calibration procedure in which the subject has to fixate only on a single point. Their method used the centers of the pupil and at least two corneal reflections that were estimated from eye images captured by at least two cameras. In addition, [Nagamatsu et al., 2011] proposed a system that is based on a binocular eye model using four synchronized cameras. A pair of stereo cameras were used for capturing the left eye, and the other stereo pair were used for the right eye. Their system enabled ~1.6° accuracy error without requiring any user calibration. Recently, there have also been interesting efforts to eliminate explicit user calibration for the purpose of more convenient and natural HCI. For instance, [Sun et al., 2014] proposed a real-time gaze estimation system with online calibration using a kinect sensor. Instead of displaying a fixed number of calibration points, they updated the eye parameters after each new point. In their methodology, the calibration process was completed as soon as the updates of eye parameters reach convergence. They reported that the system adapted to a new user by online calibration within 3 minutes and achieved an accuracy of ~2°. In addition, [Chen and Ji, 2015] proposed a 3D probabilistic gaze estimation by combining 3D model based gaze estimation with saliency maps. A Bayesian network was introduced to model the probabilistic relationships between the image, gaze, and the eye parameters, where the eye parameters and gaze were estimated by probabilistic inference. They suggested an implicit calibration, in which several images with salient objects were displayed to a user and the method adapted to the user over time. The method achieved an estimation accuracy error of <3° under natural head movements.

Regression-Based Methods

Regression based methods detect local features and learn a mapping from these to the gaze points on a monitor through the user calibration process. Contrary to *3D model based methods*, they are considered as approximation methods since they indirectly model the eye's physiology, geometry, and optical properties. So, their level of accuracy is lower. In addition, as the features non-linearly change when the user moves away from the calibration position, the major challenge is to learn a head movement invariant mapping. In this context, multiple glints based approaches have been suggested. In early efforts such as [White et al., 1993] and [Morimoto et al., 2000], polynomial regression based methods were proposed. Later, alternative regression techniques were exploited to achieve better estimation performance, such as Gaussian processes in [Hansen et al., 2002], generalized regression neural networks in [Zhu and Ji, 2004], and support vector regression in [Zhu et al., 2006]. In [Villanueva and Cabeza, 2008], the authors presented a comprehensive mathematical and geometrical investigation of the user calibration process. They explored the minimum number of hardware elements and gaze features that are needed to accurately estimate the gaze. Later, [Cerrolaza et al., 2008] presented a thorough review of polynomial based regression methods using two glints. They evaluated various models using different pupil-glint vectors and polynomial functions. Similarly, [Sesma-sanchez et al., 2012] studied how binocular information can improve the accuracy and robustness to head movements for the polynomial

based systems using single glint and two glints. Moreover, [Cerrolaza et al., 2012] demonstrated that the pattern of error due to the head movements mainly depends on the system and hardware configuration rather than the user. They suggested two calibration strategies to reduce the errors caused by head movements. Despite achieving promising results, most of the above efforts required to fix the users' head using a chin rest. Therefore, it is difficult to estimate the validity of the proposed methods under moving head conditions. On the other hand, the system proposed by [Zhu and Ji, 2007] achieved an acceptable accuracy error, ~2°, while allowing for natural head movements without a chin rest. They estimated the optical axis of user's eye in 3D by directly applying triangulation techniques on the glints and pupil center. They also suggested that 3D head pose information can be used to compensate for the bias caused by head movements. However, the main drawback of this system is that 3D information was required through a multiple camera stereo system. Their setup consisted of two synchronized cameras, and required camera and geometric scene calibration. Recently, [Xiong et al., 2014] proposed an alternative method based on 3D face structure and pupil center without requiring any glints. Despite their system sacrificed the accuracy (<4°), they eliminated the active light sources.

Cross Ratio-Based Methods

Cross ratio-based methods rely on a fundamental invariant of the projective space, called as cross ratio. They only span a small portion of studies in gaze estimation research, and share advantages from both *appearance* and *3D model based methods*. Nevertheless, their performance might be limited in accuracy and robustness due to the simplifications assumed. As these assumptions relate to the person-specific eye parameters, a user calibration is required to model the estimation bias correction to reach improved accuracy and robustness.

In the original system introduced by [Yoo et al., 2002], there was not any subject-specific bias correction. Later, they refined their method by several enhancements in feature detection and they introduced a technique to compensate for cornea's non-coplanarity using an additional light emitting diode (LED) illuminator in their hardware setup [Yoo and Chung, 2005]. Even though the calibration did not consider the correction for the axes difference, it significantly improved the estimation accuracy. Then, [Coutinho and Morimoto, 2006] proposed a method to compensate for the axes difference for the first time. Yet, their system required a fifth light source in the hardware setup similar to [Yoo and Chung, 2005]. Later, [Kang et al., 2007] introduced a homography based bias correction. They simplified the error correction using a similar calibration procedure but eliminated the need for the fifth LED. It outperformed all previous methods despite having a simpler hardware setup. Similarly, [Hansen et al., 2010] proposed a normalized homography mapping to further improve the robustness against perspective distortions.

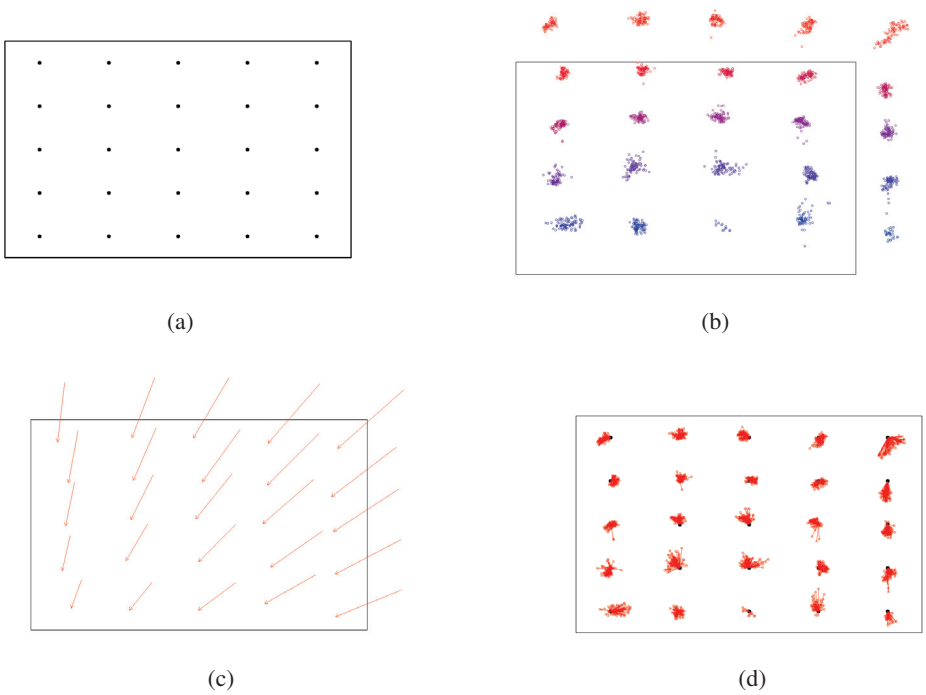
Homography-based calibration approach is widely accepted by the eye tracking community as the state-of-the-art method. The method was proven to successfully work when there is only small head movements. However, they failed to compensate for the estimation bias when there is large head movements, especially in depth, which can occur under real-world HCI

conditions. In this regard, various efforts have been made to bring explicit robustness to large head movements. Most of these efforts suggested to adapt the bias correction to the changes in head movements. For instance, [Coutinho and Morimoto, 2010] suggested a depth compensation method by dynamic correction of the displacement vector. The method accounted for only the vertical head movements. The same authors later suggested another method, *planarization of the cross-ratio features*, which accounts for both horizontal and vertical head movements [Coutinho and Morimoto, 2012]. They reported very high estimation accuracies, ~0.5°, while tolerating large head movements. However, the main limitation of their evaluation was that it required to capture high-resolution eye images by using a zoomed lens together with a chin rest to stabilize users' head pose. In addition, their method required to have an additional light source (i.e., 5 glints) to compute the compensation for the head movements.

[Zhang and Cai, 2014] suggested to use a homography-based calibration modeling with a binocular fixation constraint to jointly estimate the homography matrix from both eyes. Contrary to previous efforts, they utilized information from both eyes to improve the correction model. One drawback of their system is that the features from both eyes must be available to compute a gaze output, which constrains the estimation availability of the system due to the head pose limitations. Moreover, [Huang et al., 2014] proposed an adaptive homography calibration. They learned an offline-trained model on the simulated data by exploring the relationship between the estimation bias and varying head movements. The promising experimental results achieved both on simulated data (i.e., depth and vertical head movements up to ±25 cm) and real data (i.e., ±10 cm depth movements) indicated the efficacy of the method. Nevertheless, its main limitation is to require a chin rest in order to keep the head pose fixed. Although reporting performances using a chin rest may lead to more stable results, it causes the evaluations to discard the impact of variations in head pose. Besides, such restrictions can significantly harm the user experience and would be impractical for real-world HCI applications. Therefore, the use of a chin rest is strictly avoided in this thesis. Instead, we allowed our users to naturally move their heads during the evaluations, and developed a regression-based calibration methodology that successfully works under such head movements.

4.2 Regression-Based User Calibration

This section investigates the regression-based techniques in order to model the subject-specific estimation bias in *cross ratio-based gaze estimation*. This bias is largely introduced by the simplification assumptions of the original *cross ratio-based* method [Yoo et al., 2002]. In this context, [Kang et al., 2008] identified two major sources of estimation bias: *i*) non-coplanarity of the pupil and glints planes, and *ii*) angular offset between visual and optical axes of the eye. Firstly, *cross ratio-based* methods assume that the glints and pupil center lie on the same plane. However, there is no guarantee that they will be coplanar since the cornea has a curved surface. Secondly, it computes the PoR without considering the angular offset difference between the optical and visual axis of the eye ball. Although the real line of gaze is computed based on the visual axis, the algorithm relies on the optical axis for the PoR estimation. As the cornea



(a) (b) (c) (d)

Figure 4.1 – Sample impact of user calibration: (a) calibration stimuli points, (b) raw gaze output, (c) vector fields indicating the bias correction, (d) calibrated gaze output.

curvature and angular offset are individual-specific parameters, i.e., vary from one subject to another, a user calibration needs to be performed to reach a high gaze estimation accuracy.

The user calibration procedure is performed only once, prior to the use of the system. The users are asked to gaze at several stimuli points on the monitor. The subject-specific bias correction model, $\mathcal{F}$, can then be learned by minimizing the distances between the gazed calibration points and estimated gaze positions on the monitor as follows:

$$min \sum_{i}^{N} \sum_{j}^{K} \|\mathbf{P}_i - \mathcal{F}(\mathbf{z}_{i,j})\|, \tag{4.1}$$

where $\mathbf{P}_i$ and $\mathbf{z}_{i,j}$ are i^{th} calibration point and estimated PoRs for this point, respectively. Total number of calibration points is denoted by N, and K frames are acquired per point. A sample illustration of the impact of user calibration can be seen in Figure 4.1. The calibration analyzes the relationship between the calibration stimuli points (Figure 4.1a) and the raw gaze output (Figure 4.1b), and learns a bias correction model, as visualized in Figure 4.1c. The learned calibration model is then applied on the raw gaze data to generate the calibrated gaze output, as shown in Figure 4.1d.

As explained in Section 4.1, many techniques have been proposed to compensate for the estimation bias and significant improvements have been achieved under certain conditions. Yet, in this thesis we further investigate alternative calibration methods, motivated by the following factors: Firstly, we observe that the quality of the calibration increases, at least to a certain extent, when the amount of calibration data increases. However, increasing the data amount by displaying more stimuli points could be tedious, and thus, harms the user experience. Moreover, *cross ratio-based* methods, by their nature, are highly sensitive to the feature detection precision. For this reason, previous efforts in *feature-based* gaze estimation mostly preferred to use high resolution eye data [Yoo and Chung, 2005, Coutinho and Morimoto, 2006, Coutinho and Morimoto, 2013] and fixed head position [Coutinho and Morimoto, 2013, Zhang and Cai, 2014, Huang et al., 2014] to precisely detect the features. Differently from the majority of the previous work, our framework is designed to operate with low-resolution eye data to *i*) allow for natural head movements and a large working volume, and *ii*) enable real-time gaze processing. However, the drawback of our approach is the lower precision in feature detection due to the higher level of noise introduced. The calibration data quality is negatively affected. Hence, we explore different methods to model the estimation bias more robustly against limited amount and quality of calibration data.

In this thesis, we focus on regression analysis to implicitly model the bias correction with a high accuracy under aforementioned settings. First, we examined several regression techniques, including regularized least squares regressions (LSR), partial least squares regressions, and Gaussian process regressions. Then, we propose to utilize a weighted LSR-based method (WLSR) to further improve the calibration model. In this regard, we introduce two different weighting schemes of the calibration data: *i*) weighting of the point clusters, and *ii*) weighting of the individual samples. In addition, we investigate the calibration model convergence by iterative re-weighting schemes. Overall, we perform a comprehensive and detailed analysis of the investigated methods so as to identify their advantages and disadvantages. We compare them with widely accepted homography-based calibration methods.

4.2.1 Homography and Affine Mapping for User Calibration

Homography-based user calibration methods are widely used to model the estimation bias in *cross ratio-based* gaze estimation.An early homography-based calibration method was proposed by [Kang et al., 2007], which aimed to eliminate the need for the fifth light source as used in the previous efforts [Yoo and Chung, 2005, Coutinho and Morimoto, 2006]. Later, its variants have been proposed to bring additional benefits, e.g., [Hansen et al., 2010, Huang et al., 2014, Zhang and Cai, 2014]. As the name implies, these methods rely on a perspective homography transformation to find a mapping between the calibration stimuli points and raw estimations generated by the gaze estimators. The homographic mapping is described by a 3×3 homography matrix:

$$\mathbf{H} = \begin{bmatrix} H_{1,1} & H_{1,2} & H_{1,3} \\ H_{2,1} & H_{2,2} & H_{2,3} \\ H_{3,1} & H_{3,2} & H_{3,3} \end{bmatrix}.$$

A homography transformation has in total 8 degrees of freedom, therefore, 8 unknowns need to be recovered without any regularization considerations. In order to solve this system of equations, four points must at least be known. It is then often solved using the direct linear transformation (DLT) algorithm [Hartley and Zisserman, 2005]. Homography transformations are widely used for performing perspective projections in computer vision, such as for camera calibration, 3D reconstruction, visual metrology, stereo vision, scene understanding, etc. On the other hand, an affine homography or transformation is more constrained with 6 degrees of freedom, and can be defined as:

$$\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,1} & A_{2,2} & A_{2,3} \\ 0 & 0 & 1 \end{bmatrix}.$$

In order to model the image displacements, an affine homography would be more appropriate than the generalized homography under certain conditions, for example, when the image region in which the transformation is computed is small, or when the image has been acquired with a large focal length [Hartley and Zisserman, 2005].

Regarding the user calibration in gaze estimation, homography-based approaches have been proven to successfully model the estimation bias due to the relaxed constraints. Especially under ideal conditions, i.e., when there exists sufficient and good quality calibration data, they can learn a better mapping than those relying on transformations that have lower degrees of freedom, e.g., similarity and affine transformations. On the other hand, having less constraints does not necessarily result in a better modeling in all cases. Under non-ideal conditions, such as when the calibration data contains a lot of outliers or noisy matching pairs due to low-resolution, or when the calibration data is limited in size, homography mapping might be less appropriate to model the displacements. Since an affine transform has less degrees of freedom and model parameters, the calibration problem becomes more determined under such non-ideal conditions. Consequently, a better generalization can be expected on unseen test points. Therefore, we propose to employ an affine mapping instead of a homographic ones for the subject-specific estimation bias modeling, particularly when the calibration data is limited in size and quality. In this context, we investigate several regression-based methods, as explained in the following subsections.

4.2.2 L2-Regularized Least Squares Regression

We firstly employ a L2-regularized least squares regression (LSR), also known as Ridge regression [Hoerl and Kennard, 1970], to find an affine transform with 6 degrees of freedom. The transform $\boldsymbol{\beta}$ is defined with a 3×2 matrix, where the first column corresponds to the offset parameters. The input data $\mathbf{X}$ is a stack of the estimated PoR coordinates:

$$\mathbf{X} = \begin{bmatrix} 1 & \cdots & 1 \\ \mathbf{x}_{1,x} & \cdots & \mathbf{x}_{n,x} \\ \mathbf{x}_{1,y} & \cdots & \mathbf{x}_{n,y} \end{bmatrix}.$$

The corresponding output data $\mathbf{Y}$ stores the target coordinates for calibration. The cost function $E(\boldsymbol{\beta})$ for the regularized least squares problem is defined as:

$$E(\boldsymbol{\beta}) = \|\boldsymbol{\beta}^T\mathbf{X} - \mathbf{Y}\|^2 + \lambda\|\boldsymbol{\beta}\|_F^2, \tag{4.2}$$

where λ is the regularization shrinkage (e.g., $\lambda = 0.1$) and $\|\ \|_F$ stands for the Frobenius norm. A closed form solution can be found by setting the first order derivative of the cost function $E(\boldsymbol{\beta})$ to zero, and we obtain:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{X}\mathbf{Y}^T. \tag{4.3}$$

Using the learned model $\widehat{\boldsymbol{\beta}}$, we can predict a calibrated coordinate giving an input PoR, $\mathbf{x}$:

$$\widehat{\mathbf{y}} = \widehat{\boldsymbol{\beta}}^T\mathbf{x}. \tag{4.4}$$

In addition, we apply a kernelized Ridge regression for calibration, based on the assumption that the error we need to compensate can be nonlinear due to perspective projection. In this case, the cost function $E(\boldsymbol{\beta})$ for a kernel Ridge regression can be written as:

$$E(\boldsymbol{\beta}) = \|\boldsymbol{\beta}^T\Phi(\mathbf{X}) - \mathbf{Y}\|^2 + \lambda\|\boldsymbol{\beta}\|^2. \tag{4.5}$$

A closed form solution can then be found by setting the first order derivative of the cost function to zero, and the prediction becomes:

$$\begin{aligned} \widehat{\beta} &= (\mathbf{\Phi}\mathbf{\Phi}^T + \lambda\mathbf{I})^{-1}\mathbf{\Phi}\mathbf{Y}^T && (4.6) \\ \widehat{y} &= \widehat{\beta}^T\Phi(\mathbf{x}) \\ &= \mathbf{Y}(\mathbf{\Phi}^T\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\mathbf{\Phi}^T\Phi(\mathbf{x}) && (4.7) \\ &= \mathbf{Y}(\mathbf{K} + \lambda\mathbf{I})^{-1}\boldsymbol{\kappa}(\mathbf{x}), \end{aligned}$$

where $\boldsymbol{\kappa}(\mathbf{x}) = [k(\mathbf{x}_1, \mathbf{x}), \ldots, k(\mathbf{x}_n, \mathbf{x})]^T$, and $\mathbf{K}$ denotes the kernel matrix. In this thesis, we use a second order polynomial kernel.

4.2.3 L1-Regularized Least Squares Regression

Lasso regression, in other words, least absolute shrinkage and selection operator, is another form of regularized linear regression where the regularization is based on L1 norm [Tibshirani, 1996]. Therefore, it involves penalizing the absolute size of the regression coefficients. The cost function $E(\beta)$ for Lasso is defined as:

$$E(\boldsymbol{\beta}) = ||\boldsymbol{\beta}^T\mathbf{X} - \mathbf{Y}||^2 + \lambda||\boldsymbol{\beta}||_1, \quad (4.8)$$

where λ is the regularization shrinkage (e.g., $\lambda = 0.1$), and a large enough λ may set certain coefficients to zero.

The regularization can also be interpreted as prior in a maximum a posteriori estimation method. Under this interpretation, the Ridge and the Lasso make different assumptions to relate input and output data on the class of linear transformation. In the Ridge, the coefficients of the linear transformation are normal distributed whereas in the Lasso they are Laplace distributed. Hence, in the Lasso, it is easier for the coefficients to be zero and therefore, it is easier to eliminate some of the input variables which do not contribute to the output.

4.2.4 Partial Least Squares Regression

Partial least squares regression (PLSR) is a method that bears some relation to principal components regression, in which the regression analysis is based on principal component analysis (PCA) by finding hyperplanes of minimum variance between the response and independent variables. Instead, PLSR finds a linear regression model by projecting the predicted variables and the

observable variables to a new latent space in such a way that covariance between projected input and output vectors is maximized [Rosipal and Krämer, 2006]. It is based on partial least squares, which is used to find the fundamental relations between two matrices (**X** and **P**), i.e., a latent variable approach to modeling the covariance structures in these two spaces. More specifically, a partial least squares model tries to find the multidimensional direction in the **X** space that explains the maximum multidimensional variance direction in the **P** space. It can be formulated as follows:

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \\ \mathbf{Y} &= \mathbf{U}\mathbf{Q}^{\mathbf{T}} + \mathbf{F}, \end{aligned} \tag{4.9}$$

where **T**, **U** are the projections of **X** and **Y** in the latent space, respectively. **P**, **Q** are orthogonal loading matrices, and **E**, **F** are the error terms which are assumed to be independent and identically distributed random normal variables. The decompositions of **X** and **Y** are made in order to maximize the squares of covariance between **T** and **U** by finding weight (basis) vectors **w** and **c** such that:

$$\begin{aligned} [cov(\mathbf{T}, \mathbf{U})]^2 &= [cov(\mathbf{X}\mathbf{w}, \mathbf{Y}\mathbf{c})]^2 \\ &= max_{|\mathbf{r}|=|\mathbf{s}|=1}[cov(\mathbf{X}\mathbf{r}, \mathbf{Y}\mathbf{s})]^2, \end{aligned} \tag{4.10}$$

where $cov(\mathbf{T}, \mathbf{U}) = \mathbf{T}^{\mathbf{T}}\mathbf{U}/n$ denotes the sample covariance between score vectors **T** and **U**. Weight vectors **w** and **c** are computed by the nonlinear iterative partial least squares (also known as NIPALS) algorithm [Rosipal and Krämer, 2006] and stored into the projection matrices $\boldsymbol{W}$ and $\boldsymbol{C}$, respectively. Then, input and output data can be projected into the latent space by using these projections:

$$\begin{aligned} \hat{\boldsymbol{x}} &= \boldsymbol{W}^T\boldsymbol{x} \\ \hat{\boldsymbol{y}} &= \boldsymbol{C}^T\boldsymbol{y}. \end{aligned} \tag{4.11}$$

4.2.5 Gaussian Process Regression

A Gaussian process is a statistical distribution for which any finite linear combination of samples has a joint Gaussian distribution. Therefore, any linear functional applied to the sample function will give a normally distributed result.

Given observed samples $\mathbf{D} = \{(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n\} = (\mathbf{X}, \mathbf{Y})$, we formulate the Gaussian process regression (GPR) as follows:

$$\begin{aligned} \mathbf{y}_i &= f(\mathbf{x}_i) + \epsilon_i \\ f &\sim GP(\cdot|0, \mathbf{K}) \\ \epsilon_i &\sim \mathcal{N}(\cdot|0, \sigma^2), \end{aligned} \tag{4.12}$$

where f is the Gaussian process function which is distributed as a Gaussian process with zero mean and a squared exponential covariance function $\mathbf{K}$ [Rasmussen and Williams, 2006]:

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp(\frac{|\mathbf{x}_i - \mathbf{x}_j|}{2\ell^2}) + \sigma_n^2 \delta_{ij}, \tag{4.13}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the data points. The length scale ℓ, the signal variance σ_f^2, and the noise varinace σ_n^2 are the hyperparameters of the GPR. The exponential covariance function has been adopted since it is highly smooth and it makes it possible to account for the noise directly in the covariance function through σ_n^2. The hyperparameters are optimized by maximizing the marginal likelihood[1].

4.2.6 Weighted Least Squares Regression

In classical regression methods, each calibration sample has the same impact on the regression. However, this is not completely valid for the subject-specific calibration in eye tracking. The quality (or reliability) of the calibration data is heterogeneous over the monitor for different calibration points, as can be seen in Figure 4.2. This heterogeneity is related with several factors, including user's varying viewing angles and gazing behaviours, momentary distractions, arbitrary feature detection flaws. In fact, individual samples that belong to the same target calibration point may even have varying qualities. Thus, we propose to extend the classical least squares regression approach to a weighted least squares regression (WLSR) one. In the proposed approach, the calibration point clusters and/or individual samples have varying impacts in the overall regression

[1] A publicly available Gaussian process library is used for the implementation of GPR. The code is available at www.cs.umass.edu/~vidit/Code/GPR.tgz and it uses lapack routines for the matrix operations

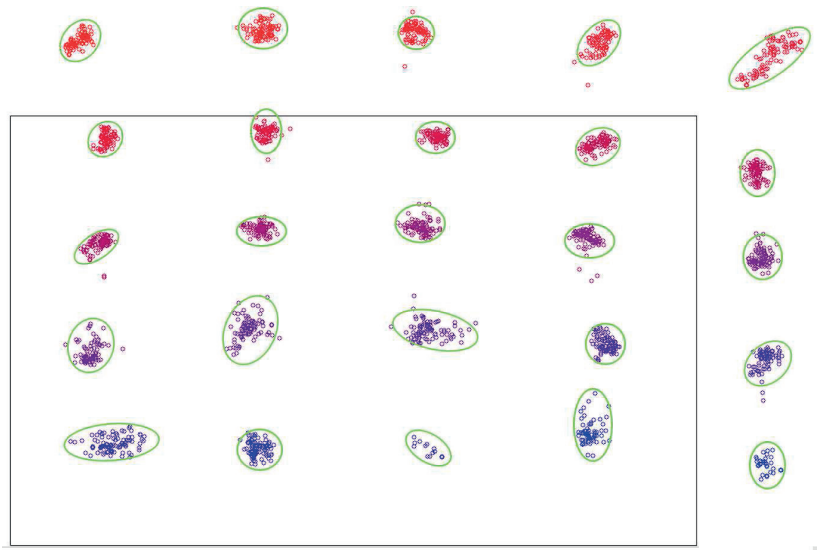


Figure 4.2 – Raw gaze calibration data obtained from a sample user.

according to the estimated quality indicators, in other words, the weights. In this thesis, we develop two weighting schemes, namely, *Point Cluster Weighting ($WLSR_{CW}$)* and *Individual Sample Weighting ($WLSR_{IW}$)*.

- *Cluster Weighting ($WLSR_{CW}$)*: We assign weights, **cw**, to the point clusters by considering the data variance within individual clusters, as defined in Equation (4.14). In this respect, if the samples of a point cluster is concentrated, i.e., a cluster with low variance, a high weight is assigned to the point cluster, and vice versa. It is important to note that individual point clusters are solely weighted in this scheme, such that individual samples of each cluster is assigned with the same weight.

  The point cluster weighting can be formulated as follows:

$$\mathbf{cw_n} = \frac{var_{total} - var_n}{var_{total}} \tag{4.14}$$

$$var_{total} = \sum_{n}^{N} var_n \tag{4.15}$$

$$var_n = \frac{\sum_{k}^{K}(x_n^k - \mu_n)^2}{K}, \tag{4.16}$$

  where $\mathbf{cw_n}$ is the computed point cluster weight for the n^{th} calibration point, x_n^k is the raw gaze estimate of k^{th} sample of the n^{th} calibration point, and μ_n is the robust mean gaze

estimate[2] of the n^{th} calibration point. Once all point cluster weights, $\mathbf{cw_n}$, are computed, they are assigned to each of the calibration sample, $\mathbf{cw_n^k}$.

- *Individual Sample Weighting ($WLSR_{IW}$)*: Differently from the previous weighting scheme, we separately assign weights, **iw**, to each of the individual calibration samples. The weights are assigned according to each sample's distance to its point cluster center as well as the cluster variance. Consequently, the samples that are in lower distances to their cluster centers are assigned with higher weights.

  The individual sample weighting can be formulated as follows:

$$\mathbf{iw}_n^k = \mathbf{cw}_n * w_n^k \tag{4.17}$$

$$w_n^k = \frac{dist_n^{total} - dist_n^k}{dist_n^{total}} \tag{4.18}$$

$$dist_n^{total} = \sum_k^K \|x_n^k - \mu_n\|. \tag{4.19}$$

  where $\mathbf{iw}_n^k$ is the computed weights of the k^{th} sample of the n^{th} calibration point, $\mathbf{cw_n}$ is the point cluster weight for the n^{th} calibration point, x_n^k is the k^{th} sample of the n^{th} calibration point, and μ_n is n^{th} calibration point's cluster center.

For both weighting schemes, once the weights are computed, a normalization is performed across all calibration samples as follows:

$$\mathbf{cw_n^k} = \frac{\mathbf{cw_n^k}}{\sum_n^N \sum_k^K \mathbf{cw_n^k}}, \tag{4.20}$$

$$\mathbf{iw_n^k} = \frac{\mathbf{iw_n^k}}{\sum_n^N \sum_k^K \mathbf{iw_n^k}}. \tag{4.21}$$

Normalized weights are then stored in a diagonal weight matrix, $\mathbf{W}$, and used for the calculation of the regression parameters by a modified version of Equation (4.3) as follows:

$$\widehat{\beta} = (\mathbf{XWX}^T + \lambda \mathbf{I})^{-1}\mathbf{XWY}^T. \tag{4.22}$$

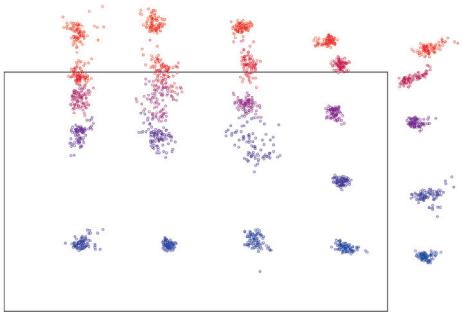
[2] We note that an outlier filtering is performed prior to the calibration mapping as explained in the next subsection.

4.2.7 Iteratively Re-weighted Least Squares Regression

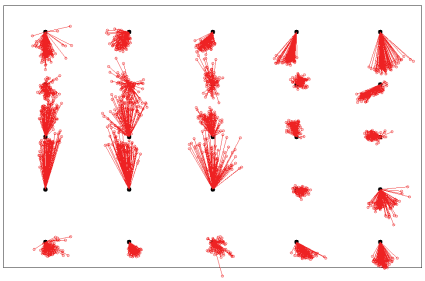
Another factor that affects the calibration quality is the outliers caused by feature detection errors and user distractions during the data acquisition. In order to handle the outliers, our system detects some of the (global) outliers prior to the subject-specific calibration. Global outlier filtering checks each individual calibration sample's distance to its point cluster center. If the distance is larger from an experimentally determined threshold distance, i.e., 4° of visual angle, the sample is classified as an outlier, and filtered out. This process is employed to detect outliers caused by momentary feature detection errors. Nonetheless, it can possibly not detect outliers caused by user distractions or long-lasting feature detection flaws. For example, if a user is distracted, i.e., change his/her gaze, for a second during data acquisition, or the feature detection fails to accurately detect the features for several frames, the system, in the meanwhile, captures several bad samples, as illustrated in Figure 4.3a. In such situations, global filtering fails to eliminate such bad samples. Therefore, a decrease in the calibration quality is experienced.

In this thesis, to overcome the aforementioned problem, we propose an iteratively re-weighted least squares regressions, in which the weights of the detected outliers are set to zero. Firstly, we perform any of the previously described calibration method. Then, instead of storing the learned model as the final calibration model, we project (calibrate) the samples using the learned model, as shown in Figure 4.3b. Then, we calculate the distances between the calibrated samples and their corresponding stimuli points. Under ideal circumstances, the calibrated samples should form dense clusters around the calibration stimuli points. However, if the calibrated sample is considerably further away from its corresponding stimuli point, e.g., 2° of visual angle, the sample is classified as an outlier and its weight is set to zero (Figure 4.3c). Once all samples are examined, we re-learn the calibration model with the remaining samples. This procedure iteratively continues until no further outliers are detected. Once the model converges, it is stored as the final calibration model.

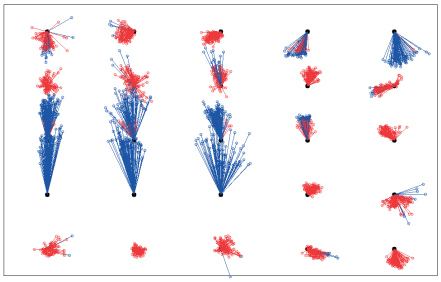
An example illustration, in which the iterative regression is greatly benefited, can be seen in Figure 4.3. The proposed iterative algorithm discards several bad samples, even all samples of a calibration point in some cases, in order to compute a more robust calibration. In this work, we investigate three iteratively re-weighted LSR methods, namely, iterative Ridge, iterative $WLSR_{CW}$, and iterative $WLSR_{IW}$. In these methods, the calibration starts with the corresponding regression methods at the first iteration, and then Ridge regression is employed in later iterations, mainly to prevent overfitting.



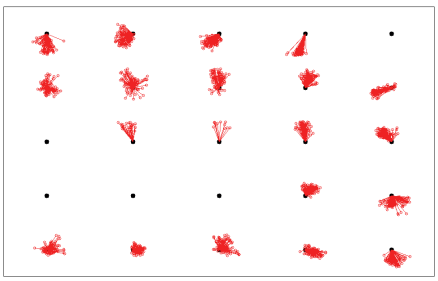
(a) Raw gaze data with many outliers.



(b) Calibrated samples with the first model.



(c) Detected outliers (in blue color).



(d) Calibrated samples with the final model.

Figure 4.3 – Sample calibration data that greatly benefits from the iterative regression approach.

4.3 Evaluations

We conducted several experiments both on simulated and real data in order to evaluate the performances of all the investigated methods. In this section, the evaluations on the simulated data are firstly presented, and then user experiments are explained. In our evaluations, the tracking performances are measured as the gaze estimation accuracy error, which is defined as the average displacement in degrees of visual angle (°) between the target stimuli points and the estimated PoRs, using all raw samples[3] as follows:

$$Error_{pixel} = \frac{\sum_i^N \sum_j^K \|\mathbf{P}_i - \mathcal{F}(\mathbf{z}^*_{i,j})\|}{NK} \tag{4.23}$$

$$Error_{mm} = \frac{Error_{pixel}}{\textit{pixel-to-mm ratio of monitor}} \tag{4.24}$$

$$Error_{\text{visual angle (°)}} = \frac{Error_{mm}}{\textit{user-to-monitor distance}} \frac{180}{\pi} \tag{4.25}$$

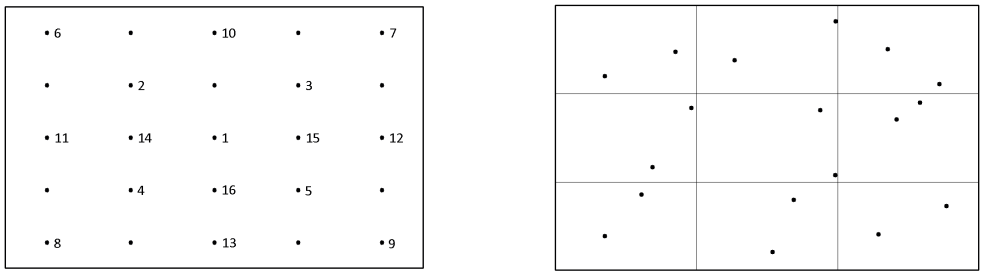
where $\mathbf{P}_i$ and $\mathbf{z}^*_{i,j}$ denote the i^{th} target stimuli point and the estimated raw PoR of the j^{th} sample for the corresponding target point, respectively. User calibration model is denoted by $\mathcal{F}$. Total number of target points is denoted by N, and K samples (frames) are acquired per point during a test session. *pixel-to-mm ratio* is obtained from the monitor specifications[4]. We note that the estimation errors are reported in degrees of visual angle (°) since it is invariant to *user-to-monitor distance*. In addition, the estimation availability is defined as the percentage of samples, which the system is able to compute an overall PoR during the whole evaluation session.

Both simulations and user experiments consist of acquiring the calibration and test data. In the calibration data acquisition, users are asked to gaze at 25 uniformly distributed target stimuli points on the monitor, as demonstrated in Figure 4.4a. The target points are displayed in a left to right and top to bottom sequence in a 5×5 grid on the monitor. Out of these 25 points, we formed 5 different calibration configurations, i.e., *5*, *9*, *13*, *16*, and *25 points calibration*, according to the number of points used for modeling the calibration. Different configurations are separately evaluated in order to examine the impact of calibration data size on the test performance. In Figure 4.4a, the numbers displayed next to the points indicate how these 5 calibration configurations were formed. For instance, the points from 1 to 5 constitute *5 points calibration*[5] configuration,

[3]Neither temporal smoothing nor post-processing is applied in our evaluations in order to demonstrate the real impact of our framework and methods. We employ temporal smoothing only in our real-time demonstration, which leads to a smoother tracking experience for the users.

[4]In our evaluations, 1 pixel is equal to 0.27 mm on our 24-inch monitor. Consequently, 1° of visual angle error indicates ~39 and ~45 pixels on the monitor when the user is at a distance of 60 cm and 70 cm, respectively.

[5]We note that an alternative set of points, i.e., $\{1, 6, 7, 8, 9\}$, could be selected for *5 points calibration* to better



(a) (b)

Figure 4.4 – Target stimuli points used during (a) the calibration data acquisition, (b) a sample test data acquisition.

whereas, the points from 1 to 9 are included in *9 points calibration*.

For the test data acquisition, we introduce a new testing protocol where we place the test points independently of the calibration points to avoid overfitting on the calibration points. In this regard, users are asked to gaze at 18 target points in a 3×3 grid covering the whole monitor. The positions of the target stimuli points in each region are randomly determined. We ensure to strictly display two points in each region so as to cover the whole monitor. The display order of the points is also randomly generated. A sample set of test points is displayed in Figure 4.4b. We believe such a testing protocol not only avoids reporting false test results due to overfitting on the calibration points, but also simulates a more natural and realistic test scenario.

In our multi-camera framework, the calibration is performed separately in each individual camera system. Therefore, we choose to evaluate the performances of the investigated calibration models on the bottom camera system only. This way we aim to highlight the direct impact of these models rather than the multi-camera impact. We also note that among all calibration configurations, higher priorities are given to the ones with lower number of points since we aim to develop a convenient and user-friendly calibration. Thus, the overall performance comparisons and conclusions rely mostly on the results of the *5 points calibration* configuration. In addition, the statistical significance of the arisen differences has been validated by means of a paired sample T-test.

The further details of the evaluation are explained in the following subsections.

cover the space of the potential inputs by including the corner points. Although this configuration may seem more appropriate as it provides more diversity of the training data, the feature detection accuracy, in practice, decreases when users gaze at the corners of the monitor. Therefore, the final *5 points calibration* configuration is formed using the points specified in Figure 4.4a to include more reliable data samples. In the later experiments as described in Chapter 5, the calibration data point locations are updated by considering the trade-off between the data diversity and feature detection accuracy.

4.3.1 Simulation Setup

Simulation data was generated using an open-source software framework developed by [Böhme et al., 2008]. The simulator allows for detailed modeling of different components of the hardware setup and user eye in 3D. It provides in overall a realistic simulation framework. On the other hand, there are few factors of interest that are not currently simulated in the software such as the non-spherical shape of the cornea, occlusion of the eye by the eyelids, the effects of glasses and contact lenses, lack of spatial extent of the light sources, lens distortions or other camera sensor imperfections. Despite these limitations, the simulator offers one of the best solutions that is publicly available[6].

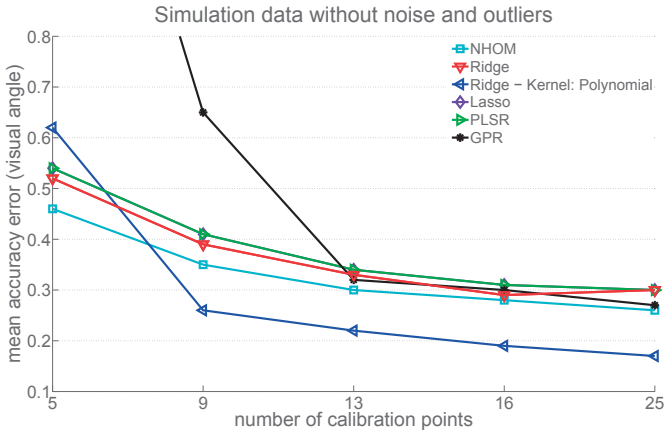
Our simulations aimed to generate the same environment and protocol that were used in the user experiments. To simulate this effect, we use a 24-inch monitor, and place four light sources on the corners of the monitor. The simulated camera is of 1280×1024 pixels resolution and is located 2 cm below the monitor. In the simulated eye, following the typical eye parameters listed in [Guestrin and Eizenman, 2006], the cornea is modeled as a sphere with a radius of 7.98 mm and the pupil center is located 6.2 mm away from the cornea center. The refractive indexes of cornea and aqueous humor are set to 1.376 and 1.3375, respectively. The visual deviations between visual axis and optical axis are 5° for horizontal angle and 1.5° for vertical angle. The eye is located 70 cm away from the monitor.

Moreover, in order to simulate realistic test conditions, we alter the noise level to examine the impact of noise-free and noisy data. In this regard, we introduce, for certain simulation setups, uniformly distributed feature position errors with a magnitude of 0.3 pixels per feature into the generated calibration data. In one of the simulation setups, we also add artificial outliers by introducing additional random feature noise on ~15% of the samples. In this context, we create three different calibration setups: i) no noise, no outliers, ii) noise, no outliers, and iii) noise and outliers. For each calibration point, we generate 100 samples. On the contrary, for each test point, we create a single sample without any noise for all setups.

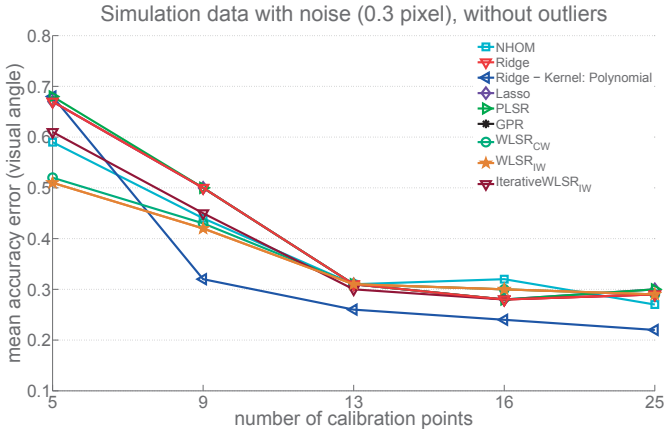
4.3.2 Simulation Results

The simulation results include the average gaze estimation accuracies obtained by employing the aforementioned regression-based calibration methods as well as a widely accepted reference method, which is the normalized homography (NHOM) method proposed by [Hansen et al., 2010]. The results obtained with varying calibration point configurations and data qualities are shown in Figure 4.5. The results of the first simulation setup, in which the calibration data contains neither noise nor outliers, are demonstrated in Figure 4.5a. Firstly, the results indicate that for all investigated methods, higher estimation accuracies are achieved when the number of calibration points are increased. Among the performances of all methods, no significant

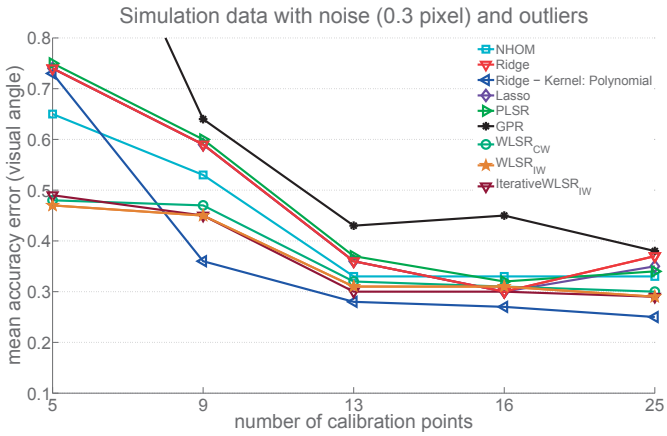
[6]The MATLAB source code of the simulation framework can be downloaded at http://webmail.inb.uni-luebeck.de/inb-toolsdemos/FILES/et-simul-1.01.zip



(a)



(b)



(c)

Figure 4.5 – Comparison of the calibration methods in case the simulation data contains (a) neither noise nor outliers, (b) feature noise, (c) feature noise and outliers.

difference is observed when using a large number of calibration points, e.g., 13, 16, and 25 points. However, the performances can significantly vary when only few points, e.g., 5 points, are used for the calibration. Secondly, the results show that non-linear regression methods, such as GPR and Ridge regression with polynomial kernel, do not perform better than linear regression methods when only few points are used. In fact, GPR requires more than 9 points to reach the accuracies obtained by all other methods. On the other hand, Ridge regression with polynomial kernel significantly outperforms all other methods when 9 or more calibration points are used. Nevertheless, its performance is significantly lower than NHOM and linear regression methods when only 5 calibration points are used. Thirdly, the results demonstrate that the conventional linear regression methods show similar performances regardless of the number of calibration points used. Lastly, and perhaps the most interestingly, NHOM method outperforms all regression-based methods, particularly for the *5 points calibration* configuration when the calibration data does not contain any noise or outliers. We note that the weighted and iterative regression methods are not plotted in this figure since their performances are exactly equal to Ridge regression method because the data variance is absent.

The results of the second simulation setup, in which we introduce certain noise in the feature positions are demonstrated in Figure 4.5b. In this setup, the performance behaviors of the non-linear and linear regression-based methods as well as the NHOM method remain the same as in Figure 4.5a. Only a small performance drop, by about 0.1°, is observed due to the introduced feature noise. On the other hand, the major difference is that the proposed weighted regression-based methods, both $WLSR_{CW}$ and $WLSR_{IW}$, outperform all the others including the NHOM. In fact, the efficacy of the proposed weighted and iterative regression-based methods can better be observed when the calibration data also contains outliers. Especially for the *5 points calibration* configuration, the performance improvement over the NHOM method is notable, as can be seen in Figure 4.5c.

The results of the conducted simulations validate the efficacy of the proposed weighted and iterative regression-based methods, particularly when the calibration data contains noise and outliers, as in real-world conditions. Nevertheless, further evaluations on real-world data are necessary for the validation. The reason is that real-world data often contains not only different levels of feature noise and outliers, but also other sources which were not taken into consideration in the simulations. These other sources include *i*) limitations due to the simulator used, such as no spherical cornea model, no refraction on the cornea surface, and no spatial extent of the light sources, fixed resolution, and *ii*) user-specific factors, such as blinks, distractions, vision disorders, and perhaps the most importantly, fixed head pose. Hence, the evaluations on user experiments are of great important to complete the validation of the investigated methods.

4.3.3 User Experiments

The following sections describe the details of the conducted user experiments together with comprehensive evaluations. As previously mentioned, user experiments were performed using

the bottom camera system to more clearly present the results. In addition, the hardware setup used in user experiments shows dissimilarities in comparison with the final setup presented in Section 3.1. The main reason is that the user calibration experiments use the previous hardware setup, which was employed in the earlier phases of this thesis. This setup was later modified to bring additional benefits, as will be discussed in Section 4.5 and in Chapter 5.

4.3.4 Hardware Setup

The hardware setup used in the user experiments consists of one PointGrey Flea3 monochrome camera, five groups of near-infrared (NIR) LEDs for the active illumination, and a micro-controller unit for the synchronization. The camera has a medium image resolution (1280×1024 pixels), and is equipped with a 12 mm manual focus lens (diagonal field-of-view (FoV) = 49°). The camera is located below the monitor and slightly closer to the user. In order to create the corneal reflections (glints), NIR LEDs with 850 nm wavelength are placed on the corners of the monitor. In addition, band-pass filters around 850 nm are mounted to the lens to filter the ambient light out. In addition, the fifth group of LEDs was placed as ring around the lens of the camera to create the bright pupil effect. Besides, a micro-controller is programmed to synchronize the cameras with the light sources, so that dark- and bright-pupil images can consecutively be obtained at 30 frames per second (fps). In this setup, the user is located approximately 70 cm away from a 24-inch monitor with a resolution of 1920×1200 pixels. The head is not fixed, therefore, the users are allowed to perform natural head movements. Figure 4.6 shows the single-camera experimental setup.

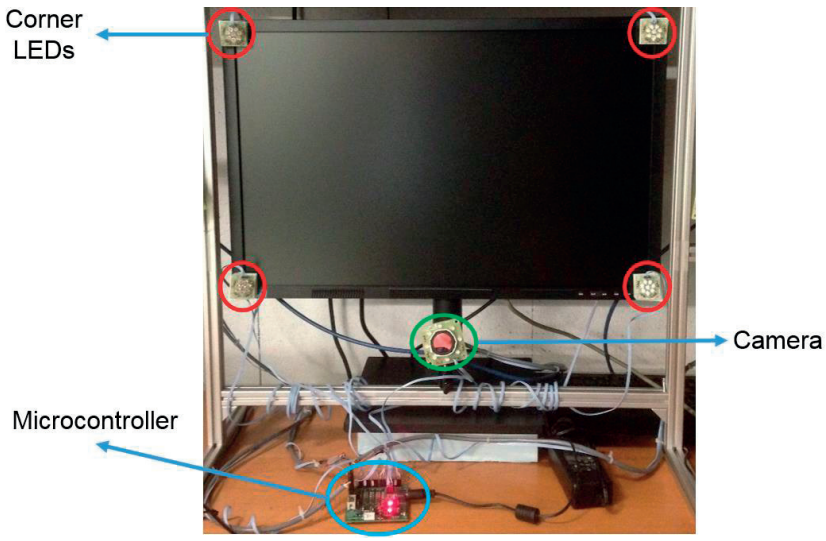


Figure 4.6 – Single-camera setup.

4.3.5 Experimental Protocol

In order to evaluate and compare the performances of the investigated subject-specific calibration methods, a dataset featuring 10 subjects, nine of whom had no previous experience with any gaze tracking system, is collected using the previously described hardware setup. The participants were from diverse origins and did not have any eye wear. Since we targeted a natural and realistic HCI environment, the subjects were asked to gaze at the target stimuli points in a natural and comfortable way. The subjects were positioned in a distance of ~70 cm to the monitor for the data acquisition. The subjects freely performed natural head movements during the experiments since a chin rest was not required to keep the subjects' head still and to keep their eye(s) within the FoV of the camera. On the other hand, as opposed to the user experiments described in the next chapter, the evaluations in this chapter did not explicitly consider large head movements since the main emphasis was given to develop efficient user calibration models under generic head movements. Yet, it included head pose variations in the range of natural human-computer interaction as no chin rest was used. Table 4.1 shows the average head pose variation statistics obtained using the pose estimation method described in [Chen et al., 2012].

The data acquisition was performed similar to the simulations as described in Section 4.3. In user experiments, however, each target stimulus point was displayed for 100 frames (3.33 seconds), and the data of both eyes was captured. In addition, the size of the circular target varied continuously from an initial radius of 30 pixels to a final radius of 20 pixels to serve as visual stimulus.

Once the dataset is collected, the proposed gaze estimation framework (Chapter 3) is employed, except the eye data is available from only a single-camera setup. More specifically, the gaze estimation process starts with face tracking on the frames where we extract eye regions of size ~130×70 pixels. Eye region extraction is followed by feature detection where we detect four glints and a pupil center using the bright-pupil effect. Then, we apply cross ratio-based gaze estimation with the detected features to calculate the initial PoR, i.e., raw gaze data. In the calibration process, we learn an estimation bias correction model on the raw gaze data acquired during the calibration session. The calibration is performed for each eye and for each user separately. In the test process, we apply the learned models to correct the raw gaze data estimated from the test session. Lastly, the corrected PoRs of each eye are combined by an adaptive fusion scheme to output an overall PoR for each frame, as described in Section 3.5.

	Calibration Data		Test Data	
	Yaw	Pitch	Yaw	Pitch
Min	-19.11	-18.51	-11.18	-19.5
Max	23.06	7.95	16.52	3.88
Mean	2.37	-6.92	2.09	-7.23
Std. Dev.	4.28	2.78	3.22	1.79

Table 4.1 – Head pose statistics (in °) obtained by the face tracker on the collected dataset.

4.3.6 Experimental Results

This section presents the results of our analysis regarding certain factors, which highly affect the performance of gaze estimation systems such as the data resolution and amount of eye data used for computing the PoR. In addition, it explains the performance comparisons between the investigated methods and the state-of-the-art methods in detail.

The Effect of Eye Data

First of all, we examine the effect of used eye data for the overall PoR estimation. Since the proposed framework and hardware setup enable to process both eyes simultaneously for a given frame, it is possible to utilize either or both eyes for the estimation of the PoR. In this regard, we obtained results by altering the used eye data, i.e., *Single eye* (either left or right eye), *Strictly both eyes*, and *Adaptive fusion*. *Adaptive fusion*, as defined in Section 3.5, corresponds to calculating the overall PoR using all the available gaze data obtained from both eyes. If the gaze data is not available for both eyes, the gaze data of the available eye is used to set the overall PoR. On the contrary, *Strictly both eyes* calculates the overall PoR only if the gaze data is available for both eyes. In this chapter, two methods are used for the adaptive fusion, namely, simple averaging (uniform weighting) and feature reliability-based weighting, as will be described in Section 5.2.

Figure 4.7 illustrates the estimation accuracies achieved under various configurations to show the impact of the used eye data for the overall PoR estimation. The results are obtained using $WLSR_{IW}$ as the calibration method. The results firstly indicate that individual eyes perform differently. In fact, this may be caused by several factors, such as illumination variations on each eye (e.g., shading, reflections of ambient light or LEDs), head pose and eyeball pose with respect

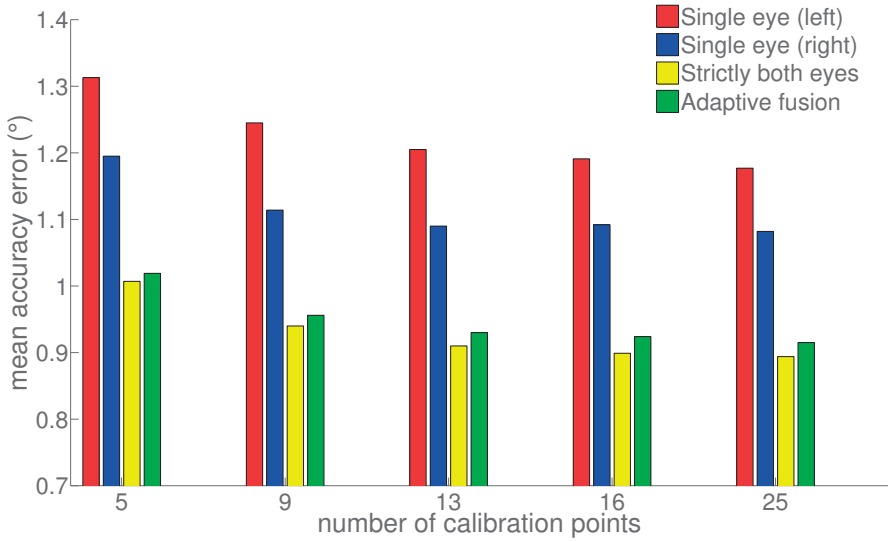


Figure 4.7 – The effect of used eye data for the overall estimation.

Eye Data	Accuracy Error (°)	Gaze Availability (%)
Single eye (left)	1.08	90.7
Single eye (right)	1.18	95.1
Strictly both eyes	0.89	87.8
Adaptive fusion by simple averaging	0.92	96.3
Adaptive fusion by weighting	0.89	96.3

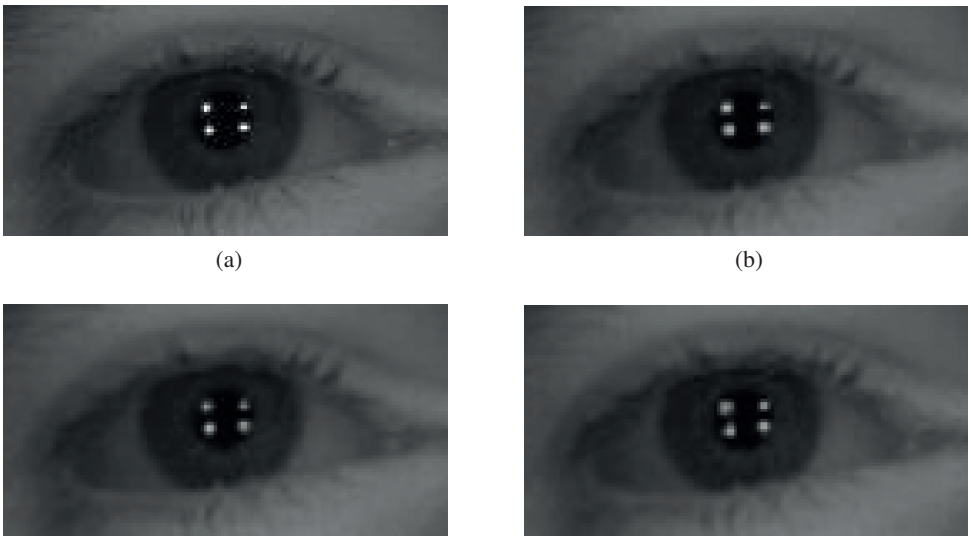
Table 4.2 – Average estimation accuracy errors and gaze availabilities when altering the used eye data and the adaptive fusion method. 25 points are used for the calibration.

to the camera and the gazed point on the monitor, user-specific vision disorders (e.g., lazy eye syndrome, strabismus). Secondly, they demonstrate that utilizing both eyes significantly improves the estimation accuracy. The reason is that the gaze data obtained from both eyes enables to output more reliable PoRs, particularly for certain target points, which require a large head pose or eyeball rotation. For those points, using single eye data may fail due to the obstructed gaze features. In fact, this is one of our main motivations to design a multi-view eye tracking system, as will be discussed in detail in Chapter 5. In addition, we observe that the results do not exhibit a notable accuracy change among the configurations that use both eyes. On the other hand, regarding the estimation availability, which is defined as the percentage of frames in which the system is able to compute an overall PoR, the results highly vary according to the configuration, as listed in Table 4.2. In this regard, the proposed adaptive fusion of both eyes achieves the best performance, such that the system outputs a PoR for 96.3% of all frames, whereas a natural eye blink is detected for 1.86% of all the frames. Therefore, the system could not output a PoR only 1.84% of all the frames due to missing or bad features. Although *Strictly both eyes* configuration notably increases the estimation accuracy in comparison to single eye configuration, the gaze availability significantly drops. The reason is that both eyes must be available to output a PoR, therefore, the system allows for a more limited head pose. Lastly, the results suggest that using *Adaptive fusion by weighting* keeps the gaze availability higher while reaching to the performance of *Strictly both eyes*. Hence, the proposed feature reliability-based weighting method enables the best performance.

Moreover, all the results consistently demonstrate that the estimation error reduces when the number of calibration points increases. However, increasing the number of calibration points has the drawback of harming the user experience.

The Effect of Data Resolution

As the second evaluation, we analyze the impact of data resolution on the estimation accuracies in order to examine the system's tolerance to data quality. Despite the proposed eye tracking system operates with relatively lower data resolution compared to most of the previous work, we further downscaled the images using bilinear interpolation in order to examine the robustness to



(a) (b) (c) (d)

Figure 4.8 – Sample eye regions extracted from (a) an original frame, and downscaled frames by (b) 75%, (c) 60%, (d) 50%.

even lower resolutions. Sample eye regions extracted from an original frame and downscaled frames are shown in Figure 4.8. The extracted eye region (Figure 4.8a) from the original frame (1280×1024 pixels) has a resolution of 130×70 pixels, and the polygon formed by the glints is around 12×7 pixels. The original frames are downscaled in each dimension by 75% (960×768), 60% (768×614) and 50% (640×512) to generate different resolution data. The same feature detection and calibration methodology are applied on the generated data. We note that no particular parameter tuning according to data resolution is performed.

Table 4.3 illustrates the resolution impacts on the overall estimation accuracies when $WLSR_{IW}$ is used as the calibration method and feature reliability-based weighting is utilized for the adaptive fusion. The results show that downscaling by up to 75% does not significantly affect the overall estimation accuracies. Towards 60% downscaling, the accuracy error starts to get higher, and more than 60% downscaling results in a very significant performance decrease. We also observe

Data Resolution	Number of Calibration Points					Gaze
	5	9	13	16	25	Availability (%)
Original frame	1.01	0.94	0.92	0.90	0.89	96.3
Downscaled by 75%	1.05	0.96	0.92	0.91	0.90	95.8
Downscaled by 60%	1.16	1.08	1.06	1.05	1.05	93.1
Downscaled by 50%	1.68	1.6	1.55	1.53	1.52	82.4

Table 4.3 – Average gaze estimation accuracy errors (in °) and gaze availabilities when altering the data resolution.

that the impact remains consistent among different calibration configurations. Hence, the results indicate that the system can tolerate a lower resolution up to 60-75% without sacrificing too much the accuracy.[7] For further downscaling, we observe that the feature detection, especially for the glints, is highly affected by low-resolution. Therefore, less precisely detected features result in lower accuracies.

Comparison of Weighted and Iterative Regression Methods

Figure 4.9 illustrates the average estimation accuracy comparison of the conventional LSR-based user calibration and the proposed weighted and iterative LSR methods when using different number of calibration points. The major observation is that the weighted LSR methods, i.e., $WLSR_{IW}$ and $WLSR_{CW}$, provide a significant performance improvement over the conventional Ridge regression-based method, particularly for the *5 points calibration* configuration. Among the weighted LSR methods, $WLSR_{IW}$ performs slightly better than $WLSR_{CW}$. However, there were no statistically significant differences according to the paired t-test, i.e., $p > 0.05$.

Furthermore, we observe that the proposed iterative LSR methods do not provide notable performance enhancement even though they require additional computations in the calibration process. In fact, the only improvement is achieved by iterative Ridge method over the traditional Ridge method. On the other hand, iterative $WLSR_{IW}$ and iterative $WLSR_{CW}$ methods perform even worse than their non-iterative versions. We believe that the effectiveness of the iterative methods greatly depends on the data, as clearly demonstrated in the evaluations on the simulated data. It is essential to emphasize once again that the iterative methods are designed to address the problem of outliers caused by user distractions and persistent feature flaws during the calibration data acquisition, as explained in Section 4.2.7. However, such situations arise rarely. In our

[7] In fact, as the results suggest that the system can tolerate lower resolution data, we later employed larger FoV lenses in our final prototype, as described in the next chapter.

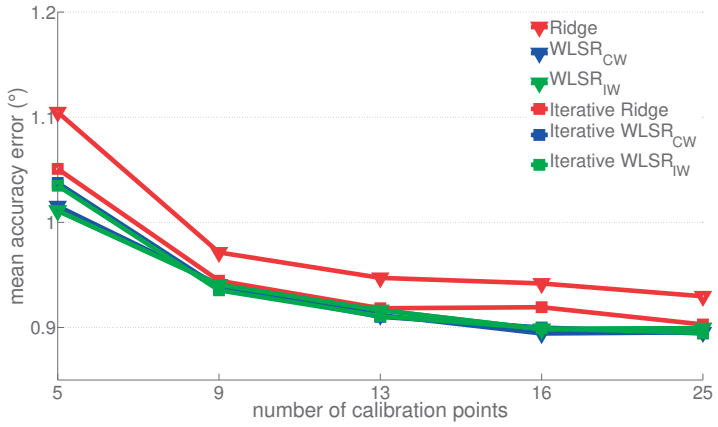


Figure 4.9 – Comparison of the proposed weighted and iterative LSR-based calibration methods.

user experiments, we have encountered only one case out of ten subjects. Even though this particular subject's results are improved by the iterative methods, the influence on the overall results is negligible. Another reason could also be that iterative regression tends to overfit the calibration data since certain samples providing the data variance are eliminated during the iterations. Considering these, we conclude that iterative regressions have the potential to learn a better calibration model for certain applications where the user data is rather noisy and contains a lot of outliers. In this thesis, among all the proposed methods we suggest to utilize $WLSR_{IW}$ as the subject-specific calibration approach since it is both effective and computationally simpler. In the following section, we only present the results of $WLSR_{IW}$ for the clarity of the presentation.

Comparison of Investigated Methods

This section presents a comparison of the investigated non-linear and linear regression-based calibration methods together with the NHOM method. First of all, as depicted in Figure 4.10, all linear regression methods notably outperform the non-linear regression methods, i.e., Ridge with polynomial kernel and GPR. The results suggest that linear regression methods are superior to non-linear methods. The main reason is that non-linear methods easily overfit on the calibration data when there is limited data, e.g., *5 points calibration*.

The results also indicate that linear regression-based methods provide significantly better generalizations than the homography-based method, especially when the calibration data is limited, such as *5 points calibration*. The main reason for this relates to the reduced model parameters and degree of freedom in affine mapping as discussed in Section 4.2.1.

Furthermore, the proposed weighted LSR method, $WLSR_{IW}$, achieves the best performance for

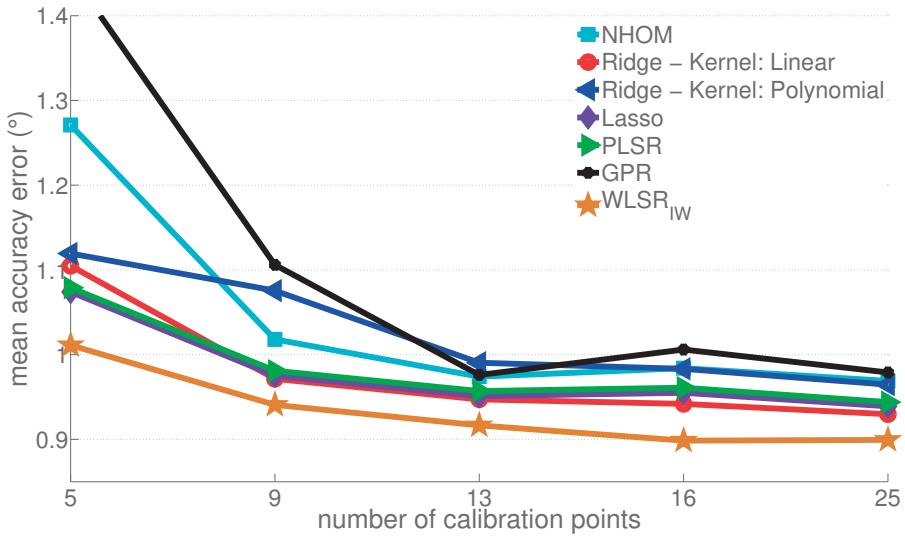


Figure 4.10 – Comparison of the investigated calibration methods.

all the calibration configurations. Particularly, the performance enhancement is noteworthy when using 5 points calibration, which validates the efficacy of the proposed methods towards obtaining a more convenient user calibration.

Moreover, the performances of the conventional linear regression methods such as Ridge, Lasso, and PLSR are all very similar. Using different regularizations or utilization of a latent space for the least squares does not seem to positively influence the quality of the regression in user calibration problem. Since the number of input variables is small in user calibration, these do not present a crucial impact on the results.

Comparison with Previous Work

In this section, we compare the performance of our best performing calibration method, $WLSR_{IW}$, with some of the recent previous efforts, including, normalized homography (NHOM) [Hansen et al., 2010], Gaussian process regression (GPR)[8] [Hansen et al., 2010], and binocular homography fusion (BHF) [Zhang and Cai, 2014], as shown in Figure 4.11. In addition, we compare the performances of all investigated methods together with the previous work in more detail in Table 4.4. It is important to note that this table does not include the earlier methods, e.g., [Yoo and Chung, 2005, Coutinho and Morimoto, 2006, Kang et al., 2007], for two reasons: i) some of these methods require special hardware material, and ii) the method that we compare have been proven, e.g., in [Hansen et al., 2010], to perform better than these earlier methods. Also, we do not include the comparison with NHOM's variants, such as [Coutinho and Morimoto, 2013] and [Huang et al., 2014], which are proposed to bring explicit robustness against large

[8]GPR was employed after the initial NHOM calibration.

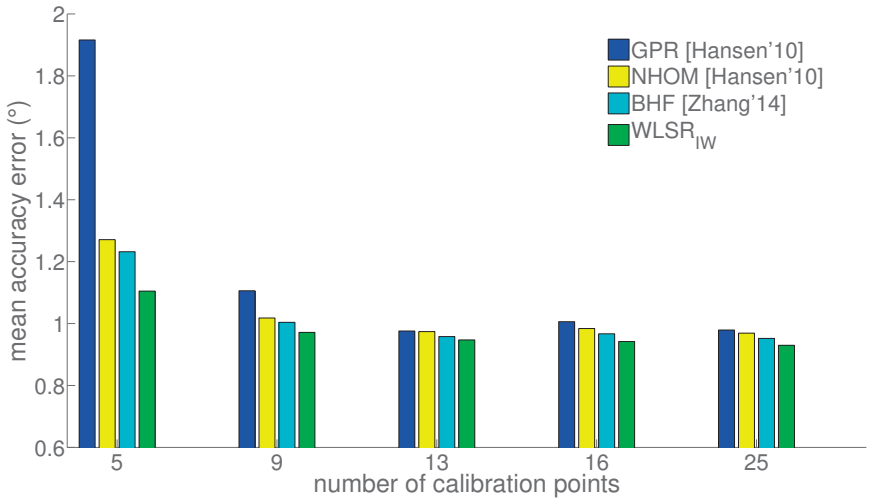


Figure 4.11 – Comparison with the state-of-the-art user calibration methods employed in cross ratio-based gaze estimation.

Method	Required Eye	Number of Calibration Points 5	9	16	25	Gaze (%) Availability
No calib [Yoo et al., 2002]	Single	6.63	-	-	-	90.7
GPR [Hansen et al., 2010]	Single	1.91	1.11	1.01	0.98	90.7
NHOM [Hansen et al., 2010]	Single	1.39	1.14	1.09	1.07	90.7
NHOM [Hansen et al., 2010]	Either	1.27	1.02	0.98	0.97	96.3
BHF [Zhang and Cai, 2014]	Both	1.23	1.00	0.97	0.95	87.8
Ridge (poly)	Either	1.12	1.08	0.99	0.96	96.3
PLSR (poly)	Either	1.10	0.99	0.97	0.96	96.3
Ridge (linear)	Either	1.10	0.97	0.94	0.93	96.3
PLSR (linear)	Either	1.08	0.98	0.96	0.94	96.3
Lasso	Either	1.07	0.98	0.96	0.94	96.3
Iterative Ridge	Either	1.05	0.95	0.92	0.9	96.3
Iter. $WLSR_{CW}$	Either	1.04	0.94	0.9	0.89	96.3
Iter. $WLSR_{IW}$	Either	1.03	0.94	0.9	0.89	96.3
$WLSR_{CW}$	Either	1.02	0.94	0.89	0.89	96.3
$WLSR_{IW}$	Either	1.01	0.94	0.9	0.9	96.3

Table 4.4 – Comparison of the investigated methods with previous work. Average estimation accuracy errors are reported in degrees of visual angle (°).

head movements. Therefore, the performance improvement over NHOM is marginal when no large head movements are considered in the evaluations. Since our user experiments for the calibration do not include large head movements, we omitted both methods from the comparison in Table 4.4.

The overall comparison of methods by altering the number of calibration points is shown in Figure 4.11. In this figure, the proposed adaptive fusion of both eyes is applied to compute the overall PoRs. The results demonstrate that the proposed calibration approach, $WLSR_{IW}$, achieves the best estimation performance in all the calibration configurations. Especially for *5 points calibration* configuration, there is a significant enhancement, about 20%, achieved by the proposed weighted regression-based method in comparison to NHOM and BHF methods. In addition, we observe that the performance of GPR [Hansen et al., 2010] is significantly poorer than the other methods when using *5 points calibration*. The results indicate that GPR requires more calibration points to achieve as good generalization as the others. In other words, as a non-linear regression method, it is more likely to fail modeling the estimation bias when the calibration data is limited. We also observe that our evaluation protocol, in which we chose the test points independently of the calibration points, is capable to avoid overfitting on the calibration points. In fact, when the calibration and test points are chosen from the same set of points, the investigated non-linear regression methods demonstrate competitive or better performances, however, this is due to the overfitting on the calibration points.

Moreover, leveraging both eyes through adaptive fusion scheme highly boosts the results, as can be seen from Table 4.4. For instance, although the improvement from NHOM to BHF does not

seem significant from Figure 4.11, there is, in fact, a significant increase compared to the original NHOM, which utilized single eye data. In order to highlight the impact of utilizing both eyes, we listed the performance of the NHOM method by using different eye data in Table 4.4. On the other hand, it is also important to note that BHF's gaze estimation availability (87.8%) is lower than the methods which require either of the eyes (96.3%), e.g., Ridge, $WLSR_{IW}$. The reason is that BHF requires both eyes to be available to output a PoR while adaptive fusion-based methods can also output even if there is only single eye available.

4.4 Discussion

A comparison of the representative eye tracking techniques in several aspects, such as hardware setup requirements, user calibration, accuracy, head pose flexibility, required data resolution, real-time property, is given in Table 4.5. Since the majority of the existing work require particular hardware and system setups, e.g., additional light sources, setup calibration, 3D or depth information requirements, we could only reproduce and validate the results of few studies. For the remaining studies, we list the reported accuracies obtained from the corresponding references. Although a direct comparison in accuracy would not be fair, the detailed information can help us to make the following inferences. First of all, we observe that the popularity of *appearance-based* methods, which have lower hardware and user calibration requirements, have been increasing recently in parallel with the recent advances in machine learning (e.g., CNNs) and in synthesizing and rendering technology. Although earlier efforts in *appearance-based* gaze estimation tended to learn person-specific gaze models through a user calibration, the recent trend is to learn person-independent gaze models, so that calibration-free gaze estimation can be obtained. Towards capturing sufficient data variation, the researchers either take advantage of CNNs trained on large-scale datasets, such as MPIIGaze [Zhang et al., 2015], GazeCapture [Krafka et al., 2016], or augment the training data using synthesized images, such as [Sugano et al., 2014, Lu et al., 2015, Wood et al., 2016a]. Nevertheless, the data variation captured using the existing approaches has not been sufficient to achieve high-accuracy eye tracking. Yet, their potential is likely to be exploited in the near future. Secondly, *feature-based* methods mostly outperform *appearance based methods* in terms of the estimation accuracy. However, they mostly require particular hardware, e.g., NIR cameras, light sources. Especially, *3D model-based* methods need fully-calibrated setups, which include multi-camera stereo vision system or a Kinect-like sensor, to accurately model the eye in 3D. The recent efforts in *3D model-based* gaze estimation aims to eliminate the explicit user calibration procedure. In this manner, implicit or online calibration strategies have been proposed, e.g., [Sun et al., 2014, Chen and Ji, 2015]. However, their accuracies are significantly affected (>2°). Thirdly, there exists only a few studies, which emphasized on improving the user calibration convenience, e.g., [Villanueva and Cabeza, 2008], in *regression-based* gaze estimation. The main focus in the majority of the recent studies has been given to improve the estimation accuracy and robustness, e.g., [Cerrolaza et al., 2012, Sesma-sanchez et al., 2012]. Similarly, most of the recent *cross ratio-based* methods, e.g., [Coutinho and Morimoto, 2013, Huang et al., 2014], have rather addressed the head movement

robustness by adapting the estimation bias models learned during the user calibration.

Furthermore, we observe that the eye data resolution plays an important role in the tracking performances regardless of the method employed. In this context, most of the previous work, in fact, used high-resolution eye data, which was captured using narrow FoV lenses, e.g., [Guestrin and Eizenman, 2006, Nagamatsu et al., 2011, Villanueva and Cabeza, 2008, Sesma-sanchez et al., 2012, Coutinho and Morimoto, 2013]. Since these systems used a narrow FoV to capture high-resolution eye images, most of them also required to use a chin rest to keep the users' head stable during the data acquisition. However, this leads to an unnatural experience for the users, and so, represents an unrealistic tracking scenario. Besides, there exists a clear performance gap between fixed-head and free-head scenarios, particularly for the approximation-based methods, such as *regression-based*, *cross ratio-based*, and *appearance-based* methods. To this effect, using a chin rest can bias the results, therefore, it remains an important limitation. Consequently, the proposed calibration framework can be considered as an effort to reduce this gap. With a minimal user calibration effort, it achieves an accuracy of $\sim 1^{\circ}$ using low-resolution eye data acquired under natural free-head scenarios.

	Method	Hardware Setup				Calib.	Accuracy (°)		Head	Further Details
		Cam(s)	Light(s)	Calib.	Resolution	Points	Rep.	Exp.	Pose	(Chin rest, eye resolution, FoV, real-time property, etc.)
Appearance	[Lu et al., 2014]	1	-	-	1280×1024	33+100	2.4	-	Fixed	Traditional regression-based method
	[Lu et al., 2015]	1	-	-	640×480	33+4	2.5	-	Free	Synthesizes images from existing calibration data
	[Zhang et al., 2015]	1	-	-	?	0	6.3	-	Free	CNN-based, trained on a large-scale dataset (MPIIGaze)
	[Wood et al., 2016b]	1	-	-	?	0	9.95	-	Free	Uses 1 million synthesized images, cross-dataset
	[Mora and Odobez, 2016]	1+Kinect	5	pre	1920×1080	0	5.7	-	Free	Mobile pose, person-independent gaze model
	[Mora and Odobez, 2016]	1+Kinect	5	pre	1920×1080	14	3.5	-	Free	Mobile pose, person-specific gaze model
	[Krafka et al., 2016]	1	-	-	?	0	~3.5	-	Free	CNN-based, trained on large-scale dataset (GazeCapture)
	[Krafka et al., 2016]	1	-	-	?	13	~2.5	-	Free	CNN-based, trained on large-scale dataset (GazeCapture)
3D Model	[Ohno and Mukawa, 2004]	2+1	1	fully	640×480	2	~1	-	Free	HR data with a pan-tilt unit
	[Hennessey et al., 2006]	1+1	3	fully	1024×768	4	~1	-	Free	HR data with a 32-mm narrow FoV lens, 15 fps
	[Guestrin and Eizenman, 2007]	2	4	fully	1280×960	1	~1	-	Free	HR data with a 35-mm narrow FoV lens, 15 fps
	[Nagamatsu et al., 2011]	4	3	fully	752×480	0	~1.6	-	Fixed	HR data with a 50-mm narrow FoV lens
	[Sun et al., 2014]	Kinect	?	pre	640×480	online	~2	-	Free	LR data with a wide FoV lens, 12 fps
	[Chen and Ji, 2015]	2	2	fully	640×480	implicit	~3	-	Free	HR data with a narrow FoV lens
Regression	[Zhu and Ji, 2007]	2	2	fully	640×480	9	~1.8	-	Free	Fully-calibrated setup, 25 fps
	[Villanueva and Cabeza, 2008]	1	2-4	-	640×480	1	~1	-	Fixed	Chin rest, HR data with a narrow FoV lens
	[Cerrolaza et al., 2008]	1	2	-	1024×768	16	~1	-	Fixed	Chin rest, HR data with a 35-mm narrow FoV lens
	[Sesma-sanchez et al., 2012]	1	2	-	1024×768	16	~1	-	Fixed	Chin rest, HR data with a 35-mm narrow FoV lens, 30 fps
	[Cerrolaza et al., 2012]	1	2	-	1024×768	16	~1	-	Fixed	Chin rest, LR data with a 16-mm lens
Cross-ratio	[Yoo et al., 2002]	1	4+1	-	640×480	0	~1.6	6.63	Free	HR data with a narrow FoV lens
	[Yoo and Chung, 2005]	1+1	4+1	-	640×480	25	~1.3	-	Free	HR data with a pan-tilt unit, 15 fps
	[Coutinho and Morimoto, 2006]	1	4+1	-	?	9	~1	-	Free	HR data with a narrow FoV lens, 30 fps
	[Kang et al., 2007]	1	4+1	-	?	9	~1	-	Fixed	Chin rest, 30 fps
	[Hansen et al., 2010]	1	4	-	800×600	4	~1	1.39	Free	LR data with a wide FoV lens
	[Coutinho and Morimoto, 2013]	1	4+1	-	640×480	9	~0.5	-	Fixed	Chin rest, HR data with a narrow FoV lens
	[Zhang and Cai, 2014]	1	8	-	1280×1024	25	~0.6	1.23	Fixed	Chin rest, LR data with a 13-mm wide FoV lens
	[Huang et al., 2014]	1	8	-	1280×1024	25	~0.8	-	Fixed	Chin rest, LR data with a 13-mm wide FoV lens
	Proposed [Arar et al., 2015b]	1	4+1	-	1280×1024	5	~1.1	1	Free	LR data with a 12-mm wide FoV lens, 30 fps
	Proposed [Arar et al., 2016a]	1	4+1	-	1280×1024	5	~1	1.01	Free	LR data with a 12-mm wide FoV lens, 30 fps

Table 4.5 – Comparison of existing eye tracking systems. "Calib." column indicates whether explicit camera and scene geometry calibrations are required prior to use: "fully" means both are required, "pre" means the sensor is pre-calibrated. In *"Accuracy"* column, *"Rep."* and *"Exp."* correspond to reported and experimented results, respectively. ″Head Pose" column indicates whether users' head pose were fixed using a chin rest or not. In ″Further Details" column, "HR" and "LR" indicate high- and low eye data resolution, respectively.

4.5 Conclusion

In this chapter, we have addressed the user calibration convenience in eye tracking. Upon carefully analyzing the previous efforts in the literature, we have proposed a novel user calibration framework that aims to ease the tedious calibration procedure. Firstly, we have identified the potential weaknesses of the state-of-the-art methods, particularly considering certain characteristics of our novel multi-camera gaze estimation framework, such as operating with low-resolution data and requiring minimal user effort. To this effect, we have carried out an extensive investigation of several regression techniques together with the widely accepted homography-based method in order to compensate for the subject-specific estimation bias. Our investigation has shown that in comparison to homography mapping, affine mapping results in a better generalization when the calibration data is limited in size and quality, owing to the reduced parameters. In addition, we have identified that the quality of the calibration data is heterogeneous due to the noise and outliers caused by various factors, such as users' viewing angles, gazing patterns, distractions, and feature detection flaws. Therefore, we have proposed novel weighted and iterative least squares regression-based methods, in which individual calibration point clusters or samples have varying impacts in the overall regression according to the determined weights, in other words, the reliability of samples/point clusters.

Both simulations and user experiments are conducted within a new evaluation scheme, where the test points are chosen independently of the calibration points, in order to avoid overfitting and increase the reliability of the results. The effectiveness of the proposed weighted regression-based calibration framework has been validated by both simulations and user experiments. The framework has been shown to outperform the state-of-the-art approaches as well as other investigated methods, especially when only a few points are used for calibration. The results on user experiments have shown that the average accuracy of the presented eye tracking framework, which requires 5 point calibration, is around 1° while allowing natural head movements.

5 Robust Eye Tracking Based on Adaptive Multi-Camera Fusion

Several works have been presented in the eye tracking literature, as comprehensively explained in Chapter 2. Among these, the main emphasis has been given mostly to the estimation accuracy improvements through introducing novel gaze estimation models (e.g., [Yoo and Chung, 2005, Guestrin and Eizenman, 2007, Zhu and Ji, 2007, Funes Mora, 2015]), and developing more effective user calibration techniques (e.g., [Kang et al., 2007, Villanueva and Cabeza, 2008, Huang et al., 2014]) to compensate for the estimation bias due to the subject-specific eye parameters, as discussed in detail in Chapter 4. As a result of such efforts, very high estimation accuracies ($<1^\circ$) have been reported, especially under controlled conditions. However, the research and validation of the robustness against real-world conditions such as head movements, varying illumination conditions, use of eye wear, and between-subject eye type variations, have been largely neglected. Thus, these remain some of the major concerns in eye tracking.

In this chapter, we emphasis on reaching high estimation accuracies while addressing the aforementioned robustness concerns. More specifically, this chapter presents the details of the adaptive fusion mechanism, which is one of the most essential processes of our multi-camera gaze estimation framework. As described in Chapter 3, differently from the previous work, we designed a multi-camera setup that allows for acquiring multiple eye appearances simultaneously from various views. Leveraging multiple appearances enables to reliably detect gaze features, even when they are obstructed in conventional single view appearance due to large head movements and occlusions caused by eye glasses. In addition, the gaze features extracted on these appearances are used for estimating multiple gaze outputs, which are then combined by an adaptive fusion mechanism to compute user's overall point of regard. Our mechanism firstly determines the estimation reliability of each gaze output according to user's general gazing behavior and momentary head pose, and then performs a reliability-based weighted fusion. We demonstrate the efficacy of our methodology with extensive simulations and user experiments on a collected dataset featuring 20 subjects. Our results show that in comparison with state-of-the-art eye trackers, the proposed methodology provides not only a significant improvement in estimation accuracy but also a notable robustness to real-world conditions, making it suitable for a wide range of applications.

In the rest of this chapter, we firstly present the related work in Section 5.1. Section 5.2 presents the proposed adaptive multi-camera fusion scheme. Evaluations on the simulated data is explained in Section 5.3, followed by the user experiments in Section 5.4. Section 5.5 discusses the acquired insights. Lastly conclusions are given in Section 5.6.

Note that the majority of the work included in this chapter has been published in the proceedings of the *Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA)* [Arar and Thiran, 2016] and submitted for a journal publication [Arar and Thiran, 2017].

5.1 Related Work

Remote video-oculography, in which users' eyes are non-intrusively captured by remote cameras, is the focus of this thesis since it provides the most natural and convenient user interaction. As described in detail in Chapter 2, among remote gaze estimation methods, *feature-based methods* extract local features from an eye image to determine the gaze. They mostly require particular hardware configuration, and leverage the local features that are formally related to the gaze points through the geometry of the system and eye physiology. Since they provide high accuracy and easy feature detection, they have become the most popular approach for gaze estimation. On the other hand, *appearance-based methods* use the image content as the input, and map the image features to the gaze points. Their system and hardware requirements tend to be simpler than those of *feature-based methods*. However, they are restricted to particular applications due to their limitations in estimation accuracy and tracking robustness.

In this thesis, we are primarily interested in *feature-based methods* as they provide in overall a better tracking performance for the targeted scenarios. Each of three subcategories of *feature-based methods*, i.e., *3D model-based*, *regression-based*, and *cross ratio-based*, has its own advantages and disadvantages in terms of system complexity, accuracy, and robustness to real-world conditions. *3D model-based* methods [Beymer and Flickner, 2003, Hennessey et al., 2006, Guestrin and Eizenman, 2007, Park, 2007, Sun et al., 2015] compute the gaze by explicitly modeling the eye in 3D. They implicitly provide the highest accuracy and head movement tolerance among *feature-based methods*. However, they require complex system setups such as particular hardware (a stereo setup or depth sensor), camera calibration, and geometric system calibration. On the other hand, *regression-based* methods [Zhu et al., 2006, Zhu and Ji, 2007, Cerrolaza et al., 2008, Sesma-sanchez et al., 2012] assume a direct mapping from the eye features to the gaze points, and *cross ratio-based* methods [Yoo and Chung, 2005, Hansen et al., 2010, Coutinho and Morimoto, 2013, Zhang and Cai, 2014, Huang et al., 2014] compute the gaze point by leveraging the cross-ratio property of the projective space. Both methods' setup complexities are much lower than those of *3D model-based* methods such that they do not require any camera or geometry calibration while allowing for certain head movement tolerance. Unfortunately, their tracking performances are worse in accuracy and head movement robustness due to the simplifications assumed. A detailed comparison of remote gaze estimation methods is presented in Table 5.1. The following subsections explain the related work on each of these

Table 5.1 – Comparison of gaze estimation methods regarding the setup complexity, hardware requirements and calibration, user calibration, estimation accuracy, and implicit tracking robustness to head movements, illumination variations, and eye wear.

	Feature-based			Appearance-based
	3D Model	**Regression**	**Cross Ratio**	
Setup Complexity	High	Medium	Medium	Low
System Calibration	Fully-calibrated	-	-	-
Hardware Requirements:				
- Cameras	2+ Infrared (stereo)	1+ Infrared	1+ Infrared	1+
- Lights	2+ Infrared	2+ Infrared	4+ Infrared	-
User Calibration	Required	Critical	Required	Optional
Gaze Accuracy Error	$< 1^\circ$	$\sim 1 - 2^\circ$	$\sim 1 - 2^\circ$	$> 2^\circ$
Implicit Robustness:				
- Head Movements	Medium-High	Low-Medium	Low-Medium	Low
- Illumination Variations	High	High	High	Low
- Eye Wear	Low	Low	Low	Medium

desired attributes in eye tracking.

5.1.1 Gaze Estimation Accuracy & Setup Complexity

As previously mentioned, the majority of the existing work focus on improving the gaze estimation accuracy. There is no doubt that the accuracy is directly proportional to the setup complexity. *3D model-based* methods (e.g., [Beymer and Flickner, 2003, Guestrin and Eizenman, 2007, Park, 2007, Hennessey et al., 2006, Lai et al., 2015]) are widely preferred as they provide high accuracy under generic head movements, owing to their explicit and fine modeling of the eye in 3D. In fact, most commercial eye tracking solutions rely on *3D model-based* methods. However, they have a significant disadvantage such that they require a fully-calibrated system setup. More specifically, for accurately modeling the eyeball in 3D, they require a complex setup, such as a stereo system or a depth sensor, which requires camera and geometric scene calibrations. Alternatively, *cross ratio-based* methods (e.g., [Yoo and Chung, 2005, Hansen et al., 2010, Coutinho and Morimoto, 2013, Huang et al., 2014, Arar and Thiran, 2016]) and *regression-based* methods (e.g., [Cerrolaza et al., 2012, Sesma-sanchez et al., 2012]) have in general lower setup complexity and often require an uncalibrated setup. Nevertheless, the drawback of both methods is that they rely on approximated models. Consequently, the performance of these methods are lower in accuracy and head movement robustness. On the contrary to aforementioned *feature-based* methods, *appearance-based* methods (e.g., [Lu et al., 2015, Zhang et al., 2015, Krafka et al., 2016, Wood et al., 2016b]) simply require an ordinary camera.

Furthermore, most of the existing eye trackers, which rely on either *feature-based* or *appearance-based* methods, use a single-camera setup. The ones with multi-camera setups rely on *feature-based*, particularly *3D model-based*, methods. These setups are mostly designed to acquire 3D stereo vision [Beymer and Flickner, 2003, Guestrin and Eizenman, 2006, Zhu and Ji, 2007, Lai

et al., 2015]. Therefore, such systems still rely on eye appearances obtained from a single view. On the other hand, the effectiveness of multi-camera setups that utilize multi-view appearances has not adequately been investigated. In this regard, very limited efforts have been made, e.g., [Utsumi et al., 2012] proposed a two-camera setup mainly to obtain a wide observation area for a gaze-reactive signboard to enable a wide range of user motions while allowing for coarse gaze estimation with an accuracy of ~11°. To the best of our knowledge, we are the first to exploit a multi-camera setup to improve the estimation accuracy for precise gaze estimation [Arar et al., 2015a]. Later, we extend our previous work to also improve the tracking robustness under challenging real-world conditions through exploring more efficient multi-camera setups and fusion mechanisms [Arar and Thiran, 2017].

5.1.2 User Calibration

In addition to hardware setup calibration, user calibration has an important role in user experience and convenience. User calibration is required for modelling the subject-specific parameters, which are crucial for the estimation bias correction, especially for *feature-based* methods. The calibration quality improves, to a certain extent, when the amount of calibration data increases. However, augmenting the data amount by increasing the number of calibration points could be tedious and thus harms the user experience. In this regard, the trade-off between the quality and convenience of the user calibration has been widely studied in the literature. Significant advancements have been made. For instance, better geometric eye models [Guestrin and Eizenman, 2006, Villanueva and Cabeza, 2008, Nagamatsu et al., 2011] and more robust bias correction models [Coutinho and Morimoto, 2006, Hansen et al., 2010, Zhang and Cai, 2014, Arar et al., 2015b, Arar et al., 2016a] were developed, and implicit calibration strategies were introduced [Sun et al., 2014, Chen and Ji, 2015]. A more detailed literature review can be found in Section 4.1.

5.1.3 Head Movement Robustness

3D-model based methods are theoretically more tolerant to the changes in head pose and location due to explicit parametrization of individual-specific eye parameters. Yet, in practice, they suffer from inaccuracy under large head movement scenarios. One of the main reasons is that most systems are faced with the dilemma of trading off between the head movement range and eye data resolution. In early efforts, e.g., [Beymer and Flickner, 2003, Ohno and Mukawa, 2004], a wide field-of-view (FoV) stereo system was employed to allow free head movement as well as a narrow FoV stereo system to capture eye images with high resolution. These systems were mostly interconnected through a pan-tilt unit which mechanically reoriented the narrow FoV camera to the users' eye according to the feedback of the wide FoV camera system. [Park, 2007] also addressed to handle head movements using a pan-tilt unit. He proposed a three-camera eye tracking system, i.e., one wide FoV camera and two narrow FoV cameras with auto-zoom and auto-focus capabilities, in which the gaze direction was estimated by the 3D pose of the pupil

using the narrow FoV stereo system. The system achieved an accuracy of ∼1° while enabling ±10 cm frontal and backward head movements with respect to the camera. Despite enabling high accuracy and robustness, the use of a pan-tilt unit increased the setup complexity and the cost. Consequently, in later efforts, researchers avoided to employ such mechanical units and focused on introducing models that were head movement robust even when having low resolution eye data. With the help of advancements in camera technology as well as more robust estimation models, the need for the narrow FoV cameras was eliminated. For instance, [Guestrin and Eizenman, 2007] introduced a method that used the centers of the pupil and at least two glints, which were estimated from the eye images captured by at least two cameras. Their system achieved <1° accuracy error by tolerating head movements in a volume of 10×8×10 cm^3[1]. Moreover, [Hennessey et al., 2006] presented a single camera non-stereo system that employed the use of ray tracing rather than depth from focus. Their system allowed an accurate (<1°) gaze estimation in a volume of 14×12×20 cm^3. Recently, [Sun et al., 2015] proposed a Kinect sensor-based technique that could handle low resolution eye data. Their system used a parametrized iris model to localize the iris center for gaze feature extraction. Thereby, the gaze direction was determined based on a 3D geometric eye model by computing the 3D position of the eyeball center and iris center. They reported 1.4−2.7° accuracy error under head movements in a volume of 20×20×8 cm^3.

Contrary to *3D model-based methods*, *regression-based methods* can be considered as approximation methods since they indirectly model the eye physiology, geometry, and optical properties. In this regard, their head movement tolerance is implicitly lower than *3D model-based methods*. The reason is that when the user moves away from the calibration position, the features non-linearly change, therefore, the calibration mapping becomes less accurate and the estimation accuracy degrades. Consequently, one of the main challenges in *regression-based* gaze estimation is to learn a head movement invariant method. In order to address this challenge, multiple glints based approaches have been suggested. First of all, [White et al., 1993] proposed to use a second light source, which permitted differentiation of head movement from eye rotation in the camera image. Using two glints as points of reference and exploiting spatial symmetries, they proposed a spatially dynamic calibration method to compensate for lateral head translation automatically. Later, a thorough review of polynomial-based regression methods using two glints was presented in [Cerrolaza et al., 2008]. They evaluated various models using different pupil-glint vectors and polynomial functions. In addition, [Sesma-sanchez et al., 2012] studied how binocular information can improve the accuracy and robustness against head movements for the polynomial based systems using one or two glints. Moreover, [Cerrolaza et al., 2012] demonstrated that the pattern of error caused by the head movements mainly depends on the system and hardware configuration rather than the user. They suggested two calibration strategies to reduce the errors caused by head movements. The results of the experiments showed that both strategies achieved a reduction in error by a factor of two when the user's head was moved ±6 cm (depth) from the calibration position. Despite achieving promising results, most of the above efforts required to fix the users' head using a chin rest. Therefore, it is difficult to determine the efficacy of the proposed methods

[1] horizontal×vertical×depth movements. Note that all the following volume ($\cdot \times \cdot \times \cdot$) measures also refer to this.

under free-head conditions. Differently from the majority of the *regression-based methods*, [Zhu and Ji, 2007] proposed a stereo vision-based system, which achieved an acceptable accuracy, ~2° while allowing for larger head movements without requiring the use of a chin rest. Their system tolerated for the head movement in a volume of 20×20×30 cm^3. They estimated the optical axis of the user's eye in 3D by directly applying triangulation techniques on the glints and pupil center. They also suggested that 3D head pose information can be used to compensate for the bias caused by head movements. However, the main drawback of this system is that a multi-camera fully-calibrated stereo setup was required to obtain 3D information.

There also have been various attempts to enhance the head movement tolerance of *cross ratio-based* methods. Most of these provided solutions by adapting the user calibration to the changes in head movements. For instance, [Coutinho and Morimoto, 2013] described two subject-specific calibration methods for improving the robustness against head movements. The first method accounted for the vertical head movement robustness through a dynamic calibration correction, whereas the second one implicitly handled both horizontal and vertical head movements since eye rotation was handled by the planarization of the gaze features. Although about 0.5° accuracy error was reported while tolerating 25×25 cm^2 head location changes, their system required high-resolution eye images (640×480 pixels) captured with a zoomed lens. Also, a chin rest was required to keep the users' eye within the FoV of the camera and to fix the users' head pose and location during the experiments. Therefore, their system's actual performance under free-head conditions may differ from the reported performance. Alternatively, [Zhang and Cai, 2014] suggested to use a homography-based calibration modeling with a binocular fixation constraint to jointly estimate the homography matrix from both eyes. Even though a small decrease in accuracy was experienced while allowing for 20×10 cm head movements, the overall estimation accuracy error of ~0.4-0.6° showed the efficacy of their method. One potential drawback of their system is that the features from both eyes must be detected to compute a gaze output, which constrains the estimation availability due to the limited head pose allowance. Moreover, [Huang et al., 2014] proposed an adaptive homography calibration. The authors learned an offline-trained model on the simulated data by exploring the relationship between the estimation bias and varying head movements. The promising experimental results achieved both on the simulated data with depth and vertical head movements up to ±25 cm and on real data with ±10 cm depth movements indicated the efficacy of the method. Nevertheless, an important limitation in [Zhang and Cai, 2014] and [Huang et al., 2014] is that they utilize a chin rest to keep the head pose fixed during their evaluation, similar to [Coutinho and Morimoto, 2013]. Using a chin re causes the evaluations to discard the impact of variations in head pose. Besides, such restrictions significantly harm the user experience and would be impractical for real-world human-computer interaction (HCI) applications. Although reporting performances using a chin rest may lead to more stable results, it causes the evaluations to discard the impact of variations in head pose. In addition, it significantly harms the user experience and would be impractical for real-world HCI applications. Therefore, it is completely avoided in our this thesis. Instead, our methodology operates with lower resolution eye data captured using small focal length lenses in order to allow

[2]horizontal×depth movements. Please note that all the following ($\cdot \times \cdot$) measures also refer to this.

for both head translations and rotations. Lower resolution data, as expected, results in a lower accuracy for individual camera-systems, yet the overall accuracy of the system is still high, owing to our adaptive fusion of the gaze outputs obtained from multiple sensors.

5.1.4 Eye Glasses Robustness

Considering that about 30 percent of young adults and more than half of elders in industrial nations need to wear eyeglasses [Morgan and Rose, 2005, Schaeffel, 2006], any intolerance to eye wear, such as eye glasses and contact lenses, significantly harms eye trackers. Eye wear robustness, especially to eye glasses, has been a challenging research problem. Especially for *feature-based* methods, the reflection and refraction from eye glasses drastically obstruct the feature detection, and therefore, cause a significant inaccuracy. Since *appearance-based* methods [Hansen and Pece, 2005, Zhang et al., 2015, Wood et al., 2016b, Krafka et al., 2016] neither rely on active illumination nor detection of individual gaze features, their performances are less affected by eye glasses. On the other hand, for *feature-based methods*, an explicit solution must be employed to address this issue. In this regard, there exists only a limited number of previous studies. [Ebisawa, 1998] introduced a robust pupil detection method by leveraging the bright-pupil effect generated with a differential lighting scheme. He also suggested a method for eliminating the reflections appearing on the glasses. His method was successfully realized by [Ji and Yang, 2002] for monitoring driver vigilance to robustly track the pupil and to roughly estimate the gaze. In addition, [Park, 2007] proposed a dual illumination technique to avoid the reflections on the glasses. In his system, when a specular reflection on the glass was detected, the system deactivated the current illuminator and activated the alternative illuminator on the opposite side, such that the reflection can be avoided.

Moreover, few other efforts [Guestrin and Eizenman, 2006, Villanueva and Cabeza, 2007] proposed more detailed models to compensate for the impact of the refraction by the glasses and cornea. They demonstrated that the gaze accuracy may differ more than 1° depending on whether refraction is accounted for. Also, [Xiong et al., 2014] proposed a gaze estimation method based on 3D face structure and pupil center without requiring any glints. Despite the fact that their system sacrificed the accuracy ($< 4^\circ$), they improved the eye glasses tolerance, owing to the elimination of the active light sources. Recently, an important effort has been made by [Kübler et al., 2016]. They proposed a method for rendering of eye glasses in synthetic eye tracking images. Their method allowed for studying the effects of refraction and reflection in connection with pupil and glint detection. From a different perspective, our work addresses the eye glasses robustness by focusing on detecting more reliable gaze features. As the occlusion and distortion of gaze features on an eye appearance depends on the relative positioning of a camera, light source, and user eye, we propose to utilize multiple cameras generating alternative eye appearances. Consequently, the benefit of our approach is that in case the gaze features are obstructed by the eye glasses effects from certain views, they can still be recovered from alternative appearances.

5.1.5 Illumination Robustness

In majority of the eye tracking applications, variations in environmental illumination (e.g., sun lighting, indoor fluorescent lighting or darkness) is inevitable. In this regard, the eye appearance in natural light-based eye trackers significantly differs when the illumination alters. As *appearance-based* methods rely on natural light, they are highly sensitive to ambient illumination changes. To address this issue, large amounts of training data under different illumination conditions must be collected, as performed by some recent studies [Zhang et al., 2015, Wood et al., 2016b, Krafka et al., 2016]. On the other hand, a simple yet effective solution is to use active near-infrared (NIR) illumination, as employed by the majority of *feature-based methods*. Under active lighting, the eye appearance remains similar while the ambient illumination alters. Yet, the feature detection can still be very challenging as the gaze features and eye appearance can be affected by illumination variations. For instance, the pupil size and opening of eyes change to adjust the amount of light entering the eye. Consequently, the features may appear more weakly, especially the pupil brightness may vary in bright-pupil settings. In this respect, feature detection mechanisms must adapt to handle such changes. To the best of our knowledge, no previous work explicitly demonstrated the illumination robustness. In this thesis, we address the effects of illumination variations on eye and gaze features' appearance and discuss their influence on the tracking performance.

5.2 Adaptive Multi-Camera Fusion

The proposed multi-camera gaze estimation framework comprises of synchronized individual single-camera systems. In this framework, each single-camera system has the ability to simultaneously track both eyes. Upon acquiring the data from all cameras, the framework proceeds with the gaze feature detection in all eye appearances. It firstly localizes the eye region. Then, it determines whether there is an eye blink or not. If there is an eye blink, no gaze output is generated. Otherwise, it checks whether there is any noisy glare, e.g., specular reflections from eye glasses, on the eye appearances, and removes them as a preprocessing to ease the actual gaze feature detection. Glint detection is performed followed by pupil detection, which relies on dark-pupil approach. Once the gaze features are obtained, cross ratio-based gaze estimation is performed to obtain an initial point of regard (PoR) estimate. Note that the details of these processes together with illustrations are presented in Chapter 3. A user calibration is then employed on the initial PoRs in order to compensate for the subject-specific estimation bias. The details of the user calibration process can be found in Chapter 4. Consequently, each camera view computes two independent PoRs, in other words, two gaze sensors, in each frame. Therefore, in a multi-camera setup with C cameras, the system is able to generate a total of 2C PoRs per frame. The overall PoR can then be computed by the fusion of available PoRs obtained from all sensors, as illustrated in Figure 5.1. Here, the adaptive fusion mechanism aims to find the most effective combination of available PoRs towards achieving higher estimation accuracy and robustness.

Throughout this thesis, several algorithms have been investigated to perform the adaptive fusion.

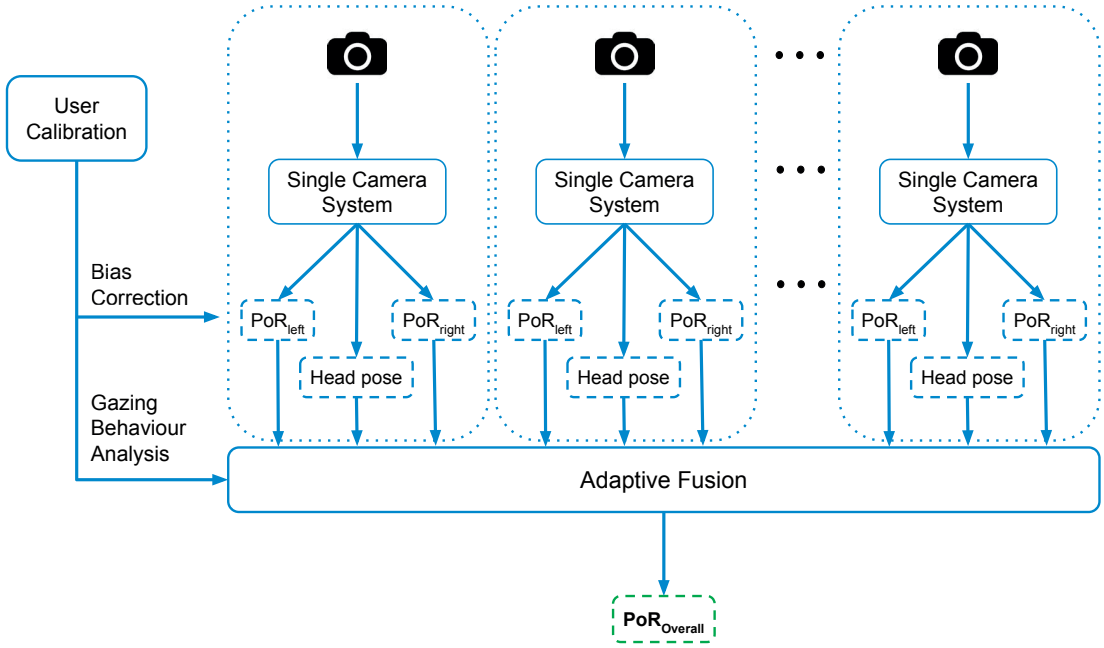


Figure 5.1 – Overview of the proposed adaptive multi-camera fusion.

Among these, the most straightforward one relies on a simple averaging. Despite its simplicity, fusion by simple averaging resulted in a significant improvement in estimation accuracy and precision in comparison with a single-camera system, especially under controlled conditions, in which the majority sensors produce reliable PoRs. In such cases, it provides a more accurate and consistent overall estimation through smoothing out the arbitrary noise. On the other hand, under more challenging scenarios, in which a higher variance exists among available PoRs, fusion by simple averaging is far from an optimal solution. To this effect, we observe that the estimation reliability of each sensor varies according to several factors, such as the viewing angles of the cameras, location of the gazed point on the monitor, eye glasses effects, person-specific gaze behaviours. Figure 5.2 illustrates sample eye appearances captured from different camera views when users gaze at different regions on the monitor. The figure clearly shows the varying quality in captured gaze features. While some views, e.g., bottom camera view in Figure 5.2a, enable to detect features reliably, some other views, e.g., left camera view in Figure 5.2c, do not even contain any available gaze features. The views capturing the best eye appearances continuously vary when users gaze at different regions. In addition, when users wear eye glasses, some reflection and refraction effects may occur on the glasses. When these effects distort or overlap with the gaze features, the estimation becomes impossible from that particular view. On the other hand, as there exists several other simultaneously captured views, the gaze features can still be recovered from these views. This, in fact, constitutes an important benefit of our multi-view approach in comparison to a single-view approach, employed by the majority of the previous work. Thus, we claim that an effective fusion, which accounts for the estimation reliability of individual PoRs, can significantly improve the overall estimation accuracy and robustness. To

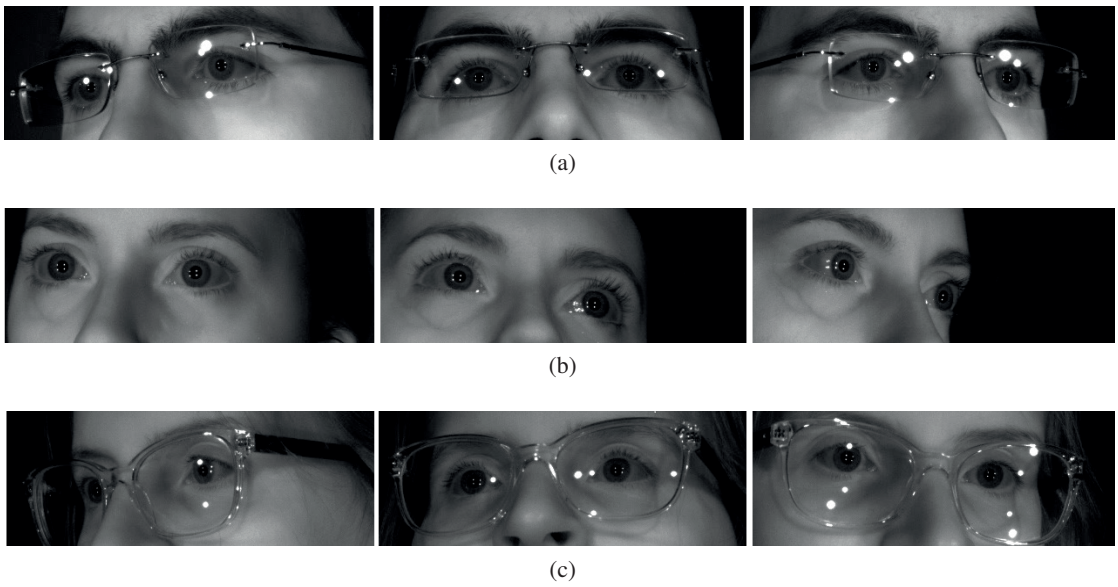


Figure 5.2 – Simultaneously captured eye appearances from three camera views: (left column) left side camera, (middle column) bottom camera, and (right column) right side camera. Each row shows a user gazing at a target stimulus point displayed on (a) central, (b) leftward, and (c) rightward region of the monitor.

this end, we propose to combine the available gaze outputs in a weighted manner, in which the weights correspond to the reliability of individual gaze outputs, as follows:

$$\mathbf{z}^* = \sum_c \sum_e \mathbf{z}_\mathbf{c}^\mathbf{e} w_c^e \tag{5.1}$$

$$\sum_c \sum_e w_c^e = 1, \quad e \in \{Left, Right\}, \quad c \in \{1, 2, .., C\},$$

where $\mathbf{z}^*$ is the overall PoR and, w_c^R and w_c^L are the weights for the right and left eye's PoRs from the c^{th} camera, respectively. In case one of the PoRs can not be calculated for a given frame, then the weight of the missing PoR is set to zero. We do not report an overall PoR in case both PoRs of all the cameras are unavailable for a given frame. It is important to note that the proposed adaptive fusion framework and algorithms are independent of the gaze estimation algorithm used. Thus, the *cross ratio-based* method used in this thesis can practically be replaced with any other previously mentioned gaze estimation methods.

In this thesis, to determine the reliability of individual gaze outputs, in other words, the weights of the PoRs, we mainly focus on two approaches, namely, head pose-based fusion and subject-specific gazing behaviour-based fusion.

5.2.1 Head Pose-Based Fusion

We observe in our experiments that the estimation accuracy achieved by each single-camera system is correlated with the locations of the target gaze points. The estimations obtained by a camera are often more accurate and reliable when the users gaze at the regions of the monitor that are closer to the camera location. For instance, when users gaze at the upper left corner of the monitor, the left side camera system generates more accurate estimations than the others. In fact, the main reason relates to the feature detection reliability with respect to user's viewing angle. We observe that prior to the fixation, most users first perform a head rotation to find the most comfortable viewing angle for the particular target gaze point. Therefore, the eye appearances change relative to the user's head pose. Since the frontal head pose (relatively to the camera) results in a more reliable feature detection, the corresponding single-camera systems output more reliable estimations.

We suggest two algorithms to determine the weights for head pose-based adaptive fusion.

Camera distance-based weighting

This method performs a camera distance-based PoR re-weighting. We firstly estimate an initial PoR, $\mathbf{z}'$, using a simple averaging of available PoRs. Then, we iteratively refine the initial estimation according to its proximity to each camera. Since the relative head pose angles increase directly proportional to the distances to each camera, we assign weights, w_c^e, inversely proportional to the distances as follows:

$$\lambda_c^e = \|\mathbf{z}' - \ell_{\mathbf{c}}\| \tag{5.2}$$

$$w_c^e = 1 - \frac{\lambda_c^e}{\sum_c \sum_e \lambda_c^e}, \tag{5.3}$$

where $\mathbf{z}'$ is the initially estimated PoR, and $\ell_{\mathbf{c}}$ is the location of the c^{th} camera. λ_c^e denotes unnormalized weights, which are later normalized using Equation (5.3). Once the initial PoR and weights are calculated, we iteratively compute the overall PoR, $\mathbf{z}^*$, using the refined weights until convergence, which often takes a few iterations. The algorithm reaches convergence when no significant change is observed in consecutively updated PoRs, i.e., $\|\mathbf{z}^* - \mathbf{z}^*_{\mathbf{old}}\| < \tau = 5$ pixels, as summarized in Algorithm 1.

Algorithm 1 Camera Distance-Based Multi-Camera Fusion

Input: $\mathbf{z_c^e}, \ell_\mathbf{c}$
if $\mathbf{z_c^e} \neq null$ **then** ▷ For any available $\mathbf{z_c^e}$
  $\lambda_c^e \leftarrow 1$ ▷ Initialize weights equally
else
  $\lambda_c^e \leftarrow 0$
end if
$w_c^e \leftarrow \frac{\lambda_c^e}{\sum_c \sum_e \lambda_c^e}$ ▷ Normalize weights
$\mathbf{z}' \leftarrow \sum_c \sum_e \mathbf{z_c^e} * w_c^e$ ▷ Get initial PoR using (5.1)
$\mathbf{z}^* \leftarrow \mathbf{z}'$
repeat
  $\mathbf{z}^*_\mathbf{old} \leftarrow \mathbf{z}^*$
  $\lambda_c^e \leftarrow \|\mathbf{z}^* - \ell_\mathbf{c}\|$ ▷ Re-weight using (5.2)
  $w_c^e \leftarrow 1 - \frac{\lambda_c^e}{\sum_c \sum_e \lambda_c^e}$ ▷ Normalize weights using (5.3)
  $\mathbf{z}^* \leftarrow \sum_c \sum_e \mathbf{z_c^e} * w_c^e$ ▷ Update the PoR using (5.1)
until $\|\mathbf{z}^* - \mathbf{z}^*_\mathbf{old}\| < \tau$
return $\mathbf{z}^*$ ▷ Return the overall PoR

Head pose angle-based weighting

Although the camera distance-based weighting performs an effective adaptive fusion, it has two limitations. Firstly, it requires the camera locations to be known to calculate the distances. Secondly, its performance is affected by the quality of the initial estimation. When the initial estimation is poor due to the outliers, which can occur especially under challenging tracking conditions, it may fail to perform an effective weighting. To address these, we suggest an alternative method which relies on users' head pose angles obtained with respect to the cameras. Once the head pose angles are estimated with respect to each camera view, the weights are assigned inversely proportional to the relative head pose angles calculated with respect to each camera view, as follows:

$$\lambda_c^e = \frac{\alpha_{max} - |\alpha_c^e|}{\alpha_{max}}, \tag{5.4}$$

$$w_c^e = \frac{\lambda_c^e}{\sum_c \sum_e \lambda_c^e}, \tag{5.5}$$

where α_c is the head pose yaw angle, α_{max} is the maximum angle allowed, e.g., 45°. We calculate the head pose angles using the method described in [Chen et al., 2012] from the landmarks obtained by the face tracker. The calculated head pose angles are applied to both eyes, and so, the same weight is assigned to both. The weights are then normalized using Equation (5.5) prior to the fusion.

5.2.2 Gazing Behaviour-Based fusion

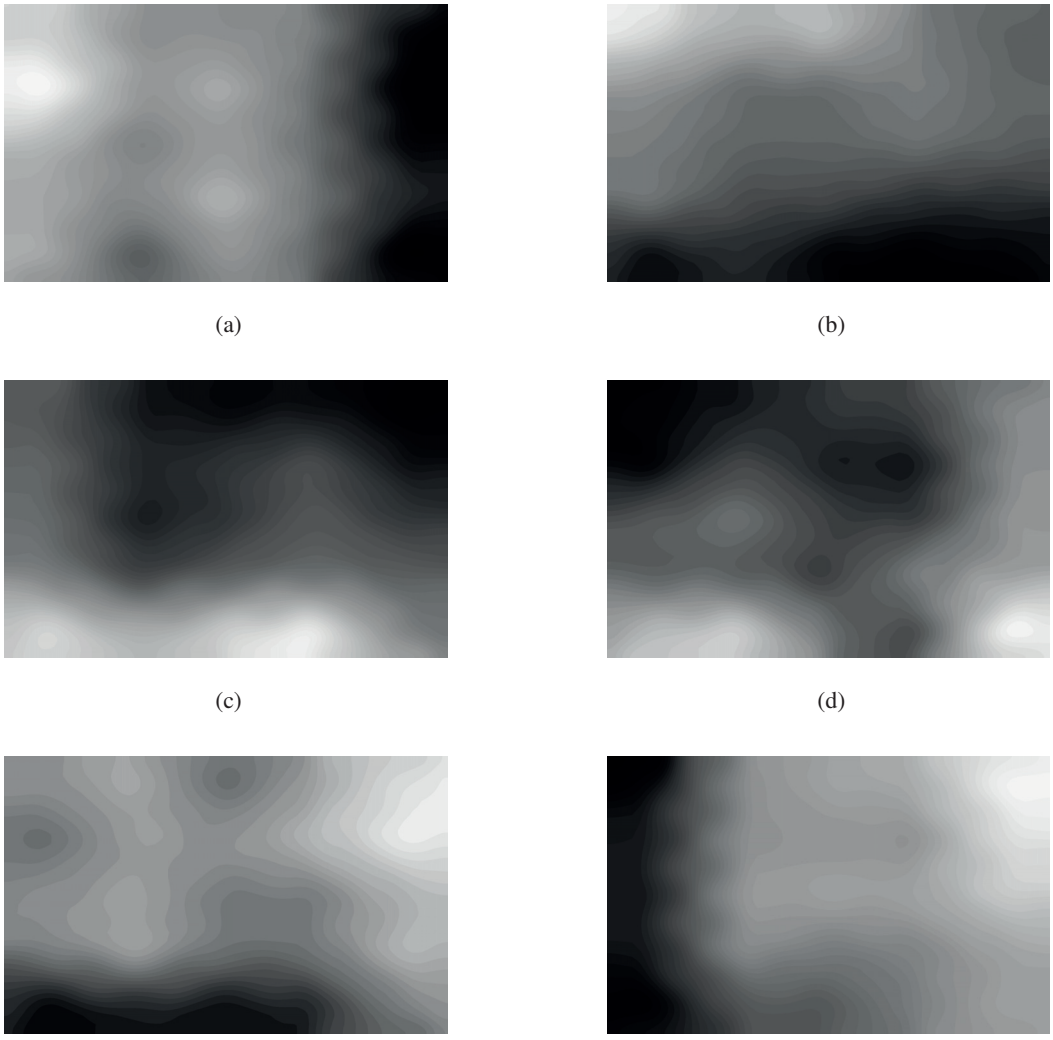
There is a wide variety of gazing behaviours among the users, due to personal preferences and habits, or due to physiological reasons (lazy eye, strabismus, etc.). Although the head pose-based weighting approach works well for the majority of the users, it does not account for the subject-specific gazing behaviours, and consequently, it may experience a performance drop when a user has a specific gazing behaviour. For instance, although the majority of the users, prior to fixation, perform head rotation to have a comfortable viewing angle (frontal eye ball pose), some users do not perform any head movements but rather rotate their eye balls (non-frontal eye ball pose). In addition, some users' vision, e.g., the ones with vision disorders, may be depended on the dominant eye. For these users, assigning equal weights to both eyes may result in a low estimation performance. Therefore, we propose an alternative weighting method, which leverages user calibration data to capture the subject-specific gazing behaviours. During the user calibration, we generate fusion weight maps in addition to learning the subject-specific bias correction model. Once the weight maps are obtained, the proposed method performs a weighted averaging of individual PoRs as follows:

$$\mathbf{z}^* = \sum_c \sum_e \mathbf{z}_\mathbf{c}^\mathbf{e}\, \mathbf{M}_\mathbf{c}^\mathbf{e}(\mathbf{z}_\mathbf{c}^\mathbf{e}.x, \mathbf{z}_\mathbf{c}^\mathbf{e}.y), \tag{5.6}$$

$$\sum_c \sum_e \mathbf{M}_\mathbf{c}^\mathbf{e}(x, y) = 1, \quad e \in \{Left, Right\}, \quad c \in \{1, 2, .., C\},$$

where $\mathbf{z}^*$ is the overall PoR, $\mathbf{z}_\mathbf{c}^\mathbf{e}$ are initial PoRs estimated using simple averaging, and $\mathbf{M}_\mathbf{c}^\mathbf{R}$ and $\mathbf{M}_\mathbf{c}^\mathbf{L}$ are the weight maps of the right and left eye of the c^{th} camera, respectively.

For generating the weight maps ($\mathbf{M}_\mathbf{c}^\mathbf{e}$), various statistics extracted from the calibration data can be leveraged. For instance, the most relevant and effective weighting indicator would be the calibration accuracy per sensor on each calibration point. The reason is that if the calibration accuracy on a point is consistently lower for a sensor than the others, that sensor's bias correction during testing is expected to be less reliable and less accurate around the same point. Hence, the calibration accuracy based weighting assigns higher weights to the sensors whose bias corrections are more reliable, so that a more accurate overall PoR can be computed. To calculate the calibration accuracy ($acc^e_{c,k}$) for each point, after learning the calibration model on the whole calibration data, we apply the learned model on the same data. Then, we measure how close the calibrated samples are to their corresponding target points. We perform this process for each calibration point of each sensor separately. As we perform calibration for each eye of each camera independently, we obtain 2C values for each calibration point. We then normalize these accuracy values to compute the sensor weights ($w^e_{c,k}$) for each calibration point as shown in Equation (5.7). Lastly, we interpolate and extrapolate the weight set ($\mathbf{W}_\mathbf{c}^\mathbf{e}$) over the whole monitor to generate the weight maps ($\mathbf{M}_\mathbf{c}^\mathbf{e}$). A set of generated weight maps is shown in Figure 5.3.



(a) (b) (c) (d) (e) (f)

Figure 5.3 – Sample generated weight maps based on the calibration accuracy and gaze availability statistics for (a) $\mathbf{M_L^R}$: right eye of left camera, (b) $\mathbf{M_L^L}$: left eye of left camera, (c) $\mathbf{M_B^R}$: right eye of bottom camera, (d) $\mathbf{M_B^L}$: left eye of bottom camera, (e) $\mathbf{M_R^R}$: right eye of right camera, (d) $\mathbf{M_R^L}$: left eye of right camera.

$$w^e_{c,k} = \frac{acc^e_{c,k}}{\sum_c \sum_e \sum_k acc^e_{c,k}}, \tag{5.7}$$

$$\mathbf{W_c^e} = \{w^e_{c,k} | e \in \{Left, Right\}, c \in \{1, 2, .., C\}, 1 \leq k \leq K\}, \tag{5.8}$$

where K is the number of calibration points.

In addition to the estimation accuracy, we use the estimation availability statistics[3] on each calibration point as a separate weighting indicator since the availability correlates to the reliability of feature detection. A low availability implies less consistent and less reliable features. Hence, a sensor with a higher availability is more likely to produce reliable PoRs. Although our current weight maps are generated using these two weighting indicators, other alternative weighting indicators can also be employed towards better modeling of users' gazing behaviours. In this context, we have also investigated the estimation precision, which is the ability to reliably reproduce consistent estimations for a target calibration point, or the histogram of the best performing sensor, indicating how often each sensor achieves the best estimation. Although these can provide complementary evidence in some cases, further evaluations are required to claim about their efficacy. We leave these as the future work.

5.3 Evaluation on Simulated Data

In our evaluations on simulated data, we primarily examine how the proposed multi-camera framework's performance is affected when the number of cameras is increased in various configurations. In a real-world setting, the number of cameras to be employed for a real-time eye tracking is limited due to the physical constraints, e.g., hardware limitations, data band-width, cost etc., unless a particular hardware optimization is performed. Therefore, we start our evaluations on the simulated data in order to analyze the efficacy and limits of the proposed multi-camera approach. In this manner, we conduct various experiments using an open-source simulator.

In our evaluations, the tracking performances are measured as the gaze estimation accuracy error, which is defined as the average displacement in degrees of visual angle (°) between the target stimuli points and the estimated PoRs, using all raw samples[4] as follows:

$$Error_{pixel} = \frac{\sum_i^N \sum_j^K \|\mathbf{P}_i - \mathcal{F}(\mathbf{z}^*_{i,j})\|}{NK} \tag{5.9}$$

$$Error_{mm} = \frac{Error_{pixel}}{\textit{pixel-to-mm ratio of monitor}} \tag{5.10}$$

$$Error_{\text{visual angle (°)}} = \frac{Error_{mm}}{\textit{user-to-monitor distance}} \frac{180}{\pi} \tag{5.11}$$

[3]The estimation availability is defined as the percentage of samples, on which the system is able to compute an overall PoR during a calibration or a test session. It is computed separately for each calibration point.

[4]Neither temporal smoothing nor post-processing is applied in our evaluations in order to demonstrate the real impact of our framework and methods. We employ temporal smoothing only in our real-time demonstration, which leads to a smoother tracking experience for the users.

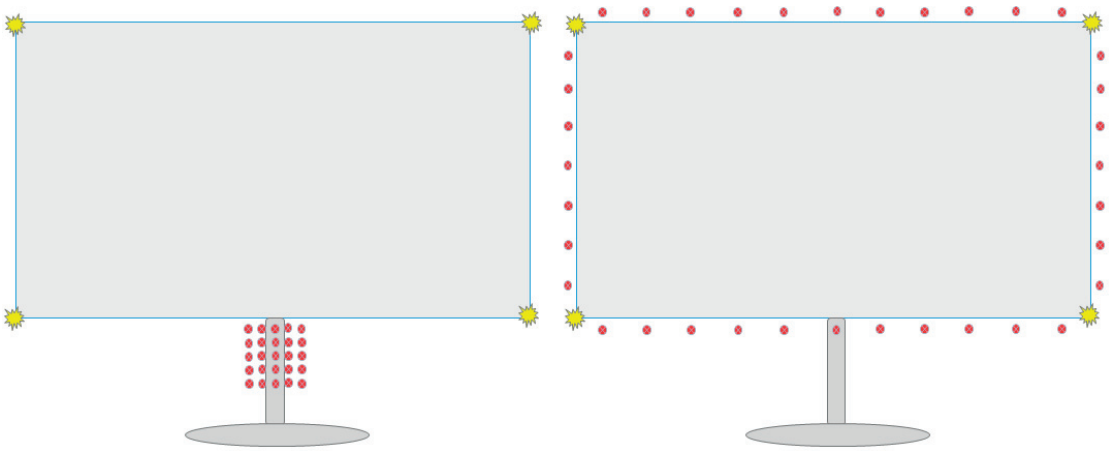


Figure 5.4 – Simulation of increased number of cameras by placing them (left) at the bottom of the monitor (case 0), and (right) uniformly around the monitor (case 1).

where $\mathbf{P}_i$ and $\mathbf{z}^*_{i,j}$ denote the i^{th} target stimuli point and the estimated raw PoR of the j^{th} sample for the corresponding target point, respectively. User calibration model is denoted by $\mathcal{F}$. Total number of target points is denoted by N, and K samples (frames) are acquired per point during a test session. *pixel-to-mm ratio* is obtained from the monitor specifications[5] We note that the estimation errors are reported in degrees of visual angle (°) since it is invariant to *user-to-monitor distance*. In addition, the estimation availability is defined as the percentage of samples, which the system is able to compute an overall PoR during the whole evaluation session. Following subsections explain the details of the conducted simulations and obtained results.

5.3.1 Simulation Setup

Simulation data was generated using an open-source software framework developed by [Böhme et al., 2008]. The simulator allows for detailed modeling of different components of the hardware setup and user eye in 3D. It provides in overall a realistic simulation framework. On the other hand, there are few factors of interest that are not currently simulated in the software such as the non-spherical shape of the cornea, occlusion of the eye by the eyelids, the effects of glasses and contact lenses, lack of spatial extent of the light sources, lens distortions or other camera sensor imperfections. Despite these limitations, the simulator offers one of the best solutions that is publicly available[6].

We started our evaluations by simulating the impact of increasing the number of cameras, which are either placed at the bottom of a monitor (case 0), or placed around a monitor (case 1), as visualized in Figure 5.4. Firstly, a realistic simulation environment is created by considering the

[5] In our evaluations, 1 pixel is equal to 0.27 mm on our 24-inch monitor. Consequently, 1° of visual angle error indicates ~39 pixels on the monitor when the user is at a distance of 60 cm.

[6] The source code of the simulation framework in Matlab can be downloaded at http://webmail.inb.uni-luebeck.de/inb-toolsdemos/FILES/et-simul-1.01.zip.

parameters of our real setup described in Section 3.1. So, the simulated monitor is of size 24-inch, and four light sources are placed on the corners of the monitor. In the simulated eye, following the typical eye parameters listed in [Guestrin and Eizenman, 2006], the cornea is modeled as a sphere with a radius of 7.8 mm. The refractive indexes of cornea and aqueous humor are 1.376 and 1.3375, respectively. The visual deviations between visual axis and optical axis are 5° for horizontal angle and 1.5° for vertical angle. The simulated cameras have resolutions of 1280×1024 and 8mm lenses (diagonal FoV=58°) are used to allow for large head movements.

The simulations consist of acquiring the calibration and test data. For the calibration, we simulate an eye looking at 9 uniformly distributed target stimuli points on the screen, whereas for the test data acquisition, we randomly generate 18 test points, which are different from the calibration points. The test points are displayed in a 3×3 grid with 2 points per region to cover the whole screen. Thus, the protocol avoids reporting false test results due to overfitting on the calibration point locations. In addition, to simulate realistic test conditions, we alter the noise level to examine the impact of noise-free and noisy data. For each calibration and test point, we collect 100 samples, and we introduce uniformly distributed feature position errors with a maximum magnitude of 0.4 pixels per feature (noise level $\in \{0, 0.1, 0.2, 0.4\}$).

We perform simulations under two different scenarios, namely, *Stationary Head (SH)* and *Moving Head (MH)*, as depicted in Figure 5.5. In *SH* scenario, the user eye is located 60 cm away from the monitor and kept at the same position during the whole simulations. In *MH* scenario, on the contrary, the user eye location is changed along three directions, X, Y, and Z. In both scenarios, the eye is calibrated at the default head position (0, 20, 60) cm.

5.3.2 Simulation Results on Stationary Head (SH) Scenario

In *SH* scenario, the main emphasis is given to the impact of increased number of cameras on estimation accuracy as there is no head movement. Figure 5.6 shows the results obtained when the proposed multi-camera approach employs various number of cameras in different configurations. For these simulations, we also introduce various levels of noise into the gaze features detection to examine the theoretical and practical impact of increased number of cameras and their positioning. In case 0 (see Figure 5.6a), when no feature detection noise is introduced, increasing the number of cameras, even up to 25 cameras, does not provide any estimation accuracy improvement. Contrarily, when a significant amount of noise is introduced, the more cameras the system employs, the higher accuracies it achieves. In fact, the results are in line with our intuitive expectations since all the cameras have similar capability in terms of estimation accuracy as well as estimation availability and robustness. When there is a noisy feature tracking scenario, e.g., low-resolution eye tracking, an improved overall estimation accuracy is expected because combining the outputs from multiple cameras can smooth out the noisy outputs.

The simulation results shown in Figure 5.6b indicate that not only the number of cameras, but also their positioning is important for improving the system performance. When there is no noise,

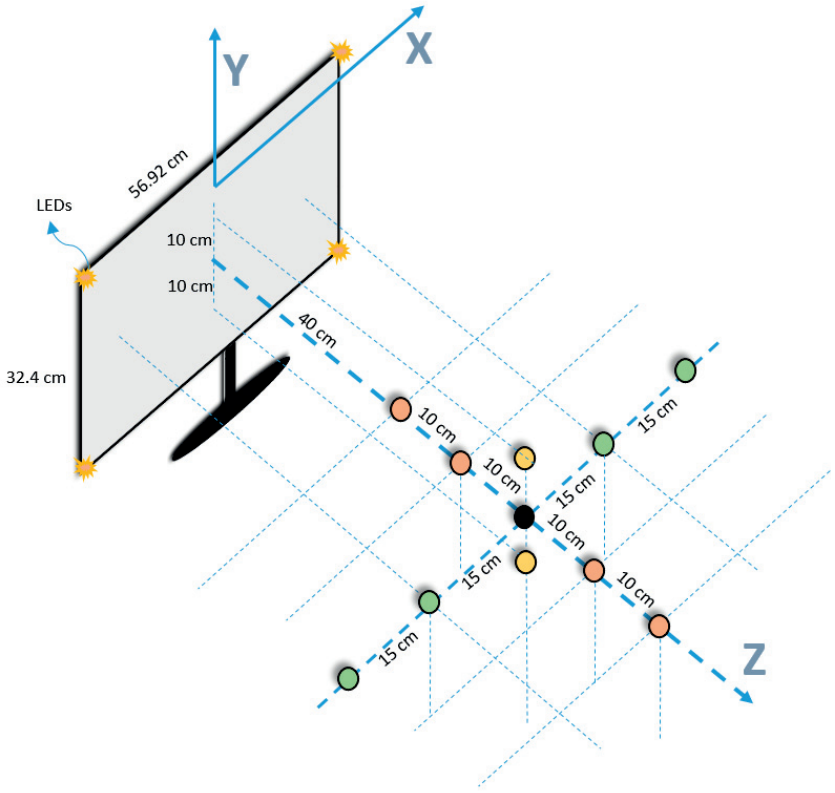
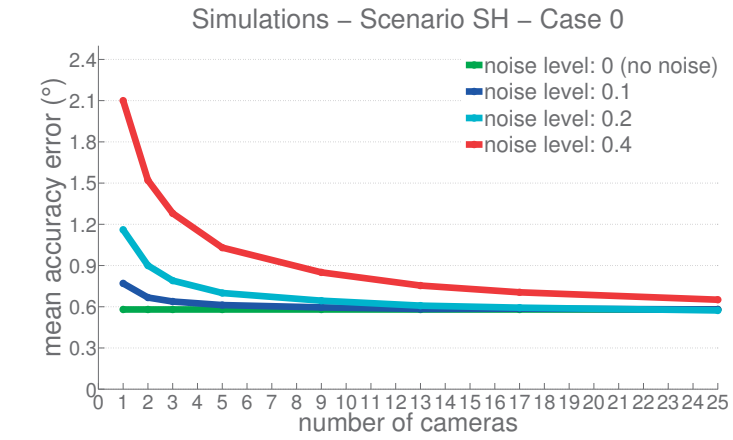
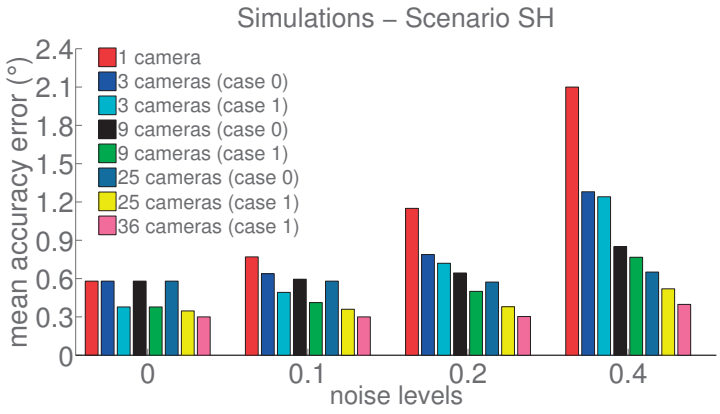


Figure 5.5 – Simulation setup. Default head position, where the calibration is performed, is at (0, 20, 60) cm, the black circle.

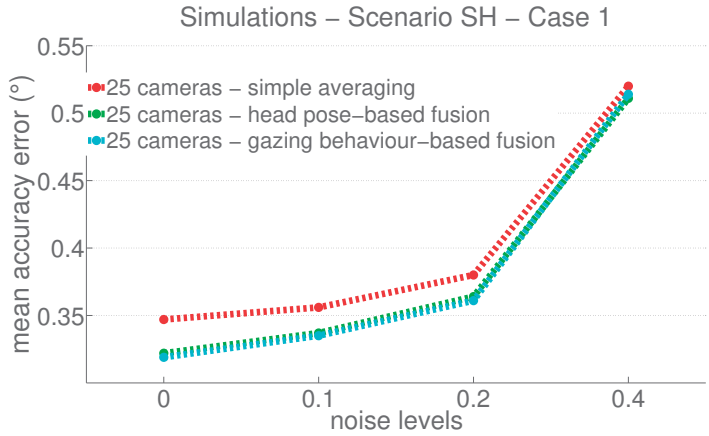
a 3-camera system with case 1 configuration, in which the cameras are placed at the bottom, left side, and right side of a monitor, outperforms even a 25-camera system with case 0 configuration, in which all the cameras are placed at the bottom of the monitor. In addition, the results suggest that when higher levels of feature detection noise is introduced, the number of cameras starts to have more impact than the positioning of the cameras since a higher number of cameras can better filter the noise out. For instance, when a very high level of noise is introduced, e.g., noise level = 0.4, the estimation accuracy increases directly proportional to the number of cameras, rather than the configuration. In case the real-world noise level is introduced, i.e., approximately 0.2 given the simulation setup, it is clear that the camera configuration is more effective than the number of cameras. The results show that case 1 configuration outperforms case 0 in both 3-camera and 9-camera setups. In fact, the 9-camera setup, in which 3 cameras are each placed in bottom, left, and right sides of the monitor (case 1), performs even better than the 25-camera setup with case 0 configuration. Furthermore, Figure 5.6c shows the achieved estimation accuracies when employing the adaptive fusion methods. The results indicate that the proposed methods, which leverage users' head pose and gazing behaviours, perform better than the simple averaging, particularly for lower noise levels. However, no significant performance difference is observed among both proposed adaptive fusion schemes.



(a)



(b)



(c)

Figure 5.6 – In static head (SH) scenario with varying feature detection noise levels, (a, b) the impact of increasing number of cameras (case 0 and 1), and (c) a comparison of the investigated adaptive fusion methods.

5.3.3 Simulation Results on Moving Head (MH) Scenario

In *MH* scenario, the goal is to examine the impact of increasing the number of cameras on the estimation availability and head movement robustness along X, Y, and Z directions, in addition to the estimation accuracy. For this scenario, the virtual eye is calibrated at the default position (0, 20, 60). The eye then is moved to various locations along three directions as shown in Figure 5.5 and the tracking is performed in these locations using the learned calibration at the default position. For the sake of simplicity, we fix the noise level to 0.2, which simulates the real-world noise level for our setup. First of all, as depicted in Figure 5.7a and 5.7c, the proposed methodology is in general highly robust against the changes in head movements along X (horizontal) and Y (vertical) directions. Increasing the number of cameras further enhances the overall estimation accuracy. Note that even the single-camera configuration is highly tolerant to horizontal head movements owing to the employed user calibration technique [Arar et al., 2016a]. On the other hand, the robustness to head movements along Z axis (depth translations) is the most challenging one among all. The main reason is that the user calibration is learned to account majorly for the angular difference between the optical and visual axes. As the user calibration is learned as an offset at a fixed head location, the learned offset does not sufficiently compensate for the bias when the user moves away from the calibrated position, especially along Z axis. Therefore, such movements cause a significant decay in estimation accuracy, as can be seen in Figure 5.7e for a single-camera system. However, the results illustrate that placing additional cameras around the monitor (case 1) yields a significant tolerance compared to a single-camera setup or a multi-camera setup with case 0 configuration. For instance, the line slopes from the calibration positions are much smaller when employing more cameras with case 1 rather than doing so with case 0 or using a single-camera setup. Hence, the proposed multi-camera setup provides an additional robustness against the depth translations.

Furthermore, an important benefit of the proposed approach is that a notable increase is observed in terms of the estimation availability when employing additional cameras around the monitor. Figure 5.7 (bottom row) demonstrates the impact of different camera configurations on the gaze availability (in %) when the user moves along X, Y, and Z directions. The results clearly show that a multi-camera setup allows for larger head movements (working volume) in all three directions than a single-camera setup. For instance, in comparison with a single-camera system, a three-camera setup with case 1 configuration provides an additional ±15 cm and ±10 cm head movement tolerance along X and Y directions, respectively. The reason is that each camera has a different viewing angle (FoV) and consequently the overall FoV of the system increases with the fusion of the FoVs of all cameras. The results also indicate that increasing the number of cameras further more (from 3 up to 36) does not drastically improve the availability as their FoVs starts to overlap after a certain number of cameras.

In this section, we do not analyze the robustness against eye wear and illumination variations since these can not be obtained by the simulator. Hence, we address them in the evaluations on real-data in the next section.

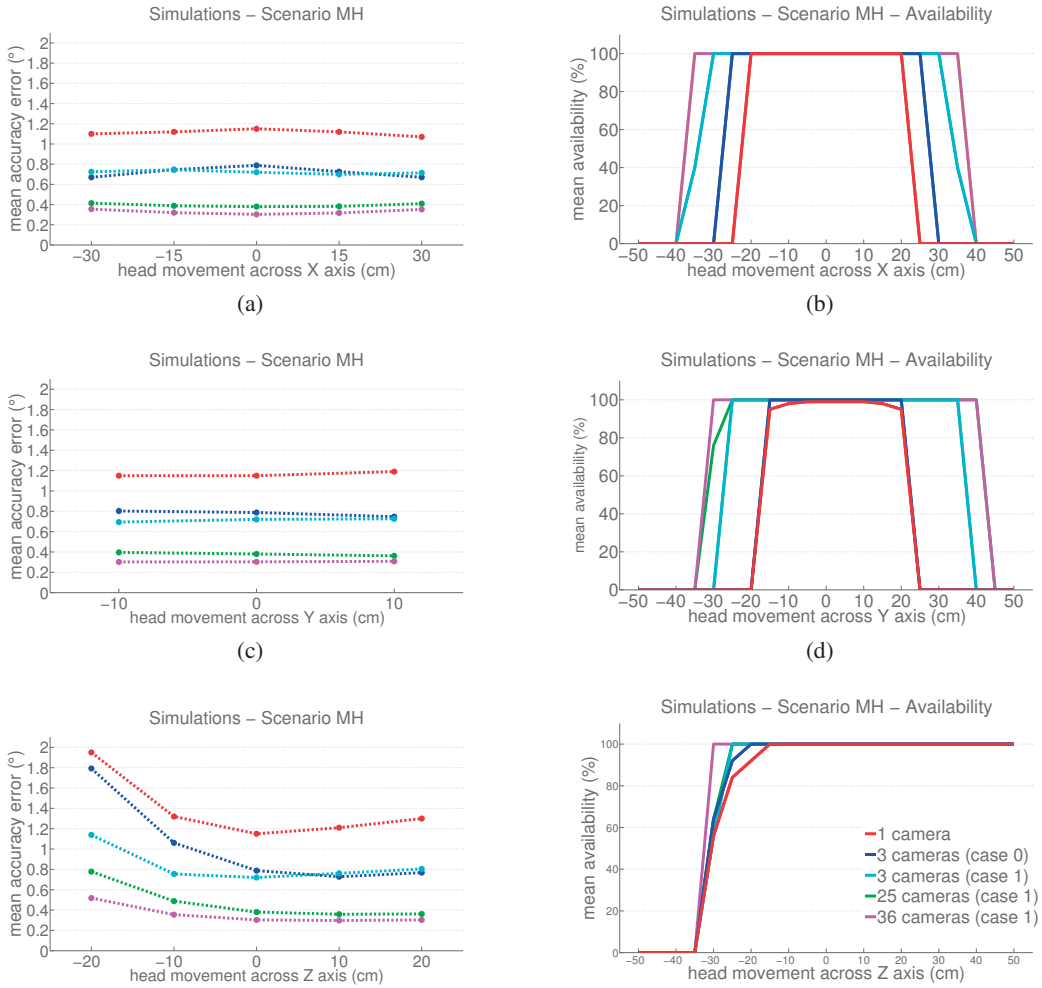


Figure 5.7 – In moving head (MH) scenario with a fixed feature detection noise level, the impact of increasing number of cameras (case 0 and 1) on the head movement robustness (top row) and gaze availability (bottom row) when user moves from the default calibration position (0, 20, 60) along X, Y, and Z directions. Please see the legend in (f) for all subfigures.

5.4 Evaluation on User Experiments

In this section, we present the evaluation of our approach on real-world data obtained from user experiments. Firstly, we describe the collected dataset and experimental protocols in Section 5.4.1. Then, we explain and discuss the results in Section 5.4.2.

5.4.1 Dataset & Experimental Protocol

We conducted a series of user experiments using the hardware setup described to comprehensively evaluate the proposed methodology regarding the estimation accuracy, availability, robustness

against head movements, varying illumination, eye wear, and between-subject variations. In total, 20 subjects, 15 males and 5 females, most of whom had no previous experience with any gaze tracking system, participated in our user experiments. 11 participants did not have any eye wear, while 5 and 4 participants wore eye glasses and contact lenses, respectively. The participants are from diverse origins, with a total of, mostly Caucasian, 12 nationalities. Therefore, the eye shapes and appearances exhibit a large variability.

Each participant was asked to follow 8 different experiments as described in Table 5.2 and Figure 5.8. Experiment #2 being the default protocol, in the first three experiments, we analyze the system's tolerance against varying illumination conditions by altering the ambient illumination (experiments #0 and #1). The remaining five experiments were conducted to evaluate the system's robustness against head movements. Four of them are conventional experiments, in which the subjects were asked to move along X (experiments #6 and #7) and Z (experiments #4 and #5) axes. The remaining one (experiments #3) stood for a novel scenario, in which the subjects were asked to continuously move their head while still fixating on the displayed gaze points. The goal of this last experiment was to analyze the system's sensitivity to continuous head movements, head pose changes and slight head translations during the fixation. Such a scenario, in fact, occurs frequently in real-world scenarios, e.g., natural course of free-head gazing, listening music, talking on the phone, etc. The subjects were provided with music of their preference to increase their motivation for such movements. As our evaluation targeted natural HCI scenarios, we tried to collect the ground truth data as natural as possible for the subjects. For instance, we did not use a chin rest to keep the subject's head still and to keep the eye within the cameras' FoV to capture high resolution eye data, as frequently performed in previous work. In addition, the subjects were asked to gaze at the target stimuli points in a natural and comfortable way. As a result, the subjects had different head pose and eye pose characteristics, facial expressions (e.g., mostly smiling and speaking), and viewing heights (along Y axis) while gazing. Table 5.3 presents sample head pose statistics of users. Since the subjects preferred to have various viewing heights for gazing, no experiment is explicitly performed to analyze the head movement robustness along Y axis. Figure 5.9 shows sample video frames from the dataset.

Table 5.2 – Experimental configurations.

Exp. No	Lighting Conditions	Head Location X	Head Location Z	Experimental Variable
0	sun	0	60	Illumination
1	dark	0	60	Illumination
2	indoor	0	60	Illumination (default protocol)
3	indoor	0	60	Continuous head movements
4	indoor	0	50	-10 cm head movements along Z axis
5	indoor	0	70	+10 cm head movements along Z axis
6	indoor	+15	60	+15 cm head movements along X axis
7	indoor	-15	60	-15 cm head movements along X axis

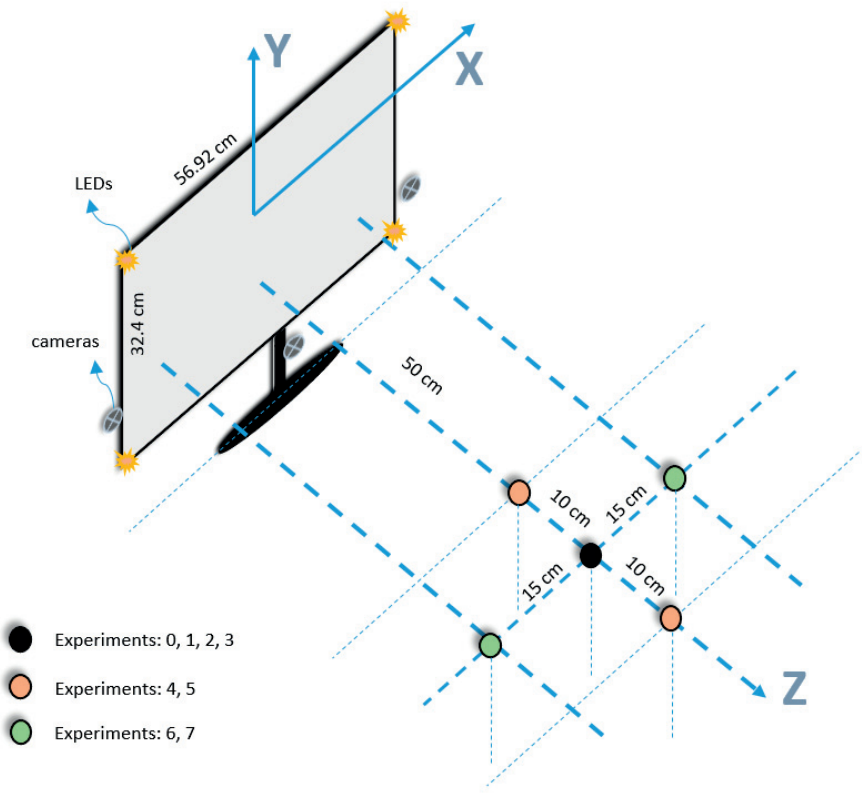


Figure 5.8 – User experiments setup. Default head position, where the calibration is performed, is at (0, 20, 60) cm, the black circle.

Table 5.3 – Head pose statistics (in °) of two subjects from the dataset. The head pose angles are estimated with respect to the bottom camera view separately on calibration and six individual test sessions relevant to head movements.

	Session	Exp No	Yaw Min	Yaw Max	Yaw Std	Yaw Mean	Pitch Min	Pitch Max	Pitch Std	Pitch Mean	Roll Min	Roll Max	Roll Std	Roll Mean
Subject #8	Calib.	#2	-13.3	2.6	4.7	-4.8	-13.6	4.9	5.3	-3.6	-6.2	0.1	1.7	-3.2
	Test	#2	-11.3	6.2	4.6	-2.6	-16.5	3.8	5.1	-5.3	-6.7	0.1	1.5	-2.7
		#3	-20.9	31.2	11.3	-1.5	-21.3	6.7	5.6	-3.4	-20.1	13.5	5.7	-3.9
		#4	-21.9	10.5	9.1	-3.4	-15.9	7.2	5.6	-1.8	-12.8	0.7	3.4	-5.3
		#5	-15.8	10.2	6.2	-3.6	-5.7	5.6	2.9	-0.2	-9.4	0.2	2.1	-4.6
		#6	-12.9	13.2	6.3	-2.4	-16.9	8.4	4.9	-2.3	-11.9	-2.6	2.2	-8.5
		#7	-15.2	9.9	5.3	-3.5	-8.2	6.4	2.9	-0.4	-3.7	4.5	1.8	0.5
Subject #18	Calib.	#2	-14.1	22.9	13.5	2.6	-21.1	0.9	6.9	-9.6	-4.4	4.4	2.4	-0.6
	Test	#2	-15.8	22.9	12.3	3.7	-23.1	-0.7	7.1	-10.2	-4.2	2.4	1.6	-1.1
		#3	-24.5	19.4	10.4	-0.9	-20.8	7.9	6.2	-7.4	-25.5	16.7	8.7	-2.7
		#4	-18.1	22.9	12.7	0.2	-24.3	-1.9	6.2	-12.6	-6.3	1.9	2.2	-1.6
		#5	-15.7	19.1	10.5	0.6	-21.3	-2.3	4.7	-10.2	-5.4	1.9	1.4	-2
		#6	-18.1	18.9	10.2	1.9	-22.9	3.6	7	-8.4	-14.2	-6.5	1.6	-9.4
		#7	-17.2	28.2	11.9	3.6	-22.6	4.9	6.8	-7.8	1.1	7.4	1.6	4.4

In the user experiments, we investigate and understand how the overall system performance is affected under varying illumination and head movements conditions, as well as eye wear and between-subject eye type variations. For the user experiments, we set the default user-to-monitor distance to 60 cm and perform the user calibration only at this distance. Then, the learned user calibration is applied during testing for all configurations with head movements. In addition, the default ambient illumination is considered as in an office environment illuminated by indoor fluorescent lighting as it corresponds to our tracker's main target scenario. Illumination robustness is examined under two alternative conditions such as total darkness and sunlight[7]. Our data acquisition and evaluation is similar to the simulations described in Section 5.3.1. The subjects were asked to gaze at 9[8] target stimuli points uniformly distributed on the screen for calibration, whereas, 18 randomly generated points were displayed for the testing. Each stimulus point was displayed for ~2 and 3 seconds for test and calibration points, respectively. In each experiment, the calibration and test sessions last for around 50 and 75 seconds, respectively. During the whole experiment, the data of both eyes was recorded from all cameras. The size of the circular target varied continuously from an initial radius of 30 pixels to a final radius of 20 pixels to serve as visual stimulus. The following section describes the results obtained from the user experiments.

5.4.2 Results

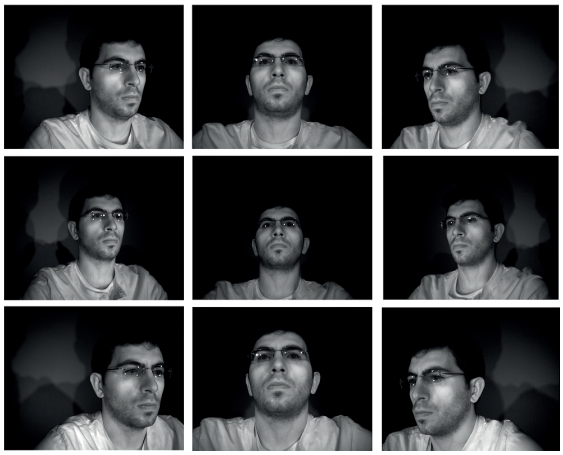
As described in Chapter 3, the proposed gaze estimation framework starts with face tracking on the captured frames, in which we extract eye regions of size ~90×50 pixels, and perform blink detection. This is followed by the removal of glares on glasses, if there exists any. Then, feature detection is performed to detect the gaze features, i.e., four glints and a pupil center (Section 3.2). The size of the polygon formed by the glints is ~9×5 pixels. Next, we apply the cross ratio-based gaze estimation (Section 3.3) with the detected gaze features to calculate the initial PoR. This procedure provides us the raw gaze output. We then apply the learned calibration models for the subject-specific bias correction on the raw gaze output (Section 4.2.6). Lastly, the calibrated PoRs obtained from each sensor are combined using an adaptive fusion mechanism to output an overall PoR per frame (Section 5.2).

The results achieved on the test data over all subjects for all experiments are shown in Table 5.6. This table lists the complete results with mean gaze estimation accuracy errors and estimation availabilities for various configurations. Prior to this, in the following subsections, we describe and discuss several subsets of these results under various emphases.

[7] The setup was placed by the windows inside a regular office at EPFL. Total darkness was obtained by closing the window blinds and turning off the indoor lights. On the other hand, the experiments with sunlight were conducted on sunny days and users were exposed to sunlight through the windows.

[8] *5 point* and *9 point* calibration configurations have been investigated. The reported results were obtained using *5 point* calibration.

(a) Illumination variations, i.e., experiments #2, #0, #1.



(b) Depth movements, i.e., experiments #2, #5, #4.

(c) Horizontal movements, i.e., experiments #2, #6, #7.

Figure 5.9 – Sample images from the collected dataset: (left column) right camera view, (middle column) bottom camera view, and (right column) left camera view.

Single-camera setup vs Multi-camera setup

We start our analysis to demonstrate the benefits of employing a multi-camera setup instead of a single-camera setup, as implemented by the majority of the previous work. In order to validate the findings of the simulations regarding different multi-camera setup configurations, i.e., case 0 vs case 1 (see Figure 5.6), we conducted additional user experiments on a small subset of the users (3 out of 20 subjects, one subject from each eye wear category) using a three-camera setup with case 0 configuration, in which all cameras are placed at the bottom of the monitor very close to each other. Besides, as the proposed approach enables to estimate gaze output for both eyes simultaneously, we also demonstrated the results by altering the used eye data such as single eye only (left or right) and both eyes for the single-camera setup. Moreover, we obtained results using the multi-camera setup with the proposed adaptive fusion mechanisms. The proposed adaptive fusion mechanisms compute the overall PoR as a weighted combination of all available gaze outputs obtained from the overall setup. In case there is no available gaze output from the setup, the overall PoR can not be computed and the gaze availability is negatively affected. Table 5.4 shows the mean estimation accuracy errors and availabilities obtained under various setup configurations for the default experimental protocol, experiment #2 listed in Table 5.2.

The results clearly demonstrate the efficacy of the proposed multi-camera setup (case 1 configuration) over the single-camera setups as well as the multi-camera setup with case 0 configuration in terms of both estimation accuracy and availability. First, a significant accuracy improvement, about 45%, is achieved compared to the best performing single-camera system. In addition, an increase in the estimation availability is obtained, nevertheless, the availability comparison is more interesting when analyzing head movements and eye wear robustness in the following subsections. Furthermore, the results indicate that the performance of a single-camera system notably increases when using both eyes rather than a single eye, regardless of which eye is used. The reason is that several factors, such as illumination effects (shading and reflection), head and eyeball pose with respect to the camera, or physiological vision disorders, may have influences on the estimation, especially when low-resolution eye data is used. In such cases, the data obtained from a single eye may not be reliable enough to output an accurate estimation. The estimation

Table 5.4 – Tracking performances for various single- and multi-camera configurations.

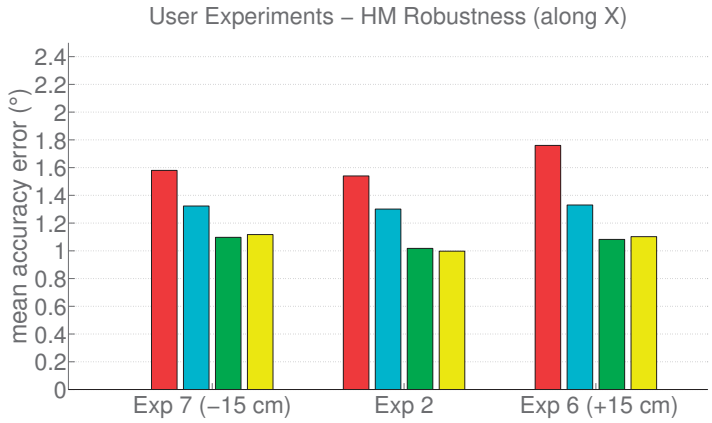
Setup configuration	Eye Data	Estimation (°)	(%)
Single-camera left eye only	1	1.62	77.6
Single-camera right eye only	1	1.6	69.6
Single-camera both eyes	max 2	1.45	93.4
Multi-camera with case 0	max 6	1.38	94.2
Multi-camera with case 1 simple averaging	max 6	0.89	99.8
Multi-camera with case 1 head pose-based fusion	max 6	0.85	99.8
Multi-camera with case 1 gazing behavior-based fusion	max 6	0.8	99.8

inaccuracy occurs more frequently for target points which require a large head pose or eyeball pose for the users. On the other hand, when both eyes are utilized, the system has a higher chance to deal with such targets since one of the eyes may have a better viewing angle for a certain camera. Consequently, utilizing both eyes enables a smoother (higher precision) and more accurate overall estimation. It also increases the estimation availability. Moreover, the results highlight the performance difference between case 0 and case 1 multi-camera setup configurations. In fact, the results are greatly in line with the findings of the simulations, such that the positioning of the cameras is crucial. Lastly, the results show the impacts of the proposed adaptive fusion mechanisms. In this respect, although the simple averaging standalone achieves a significant performance improvement in comparison with single-camera configurations, employing the proposed adaptive fusion algorithms further enhances the accuracy.

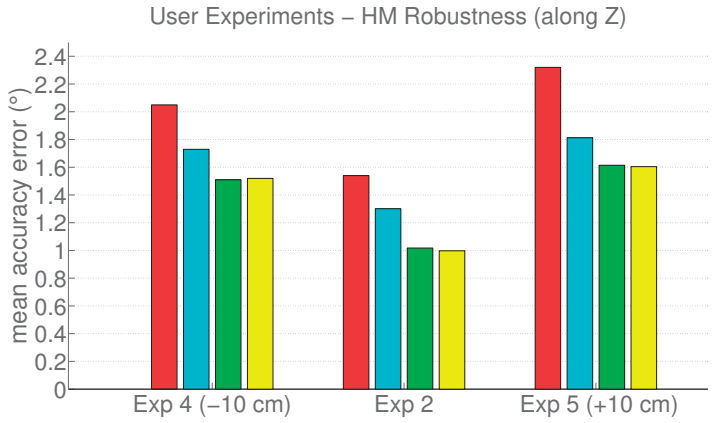
Head Movement Robustness

The proposed setup's tolerance to head movements can be examined by analyzing the accuracy and availability results of the experiments #2-7 in Table 5.2. More specifically, experiments #2, #6, and #7 account for the horizontal movements (along X axis) and experiments #2, #4 and #5 account for the depth movements (along Z axis). We note that vertical movements (along Y axis) are not explicitly experimented since the subjects, for their convenience, were asked to freely position their heights with respect to the monitor. Furthermore, we introduced a new experimental scenario (experiment #3), in which the users were asked to perform continuous head location and pose changes while still fixating on the target points. The purpose of this experiment is to measure the system's sensitivity to sudden arbitrary changes during the user interaction, which may frequently occur in real-world conditions. Figure 5.10 illustrates the results achieved on these experiments and their cross comparisons.

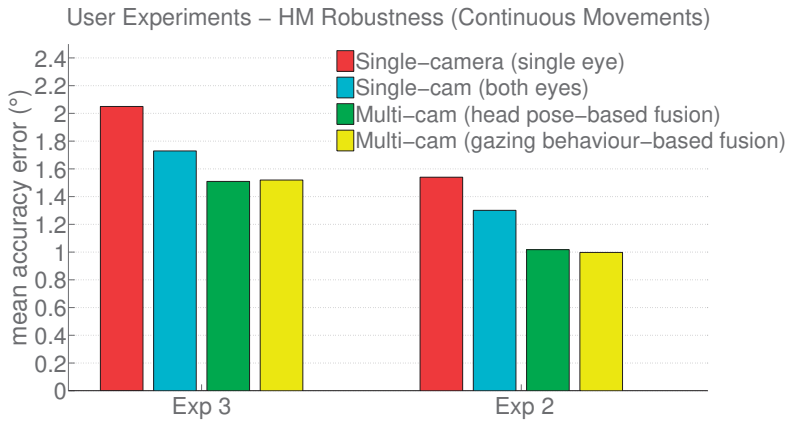
For horizontal head movement robustness, the results of the user experiments (Figure 5.10a) are highly in line with the simulation results (Figure 5.7a), such that the system is highly insensitive (1° vs 1.1°) to head movements along X axis up to ±15 cm movements. On the other hand, for head movements along Z axis (depth translations), the results partially differ from the simulation results. In simulations (Figure 5.7e), the estimation accuracy is shown to be negatively affected by the depth movements, especially when user moves towards the monitor, due to insufficient compensation for the angular difference between visual and optical axis. The same result holds for the user experiments. However, the user experiment results also show that the accuracy decreases when user moves away from the monitor, which contradicts the simulation results. In fact, the main reason relates to the physical setup. In our hardware setup, which consists of manual focus lenses, the image focus gets worse when the depth of the users vary from the default position despite the aperture adjustments to have a larger depth-of-field. This causes the features to appear more blurry. In addition, the eye image resolution gets naturally lower when the user moves away from the camera, which causes gaze features to be detected less accurately. Both reasons are not valid for the simulations since the depth-of-field and feature detection accuracy are not affected. A more detailed discussion on this limitation together with the proposed solutions are



(a)



(b)



(c)

Figure 5.10 – Performance comparison of single-camera and multi-camera setups under different head movement scenarios. Please see the legend in (c) for all subfigures.

described in Section 6.2. Despite these limitations, the multi-camera approach still provides more robustness, about 25% in accuracy and 10% in gaze availability, to depth translations than the single-camera approaches.

Moreover, the mean estimation accuracy errors obtained using the proposed head pose-based and gazing behaviour-based fusion schemes are shown in Figure 5.10. The results are inline with simulations and they illustrate that both algorithms perform similarly under different experimental scenarios. In theory, one could expect that the head pose-based fusion scheme outperforms the gazing behaviour-based one when the users move away from the default calibration position as the fusion weights, in latter one, are optimized according to the calibration position. However, no significant difference in the estimation accuracy is observed in our user experiments (p-values > 0.05, paired t-tests). Furthermore, Figure 5.10c demonstrates the system's robustness to continuous head movements, in which the users intentionally perform head rotations and translations during the fixations. The results indicate that the proposed system, as expected, experiences an accuracy drop, yet it continues to output PoRs with an acceptable accuracy ($\sim$1.4°) under such a challenging scenario. In addition, the results indicate that the multi-camera setup does not provide additional robustness to such scenarios, but rather enhances the overall estimation accuracy. We believe that this new experimental scenario constitutes an important evaluation criteria since it has an important correspondence in real-world eye tracking use case scenarios. Hence, we suggest that the future efforts consider this scenario in their validations.

Illumination Robustness

The proposed setup's robustness to varying ambient illumination conditions such as indoor lighting, darkness, and sunlight can be observed from Figure 5.11. The results illustrate that illumination variations do not significantly influence the estimation performance. Illumination by indoor fluorescent lighting slightly outperforms the others since the feature detection is mostly

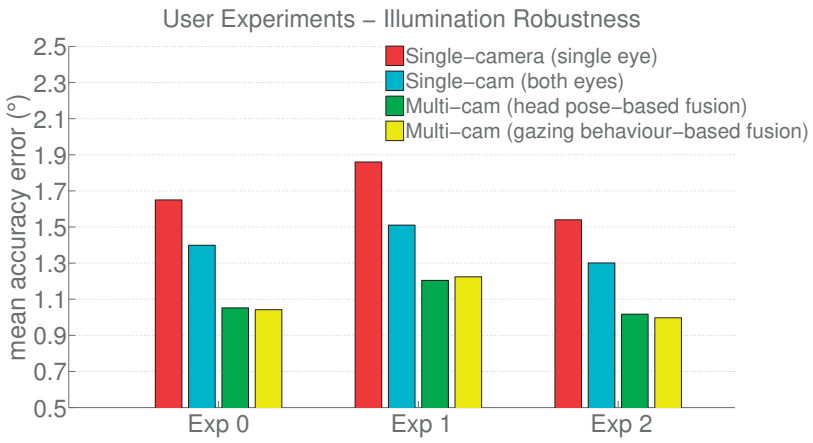


Figure 5.11 – Illumination robustness comparison of single-camera and multi-camera setups.

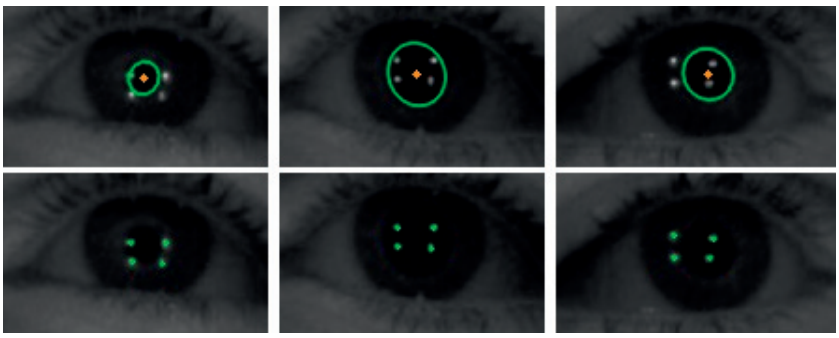


Figure 5.12 – Sample appearances of eye and gaze features (glints and pupil) under varying illumination conditions: (left) sunlight, (center) darkness, (right) indoor lighting.

optimized for this main target scenario. As mentioned previously in Section 5.1, the systems that operate under active NIR illumination, e.g., the majority of *feature-based* and few *appearance-based* ones, are implicitly more tolerant to illumination changes. Still, the feature detection mechanisms need in practice to adapt to the changes in gaze features, as illustrated in Figure 5.12. For instance, the pupil, as the aperture of our eyes, can shrink or expand to adjust the light that comes into the eye. Consequently, the pupil size gets smaller when exposed to the sun lighting, and gets larger when it is dark. In our experiments, we observe that the precision of pupil center detection is lower when the pupil size gets very large, e.g., almost as big as the iris. In addition, sunlight may bring additional side effects such as smaller eye opening and distorted glints (NIR intervention), which may negatively influence the overall estimation accuracy and availability.

Eye Wear Robustness

Eye wear robustness, as mentioned in Section 5.1.4, is undoubtedly one of the most challenging issues in eye tracking. Unfortunately, it has been neglected by the great majority of the previous studies. The main challenges stem from the reflection and refraction effects on the glasses, which can significantly affect the accuracy and precision of the gaze estimation. There are several types of glasses (e.g., bifocal, trifocal, progressive, etc.) and glass characteristics (e.g., reflective index, refraction index, filters, coatings etc.), which influence the obstructive effects of eye glasses. Sample impacts on eye appearance, such as distorted features due to the refraction and coating, lost features due to the reflection, challenging feature detection due to multiple reflections, which were encountered during our user experiments can be seen in Figure 5.13. As some of the impacts are unrecoverable, the conventional approaches with single-view eye appearances are highly likely to fail under such circumstances. On the other hand, the proposed multi-camera approach is designed in such a way to bring robustness against eye glasses. The setup leverages the eye appearances from various views so that the gaze features can be recovered from one or more of the views under challenging tracking conditions.

We evaluate the efficacy of our proposed method with two separate analysis. In the first one, we categorize the subjects into four groups according to their eye wear and vision quality, namely,

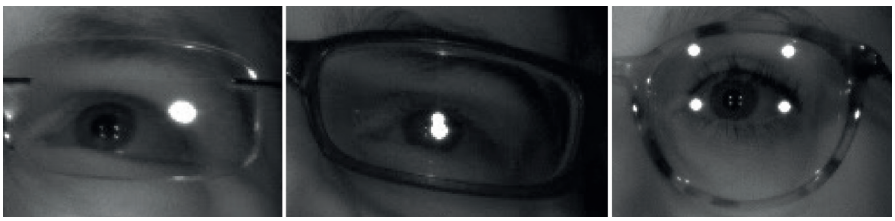


Figure 5.13 – Sample impacts of eye glasses on eye appearance: (left) weak or distorted glints, (center) glares overlapping on gaze features, and (right) multiple glares causing a challenging feature detection.

the ones who wear eye glasses, who wear contact lenses, who do not need eye wear (perfect vision), and lastly who do not wear eye glasses. The last group comprises of subjects with contact lenses, perfect vision, as well as the ones who do not use any eye wear on a daily basis due to low degree (up to 1.5 diopter) of myopia or astigmatism problems. We then examine the tracking performances for each experiment and for each group, as can be seen in Figure 5.14 and Table 5.6. We note that the multi-camera tracking performances obtained using both fusion methods are highly similar, thus, we only report the head pose based fusion results. The results clearly depict the improvements achieved using a multi-camera setup for both the default scenario (experiment #2) and over all scenarios (experiments #0-#7). Among all groups, the best performance (~0.8°) is achieved on the subjects with perfect vision (6 subjects) and contact lenses (4 subjects). For the group who do not wear eye glasses (15 subjects), a small accuracy drop is observed. The reason relates to some of the subjects' vision defects. As described in Section 5.4.1, we display the visual stimuli points to the users as varying size (20-30 pixels) circular targets with a small black dot at the center. We observed that a part of this group's subjects (3 subjects) were not able to see the black dot at the center, but rather saw a circle. Therefore, we believe that the accuracy drop for this group is expected due to the non-sharp vision of such subjects. Lastly, the

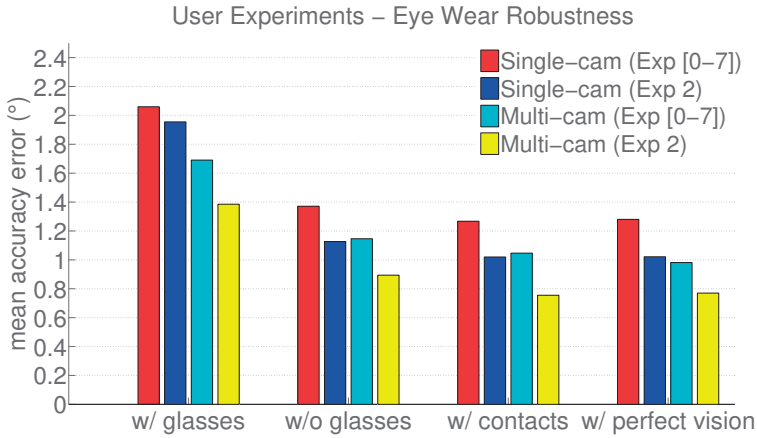


Figure 5.14 – The impact of using multi-camera system over a single-camera system when using eye wear. Average estimation accuracies obtained on Experiment #2 and over all experiments (Experiment [#2-#7]) are displayed.

Table 5.5 – Performance comparison for the same subject with eye glasses and contact lenses.

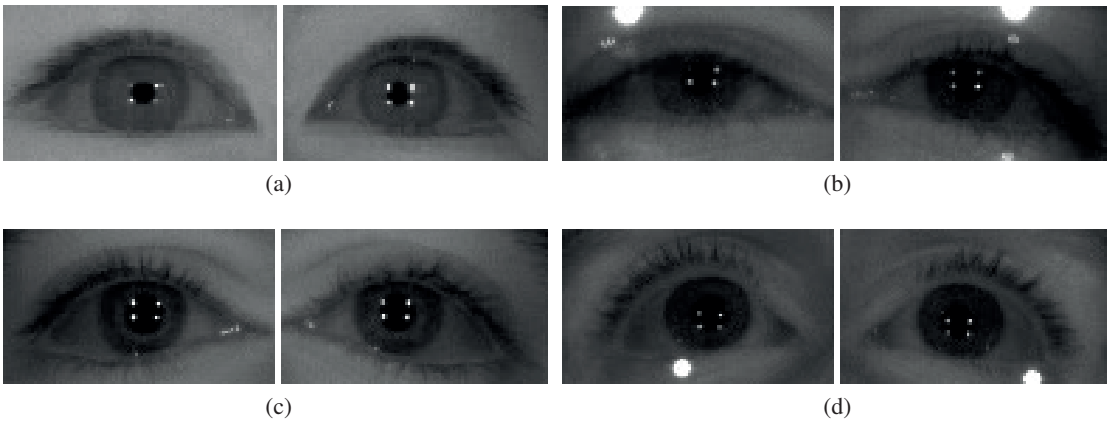
Eye Wear	Single camera				Multi camera			
	Experiment 2		Experiment [0-7]		Experiment 2		Experiment [0-7]	
	(°)	(%)	(°)	(%)	(°)	(%)	(°)	(%)
Contacts	0.99	96.1	1.18	95.1	0.76	100	0.97	99.9
Glasses	2.1	84	2.31	82	1.08	100	1.53	99

group with eye glasses (5 subjects) achieved, as expected, a lower accuracy (1.38° with 91.9% availability) in comparison with the other groups. Yet, the performance improvement, about 0.6° in accuracy and 10% in availability, compared to employing a single-camera setup validates the efficacy of the multi-camera setup.

In the second analysis, mainly to discard between-subject variations, we compared the tracking performance on a the same subject. In this evaluation, a subject, who is nearsighted with -3 diopters, completed the user experiments firstly by wearing his eye glasses and then once again by wearing his contact lenses. Table 5.5 shows the performance achieved on the default scenario as well as the average over all the scenarios. The results clearly illustrate the positive impact of a multi-camera setup over a single-camera setup in both eye wear scenarios, such that it provides a substantial improvement in accuracy about 50% and 40% for the generic scenario and all the scenarios, respectively. In addition, it brings ~17% enhancement in estimation availability. For the contact lens scenario, the single camera setup standalone yields a high accuracy and availability. Still, the multi-camera setup contributes to additional accuracy and availability gains.

Eye Type Robustness

Lastly, we analyze the proposed system's tolerance against between-subject eye type variations. Since certain eye type related factors, e.g., eye color, eye shape, pupil response etc., may affect the performance of the eye trackers [Nguyen et al., 2002], we evaluated our system's performance under varying eye types across the subjects. Figure 5.15 shows sample eye type and color variations from the dataset. Firstly, since the iris color has a great influence on both the pupil size and opening of eyelids when exposed to various illumination conditions, we categorized the subjects into two groups according to the eye color such as dark-eyed ones (10 subjects) and light-eyed (10 subjects). On average over all experiments, dark- and light-eyed groups achieve 1.15° with 96.2% availability and 1.44° with 93.3% performances, respectively. However, the results may be biased towards the dark-eyed group since most of the subjects who do not wear eye glasses are within this group. An interesting result is that the performance difference between the light-eyed (1.33°) and dark-eyed groups (0.92°) is especially large under sunlight (experiment #0). The reason is that the pupil size and eye opening are affected more for the light-eyed subjects in comparison with the dark-eyed subjects due to their higher sensitivity to the sunlight. Table 5.6 shows average estimation accuracy errors and estimation availabilities in detail for each experiment in several categories.



(a) (b)

(c) (d)

Figure 5.15 – Sample eye appearances from the dataset. (a) Asian dark eyes without glasses, (b) Asian dark eyes with glasses, (c) Caucasian dark eyes without glasses, (d) Caucasian dark eyes with glasses.

It is also important to note that the pupil detection method has an influence on the robustness to eye color variations. As discussed in Section 3.2.5, bright-pupil based method is frequently employed by the previous work as the feature detection is simpler compared to dark-pupil based one. However, in our preliminary experiments, we observed that dark-pupil based feature detection is less sensitive to the variations in eye color. In bright-pupil based method, the accuracy of the pupil detection heavily relies on the pupil response (brightness), which is highly affected by users' momentary pupil size that varies according to the eye color, ethnicity, and ambient illumination. Therefore, in our final framework suggests to employ dark-pupil based feature detection in order to become less sensitive to eye type and illumination factors. We plan to give a structured and quantitative comparison in our future work.

Furthermore, we categorized the subjects by their eye shape into two groups: Asian eyes (2 subjects) and non-Asian eyes (18 subjects) to analyze the impact of the eye shape. Our results show that Asian eyes (1.58° with 93.35% availability) perform worse than non-Asian eyes (1.25° with 97.7% availability). Yet, the system can still accurately estimate the gaze for our Asian subjects. The decrease in the availability may indicate that the feature detection for them might be more challenging due to the eye shape. Nevertheless, it is difficult to make a strong conclusion as the two sets are highly imbalanced. In addition, we note that there is a significant variation across Asian eyes [Fakhro et al., 2015, Kiranantawat et al., 2015]. The eyes may be of any shape including round, narrow, almond, hooded, triangular, prominent, or deep-set. In addition, the eyes can be a single eyelid, low/incomplete eyelid crease, and double eyelid. For some of these eye shapes (e.g., narrow, hooded), eye tracking could be highly challenging as the creation and detection of the gaze features could be exigent. Our dataset currently do not contain such challenging eye types. In our future work, we plan to recruit a higher number of Asian subjects and increase the variation in eye type to obtain a more reliable analysis and conclusions.

	Configurations		Exp 0		Exp 1		Exp 2		Exp 3		Exp 4		Exp 5		Exp 6		Exp 7		Overall	
	Camera	Eye data	(°)	(%)	(°)	(%)	(°)	(%)	(°)	(%)	(°)	(%)	(°)	(%)	(°)	(%)	(°)	(%)	(°)	(%)
single-camera	right cam	left eye	1.71	50.6	1.68	51.6	1.59	57.3	2.04	46.8	1.87	36.3	2.06	49.6	1.84	51.1	1.53	37.5	1.79	47.6
	right cam	right eye	1.50	60.4	1.86	68.1	1.52	69.6	2.16	58.6	2.27	41.7	2.26	63.2	1.83	66.7	1.55	65.1	1.87	61.6
	right cam	adaptive fusion	1.55	74.4	1.6	74.9	1.41	77.6	2	69.7	2.11	52	2.07	75.2	1.77	73.7	1.35	69.4	1.75	70.9
	left cam	left eye	1.64	62.1	2.02	68.9	1.54	65.8	2.07	57	2.1	47.4	2.18	53.1	1.88	47.6	1.68	70.4	1.89	59
	left cam	right eye	1.55	51.4	1.86	56.1	1.49	56.9	1.86	44.2	1.83	44.7	2.28	51.2	2.08	33.8	1.75	62.1	1.84	50.5
	left cam	adaptive fusion	1.5	73.1	1.8	78.3	1.36	79.4	1.88	68.3	1.79	61.8	2.07	71.9	1.77	58.7	1.51	83.2	1.71	71.8
	bottom cam	left eye	1.68	77	1.86	78.1	1.50	81.8	1.93	72.1	1.96	64.7	1.99	67.7	1.79	77.6	1.63	73.4	1.79	74.1
	bottom cam	right eye	1.65	74.6	1.86	79.5	1.54	79.2	1.90	68.3	2.05	64.8	2.32	67.4	1.76	75.5	1.58	75.1	1.83	73.1
	bottom cam	adaptive fusion	1.46	91.2	1.51	94.6	1.30	93.6	1.67	87	1.73	82.4	1.81	83.9	1.33	92.5	1.32	90.8	1.53	89.4
multi-camera	**overall**	**adaptive fusion**	**1.07**	**96.2**	**1.22**	**98.4**	**0.99**	**97.2**	**1.45**	**92.2**	**1.52**	**92.4**	**1.60**	**93.7**	**1.10**	**95.2**	**1.12**	**95.5**	**1.26**	**95.1**
	subjects with perfect vision		0.86	98	1.02	98.5	0.77	98.8	1.20	94.8	1.21	94.9	1.18	98.8	0.79	98.8	0.81	98.9	0.98	97.7
	subjects with contacts		0.83	96.8	1.05	99.7	0.76	98.9	1.23	92.5	1.41	97.5	1.17	99.1	0.88	99	1.02	98.5	1.04	97.8
	subjects without glasses		0.98	97	1.11	99.1	0.89	98.6	1.38	94.2	1.40	96.5	1.44	96.8	0.92	98.9	1.01	98.3	1.14	97.5
	subjects with glasses		1.39	93.4	1.63	95.6	1.38	91.9	1.68	84.4	1.95	76.8	2.19	82	1.77	81.4	1.51	85	1.69	86.3
	dark-eyed subjects		0.92	96.7	1.16	97.8	0.86	98.2	1.29	93.8	1.42	93.5	1.46	96.4	1.05	95.9	1.03	97	1.15	96.2
	light-eyed subjects		1.33	95.4	1.33	99.3	1.22	95.7	1.70	89.5	1.68	90.5	1.85	89.1	1.18	94	1.26	93	1.44	93.3

Table 5.6 – Average gaze estimation accuracy errors and gaze availabilities achieved by all single-camera and multi-camera configurations on each of the user experiments.

5.5 Discussion

Future directions in eye tracking research, towards becoming a pervasive technology, should not only focus on achieving high estimation accuracies, but also on having robustness against real-world settings such as natural head pose changes, large head movements, varying illumination conditions, use of eye wear, and between-subject eye type variations. Besides, having a convenient user calibration, flexible hardware setup, minimal setup calibration, low complexity, and low cost should be taken into consideration as important evaluation criteria. In this regard, in Section 5.1, we describe various eye tracking techniques, analyze their pros and cons with respect to each other, and discuss whether they satisfy some of the aforementioned criteria. Therefore, the best, in other words the most suitable, approach depends on the application type and requirements. In this work, we mainly target eye tracking scenarios that require high estimation accuracies ($\sim 1^\circ$) and robustness, e.g., gaze-based mouse controlling, navigation, typing, gaming, etc. In order to achieve our estimation and robustness goals, we design a novel multi-camera setup and methodology, which tracks users' gaze simultaneously from various views, and then combines the acquired gaze information from all sensors using an adaptive fusion mechanism to output an overall PoR. In comparison with conventional single-camera systems, simultaneously acquired multi-view eye appearances enables a reliable gaze features detection even under challenging scenarios mentioned above. Then, together with the suggested adaptive fusion mechanisms, the system achieves high estimation accuracy, availability and robustness to real-world conditions.

A comparison of existing work in several aspects such as hardware setup and calibration requirements, accuracy, robustness to real-world conditions and working volume, is given in Table 5.7. Since the majority of the existing work requires particular hardware and system setups, e.g., additional light sources, setup calibration, use of 3D or depth information, we could not reproduce and validate the estimation accuracies and robustness, instead we reported the

corresponding details from the corresponding references. Therefore, although a direct numerical comparison would not be fair, the provided information can still help us to make certain inferences. First of all, we observe that the popularity of *appearance-based* methods, which have lower hardware requirements, have been increasing recently in parallel with the recent advancements in machine learning (e.g., convolutional neural networks) and in the synthesizing and rendering technology. Although their accuracies and head movement tolerances are currently not sufficient for precise eye tracking, their potential is likely to be exploited in the foreseeable future. Secondly, *feature-based* methods, i.e., *3D model*, *regression*, and *cross ratio-based*, undoubtedly outperform *appearance-based* methods in terms of the accuracy. However, they mostly require particular hardware configuration, e.g., NIR cameras and light sources. The setup complexity is especially high for *3D model-based* systems, such that they need fully calibrated setups consisting of multiple cameras or a Kinect-like sensor to accurately model the eye in 3D. Thirdly, *cross ratio-based* systems and majority of *regression-based* systems have an important advantage over *3D model-based* systems, that they require only an uncalibrated camera to accurately operate. Despite their uncalibrated setups and less complex (2D) eye models, their accuracies are competitive with those of *3D model-based* systems. Among these systems, it is clear that there is an accuracy gap between fixed (using a chin rest) and free head pose tracking since they rely on approximated models.

Furthermore, we observe that the accuracy significantly increases when using high resolution eye data. For example, [Coutinho and Morimoto, 2013] achieved an impressive accuracy, about 0.5°, under large head movements using planarization of gaze features. However, their system required eye image resolution of 640 x 480 pixels, which was 7-fold larger than ours. In their setup, the eye data was captured using a narrow FoV lens and a chin rest was required to keep users' eye within the FoV of the camera. In addition, [Huang et al., 2014] and [Zhang and Cai, 2014] proposed two alternative methods that are effective to compensate for the head movements, while requiring relatively lower resolution eye data, i.e., 13 mm lenses were used. However, similar to [Coutinho and Morimoto, 2013], they both utilized a chin rest during their evaluation. Although the results obtained using a chin rest can be considered as more controlled and stable, the evaluations discard the head pose and continuous head movement robustness. Besides, it is unnatural for users and represents an unrealistic tracking scenario. Therefore, we believe that it remains an important limitation of these systems' evaluations. On the contrary, our methodology allows for not only head translations but also head rotations while requiring lower resolution eye data (~90×50 pixels) captured using 8 mm lenses. Lower resolution data naturally results in a lower accuracy, nevertheless, the proposed adaptive fusion mechanism successfully closes the accuracy gap by effectively combining the gaze outputs obtained by multiple sensors. Besides, our system accounts for eye wear and illumination robustness, some of the important concerns in eye tracking, which have largely been neglected by the majority of the previous efforts.

Despite our methodology achieves competitive accuracies while offering more robustness to aforementioned real-world conditions, its performance can still be improved. For instance, explicit head movement compensation techniques, such as learning an adaptive homography from simulated data [Huang et al., 2014] or planarization of cross ratio features [Coutinho

and Morimoto, 2013], can further improve the head movement tolerance. In addition, certain hardware-based solutions can alternatively be employed for further improvements. For example, auto-focus lenses or smart dynamic illumination techniques, as utilized by most of the commercial eye trackers, can greatly help to enhance the estimation accuracy and availability.

Moreover, the proposed multi-camera approach is highly flexible and can easily adapt to hardware and software modifications. First of all, alternative gaze estimation methods can be integrated towards obtaining better performance because the adaptive fusion algorithms are independent of the gaze estimation method used. In this thesis, we suggest to employ a *cross ratio-based* gaze estimation method, due to the particular advantages of the method mentioned in Section 5.1. In addition, since there is no camera or geometrical system calibration, the number of cameras and their positioning can be alternated according to the application scenario without requiring further system adjustments. For instance, the system can easily be configured to work under challenging tracking scenarios, such as in-car driving scenarios, children's eye tracking, or customized eye trackers for disabled people.

	Method	Hardware Setup			Accuracy			Robustness				FoV
		Cam(s)	Light(s)	Calib.	SH(°)	MH(°)	Eye Data	HP	Head Mov.	Eye Wear	Illum.	FL
Appearance	[Zhang et al., 2015]	1	-	-	6.3[1]	?	Both	Free	?	Yes	Yes	?
	[Wood et al., 2016b]	1	-	-	9.95[2]	?	?	Free	?	Yes	Yes	?
	[Lu et al., 2015]	1	-	-	2.5	9.65	Single	Free	(40 x ? x ?)	-	-	?
	[Mora and Odobez, 2016]	1+Kinect	5	pre	1.9[3]	3.5	Both	Free	?	-	-	6.1
	[Krafka et al., 2016]	1	-	-	~3.5[4]	?	Both	Free	?	Yes	Yes	?
3D Model	[Beymer and Flickner, 2003]	4*	2	fully	~0.6	-	Single	Free	Limited	-	-	4.8
	[Ohno and Mukawa, 2004]	2+1*	1	fully	~1	-	Single	Free	Limited	?	-	?
	[Hennessey et al., 2006]	1+1	3	fully	~1	<1	Single	Free	(14 x 12 x 20)	-	-	32
	[Guestrin and Eizenman, 2007]	2	4	fully	-	~1	Single	Free	(10 x 8 x 10)	-	-	35
	[Park, 2007]	1+2*	4	fully	-	~1	Both	Free	(? x ? x 20)	Yes	-	?
	[Lai et al., 2015]	2	2	fully	-	~1	Single	Free	(10 x 5 x 10)	-	-	37
	[Sun et al., 2015]	Kinect	?	pre	~1.5	~2	Single	Free	(20 x 20 x 8)	-	-	6.1
Regression	[Zhu and Ji, 2007]	2	2	fully	~1.1	~1.8	Single	Free	(20 x 20 x 30)	-	-	?
	[Cerrolaza et al., 2008]	1	2	-	~1	~1	Single	Fixed	(0 x 0 x 10)	-	-	35
	[Sesma-sanchez et al., 2012]	1	2	-	~1	~1	Both	Fixed	(0 x 0 x 6)	-	-	35
	[Cerrolaza et al., 2012]	1	2	-	~0.9	~1.3	Single	Fixed	(0 x 0 x 12)	-	-	16
	[Xiong et al., 2014]	2	2	fully	2.33	3.33	Single	Fixed[5]	?	Yes	?	?
Cross-ratio	[Yoo and Chung, 2005]	1+1*	4+1	-	~1.6	-	Single	Free	Limited	-	-	?
	[Coutinho and Morimoto, 2006]	1	4+1	-	~1.3	-	Single	Free	Limited	Yes	-	?
	[Kang et al., 2007]	1	4+1	-	~1	-	Single	Fixed	-	-	-	?
	[Hansen et al., 2010]	1	4	-	~1	-	Single	Free	Limited	-	-	?
	[Coutinho and Morimoto, 2013]	1	4+1	-	~0.4	~0.5	Single	Fixed	(25 x ? x 25)	-	-	> 35
	[Zhang and Cai, 2014]	1	8	-	~0.4	~0.6	Both	Fixed	(10 x ? x 20)	-	-	13
	[Huang et al., 2014]	1	8	-	~0.8	~1.6	Single	Fixed	(? x ? x 20)	-	-	13
	Proposed [Arar and Thiran, 2017]	3	4	-	**0.99**	**1.27**	**Both**	**Free**	**(30 x ? x 20)**	**Yes**	**Yes**	**8**[6]

Table 5.7 – Comparison of existing eye tracking systems. In *"Cam(s)"* column, * indicates that a pan-tilt unit is employed. *"Calib."* column indicates whether explicit camera and scene geometry calibrations are required: "fully" means both are required, "pre" means the sensor is pre-calibrated. In *"Accuracy"*, *"SH"* and *"MH"* correspond to stable and moving head scenarios, respectively. The results refer to, unless stated otherwise, person-specific scenarios on within-dataset evaluations. *"HP"* column indicates whether users' head pose were fixed, e.g., using a chin rest. In *"FoV"* column, the systems' working volume is presented by *"FL"*, focal length in mm. The smaller the focal length, the larger the FoV.

5.6 Conclusion

In this chapter, we present a multi-camera gaze estimation framework to revisit the robustness concerns in eye tracking, particularly to head movements and eye glasses. We claim that instead of computing the user gaze from a single view as performed by previous work, leveraging multiple eye appearances simultaneously acquired from various views results in improved estimation accuracy and robustness under challenging real-world conditions. The main benefit of our approach is to more reliably detect gaze features under challenging conditions, particularly when they are obstructed due to large head pose or movements, or eye glasses effects. We further propose an adaptive fusion mechanism to effectively combine the gaze outputs obtained from multi-view appearances. To this effect, our mechanism firstly determines the estimation reliability of each gaze output according to user's general gazing behavior and momentary head pose, and then performs a reliability-based weighted fusion. Under large head movements and use of eye glasses, our evaluations show that the multi-camera approach improves the estimation performance of a single-camera setup by about 0.2-0.6° in estimation accuracy and 10-20% in estimation availability. The results also demonstrate that our approach is highly tolerant to illumination and eye color variations. In addition to the improved robustness to challenging conditions, the system's overall accuracy greatly benefits from the multi-camera setup under normal conditions. The proposed methodology provides about 30% improvement in accuracy, owing to the proposed adaptive fusion mechanism and estimation reliability algorithms.

[1]Person-independent (without user calibration) within-dataset evaluation on MPIIGaze dataset [Zhang et al., 2015].
[2]Person-independent (without user calibration) cross-dataset evaluation on MPIIGaze dataset [Zhang et al., 2015].
[3]Person-specific within-dataset evaluation on Eyediap dataset [Funes-Mora and Odobez, 2014].
[4]Person-specific within-dataset evaluation on GazeCapture dataset [Krafka et al., 2016].
[5]Chin rest is only used during the user calibration.
[6]8-mm lens indicates an individual camera's property. The overall combined FoV is significantly larger in the multi-camera setup.

6 Conclusions

In this thesis, we presented an end-to-end real-time eye tracking framework. We proposed innovative solutions that address the majority of the current limitations in eye tracking research. Owing to the benefits of the developed methods, the framework enables a fast, accurate, and robust gaze estimation using a flexible setup, which makes it suitable for a large spectrum of applications ranging from diagnostics (e.g., human behavior research, aids in neurological diagnosis, marketing research) to gaze-based human-computer interfaces (e.g., typing, controlling, navigation).

We designed a non-intrusive real-time eye tracking system using multiple remote cameras. Firstly, we addressed the setup complexity and flexibility as well as the real-time gaze processing. In this regard, we proposed to take advantage of a gaze estimation method that requires neither camera nor geometric system calibration, such as *cross ratio-based* gaze estimation. We obtained a flexible and adaptable setup since the method requires an uncalibrated setup. Besides, its computational simplicity easily enabled a real-time eye tracking.

Secondly, we addressed the user calibration in order to minimize the user effort while achieving a high tracking performance. We investigated various subject-specific estimation bias correction methods and identify the advantages and disadvantages of each. We then developed a novel method, which relies on weighted least squares regression. The proposed method achieves better generalization than the state-of-the-art user calibration methods, especially when the calibration data is limited in the size and quality. The developed methods enabled the system to operate under low-resolution eye data. Hence, we equipped the cameras with large field-of-view (FoV) lenses to enhance the system's working volume and allow for large head movements.

Furthermore, we revisited the major robustness concerns in eye tracking, including head movements, illumination variations, use of eye wear, and between-subject variations in eye type. In order to improve the estimation accuracy and tracking robustness, we proposed to leverage multiple eye appearances which are simultaneously acquired from various views. This enables to reliably detect the gaze features under challenging tracking conditions, particularly when they are obstructed in conventional single camera view appearance due to large head pose and position

changes, disturbances or occlusions caused by eye glasses. We also proposed an adaptive fusion mechanism to effectively combine the gaze outputs obtained from various views. The proposed mechanism firstly determines the estimation reliability of each gaze output according to several criteria, and then performs a weighted fusion of the reliable gaze outputs. As the developed user calibration and fusion methods enabled the system to operate under low-resolution eye data, we equipped the cameras with large FoV lenses to enhance the system's working volume and allow for large head movements. In addition, we designed the system to operate under active near-infrared (NIR) illumination, and developed illumination-robust feature detection algorithms in order to bring robustness to varying ambient illumination conditions.

In the following sections, we will summarize the contributions presented in this thesis and how we addressed the aforementioned challenges. We will further discuss the limitations of our framework together with future perspectives to address these limitations.

6.1 Concluding Remarks

In Chapter 3, we introduced a remote gaze estimation framework that addresses some of the main concerns in eye tracking. First of all, differently from the existing efforts, we designed a multi-camera setup, which comprises of simultaneously operating single-camera eye trackers. To achieve high setup flexibility as well as rapid gaze estimation, we proposed to employ a *cross ratio-based* gaze estimation method in each single-camera system, due to its particular advantages over the alternatives. Our approach firstly avoids a fully-calibrated hardware setup, so neither camera nor system calibration is required. This way it enables to easily perform modifications in the setup (e.g., number of cameras, their positioning, and data resolution) without requiring re-calibration of the whole setup. Therefore, the framework is highly flexible and can effortlessly be adapted for various application scenarios, from standard personal tracking to customized ones, such as for in-car driving and disabled aid scenarios. In addition, it provides fast gaze estimation as it relies on projective transformations. Thus, the gaze can be estimated within only a few milliseconds, such that very high estimation rates can be obtained. Besides, the tracking performance is less affected by the changes in eye data resolution since the estimation relies on approximations, on the contrary to *3D model-based* methods. Therefore, we equipped the cameras with large FoV lenses to allow for large head movements. This provides a large working volume to the overall setup due to multiple cameras' combined FoVs. On the negative side, being an approximated model results in a lower estimation accuracy and head movement tolerance in comparison to *3D model-based* methods. Yet, thanks to leveraging multiple gaze outputs, which enables improved estimation accuracy and robustness, our framework's overall tracking performance is competitive with those of *3D model-based* methods. It is also very important to note that the proposed multi-camera framework is independent of the employed gaze estimation algorithm. Thus, any other gaze estimation method discussed in Chapter 2 can be utilized depending on the application scenario, desired accuracy and robustness. For instance, a multi-camera system which uses an *appearance-based* gaze estimation method could significantly enhance the estimation accuracy and robustness.

Another important advantage of the proposed multi-camera eye tracking framework is that it significantly improves the tracking robustness to real-world conditions, as emphasized in Chapter 5. Instead of estimating the user gaze from a single view as performed by previous work, it leverages multiple eye appearances simultaneously acquired from various views. In conventional single-camera setups, there exists a single appearance, on which the gaze features may be obstructed due to challenging conditions, such as large head pose or movements, or occlusions caused by eye glasses. Whereas in our design, the main benefit is to simultaneously perform feature detection on multi-view appearances. For each frame, our approach enables to compute multiple gaze outputs using the detected gaze features. Furthermore, these gaze outputs are combined by an adaptive fusion mechanism to compute user's overall point of regard. In this context, the proposed mechanism first determines the estimation reliability of each gaze output according to our defined gaze reliability indicators, such as user's general gazing behavior and momentary head poses with respect to each camera. Then, it performs a reliability-based weighted fusion. This results in an improved overall estimation accuracy and robustness to head pose variations, large head movements, and eye glasses.

In this thesis, we also addressed the robustness to varying ambient illumination conditions. Firstly, as the eye appearance is highly affected by the variations in ambient illumination, we avoided a natural-light based eye tracking system. Instead, we designed a solution that uses active NIR illumination, so that the eye appearance remains similar when the ambient illumination alters. In addition, another factor which plays an essential role in illumination tolerance is to have a robust feature detection mechanism. Under varying illumination conditions, there occur significant changes in gaze features and eye shape. For instance, the pupil size and opening of the eyelids highly vary to adjust the amount of light entering the eye. So, the feature detection mechanisms must be capable of adapting to such changes. In this respect, we put a special emphasis on our feature detection paradigm. For glint detection, we took advantage of illumination-robust image processing techniques, such as spatial adaptive thresholding. More importantly, a dark-pupil based approach was preferred to detect the pupil instead of a bright-pupil based one, as performed by most of the related work. Although the bright-pupil based approach, which leverages an optical phenomenon by placing an additional light source in the optical axis of the camera, enables a high-contrast pupil region, our experiments demonstrated that it is less tolerant to the illumination variations, not to mention the bright pupil response's high sensitivity to user's pupil size, eye color, and ethnicity.

We also observed that placing an additional light source per camera, particularly in a multi-camera setting, influences the eye glasses robustness heavily due to the additional reflections caused by the increased number of light sources. In fact, using additional light sources may also bring complications regarding users' eye safety and increase the system's power consumption [Boucouvalas, 1996]. Hence, we believe that this framework is of high value to researchers as it can help to advance the development of more accurate and robust eye tracking for diverse scenarios, including those with less constrained conditions.

In Chapter 4, we addressed the user calibration convenience, one of the major challenges in eye

tracking. We developed a novel user calibration framework that requires a lower user effort to reach high estimation accuracies. To this effect, we first identified the potential drawbacks of the state-of-the-art user calibration methods, especially in relation to our framework's certain characteristics such as operating with low-resolution data and limited calibration data. We further carried out an extensive investigation of several regression techniques together with the widely accepted homography-based method in order to compensate for the subject-specific estimation bias. Our investigation showed that in comparison to homography mapping, affine mapping results in a better generalization when the calibration data is limited in size and quality, due to the reduced parameters. Besides, we identified that the quality of the calibration data is heterogeneous due to several factors, e.g., noise and outliers caused by feature detection flaws or user distractions. Consequently, we introduced weighted least squares regression-based approaches, in which individual calibration point clusters or samples have varying impacts in the overall regression according to the estimation reliabilities of samples/point clusters.

Lastly, as one of the main contributions of this thesis, we conducted extensive simulations and user experiments. In simulations, we examined the impact of increasing the number of cameras to very large numbers as well as their relative configurations. So, the trade-off between the tracking performance and setup complexity is highlighted. Moreover, we collected a multi-camera gaze dataset with an emphasis on natural and realistic human-computer interaction (HCI) scenarios, in which the subjects were asked to follow some conventional and newly introduced experimental scenarios. The dataset consists of 20 users, which includes subjects with diverse origins, eye types, and eye wears (eye glasses, contact lenses). The users performed eight experiments under varying illumination conditions and continuous and large head movements to examine the eye tracker's robustness to unconstrained real-world conditions. The data was collected as natural as possible. No chin rest was used to allow for natural continuous head movements. The test points were randomly displayed over the monitor to eliminate the calibration bias. Hence, we targeted to systematically isolate the main variables which have an impact on gaze estimation algorithms, such as the head pose, large head movements, illumination conditions, eye wear, and person-specific eye appearance. We believe that this dataset can contribute for a more structured, objective, and rich evaluation of gaze estimation algorithms [1].

6.2 Limitations & Future Perspectives

Although our current prototype system offers a simpler and more flexible setup in terms of sensor (or system) complexity, quality, and calibration in comparison with the existing high-accuracy eye trackers, it still employs multiple cameras and multiple light sources to reach a high tracking performance. This is one of the major limitations of our approach, particularly in comparison to *appearance-based* approaches. Nevertheless, the proposed multi-camera framework is independent of the employed gaze estimation algorithm and hardware. Therefore, the framework can be utilized with other gaze estimation techniques, which have lower requirements, e.g., *appearance-*

[1] We are currently working towards making the dataset publicly available.

based gaze estimation. In this respect, one of the future directions is to employ different gaze models to further improve the setup complexity and flexibility. For instance, we plan to employ an *appearance-based* or a *3D model-based* gaze estimation method in our framework to enhance the estimation accuracy and robustness of the state-of-the-art methods.

Similarly, although our framework addresses the majority of the challenges in eye tracking, the cost of the current implementation remains a major concern due to employed NIR-sensitive sensors, c-mount lenses, and band-pass filters (in total about $700 per camera setup). However, the total cost can further be reduced by customizing ordinary low-cost webcams, i.e., removing the IR filter and setting a trigger to work with our current setup. Hence, one of our future plans is to reduce the cost by using low-cost cameras.

In fact, utilizing regular webcams will also provide a solution for another of the limitations of our setup, that is to lose the image focus (sharpness) when the user moves towards or away from the setup. As the current setup employs manual-focus lenses, the focus is affected by the depth movements despite our efforts on adjusting the aperture to have a larger depth-of-field. This causes the features to appear more blurry, and so, the tracking performance is affected negatively. In this regard, employing aforementioned customized auto-focus webcams would bring additional robustness against depth changes.

Regarding the robustness to depth movements, one of the future work would be to perform explicit head movement compensations suggested in the literature. In this context, the methods proposed by [Coutinho and Morimoto, 2013] and [Huang et al., 2014] are good candidates. Among these, the method of [Huang et al., 2014], which adapts the estimation bias correction with respect to head movements, constitutes a better alternative as it does not require any additional on-axis light source.

In order to effectively combine the gaze outputs obtained from multiple camera systems, various adaptive fusion algorithms have been proposed, which provide improved tracking performance in comparison to non-adaptive solutions. Nonetheless, we believe that the tracking performance can further be enhanced by developing better weighting schemes. In this regard, in our future work, we plan to explore alternative methods, such as using different evaluation metrics (e.g., precision, gain, etc.) or estimating the weights directly from the feature detection process. More interestingly, rather than using statistical analyses, we plan to investigate whether a more optimal weighting can be learned from the simulated and/or real data using machine learning techniques. For example, inspired from the learning from simulation data idea in [Huang et al., 2014], adapting the weights according to the head movements can notably improve the tracking performance since the current weighting schemes do not explicitly consider the changes in head translations. Hence, simulating the head movements comprehensively along all three dimensions using the proposed multi-camera framework can enable to learn the correlation between the sensor weights and the variations in head movements. Consequently, the weights can be adapted with respect to the head movement while tracking in order to achieve better tracking performances.

For gazing behaviour-based fusion, the current weight maps are generated using two weighting indicators, namely, the calibration accuracy and estimation availability, as described in Section 5.2.2. Yet, other alternatives can be employed towards better modeling of users' gazing behaviours. For example, the estimation precision, which is the ability to reliably reproduce the same estimation for a target calibration point, or the histogram of the best performing sensor, which stores the information about how often each sensor achieves the best estimation for a target calibration point, could be effective weighting indicators. Such indicators can in fact provide complementary evidence for the estimation reliability. In addition, the weights obtained from head pose-based scheme can be combined with the ones from gazing behaviour-based scheme. Hence, the overall weights can more robustly be determined using multiple indicators. We have already started to explore such alternatives, and plan to complete their validations in our future work.

Furthermore, although our framework brings robustness to eye glasses and variations in user's eye type, it still experiences inaccuracy under certain conditions. For instance, when users wear eye glasses with special coatings (e.g., NIR blocking) or thick lenses, the features are heavily disturbed or obstructed due to refraction and blocking effects. In addition, feature detection can be very challenging when the opening of eyes are very small (e.g., sleepy or tired users, Asian's eye shapes). In such cases, the gaze features are either not visible or weakly visible, particularly under low-resolution. In order to address this, at least for the latter case, one interesting solution would be to take advantage of super-resolution techniques. Since the user is viewed from many cameras, super-resolution can be achieved with multiple inputs acquired from the cameras. In fact, arrays of inexpensive cameras can be employed so as to enable high performance imaging, as in [Wilburn, 2004, Carles et al., 2014]. This enables gaze features to be detected more reliably and results in a higher tracking performance in many aspects.

One of our most essential future perspectives is to develop a new multi-camera user calibration framework. In the current framework, we model the subject-specific estimation bias correction separately in each camera and each eye. Although it currently enables a high performance, we believe that it could be significantly improved. In this regard, we plan to learn the calibration model simultaneously across multiple cameras and eyes. More specifically, we have a high confidence that the estimation bias can be very efficiently and robustly modeled by leveraging the multi-binocular constraints. In this perspective, the method proposed by [Zhang and Cai, 2014] would constitute a good starting point, and can be extended for multi-cameras.

Last but not least, our current implementation runs at 30 fps on a regular PC with Intel i7 3.2 GHz processor. We believe that higher frame rates can be achieved without much effort. In this context, we plan to replace the current face tracker, which is the most computationally expensive process in the whole framework with ~24 ms per frame, with simpler or faster face or eye region trackers, such as based on local binary features (LBF) [Ren et al., 2014], or one millisecond face alignment with an ensemble of regression trees [Kazemi and Sullivan, 2014]. As the feature extraction process does not require precisely located facial landmarks, such a modification will not significantly affect the overall performance. In this case, the system's frame rate will solely

depend on the hardware constraints, e.g., camera frame rate, total bandwidth in data transfer using USB 3.0.

Bibliography

[Alghowinem et al., 2014] Alghowinem, S., AlShehri, M., Goecke, R., and Wagner, M. (2014). *Exploring Eye Activity as an Indication of Emotional States Using an Eye-Tracking Sensor*, pages 261–276. Springer International Publishing, Cham.

[Alnajar et al., 2013] Alnajar, F., Gevers, T., Valenti, R., and Ghebreab, S. (2013). Calibration-Free Gaze Estimation Using Human Gaze Patterns. In *2013 IEEE Int. Conf. Comput. Vis.*, pages 137–144.

[Arar et al., 2012] Arar, N. M., Gao, H., Ekenel, H. K., and Akarun, L. (2012). Selection and combination of local gabor classifiers for robust face verification. In *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 297–302.

[Arar et al., 2015a] Arar, N. M., Gao, H., and Thiran, J. P. (2015a). Robust gaze estimation based on adaptive fusion of multiple cameras. In *2015 11th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*.

[Arar et al., 2015b] Arar, N. M., Gao, H., and Thiran, J. P. (2015b). Towards convenient calibration for cross-ratio based gaze estimation. In *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 642–648.

[Arar et al., 2016a] Arar, N. M., Gao, H., and Thiran, J. P. (2016a). A regression-based user calibration framework for real-time gaze estimation. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*.

[Arar et al., 2017a] Arar, N. M., Pati, P., Kashyap, A., Khartchenko, A. F., Goksel, O., Kaigala, G. V., and Gabrani, M. (2017a). Computational immunohistochemistry: Recipes for standardization of immunostaining. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.

[Arar et al., 2017b] Arar, N. M., Pati, P., Kashyap, A., Khartchenko, A. F., Goksel, O., Kaigala, G. V., and Gabrani, M. (2017b). Standardization of immunostaining using automated quantification of staining quality. *IEEE Transactions on Medical Imaging (TMI)*.

[Arar and Thiran, 2016] Arar, N. M. and Thiran, J.-P. (2016). Estimating fusion weights of a multi-camera eye tracking system by leveraging user calibration data. In *Proceedings of the*

Ninth Biennial ACM Symposium on Eye Tracking Research & Applications, ETRA '16, pages 225–228.

[Arar and Thiran, 2017] Arar, N. M. and Thiran, J.-P. (2017). Robust Real-Time Multi-Camera Eye Tracking. *(under review)*.

[Arar et al., 2016b] Arar, N. M., Thiran, J.-P., and Theytaz, O. (2016b). Eye gaze tracking system and method. US Patent 9,411,417.

[Babcock and Pelz, 2004] Babcock, J. S. and Pelz, J. B. (2004). Building a lightweight eyetracking headgear. In *Proc. Symp. Eye Track. Res. Appl. - ETRA'2004*, pages 109–114.

[Baluja and Pomerleau, 1994] Baluja, S. and Pomerleau, D. (1994). Non-Intrusive Gaze Tracking Using Artificial Neural Networks. Technical report, CMU-CS-94-102.

[Bartels and Marshall, 2012] Bartels, M. and Marshall, S. P. (2012). Measuring cognitive workload across different eye tracking hardware platforms. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, pages 161–164.

[Betke et al., 2002] Betke, M., Gips, J., and Fleming, P. (2002). The camera mouse: visual tracking of body features to provide computer access for people with severe disabilities. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 10(1):1–10.

[Beymer and Flickner, 2003] Beymer, D. and Flickner, M. (2003). Eye gaze tracking using an active stereo head. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2(2):II/451–II/458.

[Birchfield, 1998] Birchfield, S. (1998). An introduction to projective geometry (for computer vision). *Unpubl. note, Stanford Univ.*, 141(3569):1–22.

[Böhme et al., 2008] Böhme, M., Dorr, M., Graw, M., Martinetz, T., and Barth, E. (2008). A software framework for simulating eye trackers. In *Proc. Symp. Eye Track. Res. Appl. - ETRA '08*.

[Borji and Itti, 2013] Borji, A. and Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):185–207.

[Boucouvalas, 1996] Boucouvalas, A. (1996). Iec 825-1 eye safety classification of some consumer electronic products. *IET Conference Proceedings*, pages 13–13(1).

[Bradley et al., 2008] Bradley, M. M., Miccoli, L., Escrig, M. A., and Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607.

[Carles et al., 2014] Carles, G., Downing, J., and Harvey, A. R. (2014). Super-resolution imaging using a camera array. *Opt. Lett.*, 39(7):1889.

[Cerrolaza et al., 2008] Cerrolaza, J. J., Villanueva, A., and Cabeza, R. (2008). Taxonomic study of polynomial regressions applied to the calibration of video-oculographic systems. In *Proc. Symp. Eye Track. Res. Appl. - ETRA '08*.

[Cerrolaza et al., 2012] Cerrolaza, J. J., Villanueva, A., Villanueva, M., and Cabeza, R. (2012). Error characterization and compensation in eye tracking systems. *Proc. Symp. Eye Track. Res. Appl.*, 1(212):205–208.

[Chen and Ji, 2015] Chen, J. and Ji, Q. (2015). A probabilistic approach to online eye gaze tracking without explicit personal calibration. *IEEE Trans. Image Process.*, 24(3):1076–1086.

[Chen et al., 2012] Chen, S., Wu, C., Lin, S., and Hung, Y. (2012). 2d face alignment and pose estimation based on 3d facial models. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pages 128–133.

[Cootes et al., 2001] Cootes, T., Edwards, G., and Taylor, C. (2001). Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 23(6):681–685.

[Coutinho and Morimoto, 2006] Coutinho, F. and Morimoto, C. (2006). Free head motion eye gaze tracking using a single camera and multiple light sources. In *2006 19th Brazilian Symp. Comput. Graph. Image Process.*, pages 171–178.

[Coutinho and Morimoto, 2010] Coutinho, F. L. and Morimoto, C. H. (2010). A depth compensation method for cross-ratio based eye tracking. In *Proc. Symp. Eye Track. Res. Appl. - ETRA '10*, page 137.

[Coutinho and Morimoto, 2012] Coutinho, F. L. and Morimoto, C. H. (2012). Augmenting the robustness of cross-ratio gaze tracking methods to head movement. In *Proc. Symp. Eye Track. Res. Appl. - ETRA '12*, page 59.

[Coutinho and Morimoto, 2013] Coutinho, F. L. and Morimoto, C. H. (2013). Improving Head Movement Tolerance of Cross-Ratio Based Eye Trackers. *Int. J. Comput. Vis.*, 101(3):459–481.

[Duchowski, 2000] Duchowski, A. (2000). Eye-based interaction in graphical systems: Theory & practice.

[Duchowski, 2002] Duchowski, A. (2002). A breadth-first survey of eye-tracking applications. *Behav. Res. Methods, Instruments, Comput.*, 34(4):455–470.

[Duchowski, 2007] Duchowski, A. T. (2007). *Eye tracking methodology*. Springer.

[Ebisawa, 1998] Ebisawa, Y. (1998). Improved video-based eye-gaze detection method. *IEEE Trans. Instrum. Meas.*, 47(4):948–955.

[Eckstein et al., 2016] Eckstein, M. K., Guerra-Carrillo, B., Singley, A. T. M., and Bunge, S. A. (2016). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, pages –.

[Eriksson and Papanikotopoulos, 1997] Eriksson, M. and Papanikotopoulos, N. P. (1997). Eye-tracking for detection of driver fatigue. In *Proceedings of Conference on Intelligent Transportation Systems*, pages 314–319.

[Fakhro et al., 2015] Fakhro, A., Yim, H. W., Kim, Y. K., and Nguyen, A. H. (2015). The evolution of looks and expectations of asian eyelid and eye appearance. *Seminars in Plastic Surgery*, 29(3):135–144.

[Filik et al., 2017] Filik, R., Brightman, E., Gathercole, C., and Leuthold, H. (2017). The emotional impact of verbal irony: Eye-tracking evidence for a two-stage process. *Journal of Memory and Language*, 93:193 – 202.

[Fletcher and Zelinsky, 2009] Fletcher, L. and Zelinsky, A. (2009). Driver inattention detection based on eye gaze-road event correlation. *Int. J. Rob. Res.*, 28(6):774–801.

[Funes Mora, 2015] Funes Mora, K. A. (2015). *3D Gaze Estimation from Remote RGB-D Sensors*. PhD thesis, École Polytechnique Fédérale de Lausanne. Thèse EPFL, no 6680.

[Funes-Mora et al., 2014] Funes-Mora, K. A., Monay, F., and Odobez, J.-M. (2014). EYEDIAP: A Database for the Development and Evaluation of Gaze Estimation Algorithms from RGB and RGB-D Cameras. In *Proc. Symp. Eye Track. Res. Appl. - ETRA '14*, pages 255–258.

[Funes-Mora and Odobez, 2012] Funes-Mora, K. A. and Odobez, J.-M. (2012). Gaze estimation from multimodal Kinect data. In *2012 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, pages 25–30.

[Funes-Mora and Odobez, 2014] Funes-Mora, K. A. and Odobez, J.-M. (2014). Geometric Generative Gaze Estimation (G 3 E) for Remote RGB-D Cameras. In *2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1773–1780.

[Gatica-Perez et al., 2005] Gatica-Perez, D., McCowan, I. A., Zhang, D., and Bengio, S. (2005). Detecting group interest-level in meetings. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*.

[Guestrin and Eizenman, 2006] Guestrin, E. and Eizenman, M. (2006). General Theory of Remote Gaze Estimation Using the Pupil Center and Corneal Reflections. *IEEE Trans. Biomed. Eng.*, 53(6):1124–1133.

[Guestrin and Eizenman, 2007] Guestrin, E. D. and Eizenman, M. (2007). Remote Point-of-Gaze Estimation with Free Head Movements Requiring a Single-Point Calibration. In *2007 29th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pages 4556–4560.

[Güney et al., 2013] Güney, F., Arar, N. M., Fischer, M., and Ekenel, H. K. (2013). Cross-pose facial expression recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6.

[Hansen et al., 2002] Hansen, D., Hansen, J., Nielsen, M., Johansen, A., and Stegmann, M. (2002). Eye typing using Markov and active appearance models. In *Sixth IEEE Work. Appl. Comput. Vision, 2002. (WACV 2002). Proceedings.*, volume 2002-Janua, pages 132–136. IEEE Comput. Soc.

[Hansen et al., 2010] Hansen, D. W., Agustin, J. S., and Villanueva, A. (2010). Homography normalization for robust gaze estimation in uncalibrated setups. In *Proc. Symp. Eye Track. Res. Appl. - ETRA '10*.

[Hansen and Hansen, 2006] Hansen, D. W. and Hansen, J. P. (2006). Eye typing with common cameras. In *Proc. 2006 Symp. Eye Track. Res. Appl. - ETRA '06*, page 55.

[Hansen and Ji, 2010] Hansen, D. W. and Ji, Q. (2010). In the eye of the beholder: a survey of models for eyes and gaze. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 32(3):478–500.

[Hansen and Pece, 2005] Hansen, D. W. and Pece, A. E. C. (2005). Eye tracking in the wild. *Comput. Vis. Image Underst.*, 98(1):155–181.

[Hartley and Zisserman, 2005] Hartley, R. and Zisserman, A. (2005). *Multiple view geometry in computer vision*. Cambridge University Press.

[Hennessey et al., 2006] Hennessey, C., Noureddin, B., and Lawrence, P. (2006). A single camera eye-gaze tracking system with free head motion. *Measurement*, 1(March):27–29.

[Hoerl and Kennard, 1970] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

[Hortensius et al., 2014] Hortensius, R., van Honk, J., de Gelder, B., and D., T. (2014). Trait dominance promotes reflexive staring at masked angry body postures. *PLoS ONE*, 9(12).

[Huang et al., 2014] Huang, J.-B., Cai, Q., Liu, Z., Ahuja, N., and Zhang, Z. (2014). Towards accurate and robust cross-ratio based gaze trackers through learning from simulation. In *Proc. Symp. Eye Track. Res. Appl. - ETRA '14*, pages 75–82.

[Huey, 1908] Huey, E. (1908). *The Psychology and Pedagogy of Reading*. "The MIT Press.

[Hutchinson et al., 1989] Hutchinson, T. E., White, K. P., Martin, W. N., Reichert, K. C., and Frey, L. A. (1989). Human-computer interaction using eye-gaze input. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(6):1527–1534.

[Javal, 1878] Javal, E. (1878). Essai sur la physiologie de la lecture. *Annales d'Ocilistique*, 80:61–73.

[Ji and Yang, 2002] Ji, Q. and Yang, X. (2002). Real-Time Eye, Gaze, and Face Pose Tracking for Monitoring Driver Vigilance. *Real-Time Imaging*, 8(5):357–377.

[Just and Carpenter, 1980] Just, M. A. and Carpenter, P. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

[Kang et al., 2008] Kang, J., Eizenman, M., Guestrin, E., and Eizenman, E. (2008). Investigation of the Cross-Ratios Method for Point-of-Gaze Estimation. *IEEE Trans. Biomed. Eng.*, 55(9):2293–2302.

[Kang et al., 2007] Kang, J. J., Guestrin, E. D., Maclean, W. J., and Eizenman, M. (2007). Simplifying the Cross-Ratios Method of Point-of-Gaze Estimation. In *30th Can. Med. Biol. Eng. Conf.*, pages 1–4.

[Kar-Han Tan et al., 2002] Kar-Han Tan, Kriegman, D., and Ahuja, N. (2002). Appearance-based eye gaze estimation. In *Sixth IEEE Work. Appl. Comput. Vision, 2002. (WACV 2002). Proceedings.*, pages 191–195. IEEE Comput. Soc.

[Kazemi and Sullivan, 2014] Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1867–1874.

[Kiranantawat et al., 2015] Kiranantawat, K., Suhk, J., and Nguyen, A. (2015). The asian eyelid: Relevant anatomy. *Seminars in Plastic Surgery*, 29(3):158–164.

[Krafka et al., 2016] Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., and Torralba, A. (2016). Eye Tracking for Everyone. In *2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2176–2184.

[Kübler et al., 2016] Kübler, T. C., Rittig, T., Kasneci, E., Ungewiss, J., and Krauss, C. (2016). Rendering refraction and reflection of eyeglasses for synthetic eye tracker images. In *Proc. Symp. Eye Track. Res. Appl. - ETRA '16*, pages 143–146.

[Kuo et al., 2014] Kuo, Y.-L., Lee, J.-S., and HsiehLee, M.-C. (2014). Video-based eye tracking to detect the attention shift: A computer classroom context-aware system. *International Journal of Distance Education Technologies*, 12(4):66–81.

[Lai et al., 2015] Lai, C. C., Shih, S. W., and Hung, Y. P. (2015). Hybrid method for 3-D gaze tracking using glint and contour features. *IEEE Trans. Circuits Syst. Video Technol.*, 25(1):24–37.

[Levine, 1981] Levine, J. (1981). *An Eye-controlled Computer*. RC 8857. IBM Research Division, T.J. Watson Research Center.

[Levy et al., 2010] Levy, D. L., Sereno, A. B., Gooding, D. C., and O'Driscoll, G. A. (2010). Eye tracking dysfunction in schizophrenia: Characterization and pathophysiology. *Current Topics in Behavioral Neurosciences*, 4:311–347.

[Lu et al., 2011] Lu, F., Sugano, Y., Okabe, T., and Sato, Y. (2011). Inferring human gaze from appearance via adaptive linear regression. In *2011 Int. Conf. Comput. Vis.*, pages 153–160.

[Lu et al., 2012] Lu, F., Sugano, Y., Okabe, T., and Sato, Y. (2012). Head pose-free appearance-based gaze sensing via eye image synthesis. In *2012 20th Int. Conf. Pattern Recognit.*, pages 1008–1011.

[Lu et al., 2014] Lu, F., Sugano, Y., Okabe, T., and Sato, Y. (2014). Adaptive Linear Regression for Appearance-Based Gaze Estimation. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, 36(10):2033–2046.

[Lu et al., 2015] Lu, F., Sugano, Y., Okabe, T., and Sato, Y. (2015). Gaze Estimation From Eye Appearance: A Head Pose-Free Method via Eye Image Synthesis. *IEEE Trans. Image Process.*, 24(11):3680–3693.

[Majaranta, 2011] Majaranta, P. (2011). *Gaze Interaction and Applications of Eye Tracking: Advances in Assistive Technologies: Advances in Assistive Technologies*. IGI Global.

[Merchant et al., 1974] Merchant, J., Morrissette, R., and Porterfield, J. L. (1974). Remote Measurement of Eye Direction Allowing Subject Motion Over One Cubic Foot of Space. *IEEE Trans. Biomed. Eng.*, BME-21(4):309–317.

[Mora and Odobez, 2016] Mora, K. A. and Odobez, J.-M. (2016). Gaze Estimation in the 3D Space Using RGB-D Sensors. *Int. J. Comput. Vis. (IJCV)*, 118(2):194–216.

[Morgan and Rose, 2005] Morgan, I. and Rose, K. (2005). How genetic is school myopia? *Prog. Retin. Eye Res.*, 24(1):1–38.

[Morimoto et al., 2000] Morimoto, C., Koons, D., Amir, A., and Flickner, M. (2000). Pupil detection and tracking using multiple light sources. *Image Vis. Comput.*, 18(4):331–335.

[Morimoto and Mimica, 2005] Morimoto, C. H. and Mimica, M. R. M. (2005). Eye gaze tracking techniques for interactive applications. *Comput. Vis. Image Underst.*, 98(1):4–24.

[Nagamatsu et al., 2011] Nagamatsu, T., Sugano, R., Iwamoto, Y., Kamahara, J., and Tanaka, N. (2011). User-calibration-free gaze estimation method using a binocular 3D eye model. *IEICE Trans. Inf. Syst.*, E94-D(9):1817–1829.

[Nguyen et al., 2002] Nguyen, K., Wagner, C., Koons, D., and Flickner, M. (2002). Differences in the infrared bright pupil response of human eyes. In *Proc. Symp. Eye Track. Res. Appl. - ETRA '02*, page 133.

[Noris et al., 2011] Noris, B., Keller, J.-B., and Billard, A. (2011). A wearable gaze tracking system for children in unconstrained environments. *Comput. Vis. Image Underst.*, 115(4):476–486.

[Ohno and Mukawa, 2004] Ohno, T. and Mukawa, N. (2004). A free-head, simple calibration, gaze tracking system that enables gaze-based interaction. In *Proc. Symp. Eye Track. Res. Appl. - ETRA'2004*, pages 115–122.

[Palinko et al., 2010] Palinko, O., Kun, A. L., Shyrokov, A., and Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, ETRA '10, pages 141–144.

[Park, 2007] Park, K. R. (2007). A real-time gaze position estimation method based on a 3-D eye model. *IEEE Trans. Syst. Man. Cybern. B. Cybern.*, 37(1):199–212.

[Pärnamets et al., 2015] Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J., and Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences*, 112(13):4170–4175.

[Rasmussen and Williams, 2006] Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*. "The MIT Press.

[Reale et al., 2011] Reale, M. J., Canavan, S., Yin, L., Hu, K., and Hung, T. (2011). A Multi-Gesture Interaction System Using a 3-D Iris Disk Model for Gaze Estimation and an Active Appearance Model for 3-D Hand Pointing. *IEEE Trans. Multimed.*, 13(3):474–486.

[Ren et al., 2014] Ren, S., Cao, X., Wei, Y., and Sun, J. (2014). Face Alignment at 3000 FPS via Regressing Local Binary Features. In *2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1685–1692.

[Rosipal and Krämer, 2006] Rosipal, R. and Krämer, N. (2006). Overview and recent advances in partial least squares. In *Conference on Subspace, Latent Structure and Feature Selection*.

[Salvucci and Anderson, 2000] Salvucci, D. D. and Anderson, J. R. (2000). Intelligent gaze-added interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pages 273–280.

[Saragih et al., 2011] Saragih, J., Lucey, S., and Cohn, J. (2011). Deformable model fitting by regularized landmark mean-shift. *Int. J. Comput. Vis. (IJCV)*, 91(2):200–215.

[Schaeffel, 2006] Schaeffel, F. (2006). Myopia: The Importance of Seeing Fine Detail. *Curr. Biol.*, 16(7):R257–R259.

[Sesma-sanchez et al., 2012] Sesma-sanchez, L., Villanueva, A., and Cabeza, R. (2012). Gaze Estimation Interpolation Methods Based on Binocular Data. *IEEE Trans. Biomed. Eng.*, 59(8):2235–2243.

[Shih and Liu, 2004] Shih, S.-W. and Liu, J. (2004). A Novel Approach to 3-D Gaze Tracking Using Stereo Cameras. *IEEE Trans. Syst. Man Cybern. Part B*, 34(1):234–245.

[Smith et al., 2013] Smith, B. A., Yin, Q., Feiner, S. K., and Nayar, S. K. (2013). Gaze locking: Passive eye contact detection for human-object interaction. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, pages 271–280.

[Sugano et al., 2010] Sugano, Y., Matsushita, Y., and Sato, Y. (2010). Calibration-free gaze sensing using saliency maps. In *2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 2667–2674.

[Sugano et al., 2014] Sugano, Y., Matsushita, Y., and Sato, Y. (2014). Learning-by-Synthesis for Appearance-Based 3D Gaze Estimation. In *2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1821–1828.

[Sun et al., 2015] Sun, L., Liu, Z., and Sun, M. T. (2015). Real time gaze estimation with a consumer depth camera. *Inf. Sci. (Ny).*, 320:346–360.

[Sun et al., 2014] Sun, L., Song, M., Liu, Z., and Sun, M.-t. (2014). Real-Time Gaze Estimation with Online Calibration. *IEEE Multimed.*, 21(4):28–37.

[Terburg et al., 2011] Terburg, D., Hooiveld, N., Aarts, H., Kenemans, J. L., and van Honk, J. (2011). Eye tracking unconscious face-to-face confrontations: Dominance motives prolong gaze to masked angry faces. *Psychological Science*, 22(3):314–319.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

[Ting et al., 2014] Ting, W. K.-C., Perez Velazquez, J. L., and Cusimano, M. D. (2014). Eye movement measurement in diagnostic assessment of disorders of consciousness. *Frontiers in Neurology*, 5:137.

[Tinsley et al., 2016] Tinsley, J. N., Molodtsov, M. I., Prevedel, R., Wartmann, D., Espigulé-Pons, J., Lauwers, M., and Vaziri, A. (2016). Direct detection of a single photon by humans. *Nature Communications*, 7.

[Topal et al., 2014] Topal, C., Gunal, S., Kocdeviren, O., Dogan, A., and Gerek, O. N. (2014). A Low-Computational Approach on Gaze Estimation With Eye Touch System. *IEEE Trans. Cybern.*, 44(2):228–239.

[Tseng et al., 2013] Tseng, P.-H., Cameron, I. G. M., Pari, G., Reynolds, J. N., Munoz, D. P., and Itti, L. (2013). High-throughput classification of clinical populations from natural viewing eye movements. *Journal of Neurology*, 260(1):275–284.

[Underwood, 2005] Underwood, G. (2005). *Cognitive processes in eye guidance*. Oxford University Press.

[Utsumi et al., 2012] Utsumi, A., Okamoto, K., Hagita, N., and Takahashi, K. (2012). Gaze tracking in wide area using multiple camera observations. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*.

[Valenti et al., 2012] Valenti, R., Sebe, N., and Gevers, T. (2012). What Are You Looking at? *Int. J. Comput. Vis.*, 98(3):324–334.

[Valuch et al., 2015] Valuch, C., Pflüger, L. S., Wallner, B., Laeng, B., and Ansorge, U. (2015). Using eye tracking to test for individual differences in attention to attractive faces. *Frontiers in Psychology*, 6:42.

[Villanueva and Cabeza, 2007] Villanueva, A. and Cabeza, R. (2007). Models for gaze tracking systems. *Eurasip J. Image Video Process.*, 2007:1–16.

[Villanueva and Cabeza, 2008] Villanueva, A. and Cabeza, R. (2008). A novel gaze estimation system with one calibration point. *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, 38(4):1123–1138.

[Viola and Jones, 2004] Viola, P. and Jones, M. (2004). Robust real-time face detection. *Int. J. Comput. Vis. (IJCV)*, 57:137–154.

[Wells et al., 2016] Wells, L. J., Gillespie, S. M., and Rotshtein, P. (2016). Identification of emotional facial expressions: Effects of expression, intensity, and sex on eye gaze. *PLOS ONE*, 11(12):1–20.

[White et al., 1993] White, K., Hutchhinson, T., and Carley, J. (1993). Spatially Dynamic Calibration of an Eye-Tracking System. *IEEE Trans. Syst. Man Cybern.*, 23(4):1162–1168.

[Wilburn, 2004] Wilburn, B. (2004). *High Performance Imaging Using Inexpensive Arrays of Cameras*. PhD thesis, Stanford University.

[Wood et al., 2016a] Wood, E., Baltrušaitis, T., Morency, L.-P., Robinson, P., and Bulling, A. (2016a). A 3D Morphable Eye Region Model for Gaze Estimation. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *ECCV*, volume 9905 of *Lecture Notes in Computer Science*, pages 297–313. Springer International Publishing, Cham.

[Wood et al., 2016b] Wood, E., Baltrušaitis, T., Morency, L.-P., Robinson, P., and Bulling, A. (2016b). Learning an appearance-based gaze estimator from one million synthesised images. In *ACM Symp. Eye Track. Res. Appl.*, pages 131–138.

[Xiong et al., 2014] Xiong, C., Huang, L., and Liu, C. (2014). Gaze Estimation Based on 3D Face Structure and Pupil Centers. In *2014 22nd Int. Conf. Pattern Recognit.*, pages 1156–1161.

[Xiong and De la Torre, 2013] Xiong, X. and De la Torre, F. (2013). Supervised Descent Method and Its Applications to Face Alignment. In *2013 IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 532–539.

[Xu et al., 1998] Xu, L.-Q., Machin, D., and Sheppard, P. (1998). A Novel Approach to Real-time Non-intrusive Gaze Finding. In *Procedings Br. Mach. Vis. Conf. 1998*, pages 43.1–43.10. British Machine Vision Association.

[Yoo and Chung, 2005] Yoo, D. H. and Chung, M. J. (2005). A novel non-intrusive eye gaze estimation using cross-ratio under large head motion. *Comput. Vis. Image Underst.*, 98(1):25–51.

[Yoo et al., 2002] Yoo, D. H., Kim, J. H., Lee, B. R., and Chung, M. J. (2002). Non-contact eye gaze tracking system by mapping of corneal reflections. In *Proc. Fifth IEEE Int. Conf. Autom. Face Gesture Recognit.*, pages 101–106.

[Young and Sheena, 1975] Young, L. R. and Sheena, D. (1975). Survey of eye movement recording methods. *Behav. Res. Methods Instrum.*, 7(5):397–429.

[Yüce et al., 2013] Yüce, A., Arar, N. M., and Thiran, J.-P. (2013). Multiple local curvature gabor binary patterns for facial action recognition. In *Human Behavior Understanding*, pages 136–147.

[Zhang et al., 2015] Zhang, X., Sugano, Y., Fritz, M., and Bulling, A. (2015). Appearance-based gaze estimation in the wild. In *2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, volume 07-12-June, pages 4511–4520.

[Zhang and Cai, 2014] Zhang, Z. and Cai, Q. (2014). Improving cross-ratio-based eye tracking techniques by leveraging the binocular fixation constraint. In *Proc. Symp. Eye Track. Res. Appl. - ETRA '14*, pages 267–270.

[Zhu and Ji, 2004] Zhu, Z. and Ji, Q. (2004). Eye and gaze tracking for interactive graphic display. *Mach. Vis. Appl.*, 15(3):139–148.

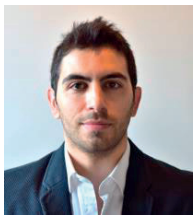
[Zhu and Ji, 2007] Zhu, Z. and Ji, Q. (2007). Novel Eye Gaze Tracking Techniques Under Natural Head Movement. *IEEE Trans. Biomed. Eng.*, 54(12):2246–2260.

[Zhu et al., 2006] Zhu, Z., Ji, Q., and Bennett, K. P. (2006). Nonlinear eye gaze mapping function estimation via support vector regression. *Proc. - Int. Conf. Pattern Recognit.*, 1:1132–1135.

Nuri Murat Arar

Avenue de Florimont 13, 1006 Lausanne
+41 (0)78 944 02 64
muratnarar@gmail.com
ch.linkedin.com/in/nuri-murat-arar

21.01.1988
Single
Turkish



I am a PhD candidate with a strong R&D experience. My research interests include computer vision, image processing, and machine learning. I am a passionate researcher and an adaptable engineer with strong technical skills.

PROFESSIONAL EXPERIENCE

École polytechnique fédérale de Lausanne (EPFL) & Logitech Europe SA — **Lausanne, Switzerland**
Research Assistant @ Signal Processing Lab (LTS5) — 09/2012-Present

- Designed & developed (in C++) a robust real-time multi-camera eye tracking system using computer vision & machine learning techniques, in close collaboration with **Logitech**.
  - Led and managed the project, conducted research, developed a complete C++ library.
  - Granted one patent & published several articles in prestigious conferences and journals.
- Involved actively in an R&D project in collaboration with **PSA Peugeot-Citroën** for improving human-computer interaction in cars such as driver's facial expression, fatigue, and stress detection.
- Supervised various student projects related to gaze estimation and facial image analysis.

IBM Research Zurich — **Zurich, Switzerland**
Research Intern @ Cognitive Computing Group — 06/2016-12/2016

- Designed & developed (in Python) a novel framework for digital pathology to quantitatively assess the quality of IHC-stained human tissue images using machine learning & image processing techniques.
  - Led the project, conducted research, applied and validated the framework on IHC-stained breast cancer tissues, developed a Python library, supervised an ETHZ master thesis.
  - The framework is currently in active use at **IBM Research**.
  - Filed one patent application & authored two papers in top-tier biomedical conferences and journals.

Karlsruhe Institute of Technology (KIT) — **Karlsruhe, Germany**
Research Intern @ Facial Image Processing and Analysis (FIPA) Group — 06/2011-10/2011

- Designed & developed (in C++) an illumination and occlusion robust face recognition framework.
  - Achieved the highest verification rate (94.3% at 0.1% FAR) at the time in the FRGC challenge.
  - Published a paper in a top biometrics conference.

Bogazici University — **Istanbul, Turkey**
Research Assistant @Perceptual Intelligence Lab (PILAB) — 06/2010-08/2012

- Designed & developed (in C++) a novel robust face recognition framework, a real-time face swapping application for the industry. Involved actively in a cross-view facial expression recognition project.
  - Published several scientific papers in prestigious international and national conferences.

AYESAS Incorporation — **Ankara, Turkey**
Software Development Intern — 06/2009-09/2009

- Developed embedded avionics software (in C) for an aircraft modernization project.

EDUCATION

École polytechnique fédérale de Lausanne (EPFL) — **Lausanne, Switzerland**
PhD in Computer Vision — 09/2012-Present

Karlsruhe Institute of Technology (KIT) — **Karlsruhe, Germany**
Erasmus MSc Program — 06/2011-10/2011

Bogazici University — **Istanbul, Turkey**
MSc in Computer Engineering (GPA: 3.81/4.00, Class Rank: Top 5%) — 06/2010-08/2012

Bilkent University — **Ankara, Turkey**
BSc in Computer Science (GPA: 3.61/4.00, Class Rank: Top 10%) — 09/2005-06/2010

SKILLS

Languages: English (advanced - C2), French (intermediate - B1), Turkish (native)

Programming: C, C++ (8+ years), MATLAB (6+ years), Python & Java (3+ years)

Tools, Libraries: STL, OpenMP, OpenCV, Numpy, Scipy, Scikit-learn, CMake, Bash, LaTeX

Deep Learning: Good theoretical knowledge and understanding, basic hands-on experience on TensorFlow & Theano

AWARDS & HONORS

- UCLA-IPAM Graduate Summer School Grant — University of California Los Angeles, 2013
- Erasmus Research Internship Grant — Erasmus Programme, 2011
- Dean's High Honor List — Bilkent & Bogazici University, 2006-2012
- Graduate Scholarship (tuition + monthly stipend) — TUBITAK, 2010-2012
- Undergraduate Scholarship (tuition + monthly stipend) — Bilkent University, 2005-2010
- Ranked 301th among 1.8 million students — Turkish University Entrance Exam, 2005
- Silver Medal — National Mathematics Olympics, 2002

PATENTS

- **"An automated method for process parameter optimization for tissue section immunostaining"** (primary inventor) *Patent(s) to be filed in US Patent Office*, 2017.
- **"Eye gaze tracking system and method"** (primary inventor) *US Patent #9,411,417* Granted in August 9, 2016.

SELECTED RECENT PUBLICATIONS

- **NM Arar** and J-Ph Thiran, **"Robust Real-Time Multi-Camera Eye Tracking"**, submitted for a journal publication*, 2017. (under review)*
- **NM Arar**, P Pati, A Kashyap, AF Khartchenko, O Yuksel, GV Kaigala and M Gabrani, **"Computational Immunohistochemistry: Recipes for Standardization of Immunostaining"**, *in Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2017.*
- **NM Arar**, H Gao and J-Ph Thiran, **"A Regression Based User Calibration Framework for Real-time Gaze Estimation"**, *in IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 2016.*
- **NM Arar**, H Gao and J-Ph Thiran, **"Robust Gaze Estimation Based on Adaptive Fusion of Multiple Cameras"**, *in IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2015.*
- **NM Arar**, H Gao and J-Ph Thiran, "**Towards Convenient Calibration for Cross-Ratio based Gaze Estimation"**, *in IEEE Winter Conference on Applications of Computer Vision (WACV), 2015.*

*Please check my homepage for the full list of publications.
https://people.epfl.ch/murat.arar

INTERESTS & HOBBIES

Basketball player in Phoenix Mediterranean Club (Lausanne)
Passionate basketball player since age 7, played for many clubs and won many medals and titles. Basketball has been helping me to develop my social and professional skills such as communication, leadership and being a team player.

Sports and travelling enthusiast
Actively interested in skiing, sailing and hiking. Love backpacking, been to 30+ countries in different continents.

REFERENCES

Available upon request

List of Publications

Patents

- **Arar N.M.**, Pati P., Kashyap A., Fomitcheva Khartchenko A., Kaigala G.V., and Gabrani M., "An Automated Method for Process Parameter Optimization for Tissue Section Immunostaining", 2017. Patent(s) to be filed at the US Patent Office.
- **Arar N.M.**, Thiran J.P., and Theytaz O., "Eye Gaze Tracking System and Method", US Patent 9,411,417 Granted in August 9, 2016.

Journal Articles

- **Arar N.M.**, Pati P., Kashyap A., Fomitcheva Khartchenko A., Yuksel O., Kaigala G.V., and Gabrani M., "Towards Standardization of Immunostaining Using Automated Quantitative Analysis of Staining Quality", in preparation for IEEE Transactions on Medical Imaging (TMI'17), 2017. (in preparation)
- **Arar N.M.** and Thiran J.-P., "Robust Real-Time Multi-Camera Eye Tracking", submitted to IEEE Transactions on Cybernetics, 2017. (under review)
- **Arar N.M.**, Gao H., and Thiran J.-P., "A Regression based User Calibration Framework for Real-time Gaze Estimation", In IEEE Transactions on Circuits and Systems for Video Technology (TCSVT'16), 2016.

Conference Papers

- **Arar N.M.**, Pati P., Kashyap A., Fomitcheva Khartchenko A., Yuksel O., Kaigala G.V., and Gabrani M., "Computational Immunohistochemistry: Recipes for Standardization of Immunostaining", In Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI'17), Quebec City, Quebec, Canada, 2017.
- **Arar N.M.** and Thiran J.-P., "Estimating Fusion Weights of a Multi-Camera Eye Tracking System by Leveraging User Calibration Data", In Proceedings of the ACM Symposium on Eye Tracking Research & Applications (ETRA'16), pp 225–228, Charleston, SC, USA, 2016.

- **Arar N.M.**, Gao H., and Thiran J.P., "Robust Gaze Estimation Based on Adaptive Fusion of Multiple Cameras", In Proceedings of 11th IEEE International Conference on Automatic Face and Gesture Recognition (FG'15), Ljubljana, Slovenia, 2015.
- **Arar N.M.**, Gao H., and Thiran J.P., "Towards Convenient Calibration for Cross-Ratio based Gaze Estimation", In Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV'15), pp 642–648, Waikoloa Beach, Hawaii, USA, 2015.
- Yüce A., **Arar N.M.** and Thiran J.P., "Multiple Local Curvature Gabor Binary Patterns for Facial Action Recognition", In Proceedings of 4th International Workshop on Human Behavior Understanding (HBU'13), in conjunction with ACM Multimedia (ACM MM'13), pp 136–147, Barcelona, Spain, 2013.
- Guney F., **Arar N.M.**, Fischer M., and Ekenel H.K., "Cross-pose Facial Expression Recognition", In 2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE), in conjunction with 10th IEEE International Conference on Automatic Face and Gesture Recognition (FG'13), Shanghai, China, 2013.
- **Arar N.M.**, Gao H., Ekenel H.K., and Akarun L., "Selection and Combination of Local Gabor Classifiers for Robust Face Verification", In Proceedings of 5th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS'12), Washington DC, USA, 2012.
- **Arar N.M.**, Bekmezci N.K., Guney F., Gao H., and Ekenel H.K., "Real-time Face Swapping in Video Sequences: Magic Mirror", In Proceedings of International Conference on Computer Vision Theory and Applications (VISAPP'12), Rome, Italy, 2012.
- **Arar N.M.**, Gao H., Ekenel H.K., and Akarun L., "Face Recognition Using Curvature Gabor Features", In Proceedings of 20th IEEE Signal Processing and Communications Applications Conference (SIU'12), Oludeniz, Turkey, 2012.
- Guney F., **Arar N.M.**, and Ekenel H.K., "Open-set Face Recognition System", In Proceedings of 20th IEEE Signal Processing and Communications Applications Conference (SIU'12), Oludeniz, Turkey, 2012.
- **Arar N.M.**, Bekmezci N.K., Guney F., and Ekenel H.K., "Real-time Face Swapping in Video Sequences: Magic Mirror", In Proceedings of 19th IEEE Signal Processing and Communications Applications Conference (SIU'11), Antalya, Turkey, 2011.

Conference Presentations & Posters

- Kashyap A., Fomitcheva Khartchenko A., **Arar N.M.**, Gabrani M., and Kaigala G.V., "Making Immunohistochemistry Quantitative for Diagnostic Pathology and Patient Stratification", in NanoBioTech-Montreux, Montreux, Switzerland, November 9, 2016.
- Fomitcheva Khartchenko A., Kashyap A., **Arar N.M.**, Pati P., Gabrani M., and Kaigala G.V., "Micro-immunohistochemistry meets machine learning: towards standardization", in 3rd

NordiQC Conference on Applied Immunohistochemistry, Aalborg, Denmark, June 6th – 9th 2017.

Master Thesis

- "Fusing of Local Appearance Models for Face Recognition", M.Sc. Thesis, Bogazici University, Istanbul, Turkey, 2012.