Towards Energy Efficient Mobile Eye Tracking for AR Glasses through Optical Sensor Technology

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Johannes Meyer

aus Haselünne

Tübingen

2022

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 23.11.2022
Dekan: Prof. Dr. Thilo Stehle
1. Berichterstatter/-in: Prof. Dr. Enkelejda Kasneci
2. Berichterstatter/-in: Prof. Dr. Andreas Zell

Acknowledgements

Acknowledgements

topics.

Furthermore, I had the opportunity to work with highly motivated bachelor's and master's students during the supervision of their theses. Many thanks to Alexander Zimmer, Adrian Frank, Mario Weckerle, Sarah Dagne, and Michael Mühlbauer for their commitment and dedication to this project. It was a pleasure working with you.

I would also like to gratefully acknowledge Dr. André Kretschmann and Dr. Anne Serout from the microsystem technologies department for allowing me to work in their department with extensive freedom on this exciting topic. Thank you, Anne, for proofreading my manuscripts over the last three years.

Writing a Ph.D. thesis during a global pandemic was not an easy task at all. Therefore, I also want to thank my former roommates Philip, David, Patrick, Natalie, Laura, and José and my friends Christian, Daniel, Patrick, René, Michael, Maik, Markus, Carmen, Hendrik, Jason, and Marcel, who supported me aside of the work and reminded me of the joy of life.

Also, I want to thank Prof. Dr. Gerd von Cölln, who encouraged me already during my master's study to pursue a Ph.D.

Further, I am very grateful to Itis for proofreading the thesis.

Finally, I want to thank my parents, Karin and Hans-Jörg, and my brothers Daniel and Christian for their understanding and continuous support throughout my life.

Renningen, December 6, 2022 J. M.

Abstract

After the introduction of smartphones and smartwatches, Augmented Reality (AR) glasses are considered the next breakthrough in the field of wearables. While the transition from smartphones to smartwatches was based mainly on established display technologies, the display technology of AR glasses presents a technological challenge. Many display technologies, such as retina projectors, are based on continuous adaptive control of the display based on the user's pupil position. Furthermore, head-mounted systems require an adaptation and extension of established interaction concepts to provide the user with an immersive experience. Eye-tracking is a crucial technology to help AR glasses achieve a breakthrough through optimized display technology and gaze-based interaction concepts. Available eye-tracking technologies, such as Video Oculography (VOG), do not meet the requirements of AR glasses, especially regarding power consumption, robustness, and integrability. To further overcome these limitations and push mobile eye-tracking for AR glasses forward, novel laser-based eye-tracking sensor technologies are researched in this thesis. The thesis contributes to a significant scientific advancement towards energy-efficient mobile eye-tracking for AR glasses.

In the first part of the thesis, novel scanned laser eye-tracking sensor technologies for AR glasses with retina projectors as display technology are researched. The goal is to solve the disadvantages of VOG systems and to enable robust eye-tracking and efficient ambient light and slippage through optimized sensing methods and algorithms.

The second part of the thesis researches the use of static Laser Feedback Interferometry (LFI) sensors as low power always-on sensor modality for detection of user interaction by gaze gestures and context recognition through Human Activity Recognition (HAR) for AR glasses. The static LFI sensors can measure the distance to the eye and the eye's surface velocity with an outstanding sampling rate. Furthermore, they offer high integrability regardless of the display technology.

In the third part of the thesis, a model-based eye-tracking approach is researched based on the static LFI sensor technology. The approach leads to eye-tracking with an extremely high sampling rate by fusing multiple LFI sensors, which enables methods for display resolution enhancement such as foveated rendering for AR glasses and Virtual Reality (VR) systems.

The scientific contributions of this work lead to a significant advance in the field of mobile eye-tracking for AR glasses through the introduction of novel sensor technologies that enable robust eye tracking in uncontrolled environments in particular. Furthermore, the scientific contributions of this work have been published in internationally renowned journals and conferences.

Zusammenfassung

Nach der Einführung von Smartphones und Smartwatches gelten Augmented Reality (AR)-Brillen als der nächste Durchbruch im Bereich der Wearables. Während der Übergang von Smartphones zu Smartwatches weitgehend auf etablierten Displaytechnologien beruhte, stellt die Displaytechnologie von AR-Brillen eine technologische Herausforderung dar. Viele Display-Technologien, wie z. B. Retina-Projektoren, basieren auf einer kontinuierlichen adaptiven Steuerung des Displays in Abhängigkeit der Pupillenposition des Nutzers. Weiterhin erfordern kopfgetragene Systeme eine Anpassung und Erweiterung etablierter Interaktionskonzepte, um dem Nutzer ein immersives Erlebnis zu ermöglichen. In beiden Fällen stellt Eye-Tracking eine Schlüsseltechnologie dar, um AR-Brillen durch optimierte Displaytechnologie und blickbasierte Interaktionskonzepte zum Durchbruch zu verhelfen. Verfügbare Eye-Tracking-Technologien, wie z.B. die Video- Okulographie (VOG), erfüllen die Anforderungen von AR-Brillen insbesondere in Bezug auf Stromverbrauch, Robustheit und Integrierbarkeit nicht. Um diese Einschränkungen zu beheben und mobiles Eye-Tracking für AR-Brillen weiter voranzutreiben, werden in dieser Arbeit neuartige laserbasierte Eye-Tracking Sensortechnologien erforscht. Die Beiträge dieser Arbeit tragen zu einem bedeutenden wissenschaftlichen Fortschritt in Richtung energieeffizientes mobiles Eye-Tracking für AR-Brillen bei.

Im ersten Teil der Arbeit werden neuartige gescannten Laser Eye-Tracking Sensortechnologien für AR-Brillen mit Retina-Projektoren als Displaytechnologie erforscht. Ziel ist es, die Nachteile von VOG-Systemen zu lösen und energieeffizientes, Umgebungslicht- sowie gegenüber Verrutschen der Brille robustes Eye-Tracking durch die Einführung von optimierten Messmethoden und Algorithmen zu ermöglichen.

Der zweite Teil der Arbeit erforscht den Einsatz von statischen Laser Feedback Interferometrie (LFI) Sensoren als stromsparende kontinuierlich verfügbare Sensormodalität für die Detektion von Benutzerinteraktion auf Basis von Blickgesten sowie Kontexterkennung durch die Erkennung von menschlichen Aktivitäten für AR Brillen. Die statischen LFI-Sensoren sind in der Lage, den Abstand zum Auge und die Geschwindigkeit der Augenoberfläche mit einer herausragenden Abtastrate zu messen. Weiterhin weisen sie unabhängig von der Displaytechnologie eine hohe Integrierbarkeit auf.

Im dritten Teil der Arbeit wird basierend auf der statischen LFI-Sensortechnologie ein modellbasierter Eye-Tracking-Ansatz erforscht. Der Ansatz führt durch die Fusion mehrerer LFI-Sensoren zu einem Eye-Tracking System mit einer äußerst hohen Abtastrate, was Methoden zur Verbesserung der Displayauflösung wie z.B. foveated rendering für AR-Brillen und Virtual Reality (VR) Systeme ermöglicht.

Zusammenfassung

Die wissenschaftlichen Beiträge dieser Arbeit tragen durch die Einführung neuartiger Eye-Tracking Sensortechnologien, die insbesondere robustes Eye-Tracking in unkontrollierten Umgebungen ermöglichen, zu einem wesentlichen Fortschritt auf dem Gebiet des mobilen Eye-Trackings für AR Brillen bei. Die wissenschaftlichen Beiträge dieser Arbeit wurden in international renommierten Fachzeitschriften und Konferenzen veröffentlicht.

List of Figures

List of Tables

List of Abbreviations

AR   Augmented Reality

ADC   Analog Digital Converter

AM   Amplitude Modulation

ASIC   Application Specific Integrated Circuit

CAD   Computer Aided Design

CIE   International Commission on Illumination

CMOS Complementary Metal-Oxid Semiconductor

CNN   Convolutional Neural Network

COGAIN Communication by Gaze Interaction

CNN   Convolutional Neural Network

DAC   Digital Analog Converter

DC   Direct Current

DBR   Distributed Bragg Reflector

EOG   Electro Oculography

EEL   Edge Emitting Laser

FMCW Frequency Modulated Continuous Wave

FOV   Field of View

FSM   Finite State Machine

FFT   Fast Fourier Transform

HAR   Human Activity Recognition

HCI   Human Computer Interaction

List of Tables

HOE   Holographic Optical Element

HMD   Head Mounted Device

HMM   Hidden Markov Model

IEC   International Electrotechnical Commission

IMU   Inertial Measurement Unit

IR   Infra Red

IPD   Intra Pupillary Distance

LFI   Laser Feedback Interferometry

LIDAR   Light Detection and Ranging

LED   Light Emitting Diode

LOPOCV   Leave one Participant out Cross Validation

MEMS   Micro-Electro-Mechanical System

NPU   Neural Processing Unit

POG   Point of Gaze

PSOG   Photo Sensor Oculography

PCCR   Pupil Center Corneal Reflection

RGB   Red Green Blue

RANSAC   Random sample consensus

SNR   Signal to noise ratio

TPU   Tensor Processing Unit

TIA   Trans Impedance Amplifier

UI   User Interface

VR   Virtual Reality

VOG   Video Oculography

VAC   Vergence Accomodation Conflict

VCSEL   Vertical Cavity Surface Emitting Laser

Contents

Contents

Contents

1 List of Publications

The research conducted within this thesis was published in renowned international peer-reviewed conferences (such as ETRA or CHI) and high-impact journals (such as IMWUT or IEEE Sensors) and paves the way for novel energy-efficient and highly integrated sensor solutions for human-computer interaction as well as eye-tracking within AR glasses. The full-text publications are included in the appendix of the thesis. Furthermore the technology developed in this thesis lead to granted patents[10, 11].

Accepted Articles

I. **Johannes Meyer**, Tobias Wilm, Reinhold Fiess, Thomas Schlebusch, Wilhelm Stork, Enkelejda Kasneci. "A holographic single-pixel stereo camera eye-tracking sensor for calibration-free eye-tracking in retinal projection AR glasses". In 2022 Symposium on Eye Tracking Research and Applications (2022) 🏆 Best Short Paper

II. **Johannes Meyer**, Adrian Frank, Thomas Schlebusch, Enkelejda Kasneci. "U-HAR: A Convolutional Approach to Human Activity Recognition Combining Head and Eye Movements for Context-Aware Smart Glasses". Proc. ACM Hum.-Comput. Interact. 6 (2022)

III. **Johannes Meyer**, Thomas Schlebusch, Enkelejda Kasneci. "A Highly Integrated Ambient Light Robust Eye-Tracking Sensor for Retinal Projection AR Glasses Based on Laser Feedback Interferometry". Proc. ACM Hum.-Comput. Interact. 6 (2022)

IV. **Johannes Meyer**, Adrian Frank, Thomas Schlebusch, Enkelejda Kasneci. "A CNN-based Human Activity Recognition System Combining a Laser Feedback Interferometry Eye Movement Sensor and an IMU for Context-aware Smart Glasses". Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5 (2021)

V. **Johannes Meyer**, Thomas Schlebusch, Hans Spruit, Jochen Hellmig, Enkelejda Kasneci. "A compact low-power gaze gesture sensor based on laser feedback interferometry for smart glasses". Proc. SPIE 11788, Digital Optical Technologies 2021 (2021)

VI. **Johannes Meyer**, Thomas Schlebusch, Hans Spruit, Jochen Hellmig, Enkelejda Kasneci. "A Novel Gaze Gesture Sensor for Smart Glasses Based on Laser Self-Mixing". In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (2021)

VII. **Johannes Meyer**, Thomas Schlebusch, Wolfgang Fuhl, Enkelejda Kasneci. "A novel camera-free eye tracking sensor for augmented reality based on laser scanning". In IEEE Sensors Journal (2020) 🏆 VDE GMM Publication Award

VIII. **Johannes Meyer**, Thomas Schlebusch, Hans Spruit, Jochen Hellmig, Enkelejda Kasneci. "A novel-eye-tracking sensor for ar glasses based on laser self-mixing showing exceptional robustness against illumination". In ACM Symposium on Eye Tracking Research and Applications (2020)

IX. **Johannes Meyer**, Thomas Schlebusch, Thomas Kuebler, Enkelejda Kasneci. "Low Power Scanned Laser Eye Tracking for Retinal Projection AR Glasses". In ACM Symposium on Eye Tracking Research and Applications (2020)

Submitted Articles

I. **Johannes Meyer**, Stefan Gehring, Enkelejda Kasneci. "Static Laser Feedback Interferometry based Gaze Estimation for Wearable Glasses". Submitted to IEEE Transactions on Systems, Man, and Cybernetics: Systems (2022)

Granted Patents

I. Andreas Petersen, Thomas Schlebusch, **Johannes Meyer**, Hans Spruit, Jochen Hellmig. "Method for detecting a gaze direction of an eye". US 11,435,578 B2, (2022)

1.1 Scientific Contribution

This section summarizes this thesis's contributions and most essential results to enable energy-efficient mobile eye-tracking for AR glasses. Contributions within this thesis focus mainly on the mobile eye-tracking sensor technology field. They are split into three parts, which are summarized in the following.

Scanned Laser Eye Tracking

The main contribution in this part of the thesis is a series of novel methods to a scanned laser eye-tracking sensor system in which 2D images of the eye are captured via a single-pixel camera sensor. The scanned laser beam is directed by an Holographic Optical Element (HOE), integrated into the glasses lens, to the eye, resulting in a broad coverage of the user's Field of

View (FOV). The single-pixel sensor used to capture the 2D image is fully integrated into the frame temple, leading to outstanding integrability of the sensor system. With redirection of the laser beam over the HOE, the virtual camera observes the eye from a frontal perspective through the glasses lens, which leads to a superior perspective compared to mobile VOG-systems. To solve the limitations of VOG-systems in terms of robustness to ambient light and power consumption, the integration of an LFI-sensor is proposed as the coherent detection scheme of the LFI sensor is immune to ambient light. In addition, the LFI sensor detects a bright pupil signal, which eliminates the pupil segmentation stage known from VOG algorithms, reducing the power consumption due to the algorithm's execution on an embedded platform. The sensor operates robustly regardless of iris color or partially occluded pupils, e.g., by the eyelid or eyelashes, leading to an outstanding pupil detection and thus eye-tracking robustness compared to VOG systems. Finally, an approach to limit the degradation of gaze accuracy due to glasses slippage known from VOG systems is presented. The HOE is spatially multiplexed to generate images from two perspectives, leading to a highly integrated virtual single-pixel stereo camera system. The stereo camera system enables 3D reconstruction of the pupil disc, which significantly reduces the calibration effort of the system and thus improves usability, especially for consumer AR glasses, as well as solves the glasses slippage issue, which is mandatory for everyday consumer devices. To summarize contributions within this part of the thesis, novel scanned laser eye-tracking sensor approaches are derived, which solve well-known limitations of VOG systems concerning sensor integration, power consumption, pupil signal robustness, ambient light robustness as well as slippage robustness. Therefore, this part of the thesis paves the way for integrating eye-tracking sensor technology in consumer-grade retinal projection AR glasses.

Static LFI Human Computer Interaction

In this second part of the thesis, static LFI sensors are introduced as a novel sensing modality in a near-to-eye setting. The sensors allow measuring the distance between the sensor and the eye and the eye's surface velocity with an outstanding update rate of 1 kHz by operating the laser sensors in a Frequency Modulated Continuous Wave (FMCW) modulation scheme. Compared to VOG systems, the static LFI sensors require only a fraction of the power consumption. Therefore, they enable novel applications for consumer-grade AR glasses, which always require eye-tracking. As the first application, gaze gesture interaction for hands-free control of the User Interface (UI) is presented. Due to the high update rate, a negative latency between gaze gesture execution and classification of the gaze gesture is achieved, allowing rendering systems to react to user inputs before it is finished. The second application presents context awareness for AR glasses by recognition of human activities. Within this application, eye- and head movements were fused to recognize a wide range of activities from the physical as well as the cognitive domain. Two data sets are collected during this part of the thesis, one data set with a system containing the LFI sensor and an Inertial Measurement Unit (IMU) sensor, and one data set containing a VOG sensor and an IMU sensor. The second data set is published as an additional contribution to emphasize research within the area of context-aware AR glasses.

Static LFI Eye Tracking

The third and final part of this thesis, six static LFI-sensors are combined into a high-speed gaze estimation sensor system. The approach outperforms VOG systems with its outstanding update rate of 1 kHz by magnitudes while consuming less power, showing higher integrability, and being robust to ambient light. Therefore the sensor technology introduced in this part of the thesis enables applications such as foveated rendering or saccadic endpoint prediction for mobile AR and VR devices. Furthermore, since the sensors work independently of the display technology, this approach contributes to mobile eye-tracking in general and thus is applicable for AR-glasses with, e.g., waveguide or micro-LED displays as well as VR-glasses. The essential contribution besides the sensor system is the sensor fusion algorithm used to estimate the user's gaze from the LFI sensor readings. A hybrid model-based sensor fusion algorithm that combines a geometric model of the eye tailored to the LFI sensor modalities with a machine learning approach is proposed. The algorithm is robust to glasses slippage. Furthermore, the system works purely based on distance and velocity information. It thus does not capture privacy-related information such as images from the iris region, thus paving the way for privacy-preserving high-speed gaze estimation for AR as well as VR.

2 Introduction

The launch of the first iPhone in 2010 opened the market for smart wearables, such as smartwatches and smart headphones. After the success of smartphones and smartwatches, the next smart wearables to be considered are smart glasses. The term smart glasses summarizes different types of head-worn intelligent wearables, ranging from simple audio glasses with integrated speakers to AR glasses with lightweight, transparent HMD. Unlike VR, whose display technology has mainly been adopted from smartphones, AR glasses require novel display technologies such as retinal projection displays to provide high visual comfort while minimizing the system's weight. The main disadvantage of retinal projection displays is the small eye box, as there is only a single exit pupil from which light must hit the retina to display an image. Therefore mobile eye-tracking is a crucial technology for retinal projection AR glasses to steer the exit pupil based on the pupil position in a closed loop.

Part one of the thesis presents novel sensor technology approaches to enable mobile eye-tracking within retinal projection AR glasses. To better understand the integration of the presented sensor technology approaches, the building blocks and the system design of a retinal projection display are introduced in Section 2.1. To further motivate the research towards novel sensor technologies for mobile eye-tracking, the subsequent section, Section 2.2, gives an overview of eye-tracking applications for AR glasses. The applications and the literature requirements for mobile eye-tracking sensors are derived in Section 2.2.5.

In order to place the contributions of this thesis in state-of-the-art, Section 2.3 gives an overview of mobile eye-tracking sensor technologies with a focus on established VOG sensor technology and algorithms. In addition to VOG sensor systems, related scanned laser eye-tracking sensors, and other emerging non-intrusive mobile eye-tracking sensor approaches are introduced. Finally, the limitations of the different sensor technologies are discussed in the section.

Part two and three of this thesis's main contribution introduces the LFI sensor technology as an eye-tracking sensor. For this purpose, the sensing principle of the sensor is introduced in Section 2.4 together with a brief description of the building blocks of the sensor system.

2.1 Retinal Projection AR Glasses

Although the retinal projection display technology itself was already described by Viirre et al.
[12] back in 1998, the technology was first introduced by Sugawara et al. in 2017 [13, 14] in the
domain of AR glasses. Figure 2.1 shows a sketch of a retinal projection HMD.
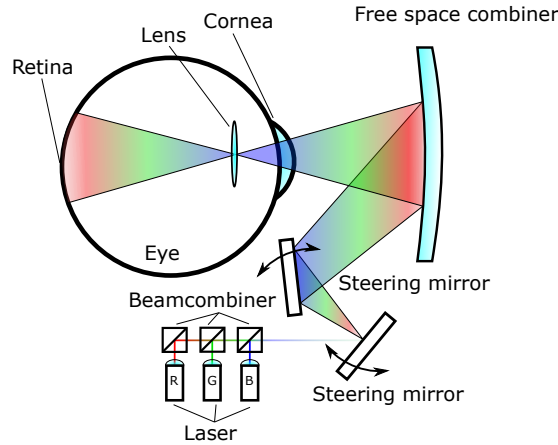
Figure 2.1: Working principle of a retinal projection HMD.

Retinal projection HMDs writes an image directly onto the human eye's retina. The image's
pixels are created by a set of Red Green Blue (RGB) laser diodes, which are combined using a
beam combiner and collimated to a tiny light source. The optical power of the laser diodes is
modulated in their amplitude at a high frequency to control the color and brightness of the
individual pixels. A set of two one-dimensional steering mirrors or a single two-dimensional
steering mirror are used to deflect the light source vertically and horizontally to form a two-
dimensional image. The two-dimensional image is combined to form a uniform ray bundle,
the so-called exit pupil, using a free space combiner. The exit pupil is deflected towards the eye,
and after all the beams crossed in the eye's lens, an image is built up at the retina [14]. Current
retinal projection systems use Micro-Electro-Mechanical System (MEMS) micro mirrors to
direct the laser beam and HOEs as free-space combiners because they provide high visual
comfort. In addition, system efficiency is very high as almost all the light from the laser diodes
is directed to the human eye, reducing thus power consumption compared to waveguide
combiners. The direct projection onto the retina also results in a focus-free image projection
with a large FOV, which is independent of the accommodation state of the eye lens. Despite
these advantages, the tiny exit pupil is a significant drawback of retinal projection HMDs. A
small deviation of the eye's position, e.g., due to a slipping of the glasses or a large rotation of
the eyeball, leads to a complete disappearance of the projected image, as the laser beams do
not enter the pupil anymore. To solve this limitation, the exit pupil must be adjusted according
to the current pupil position. This limitation motivates using mobile eye-tracking sensors
for AR glasses to track the pupil and steer the exit pupil in a closed loop to ensure high visual
comfort.

2.2 Eye Tracking for AR glasses

The term eye-tracking in AR glasses summarizes different approaches to recognizing a user's visual state. These approaches mainly rely on representing the Point of Gaze (POG) in a corresponding coordinate system. For example, in the exit pupil control approach, the pupil position is mapped in head-fixed coordinates onto the display coordinate system to control the exit pupil and ensure that the image is projected correctly onto the retina.
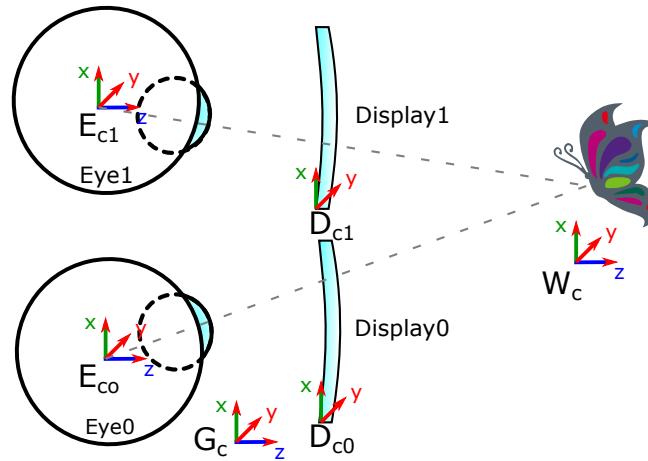
Figure 2.2: Overview of the different coordinate spaces in a mobile eye-tracking AR glasses

Figure 2.2 gives an overview of the different coordinate systems in mobile eye-tracking AR glasses. Each eye has its local coordinate system E_{c0}, E_{c1} typically originated at the eye's rotational center. These coordinate systems are referred to as the head-fixed coordinate system, and the gaze vector per eye is described as a head-fixed gaze vector. The second coordinate system G_c originated in the glasses frame. It is loosely coupled to the eye's coordinate system E_C determined by the positioning of the glasses on the head. The correspondence between the coordinate systems is individual for each user as the Intra Pupillary Distance (IPD) and the head geometry vary across different users. Therefore dependent on the application, a mapping between the individual coordinate systems is necessary.

2D eye tracking applications map the POG from the corresponding eye coordinate system E_c to the corresponding display coordinate system D_c. The POG is described by the intersection of the individual gaze vector with the 2D display plane.

3D eye tracking applications map the eye coordinate system of both eyes (E_{c0}, E_{c1}) to the glasses coordinate system G_c and thus the POG is described by the intersect of both gaze vectors.

3D gaze tracking applications map the POG to the world coordinate systems W_c by adding a reference sensor to the glasses linked to the glasses coordinate system G_c.

2.2.1 Eye Movement Pattern Recognition

Eye movement pattern recognition applications rely on relative eye movement or position sequences, and therefore, a complete description of the relationship between G_c and E_c is not required. Eye movement pattern recognition is mainly used to derive contextual information of the user by detection of specific eye movement patterns. Most of the applications in this context belong to gaze-based Human Computer Interaction (HCI) [15], such as gaze-based control of the UI of the AR glasses [16, 17, 18], gaze-based HAR [19, 20, 21], expertise identification [22, 23, 24, 25, 26, 27, 28], or user identification [29, 30, 31]. In addition, medical features can be derived from a sequence of pupil dilation measurements, which can be used to estimate the user's cognitive load [32, 33, 34, 35, 36] or to detect mental disorders [37].

2.2.2 2D Eye Tracking

2D eye-tracking applications cast the individual gaze vector to its corresponding display coordinate system and thus describe the POG in 2D display coordinates. 2D eye-tracking applications are used to enhance the display quality based on the users POG on the display. Applications are mainly divided into approaches to improving the image resolution and applications to increase the displays FOV. Foveated rendering is a common application enabled by 2D eye-tracking to improve image resolution. Foveated rendering adapts the display resolution based on the POG on the display by rendering content with high image quality around the POG on the display. This dynamic rendering scheme increases the image quality of AR systems or reduces the displays power consumption while keeping the image resolution [38, 39, 40]. Further 2D eye-tracking applications like pupil duplication [41] or exit pupil steering [42, 43] are used to increase the displays FOV of AR glasses [44]. In pupil duplication systems, the exit pupil is duplicated to cover a wide range of pupil positions to project display content across a large FOV. To prevent image degradation, e.g., by double images formed by two exit pupils reaching the retina, 2D eye-tracking is required to disable not matching exit pupils. In pupil steering systems, content is projected through a single exit pupil into the eye. This exit pupil is steered, e.g., by a mirror in a way that the exit pupil of the display follows the pupil to provide image content over a wide FOV. To correctly steer the exit pupil, 2D eye-tracking information is necessary.

2.2.3 3D Eye Tracking

In 3D eye-tracking applications, the gaze vectors of both eyes are calculated and mapped to the glasses coordinate system G_c to estimate the POG based on the intersect of both gaze vectors. 3D eye-tracking allows the estimation of the depth of focus of a user, which allows to estimate the depth of field of AR content accurately and thus enables gaze-contingent stereo rendering [45]. Furthermore, 3D eye-tracking is used to solve the Vergence Accomodation Conflict (VAC) of head-mounted displays by estimation of the vergence [46] to allow adaption of the display focus plane e.g., in a multi-focal display [47], or an auto-focus display [48].

2.2.4   3D Gaze Tracking

In 3D gaze tracking applications, the POG is mapped to the world coordinate system W_c. The relationship between G_c and W_c is known by an additional visual sensor such as a world camera sensor or Light Detection and Ranging (LIDAR) sensor attached to the glasses. By casting the POG to W_c and gathering additional information by the world camera sensor, contextual information such as the object (e.g., the butterfly) the user is fixating on can be derived. The contextual information can be used, i.e., for advertisement [49]. In addition, 3D gaze tracking enables applications like gaze adaptive AR. Gaze adaptive augmented reality summarizes approaches to adapt the projected content of AR glasses based on the POG and an object in the real world [50].

2.2.5   Requirements for Mobile Eye Tracking Sensors in AR glasses

As discussed previously, mobile eye-tracking enables a wide range of applications in AR glasses and is therefore mobile eye-tracking sensors are a key technology for AR glasses [51].

However, the adaption of eye-tracking into AR glasses adds new requirements to eye-tracking sensors such as robust operation for a large portion of the population e.g. with different head and face geometries or different iris colors, robust operation under a variety of lightning conditions as well as robust operation during different scenarios while wearing the glasses e.g. while performing physical activities like cycling [52].

Derived from the application, which the eye-tracking sensor signal should support, further requirements arise for the sensor update rate and gaze accuracy [52]. In addition, the lightweight system design of AR glasses further adds requirements to the sensor integration to be highly integrated into the space constraint glasses frame without interfering with the user's FOV [51].

Finally, like other wearable devices, AR glasses must operate for a whole day. Therefore a low power consumption of the eye-tracking sensor is mandatory as battery capacities are limited due to space and weight constraints of the glasses [51]. These different requirements are addressed in more detail during the next subsections.

Gaze Accuracy

The gaze accuracy of an eye-tracking sensor describes the average angular offset between a fixation location and the corresponding location of the reference target of fixation [53]. The required accuracy is given by the application the eye-tracking sensor needs to fulfill e.g. gaze gesture recognition applications have rather low accuracy requirements. In contrast, display enhancement applications like foveated rendering or exit pupil steering require a gaze accuracy of 1° [39].

9

Power Consumption

AR glasses are battery-powered like other wearables, and the weight of the glasses is directly related to the wearing comfort, which is why the battery capacity is limited. For this reason, the power consumption of the eye-tracking sensor must be as low as possible. Hong et.al. [54] stated that the power consumption of eye-tracking sensors for battery-powered AR- and VR glasses should not exceed 100 mW.

Update rate

The update rate describes the number of successive measurements of the eye pose per second. A high update rate above 300 Hz is required to precisely reconstruct the eye position during fast eye movements [55], which is especially required for display enhancement applications like foveated rendering or exit pupil switching [54].

Robust Operation

AR glasses are everyday devices where eye-tracking functionality needs to operate robustly for various users and in various environmental conditions [56, 57]. They need to cover a wide range of users with their eye and face properties and thus be robust to different iris color [58], eyelashes covering the eye [59], mascara [59], partially occluded pupils by the eyelids [59], as well as specular reflections on the eye surface caused by contact lenses or the glasses lens. Furthermore, the eye-tracking sensors are exposed to uncontrolled ambient lighting conditions ranging from bright sunlight to darkness [15].

Sensor Integration

The eye-tracking sensor must be integrated into the temple of the spectacle frame or the spectacle lens without restricting the user's field of vision and thus not distracting the user. From an integration point of view, the sensor should be fully integrated into the temple so that ideally, no components are integrated into the glasses lens, as components in the glasses lens require more difficult wiring via the glasses hinge. A low number of required components facilitates the integration of the eye-tracking sensor into the glasses, which is particularly advantageous for a high-volume product such as AR glasses.

2.3 Mobile Eye Tracking Sensor Technologies

To motivate this thesis, an overview of state-of-the-art eye-tracking sensor technology concerning the requirements introduced in Section 2.2.5 is given. Before diving into the details of eye-tracking sensor technology, a brief introduction to anatomical and optical features of the eye is given to ease understanding of the working principle of the different approaches.

2.3.1 Anatomical and Optical Features of the Eye
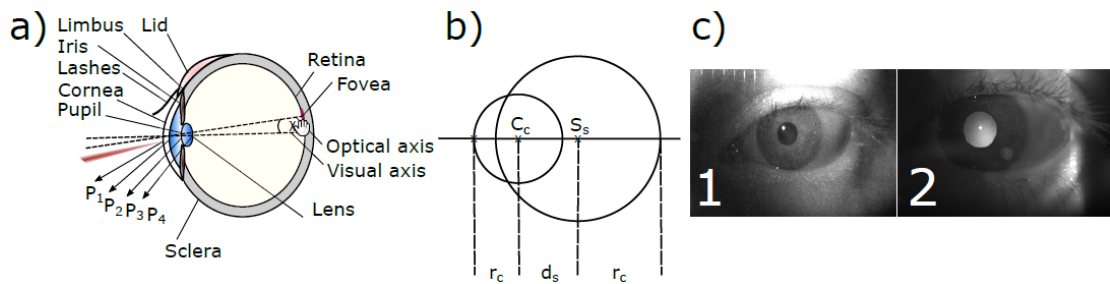
Figure 2.3: a) Section through the human eye to describe the anatomical structure in a simplified form. b) Emsley's reduced eye model according to [60] c) Visualizing the bright- and dark pupil effect in the presence of IR illumination.

Figure 2.3 a) shows a section through the human eye to describe the anatomy and physiology in a reduced form. The human eye consists of the retina, a light-sensitive layer of tissue that converts incident light into a two-dimensional image, and an optical system that focuses the incident light onto the retina using an adjustable lens. Inside the retina, the fovea is embedded, a region with a high density of cones responsible for sharp central vision. The fovea has an individual angular offset χ of ~5° between the visual and the optical axis of the eye [61].

The optical system of the eye consisting of the lens and the pupil, is protected by the transparent cornea. The eye's depth of focus is adjusted by a set of muscles attached to the lens. The iris covers these muscles. In the iris, the pupil is centered. The pupil serves as an aperture of the optical system to control the amount of ambient light entering the optical system. The human eye can operate in a high dynamic range by variation of the pupil diameter between 7.5 mm in the dark and 2 mm in bright sunlight. The outer structure of the iris, also referred to as limbus, connects the iris with the white eyeball, the sclera [61].

The normal vector of the pupil plane or the limbus plane directly correlates with the eye's visual axis and thus with the user's gaze. Therefore limbus and pupil features are used in most imaging applications to inferring gaze information from images.

The advantage of the limbus feature is that the diameter does not vary in the presence of varying ambient light and that, especially for dark-colored eyes, the limbus is characterized by a stark contrast to the white sclera, which eases its detection in images. The main drawback of the limbus as a feature of the human eye is that it is often partially occluded by the eyelid or lashes. As the pupil is centered inside the iris, it is less prone to occlusion. Especially with dark eye colors, the contrast between iris and pupil is rather low, which hamper pupil detection in images. To overcome this drawback, the dark- and bright pupil effect is used to increase its contrast. Figure 2.3 c) shows images of an eye taken by an IR camera sensor while the eye was illuminated with active IR illumination. Suppose the illumination source is off-axis w.r.t the camera sensors. In that case, the pupil appears dark (Figure 2.3 c) 1) while the pupil

appears bright when the axis of the camera sensor and IR illumination are aligned (Figure 2.3 c) 2) [62]. Besides this effect, the use of active IR illumination has several advantages as IR illumination is invisible to the human eye and therefore does not influence the pupil diameter while improving pupil detection based on images in low lighting conditions.

Aside from illuminating the eye, IR illumination sources are also used to create bright spots on the eye's surface, also known as glints, which are used as a feature in various eye-tracking sensor technologies to infer the user's gaze. To illustrate the occurrence of glints, Figure 2.3 a) shows a red laser beam that hits the cornea as well as the lens of the eye. At each boundary surface of the different tissues characterized by a change in refractive index n, a portion of the incident light is reflected according to the Fresnel equation, resulting in four reflections (P_1-P_4), also called Purkinje images. The first reflection P_1 is brightest due to the refractive index jump at the cornea and is further referred to as a glint.

Finally, especially model-based eye-tracking methods incorporate the geometrical structure of the eye. A commonly used eye model e.g., by Gustrin et al., is the Emsley's reduced eye shown in Figure 2.3 b). The eye model is a first-order approximation based on a sclera sphere and a cornea sphere with its centers S_c and C_c and its corresponding radii r_s and r_c. An offset d_s shifts the centers of the spheres. According to Guestrin et.al. [63], the cornea sphere rotates around the sclera sphere with an offset d_s of 6.1 mm and the radius of the cornea sphere r_c is 7.8 mm while the curvature of the sclera sphere r_s is 12 mm.

According to [60] the Emsley's reduced eye, which is shown in Figure 2.3 b) is widely used to describe the human eye in a first-order approximation not including optical properties. The eye model is constructed based on two spheres describing the sclera with its center at S_c and the cornea C_c. According to Guestrin et.al. [63], the cornea sphere rotates around the sclera sphere with an offset d_s of 6.1 mm and the radius of the cornea sphere r_c is 7.8 mm while the curvature of the sclera sphere r_s is 12 mm.

2.3.2 Videooculography Mobile Eye Tracking

State of the art in mobile eye-tracking systems are VOG systems. They rely on IR-sensitive camera sensors attached to the eyeglass frame to capture eye images to improve the image quality and allow operation in low lighting conditions, VOG systems are equipped with IR Light Emitting Diodes (LEDs) to illuminate the eye region. The user's gaze is estimated from the appearance of the eye in the captured images utilizing the anatomical and optical features as introduced in Section 2.3.1.

In general VOG systems incorporate a per-user calibration to gather a mapping between the different coordinate systems, capturing camera images to extract features, and a gaze estimation step [64]. Feature extraction from images and the gaze estimation step are active areas of research, and thus a brief overview of state-of-the-art methods is given in the upcoming Sections [65, 64].

2.3.3 2D Regression based Approach

A standard method of estimating the gaze is using a 2D regression-based approach is the Pupil Center Corneal Reflection (PCCR) method. In this method, a set of IR LEDs is used to produce glints on the surface of the cornea as described in Figure 2.3 a). Afterward, images are captured from the eye, and the glint centers, as well as the pupil center, are extracted. For each glint, the glint vector between the pupil center and the corresponding glint center is calculated [63]. This set of glint vectors is used as input features for a 2D regression approach to map the input features to a gaze vector to 2D display coordinates D_c using a polynomial mapping function and a set of calibration markers within the display coordinate space [63].

2.3.4 3D Model based Approach

In model-based approaches a 3D model of the eye (e.g. Figure 2.3 b)) is constructed based on the observations of features of the eye. Dependent on the utilized features, either a corneal reflection model or a geometric eye model is used to construct an eye model and derive the gaze vector.

Corneal Reflection Model
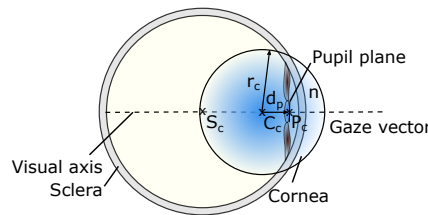
Figure 2.4: Eye model to calculate the POG using corneal reflection methods according to [8]. The model parameters are the radius of the cornea r_c, the distance between the pupil plane and the center of the cornea d_p and the refractive index n of the aqueous humor.

The corneal reflection model estimates the point of gaze-based on the vector between the cornea center C_c and the center of the pupil P_c based on observations of the pupil center and glints created by IR LEDs through an IR camera sensor. Under the assumption that the glasses frame is a rigid body, the pose of the IR LEDs w.r.t. the camera center is fixed. Furthermore, the refractive index n of the cornea as well as the radius of the cornea r_c, and the distance between the cornea center and the pupil plane d_p are fixed optical and geometric properties of the human eye, which can be derived e.g. from Emsley's reduced eye model as presented in Figure 2.3 b).

Based on the known geometry of the system and the known optical and geometric properties of the human eye, Hennessey et al. [8] showed that two glints are sufficient to construct an over-determined system of equations to calculate the corneal center C_c in 3D space. To

construct the gaze vector, they extract the pupil center P_c as a second reference point from the camera image. To increase robustness of the pupil extraction they apply the bright pupil effect as shown in Figure 2.3 c) by adding an additional on-axis IR LED to their system [8].

Geometric Eye Model
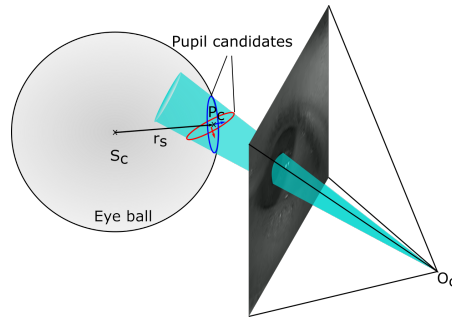
Figure 2.5: Construction of a geometric eye model by projection of the pupil as cone from the 2D image plane to the 3D space according to [9].

Geometric eye models use a single-camera VOG system without additional glint LEDs. The 3D eye model is derived from a sequence of pupil observations in the 2D image plane.

Considering a calibrated camera sensor and assuming a pinhole camera model, in a first step the 2D projection of the pupil in the image coordinate space is unprojected to its 3D representation by constructing a cone using the 2D pupil ellipse as a base and the focal point of the pinhole camera model as the vertex, as shown in Figure 2.5. The cone is propagated towards the eye in the 3D space. Considering a fixed pupil diameter, two possible pupil candidates (red and green) along the cone remain valid solutions, leading to a disambiguity. By adding additional observations from the sequence of images and adding the constraint that the normal vector of the pupil must be directed towards the camera sensor, the disambiguity can be resolved. From a set of observed pupils and their corresponding unprojections, the center of the eye S_c in 3D space is estimated assuming a fixed eyeball radius r_s. Based on the eyeball center and the pupil observations, the gaze vector is calculated as the vector from the eyeball center towards the pupil center [9]. To further improve the accuracy of geometric eye models, Dierkes et al. [66] extend the initial geometric eye model by Swirski et al. [9] to handle refraction effects to estimate the pupil contour correctly.

2.3.5 Pupil Detection Pipeline

A key challenge for all mentioned gaze vector estimation approaches is a robust extraction of pupil features from camera images, either the center of the pupil, e.g., for 2D regression approaches, or the entire pupil contour. Thus the pupil ellipse, e.g., for the geometric eye model approach. Figure 2.6 a) visualizes the different steps of the pupil detection pipeline using classic computer vision methods while Figure 2.6 b) visualizes the pupil detection
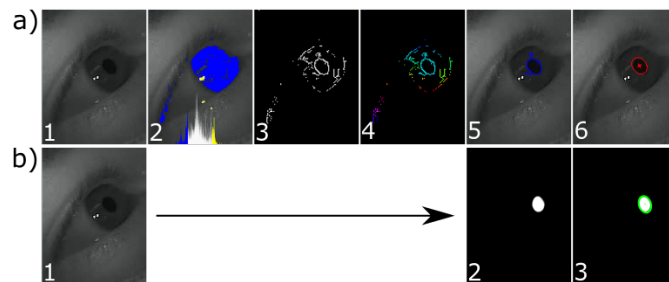
Figure 2.6: Exemplary pupil detection pipeline of different VOG systems. a) shows a pupil detection pipeline comparable to the pupil labs approach using classic computer vision methods [56] while b) shows the pupil detection approach using a CNN model for pupil segmentation as proposed by [67]

pipeline of deep learning-based pupil detection approaches.

Classic computer vision-based pupil detection methods perform several steps to extract the pupil features. In the first step, a coarse pupil region is estimated, and the captured image is cropped to the pupil region to reduce distortions, e.g., by eyelashes [68, 53]. Afterward, the pupil candidate pixels are estimated, e.g., by using histogram-based methods. The pupil candidate pixels are either the dark pixels if dark pupil tracking is used or the bright pixels if the bright pupil effect is exploited to increase the contrast between pupil and iris as described in Section 2.3.1. In the next step, possible pupil edges are extracted from an image, e.g., by using the canny edge detector as shown in Figure 2.6 a) 3 [56]. Next, the detected edges are converted into pupil contours using, for example, connected component analysis. Then, the pupil contour candidates are filtered based on a set of extracted features such as straightness, the intensity value, or elliptical features such as ellipse aspect ratio or ellipse outline contrast of the pupil contour candidates [69, 70, 71, 72]. Finally, based on this extracted pupil, the most likely pupil contour is selected as shown in Figure 2.6 a) 5. Next, the candidate pupil pixels are used to estimate the pupil ellipse, using, e.g., the Random sample consensus (RANSAC) algorithm, and calculate the pupil center as shown in Figure 2.6 a) 6.

To get rid of handcrafted features e.g., for pupil candidate ellipse selection, and thus improve the robustness of extracting pupil features from images, most recent pupil detection approaches apply machine-learning methods [73, 74, 75, 67, 76, 77]. Figure 2.6 b) shows the pupil detection pipeline of the *DeepVOG* method [67]. Most of the classical pupil detection steps are replaced by a Convolutional Neural Network (CNN) model to segment the pupil and robustly estimate the pupil contour afterward. The segmented pupil contour is used as an input to a geometric eye model, e.g., by Swirski et al.. Kim et al. [78] replace the geometric eye model and present *NVGaze*, an end-to-end gaze estimation approach including gaze vector regression. Most recently, Harsimran et al. [79] proposed to further incorporate learned eye model parameters to improve the accuracy of end-to-end gaze estimation solutions.

2.3.6 Technological Challenges of VOG Systems for AR Glasses
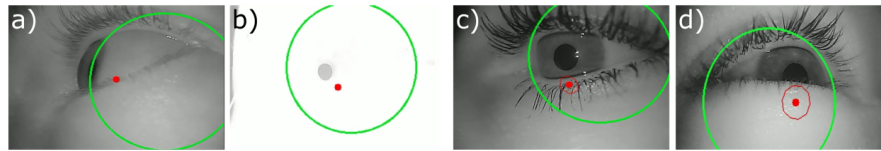
Figure 2.7: Visualization of technological challenges of VOG systems for mobile eye-tracking. a) Limited FOV coverage due to highly off-axis camera integration. b) Loss of pupil detection due to ambient light and limited dynamic range of camera sensors c) False pupil contour edge detection caused by mascara d) False pupil detection due to partial pupil occlusion by the lower eyelid. The images were taken with a Pupil Core eye tracker.

Figure 2.7 highlights some of the technological challenges of VOG systems for use in unconstrained mobile eye-tracking scenarios, which arise from the sensing technology (Complementary Metal-Oxid Semiconductor (CMOS) image sensors) and the pupil detection. The following discusses the main technological challenges of VOG systems and possible solutions.

Power Consumption

A state of the art IR CMOS image sensor (Python300 series [80]) with a in VOG systems commonly used image resolution of 0.3 MP, consumes ≈ 90 mW at a frame rate of 120 Hz. As CMOS image sensors mainly consist of digital logic, the power consumption scales linearly with the frame rate. It thus will further increase as according to Juhola et al. [55] for detecting saccadic eye movements, at least 300 fps are required. To reduce the sensor's power consumption, Mayberry et al. [81] proposed to capture only a subset of pixels selected by a neural network and infer gaze information from the particular subset.

Additional power consumption is added by the pupil detection algorithm as well as the gaze estimation algorithm to VOG systems as these algorithms need to be executed on an embedded system on the glasses. Furthermore, to achieve a robust pupil detection and thus a high gaze estimation accuracy, pupil detection algorithms robust against errors due to mascara or partially occluded pupils, as shown in Figure 2.7 c) and d), are mandatory. These robustness requirements lead to the need for machine learning-based pupil detection algorithms, as discussed in Section 2.3.5, which significantly impacts the power consumption of VOG systems. Approaches to reduce the power consumption added by the algorithms are either in the direction of using optimized hardware like Tensor Processing Units (TPUs) or Neural Processing Units (NPUs) to execute the algorithms [82, 83] or by optimization of the pupil detection models as proposed by Fuhl et al. [84].

Ambient Light Robustness

Figure 2.7 b) illustrates the effects of bright ambient light, e.g., sunlight during outdoor activities, on VOG systems. Due to the limited dynamic range of CMOS image sensors, overexposure makes pupil detection and thus gaze estimation impossible [56] and has been considered as a limiting factor in many outdoors studies [85, 86, 87]. According to Geissler et al. [88] this effect also appears for glint-based methods as the glints are not reliably detected in bright sunlight.

Camera Sensor Integration

A highly off-axis integration of the camera sensor into the eyeglass frame restricts the FOV in which the pupil can be reliably detected. Therefore, gaze estimation deteriorates if the camera sensor and the eye's optical axis are far apart, as shown in Figure 2.7 a). Narcizo et al. [89] shows that the highest gaze estimation accuracy of VOG systems is achieved if the camera sensor is centered in front of the eye. However, this camera sensor position infers with the user's FOV and therefore leads to user distraction and reduces the comfort of VOG systems. To overcome user distraction by camera integration and still cover a wide FOV, Tonsen et al. [90] propose using several low-resolution camera sensors arranged around the lens frame and fuse information of all cameras for gaze estimation.

A second aspect to consider is the size of the camera sensors and imaging optics in front of the image sensors, which limit sensor integration options and increase the weight of AR glasses, which also results in reduced user comfort [91].

Glasses slippage

Especially during physical activities, glasses tend to slip on the user's head, resulting in a change in the relationship between the head-fixed coordinate system and the glasses coordinate system and thus the camera coordinate system. Since the relationship between the head-fixed coordinate system and the glasses coordinate system in 2D regression-based methods is learned during initial calibration, as described in Section 2.3.3, the 2D regression-based methods are prone to glasses slippage as the learned relationship between the coordinate systems mismatches after slippage. Thus the gaze estimation is erroneous [92], [93].

Geometric model-based approaches, as described in Section 2.3.4, derive the relationship between the head-fixed coordinate system and the glasses coordinate system from a set of observations of the pupil ellipse. Therefore they are, to some extent, vulnerable to slippage as the coordinate system mismatch needs to be detected, and a new mapping needs to be learned from new observations [93].

To approach the degradation of gaze accuracy due to glasses slippage, several methods have been applied to estimate the relationship between the coordinate systems, e.g., by reconstruction from landmarks like eye corners [94] or the eyelids [95]. More recently, Santini et al. [92]

proposed the *Get a Grip* method to extract slippage robust features for geometric model-based gaze estimation approaches to minimize the impact of slippage after calibration.

Other solutions to improve this technological challenge are proposed by Kohlbecher et al. [96], and Tobii [97], which address the problem by adding additional camera sensors to construct a stereo camera setup to reconstruct the pupil in 3D space.

Update Rate

The update rate or frame rate of VOG system is limited by the trade-off between image resolution, system bandwidth, and power consumption of the camera sensor and the pupil detection pipeline. Therefore update rates above 120 Hz is rare for highly integrated mobile eye-tracking sensors. Limitations in update rate further have a negative influence on the system latency.

Calibration

Depending on the chosen gaze estimation method and the hardware equipment of a VOG system, a more or less complex calibration procedure is required to achieve a high gaze accuracy of the system. 2D regression-based methods require one calibration per session, limiting the user experience and adoption of VOG systems for AR glasses [98]. For model-based approaches, either glint-based models or geometric models, the calibration is reduced to one calibration per user since only the offset between the visual and the optical axis of the user's eye has to be calibrated once [99].

Calibration-free approaches e.g., as proposed by Tonsen et al. [100], avoid calibrating the system at the cost of lower gaze accuracy.

2.3.7 Scanned Laser Mobile Eye Tracking Technologies

Due to the technological challenges of VOG systems as summarized in Section 2.3.6, particularly in the areas of power consumption and update rate, laser eye-tracking approaches have emerged in recent years. These methods are characterized by using a scanning unit, e.g., a set of MEMS micro mirrors to deflect a laser beam towards the eye and a photodetector to detect reflected IR light to estimate the gaze vector. In general, the approaches can be divided into corneal reflection tracking methods and retinal tracking methods. Figure 2.8 visualize the system design of both approaches.

Corneal Reflection Tracking

Corneal reflection tracking methods utilize the glint feature similar to VOG systems. Instead of a divergent light source and a 2D array of photodetectors (e.g. a CMOS sensor), they scan a focused laser beam in a 2D pattern over the surface of the eye and use a single photodetector
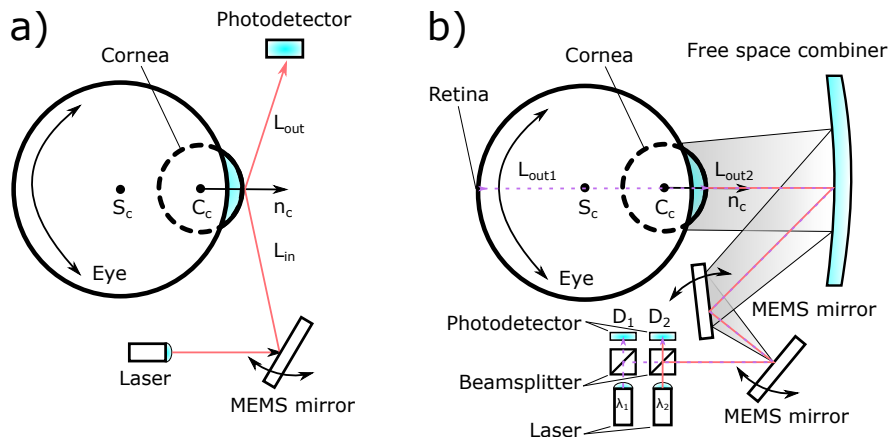
Figure 2.8: System concepts of scanned laser mobile eye-tracking approaches. a) Corneal reflection tracking method with an off-axis alignment of laser and photo diode b) Retinal reflection tracking method with an on-axis alignment of laser sources and photo diodes

to detect the corneal reflection, which occurs for a mirror poistion [101]. Figure 2.8 a) shows a corneal reflection tracking method as proposed by Sarkar et al. [101]. They integrated a resonant 2D MEMS micro mirror into the glasses frame to scan an IR laser beam with a scan frequency of 5 kHz over the cornea and track the corneal reflection with a photodiode integrated into the nose pad. Suppose the photodiode detects a peak due to a corneal reflection. In that case, the corresponding mirror deflection angles are captured and used as input features for a regression-based approach to estimate the gaze vector [102]. As regression-based methods tend to be vulnerable to slippage, as discussed in Section 2.3.6, Sarkar et al. added three additional photodiodes to their system in more recent work reconstructed the position and orientation of the cornea from four corneal reflections. This approach led to an accuracy of $\approx 1\,°$ at an output rate of 500 Hz [103]. Instead of additional photodiodes, Holmqvist et al. [104] added a stereo camera to correct for glasses slippage. This hybrid eye tracker combines the high update rate of a corneal reflection MEMS-based tracker and the absolute reference of a stereo camera setup.

Retinal Reflection Tracking

Retinal reflection tracking methods utilize the high reflectivity of the retina under IR illumination, which also causes the bright pupil effect, as mentioned in Figure 2.3, to construct a feature that somehow corresponds to the pupil center. To detect this feature, a collimated IR laser beam is scanned in a 2D pattern over the eye's surface. If the laser beam hits the pupil, it enters the eye and gets backscattered from the retina, leading to a strong on-axis backscattering of IR light L_{out1}. This is detected by a photodetector, which is integrated into the IR laser beam path [105]. Figure 2.8 b) shows a retinal tracking method proposed by Greenberg et al. [105] according to [106]. They used a set of two lasers with different wavelengths λ_1 and λ_2 to simultaneously track the bright pupil response from retinal center L_{out1}, corresponding

to the pupil center, and the glint on the vertex point of the cornea L_{out2}. To detect the two reflections, two photodiodes D_1 and D_2 are integrated into the system on-axis w.r.t. to the laser beam via a beamsplitter. Under the assumption of a known and fixed system geometry, the gaze vector can be estimated from the detected pupil center and the vertex point [105].

Technological Challenges

While scanned laser eye-tracking approaches solve issues of VOG systems regarding the update rate and reduce the power consumption, technological challenges remain mainly in sensor integration. For example, the final approach of Sarkar et al. [103] requires extensive wiring for the four photodiodes across the glasses hinge while the approach of Greenberg et al. [105] adds complexity due to the high amount of required optical components. In addition, ambient light could lead to an accuracy degradation of the system, especially for cornea reflection tracking methods, since the photodiodes are mounted outside the temple and are directly exposed to ambient light. However, according to Sarkar et al. and Greenberg et al., both methods are on par with VOG systems in terms of slippage robustness.

2.3.8 Other Mobile Eye Tracking Technologies

Aside from VOG systems and scanned laser eye-tracking systems described in Section 2.3.2 and Section 2.3.7 respectively, several other eye-tracking sensor technologies have been introduced in the past. The following Photo Sensor Oculography (PSOG) and neuromorphic camera-based eye-tracking methods are introduced as they fit to some extent the requirements of AR glasses as mentioned in Section 2.2.5. Other mainly intrusive approaches such as the sclera search coil technology [107] or Electro Oculography (EOG) [108] are not described here as they are not applicable for unobtrusive mobile eye-tracking.

Photosensor Oculography

PSOG systems, initially introduced by Torok et al. [109], use a set of photodetectors and IR illumination sources to estimate gaze by examining the scattering behavior and reflectivity of ocular tissues, particularly of the sclera and iris. The IR illumination sources are arranged around the glasses lens to illuminate different parts of the ocular tissue. Similarly, photodetectors are arranged around the glasses lens to receive the backscattered light of the corresponding illuminated part. The measured reflectivity is used as an input feature for 2D regression-based eye-tracking methods [110]. Since this technology is based on 2D regression methods, it has similar weaknesses in terms of glasses slippage as the 2D regression methods of VOG systems. Rigas et.al. [111] propose combining a camera sensor at a low sampling rate with a PSOG system to address the limitation of glasses slippage. Katrychuk et al. address the limitations of the PSOG system concerning glasses slippage by simulating the PSOG outputs of a set of 15 pairs of IR illumination sources and corresponding photodetectors aligned in a

2D array over the entire eye region. They used the simulated data combined with machine learning methods to correct for glasses slippage after an initial system calibration [112].

While PSOG systems provide a high update rate and low power consumption due to their simplicity, the sensor integration into an AR glasses system is a technical challenge. In addition, the PSOG systems are sensitive to ambient light because the reflected light detected by the photodetectors, and thus the detected ambient light, is directly related to gaze estimation. Finally, a PSOG system requires a per session calibration because a 2D regression-based method is used to estimate gaze, which also induces a reduction of gaze estimation accuracy due to glasses slippage.

Neuromorphic Camera

Neuromorphic cameras, also called event-based cameras, are a new bio-inspired camera sensor technology that mimics the behavior of the human retina to overcome the limitations of classic CMOS camera sensors. Instead of capturing images with a fixed global shutter, pixels in neuromorphic camera sensors act independently to brightness changes and send out digital events. As the events operate asynchronously, event cameras reach equivalent update rates above 10 kHz while consuming only a fraction of power (about 10 mW on die level) compared to CMOS sensors. Furthermore the dynamic range of neuromorphic camera sensors is very high (above 120 dB) compared to CMOS sensors (60 dB) [113].

From a sensor point of view, event cameras overcome drawbacks with respect to power consumption, update rate, and ambient light robustness of classic VOG systems as discussed in Section 2.3.6. Angelopoulos et al. [114] were the first to adopt this technology to the eye-tracking domain. They used a classic camera sensor to initialize a geometric eye model consisting of a parametric ellipse, a parabola, and a circle model for the pupil, eyelid, and a single corneal glint. Afterward, the events of the event camera are used to interpolate the geometric model with an update rate of 10 kHz between successive full camera frames. Finally, a 2D regression-based method is proposed to estimate the gaze vector with the pupil and the glint model features as input features. Stoffregen et al. [115] advances the approach of Angelopoulos et al. by removing the classic camera sensor and relying solely on an event camera sensor combined with a set of IR LEDs to produce a circle of glints on the cornea. Toggling the IR LEDs with a predefined pattern leads to an event pattern in the neuromorphic camera sensor. These events correspond to corneal glints, which can be tracked over time to estimate the center c_c as well as the orientation of the cornea in space using a regression-based approach.

While neuromorphic cameras solve some limitations of classic camera sensors used in most VOG systems, especially in the area of power consumption and update rate, algorithm approaches to robustly estimate the gaze vector from a set of events are still not fully solved. Additionally, the same technological challenges from VOG systems arise with respect to sensor integration as discussed in Section 2.3.6.

2.3.9 Summary on Eye Tracking Sensor Technologies

The different eye-tracking sensor technologies are rated according to the requirements for mobile eye-tracking sensors for AR glasses as summarized in Section 2.2.5 to summarize the different technology approaches introduced within this part of the thesis.

Since most of the works lack an accurate description of the system-level power consumption, as the power consumption of the gaze estimation algorithms is usually not considered, power consumption is compared on sensor level including all components required to track a single eye. To allow a fair comparison between the setups, the power consumption is set similar to [115] of IR LEDs to 5 mW and the power consumption of photodiodes including the TIA to 15 mW, which is reasonable for state of the art integrated circuit solution like the THS4567 [116].

The sensor integration requirement is evaluated based on the number of components that need to be integrated into the glasses lens as they restrict the user's field of view and require more complex wiring than components integrated into the frame temple. In addition, the robustness against ambient light is compared by the dynamic range of the receiver circuit (single photodiode circuit or CMOS sensor).

As a robust operation in the presence of glasses slippage is mandatory for AR glasses, only approaches that tend to be robust against glasses slippage are taken into account in the following comparison.

VOG system

The latest TobiiPro Glasses 3 is chosen as a state-of-the-art reference to compare other VOG systems. The eye-tracking hardware consists of a stereo camera setup and eight IR LEDs, which are integrated into the glasses lens. The setup allows slippage robust gaze estimation with up to 100 Hz and a gaze accuracy of $0.6°$ [97].

As no information regarding the power consumption of the camera sensors is available, the power consumption is estimated from the Python 300 CMOS camera sensor [80] by scaling the dynamic power linearly according to the frame rate. In addition, we scale down the dynamic power consumption linearly by the pixel difference to match the sensor size of the cameras used by the Tobii 3 glasses. For a single camera sensor, this leads to

$$
\begin{aligned}
P_{Camera} &= P_{static} + P_{dynamic} \cdot f \cdot pixel_{diff} \\
&= 50\,mW + 0.3312\frac{mW}{Hz} * 100Hz * \frac{240 \times 960}{640 \times 480} = 74.85\,mW.
\end{aligned}
\tag{2.1}
$$

As two cameras are required per eye and in addition 8 LEDs are used to illuminate the eye and generate glints, in total 10 components are integrated into the glasses lens.This sums up to a total power consumption of 189.69 mW, not including the gaze tracking algorithm. The

ambient light robustness of VOG systems is derived from the Python 300 CMOS sensor which states a dynamic range of 60 dB [80].

Scanned Laser System

The Adhawk Mindlink [103] is chosen as a reference for scanned laser eye-tracking systems because it is the only system that is robust against slippage and achieves product maturity. The hardware consists of a 2D MEMS micro mirror, which is integrated into the glasses nose pad, and four photodiodes integrated into the glasses lens frame. The power consumption is split between the power consumption of the micro mirror (15 mW [101]) and the four photodiodes, including the associated TIA, resulting in estimated total power consumption of 75 mW, not including the gaze tracking algorithm. Adhawk reports a gaze accuracy below 1° and an update rate of 500 Hz of their scanned laser eye-tracking system. According to Adhawk [103] robust tracking in the presence of sunlight is possible thus, the ambient light robustness exceeds the ambient light robustness of VOG systems. This is mainly because scanned laser systems rely on coherent light sources with a small bandwidth variation around the central wavelength, which allow efficient filtering of ambient light on the detector side e.g. by using narrow-band optical filters in front of the photodiode as mentioned by Sarkar et. al. [101].

PSOG system

As a PSOG eye-tracking sensor reference system, the latest work of Katrychuk et al. [112] is considered. They achieve a slippage robust design by using an array of 15 photodiodes integrated into the glasses lens and at least one IR LED as an illumination source to illuminate the eye region. This leads to 16 components that consume power on a sensor hardware level of 230 mW. The authors reported a gaze accuracy of 1.09° with a maximum update rate of 1 kHz using a CNN. The dynamic range of PSOG systems of ≈ 30 dB is derived from the reference Trans Impedance Amplifier (TIA) [116] gain bandwidth product at a frequency of 1 kHz.

Neuromorphic Camera

Recently, Stoffregen et al. [115] presented a neuromorphic camera eye-tracking system. They used ten blinking IR LEDs integrated into the eyeglass frame to generate corneal glints, leading to a corneal glint event stream. Based on the event stream, the cornea is tracked at an update rate of 1 kHz. By reconstructing the cornea from the glint events, the authors can perform slippage robust gaze estimation with a total power consumption of the hardware components of 85 mW. Since the authors only report the pixel error (0.5 pixel) and not gaze accuracy directly, this is derived from a comparable work by Angelopoulos et al. [114], which achieved a gaze accuracy of 0.46°. According to Gallego et al. [113] the dynamic range of the neuromorphic camera technology is 120 dB.

Conclusion

The radar chart in Figure 2.9 summarizes the different eye-tracking sensor technologies with respect to the key requirements for mobile eye-tracking sensors for AR glasses. A large coverage of the chart corresponds to a better fulfillment of the requirements. As the power consumption
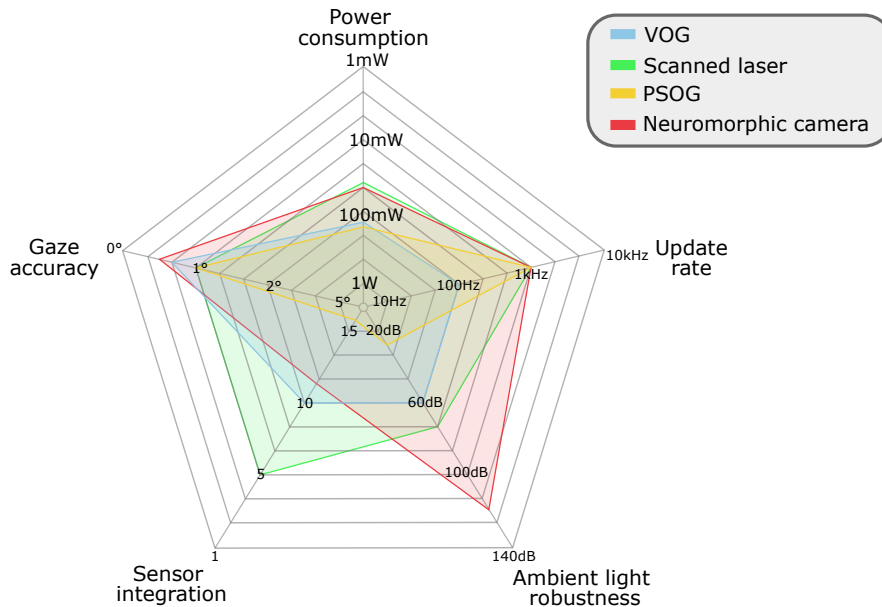
Figure 2.9: Radar chart to compare different eye-tracking sensor technologies for AR glasses

is limited to the hardware power consumption, excluding the power consumption of tracking algorithms, this will increase, especially for systems with a high update rate. To get a sense of the impact of gaze estimation algorithms on system power consumption, the authors of the PSOG system [112] estimated the power consumption of their CNN model optimized for low power consumption between 70 mW to 5.75 W. Therefore this is the most critical requirement as current systems already have a significant power consumption on the hardware level.

The scanned laser technology has the best coverage among all eye-tracking sensor technologies compared in the chart. The technology has advantages with respect to sensor integration and power consumption.

2.4 Laser Feedback Interferometry

As AR glasses with retinal projection are already equipped with an integrated laser scanner and scanned laser eye-tracking sensors seem to fit best the requirements for AR glasses, this thesis focuses on laser-based eye-tracking sensors. In particular, the use of LFI sensors for eye tracking is investigated. To better understand the contributions in the domain of LFI sensors of this thesis, the basic functionality of semiconductor lasers and, in particular, the measurement method of LFI will be introduced. The following Sections are mainly based on

the ground laying work of Coldren et.al. [1] about diode lasers as well as the works of Michalzik et.al. [117] about Vertical Cavity Surface Emitting Lasers (VCSELs) and Taimre et.al. [6] about LFI.

2.4.1 Semiconductor Laser

The main component of semiconductor lasers is the gain medium with a crystalline structure. In this medium, electrons can occupy several energy levels, so-called energy bands. Similar to other semiconductor devices, these structures are divided into a valence band with energy level E_v and a conduction band with energy level E_c. If the electrons are not excited, all electrons are in the valence band. However, suppose an external current is applied to the active medium. In that case, electrons get excited and travel to the conduction band, which results in holes in the valence band and free electrons in the conduction band [1].

Figure 2.10 shows different electron transitions between the valence band and the conduction band, which are essential for the absorption or emission of photons.
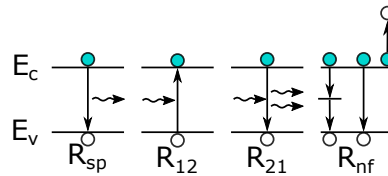
Figure 2.10: Different electron transitions that are important, emphasizing those that involve the absorption or emission of photons [1].

The four recombination or generation mechanisms of photons are spontaneous recombination R_{sp}, stimulated generation R_{12}, stimulated recombination R_{21} and non-radiative recombination R_{nf}.

R_{sp} describes spontaneous recombination of an electron in the conduction band with a hole (missing electron) in the valence band. They do not contribute to a coherent emission [1].

R_{12} describes photon absorbance. A photon is absorbed by an electron in the valence band and stimulates the generation of an electron in the conduction band [1].

R_{21} describes a photon that perturbs the system stimulating recombination of an electron and a hole and simultaneously generating a new photon. This state is also called stimulated emission and is the effect that allows the laser to operate and output coherent light [1].

The last effect is non-radiative recombination, which means energy is dissipated as heat in the semiconductor crystal structure [1]. These losses are material dependent and can be summarized as internal material loss α_i [1].

2.4.2 Vertical Cavity Surface Emitting Lasers

Figure 2.11 shows the basic structure of an VCSEL. The excitation current flows from the
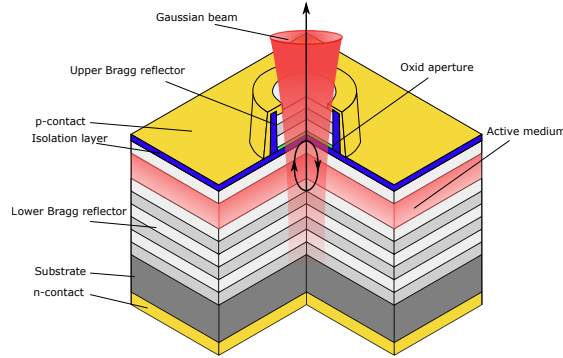
Figure 2.11: Schematics of a vertical-cavity surface-emitting lasers illustrating the structure and the basic operation of an VCSEL [117]

p-contact through the active medium towards the n-contact of the semiconductor surpassing the resonator of the VCSEL. The resonator builds up from two DBRs on top and bottom of the active medium. Inside the resonator, the emission and recombination process occurs, which finally leads to a coherent Gaussian laser beam emission. The diameter and the beam's divergence angle are controlled by the oxide aperture applied on top of the active medium. Compared to Edge Emitting Lasers (EELs), the VCSELs has a lower threshold current and a high-power conservation efficiency η_i, which allows a low power operation of the laser, a circular beam profile with lower divergence angles, which allows a simple and smaller optical design. VCSELs further can be designed to work in a single transversal mode condition to emit only a single-mode electromagnetic wave, which is beneficial for an LFI sensor [117].

Distributed Bragg Reflectors
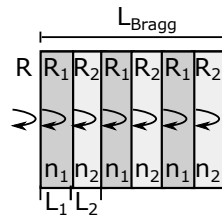
Figure 2.12: Schematics of a DBR mirror of a VCSEL [1].

The DBRs in Figure 2.11 consist of an alternating sequence of high- and low refractive index layers (n_1, n_2) with thickness (L_1, L_2) of one quarter of the material wave length as shown in Figure 2.12. With these conditions, the reflection of each reflector pair will add up to an overall effective reflectivity R of the mirror.

If the oscillating electric field $E(z, t)$ propagates along the z-axis inside the resonator, it is

reflected by one of the two mirrors. Losses occur inside the mirrors as the mirrors are no ideal reflectors and a part of the field is transmitted and coupled out of the cavity. The mirror losses of a DBR are calculated according to Michalzik et. al. [117] by

$$\alpha_m = \frac{1}{L_{Bragg}} ln \left[\frac{1}{\sqrt{R_1 R_2}} \right]. \tag{2.2}$$

Optical Resonator
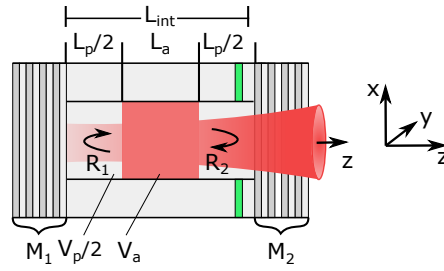
Figure 2.13: Schematics of the resonator of a diode laser [1].

The optical resonator shown in Figure 2.13 is the main component of a laser. It can be divided into two parts, an active volume V_a with length L_a including the active medium and a passive volume V_p with length L_p. Together both volumes form the whole resonator volume V. The fraction between the active volume and the overall volume is described by the confinement factor Γ. In the active volume V_a photons are generated, which propagate with the group velocity v_g along the z-axis of the resonator until they hit one of the two DBRs (M_1, M_2) and are back reflected leading to a harmonic oscillating field $E(z, t)$ inside the resonator. The group velocity inside the resonator is given by

$$v_g = \frac{c_0}{\overline{n}} \tag{2.3}$$

with c_0 describes the speed of light and $\overline{n}$ the effective refraction index of the materials [1].

The harmonic oscillating field $E(z, t)$ inside the resonator is described by

$$E(z, t) = E_0 e^{-j\omega t - \tilde{\beta} z} \tag{2.4}$$

with the angular frequency ω of the electric field E and the complex propagation constant $\tilde{\beta}$. The angular frequency ω of E in Equation (2.4) is determined by the laser wavelength λ and the speed of light as follows

$$\omega = 2\pi f = 2\pi \frac{\lambda}{c_0}. \tag{2.5}$$

The complex propagation constant is according to Coldren et. al. [1] given by

$$\tilde{\beta} = \beta + \frac{1}{2} j(\Gamma g - \alpha_i) \tag{2.6}$$

27

where β summarizes the real part of $\tilde{\beta}$ and the modal gain g. α_i summarizes the internal losses as described in Section 2.4. Besides the gain, there are internal losses α_i inside the resonator which are mainly related to non-radiative recombination R_{nf} of charge carries [1].

The real part β is composed of the laser wavelength λ and the effective refractive index $\overline{n}$ of the active medium

$$\beta = \frac{2\pi\overline{n}}{\lambda}. \tag{2.7}$$

The propagation of a mode inside the resonator requires that stimulated emission of photons exceed the internal losses α_i as well as the mirror losses α_m. At this point the modal gain has reached g_{th}, the threshold gain. At this condition the electric field $E(z,t)$ will replicate itself after a round trip through the resonator which yields $E(z = 2L_{int}, t) = E(z = 0, t)$ [1]. Inserting these boundaries and the DBR reflectivity R_1 and R_2 into Equation (2.4) together with Equation (2.6) gives the threshold condition which needs to be fulfilled for stable oscillation of $E(z,t)$ inside the resonator

$$\sqrt{R_1} \cdot \sqrt{R_2} \cdot e^{-j\beta_{th}L + (\Gamma g_{th} - \alpha_i L)} = 1. \tag{2.8}$$

The subscript th denotes that Equation (2.8) defines β and g at threshold operation and above. The amplitude condition is therefore the real part of Equation (2.8) given by

$$\sqrt{R_1} \cdot \sqrt{R_2} e^{(\Gamma g_{th} - \alpha_i)L} = 1. \tag{2.9}$$

Resolving Equation (2.9) for the threshold gain leads to

$$g_{th} = \frac{1}{\Gamma}\left(\alpha_i + \frac{1}{L_{int}} ln\left(\frac{1}{\sqrt{R_1} \cdot \sqrt{R_2}}\right)\right) \tag{2.10}$$

where L_{int} denotes the length of the resonator. Recalling Equation (2.2), Equation (2.10) can be rewritten to

$$\Gamma g_{th} = \frac{1}{v_g \tau_p} = v_g(\alpha_i + \alpha_m) \tag{2.11}$$

where τ_p is referred to as photon decay lifetime containing all losses inside the resonator. In Equation (2.11) the losses are described in a more general way by the photon lifetime τ_p, leading to

$$\frac{1}{\tau_p} = \alpha_i + \alpha_m. \tag{2.12}$$

The photon lifetime therefore describes the mean lifetime of a photon traveling inside the resonator until the internal losses stops it [1].

The duration a photon takes to fulfill a full round trip inside the resonator is given by the laser round trip time τ_{int}. The round-trip time is calculated from the group velocity and the cavity length

$$\tau_{int} = \frac{2 \cdot L_{int}}{v_g}. \tag{2.13}$$

The length of the cavity further defines the wavelength of the laser. This can be shown by analyzing the complex part of Equation (2.8) which states the following condition

$$e^{-2j\beta_{th}L_{int}} = 1. \tag{2.14}$$

A solution which fulfills this condition is according to Coldren et. al. [1]

$$\beta_{th}L_{int} = m\pi \tag{2.15}$$

with m describing the longitudinal mode number of the laser, which is 1 for a single mode laser used in this work. Inserting Equation (2.7) into Equation (2.15) the wavelength of the laser is given by

$$\lambda = \frac{2}{m}\overline{n}L_{int}. \tag{2.16}$$

Threshold Current and Optical Power

One important parameter of a VCSEL is the threshold current. If this current is applied to the VCSEL, the laser condition, which is described in Equation (2.8), is fulfilled and stimulated emission takes place.

The threshold current is given by

$$I_{th} = \frac{qV}{\eta_i \tau_{sp}} \cdot N_{th} \tag{2.17}$$

with q describing the elementary charge and τ_{sp} the spontaneous emission rate which does not contribute to the stimulated emission of photons, as described in Section 2.4.1 [1]. At threshold the carrier density N_{th} inside the cavity reaches its maximum and a further increase of the current I will not further increase the carrier density.

Based on this observation, above threshold only the photon density N_p will increase and contribute to the emission of photons

$$N_p = \frac{\eta_i(I - I_{th})}{q \cdot v_g \cdot g_{th} \cdot V} \tag{2.18}$$

which is equal to an increase of optical power. This leads us to the basic equation to describe the optical power of a laser

$$P_0 = \eta_d \frac{h \cdot v}{q}(I - I_{th}) \tag{2.19}$$

with the differential efficiency

$$\eta_d = \frac{\eta_i \cdot \alpha_m}{\alpha_m + \alpha_i}. \tag{2.20}$$

Equation (2.19) is only valid if both mirrors are symmetric and have the same reflectivity R.
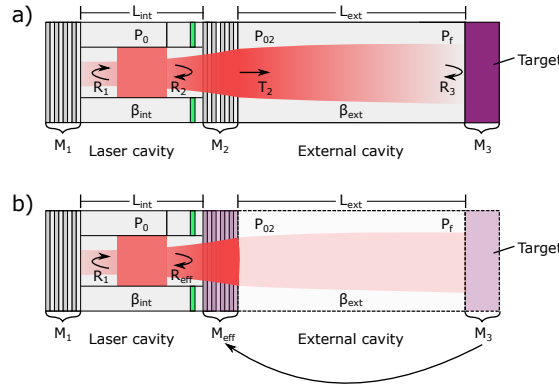
Figure 2.14: a) shows the active laser cavity coupled with a passive external cavity formed by an external measurement target. The external target is modeled as an additional Mirror M_3 with a reflectivity R_3. b) shows the three-mirror model according to Coldren et. al. [1]. In this model the out-coupling mirror M_2 is replaced by an effective mirror M_{eff}, which describes the coupling of M_2 and M_3 obtained by an S-parameter analysis.

Otherwise a correction factor F_2 is applied, yielding

$$P_{02} = F_2 \cdot \eta_d \cdot \frac{h\nu}{q} \left(I - I_{th}\right) \tag{2.21}$$

with

$$F_2 = \frac{T_2}{\left(1 - \sqrt{R_2}\right)^2 + \sqrt{\frac{R_2}{R_1}} \cdot \left(1 - \sqrt{R_1}\right)^2}. \tag{2.22}$$

2.4.3 Semiconductor Laser under Feedback

LFI sensors are VCSELs working on the self-mixing phenomena. Self-mixing describes an operation mode of a laser where a part of the emitted light is back injected into the laser cavity and thus leads to direct feedback to the laser source [118]. The operation of a laser under feedback can be described by extending the laser cavity as shown in Figure 2.13 by addition of an external cavity with an additional mirror M_3, which scatters back light into the laser cavity, as shown in Figure 2.14 a). This leads to the well-known three-mirror model of a semiconductor laser under feedback. A schematic of the three-mirror model is shown in Figure 2.14 b). An analytical description of the physical behavior of a laser under feedback was introduced by Lang and Kobayashi [119].

The cavity of the laser with length L_{int}, propagation constant β_{int} and the two mirrors M_1 and M_2 with reflectivity R_1 and R_2 are coupled with an external passive cavity with length L_{ext}, propagation constant β_{ext} and a mirror M_3 with a reflectivity R_3 [1].

To simplify the further analysis of the system, an effective mirror model is introduced. The mirrors M_2 and M_3 of the laser cavity and the external cavity are replaced by an effective mirror

M_{eff} with an effective reflectivity R_{eff}. Based on the S-parameter analysis, the reflectivity of the effective mirror is according to Coldren [1] given by

$$\sqrt{R_{eff}} = \frac{\sqrt{R_2} + (T_2) \cdot \sqrt{R_3} \cdot e^{-j2 \cdot \tilde{\beta}_{ext} \cdot L_{ext}}}{1 + \sqrt{R_2} \cdot \sqrt{R_3} \cdot e^{-j2 \cdot \tilde{\beta}_{ext} \cdot L_{ext}}}.$$
(2.23)

Considering a weak reflectivity of the external mirror $R_3 << 1$ the effective mirror reflection, Equation (2.23) can be reduced to

$$\sqrt{R_{eff}} = \sqrt{R_2} + (T_2) \cdot \sqrt{R_3} \cdot e^{-j2 \cdot \tilde{\beta}_{ext} \cdot L_{ext}}.$$
(2.24)

Thus, a change of feedback due to a change of the reflectivity ΔR can impact both the amplitude and the phase of the effective mirror. A change of external reflectivity is expressed by $\Delta R = R_{eff} - R_2$. Inserted into Equation (2.24), the impact of changing reflectivity can be described by a complex vector, which can, according to Coldren et.al. [1], be partitioned into a real in-phase part ΔR_{ip}

$$\Delta R_{ip} = (T_2) \cdot \sqrt{R_3} \cdot \cos\left(2 \cdot \beta_{ext} \cdot L_{ext}\right)$$
(2.25)

and a imaginary quadrature part

$$\Delta R_q = -(T_2) \cdot \sqrt{R_3} \cdot \sin\left(2 \cdot \beta_{ext} \cdot L_{ext}\right).$$
(2.26)

Recalling Equation (2.15), the term $2 \cdot \beta_{ext} \cdot L_{ext}$ describes the external phase ϕ_{ext} of the external cavity and thus equals $\omega_{ext} \cdot \tau_{ext}$.

The in-phase component Equation (2.25) affects the laser by changing the reflectivity of the second mirror and therefore the mirror losses $\Delta \alpha_i$ given by Equation (2.2). This leads to a new photon life time (Equation (2.12))

$$\frac{1}{\tau_p'} = \frac{1}{\tau_p} + v_g \cdot \Delta \alpha_m = \frac{1}{\tau_p} + \frac{v_g}{L_{int}} \frac{\sqrt{R_1 \cdot \Delta R_{ip}}}{\sqrt{R_1 \cdot R_2}}.$$
(2.27)

By plugging in the laser round trip time, given in Equation (2.13), and the in-phase component of the effective reflectivity, Equation (2.25) leads to

$$v_g \cdot \Delta \alpha_i = -2 \cdot T_2 \cdot \sqrt{\frac{R_3}{R_2}} \cdot \frac{1}{\tau_{int}} \cdot \cos\left(2 \cdot \beta_{ext} \cdot L_{ext}\right) = -2 \cdot k_f \cdot \cos\left(2 \cdot \beta_{ext} \cdot L_{ext}\right).$$
(2.28)

$$k_f = T_2 \cdot \sqrt{\frac{R_3}{R_2}} \cdot \frac{1}{\tau_{int}}$$
(2.29)

describes the coupling strength between the laser cavity and the external cavity and influences the signal quality.

The cosine modulation of the photon life time leads also to a modulation of the threshold gain g_{th}, introduced in Equation (2.11), and therefore to a shift in threshold carrier density N_{th}, which leads to a variation in the threshold current I_{th} and therefore a modulation of the optical power (Equation (2.19)). In addition the differential efficiency η_d (Equation (2.20)) and the symmetry correction factor F_2 given by Equation (2.22) are modified by the change in mirror reflectivity. This effects lead to a new optical power equation for a perturbed laser [1]

$$P_f = P_{02}\left(1 - m \cdot \cos\left(2 \cdot \beta_{ext} \cdot L_{ext}\right)\right) \tag{2.30}$$

which is equal to the more common notation of the equation

$$P_f(\omega_{ext}) = P_{02}\left(1 - m \cdot \cos\left(\omega_{ext} \cdot \tau_{ext}\right)\right) \tag{2.31}$$

as $2 \cdot \beta_{ext} \cdot L_{ext}$ equals $\omega_{ext}\tau_{ext}$ as well as ϕ_{ext}. The modulation factor m describes the power variation of the perturbed laser due to coupling effects between the two cavities for robust operation of the laser above the threshold [1]

$$m = 2 \cdot k_f \cdot \tau_p\left(\frac{\eta_i}{\eta_d} - 1\right) + k_f\tau_{ext}(1 - F_2)\left(\frac{1 + R_2}{T_2}\right). \tag{2.32}$$

Variation of the in-phase component ΔR_{ip} affects the threshold level and the optical power of the laser leading to an Amplitude Modulation (AM) of the laser. The quadrature component ΔR_q in contrast modifies the round-trip phase angle ϕ_{eff} at mirror R_2

$$\phi_{eff} = \frac{\Delta R_q}{R_2} \tag{2.33}$$

modifying the resonant wavelength of the cavity given by Equation (2.16) and thus leading to an angular frequency shift, which is given according to Coldren et.al. [1] by

$$\Delta\omega_\phi = -k_f \cdot \sin\left(2 \cdot \beta_{ext} \cdot L_{ext}\right). \tag{2.34}$$

Besides the round trip phase angle, an additional change of the angular frequency is introduced by the change of the threshold carrier density N_{th} described by Henry's line-width enhancement factor α [120]

$$\Delta\omega_N = -\alpha k_f \cdot \cos\left(2 \cdot \beta_{ext} \cdot L_{ext}\right). \tag{2.35}$$

Adding both frequency shifts $\Delta\omega_\phi$ and $\Delta\omega_N$ using the trigonometric identity, the total frequency shift $\Delta\omega$ is given by

$$\Delta\omega = -k_f \cdot \sqrt{1 + \alpha^2} \cdot sin(\phi_{ext} + \arctan(\alpha)). \tag{2.36}$$

The angular frequency shift describes the difference of the angular frequency ω_0, experienced

by the laser during the round trip through the external cavity without optical feedback. The angular frequency ω_{ext} corresponds to the actual frequency response including optical feedback [6]. As linkage between phase and angular frequency is given by $\phi = \omega \cdot \tau$, $\Delta\omega$ can be decomposed as follows

$$\Delta\omega = \omega_{ext} - \omega_0 = \frac{\phi_{ext}}{\tau_{ext}} - \frac{\phi_0}{\tau_{ext}} = \left(\phi_{ext} - \phi_0\right)\frac{1}{\tau_{ext}}. \tag{2.37}$$

With this expression and the introduction of Acket's feedback parameter C [118]

$$C = k_f\sqrt{1 + \alpha^2}\,\tau_{ext} \tag{2.38}$$

Equation (2.36) can be rephrased, and the well-known excess phase equation is obtained

$$\phi_{ext} - \phi_0 + C \cdot \sin\left(\phi_{ext} + \arctan\left(\alpha\right)\right) = 0 \tag{2.39}$$

which is the fundamental equation to describe the LFI sensing principle found in most of the literature dealing with LFI sensors e.g., by Giuliani et.al. [121] or Taimre et.al. [6]. As Equation (2.39) is a transcendental equation, a stable phase contingent operation of the laser is only possible if $C < 1$ otherwise more than one possible solution of Equation (2.39) exists, which leads to phase jumps and thus an unstable behavior of the laser [6]. Thus, in the further descriptions a stable operation of the system in the low feedback regime ($C < 1$) is assumed.

2.4.4 Modulation of Semiconductor Lasers under Feedback

Modulation of the coupled cavity is required to observe a change of phase response and thus measure physical quantities. In general, modulation of the coupled cavity can be divided into modulation of the internal cavity and modulation of the external cavity. Modulation of the internal cavity affects the effective mirror M_{eff} of the coupled system by changing the output mirror M_2 of the laser by variation of the internal laser parameters. Modulation of the external cavity affects the effective mirror M_{eff} by variation of the external cavity mirror M_3 by variation of the parameters of the external cavity. In the following, both modulation schemes and the corresponding measurement quantities are discussed.

External Cavity Modulation

Based on Equation (2.30), the modulation of optical power is strongly dependent on the cosine term and therefore to $\omega_{ext} \cdot \tau_{ext} = \phi_{ext}$, which is linked via Equation (2.39) with the phase ϕ_0 of the unperturbed laser. Examining Equation (2.39), two parameters influence ϕ_{ext}. One is Acket's feedback parameter C and the other is the phase ϕ_0 of the unperturbed laser. The linewidth enhancement factor α as third possible parameter is omitted, as according to Taimre et al. [6] this parameter can modeled as constant.

Analyzing the remaining parameters, the most straightforward way to modulate ϕ_{ext} is a mod-

ulation of the round trip phase ϕ_0 of the unperturbed laser [6]. Recalling the correspondence between phase ϕ_0 and the angular frequency ω_0 ($\phi_0 = \omega_0 \cdot \tau_0$) as well as Equation (2.16) and the relationship between wavelength and frequency $\lambda = c_0/f$, the following equation shows which parameters affect the round trip phase ϕ_0 of the unperturbed laser

$$\phi_0 = 2 \cdot \pi \cdot f_0 \cdot \tau_{ext} = \frac{4 \cdot \pi \cdot n_{ext} L_{ext}}{\lambda}. \tag{2.40}$$

Considering operation in free space ($n_{ext} = 1$) and a constant wavelength λ of the laser, the round-trip phase is directly proportional to the length of the external cavity $\phi_0 \approx L_{ext}$. A periodic variation of ΔL_{ext} of at least $\lambda/2$ leads to a periodic variation of the optical power P_f. Differentiating Equation (2.40) with respect to time t leads to

$$\frac{d\phi_0}{dt} = \frac{4\pi}{\lambda}\frac{dL_{ext}}{dt}. \tag{2.41}$$

As the derivation of the phase with respect to time equals the frequency and the derivation of the distance with respect to the time equals the velocity v of displacement in beam direction. This leads to the well-known Doppler equation

$$f_d = \frac{2 \cdot |v_T|}{\lambda} \tag{2.42}$$

which equals a Doppler shift by the superposition of the incident wave and the back scattered wave. Equation (2.42) is limited to target movements which are transversal to the propagation axis denoted by v_T. To overcome this limitation, the Doppler frequency equation is expressed in a more general way

$$f_d = \frac{2 \cdot |v| \cdot \cos(\Psi)}{\lambda} \tag{2.43}$$

where Ψ denotes the angle between the wave vector k and the velocity vector v [122]. To measure absolute velocity, the internal parameters of the cavity need to be modulated as well.

Internal Cavity Modulation

The modulation of internal cavity parameters of a laser under feedback is carried out by a modulation of the drive current of the laser [117]. A variation in drive current affects the internal cavity by two different mechanisms. First, as the current increases, the cavity temperature rises, which affects the effective cavity length L_{int} and the gain curve in the material. Secondly, increasing the current increases the carrier and photon density in the material, which influences the material's refractive index n_{int}. This second effect is also referred to as the plasma effect, which tends to be the dominating effect at high modulation frequencies [123]. Recalling Equation (2.16), both effects influence the wavelength of the laser. For low modulation frequencies, the temperature effect dominates the modulation of the cavity [124] which leads to

$$\lambda(t) = 2 \cdot \overline{n} \cdot L_{int}(t). \tag{2.44}$$

34

Recalling Equation (2.40) and plugging in the variation of the wave length leads to a modulation of ϕ_0 and therefore a modulation of the optical output power P_f. The modulation of the wavelength can be interpreted in Equation (2.40) as a slight movement of the laser towards the target, if the driving current is increased and a movement away from the target as the drive current is reduced respectively. One important observation is the proportionality between drive current modulation, active cavity modulation and the modulation of the wavelength

$$\frac{d\lambda}{dt} \sim \frac{dL_{int}}{dt} \sim \frac{dI}{dt}. \tag{2.45}$$

Plugging in Equation (2.44) into Equation (2.40) leads to a modulation of ϕ_s even if L_{ext} does not vary, which allows the measurement of the distance L_{ext} towards still standing targets.

Time differentiating Equation (2.41) with respect to the wavelength λ as a modulation of the drive current leads to a variation of the wavelength and gives the following equation

$$f_b = \frac{2 \cdot L_{ext}}{\lambda^2} \cdot \frac{d \cdot \lambda}{dt}. \tag{2.46}$$

with f_b as beat frequency [125]. The beat frequency is directly proportional to the target distance [125] .

Simultaneous Measurement of Distance and Velocity

As both frequencies, f_d and f_b lead to a modulation of the optical power P_f either induced by a moving target or by a change of the wavelength caused by the laser drive current modulation, both effects superimpose and thus cannot be observed independently. Therefore, a known current modulation scheme with two distinguishable states is required to resolve this singularity. A widely applied modulation scheme with two states is a triangular modulation as used in FMCW LIDAR systems [126].

The drive current of the laser is modulated by a symmetric Direct Current (DC)-free triangle signal with a frequency f_{mod}, a period duration T_{mod} and an amplitude I_{mod}.

Figure 2.15 a) shows the effect of triangular modulation of the wavelength λ by the drive current of the laser on the transmitted signal (solid triangular line) and the back-reflected echo signal (dashed line) from the target. The echo is delayed by τ_{ext} and shifted by the Doppler frequency. f_{up} corresponds to the difference frequency measured during the up ramp, and f_{down} corresponds to the difference frequency measured during the down ramp of the triangular signal. The amplitude of the triangular signal is controlled by the modulation current I_{mod} and the duration of a triangular wave by T_{mod}. Figure 2.15 b) shows the resulting frequency inside the laser cavity as the sum of transmitted and back-reflected signal. The mean resulting frequency corresponds to the distance-related beat frequency f_b and the difference between up ramp frequency f_{up} and the down ramp frequency f_{down} corresponds to twice the frequency shift induced by the Doppler effect and thus to the target velocity.

The sum frequency finally modulates the optical power inside the laser cavity according to Equation (2.31)
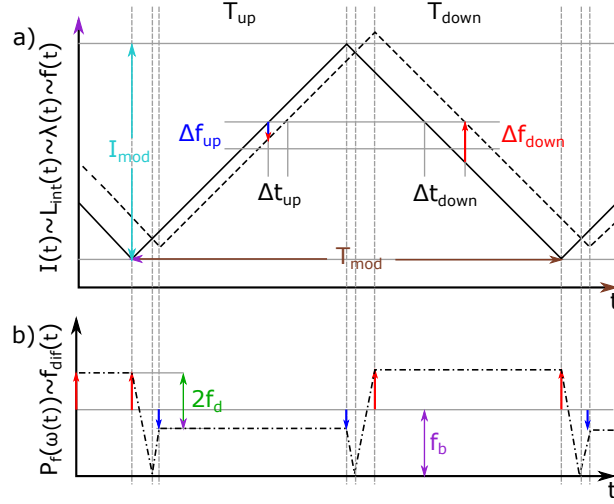
Figure 2.15: a) Effect of triangular modulation of the wavelength λ by the laser's drive current on the transmitted signal (solid triangular line) and the back-reflected echo signal (dashed line) from the target. b) the resulting frequency inside the laser cavity as a sum of transmitted and back-reflected signal.

The current modulation leads to a change in the internal cavity length, and thus the wavelength of the cavity as described by Equation (2.45). This change leads to a modulation of the frequency of the optical wave transmitted from the laser. With a delay of τ_{ext}, corresponding via Equation (2.13) to the target distance, the modulated wave is back-reflected from the target, and the echo signal is back injected into the laser cavity. Due to the velocity of the target, the optical wave is shifted by the Doppler frequency. Thus the cavity acts as an optical mixer constructing the sum and difference of both wave frequencies[125]. Figure 2.15 b) shows the frequency difference of both wave frequencies. While the laser operates in the Terahertz Regime, the optical mixer converts the difference signal to the Megahertz regime, which eases signal processing. As depicted in Figure 2.15, the mean of both frequencies f_{up} and f_{down} corresponds to the beat frequency f_b and half of the difference between them to the Doppler frequency f_d.

Due to the symmetric shift of the signal wave during a full modulation swing T_{mod}, the distance related frequency f_b is given by

$$f_b = \frac{f_{up} + f_{down}}{2} \tag{2.47}$$

and the target velocity related frequency f_d is given by

$$f_d = \frac{f_{up} - f_{down}}{2}. \tag{2.48}$$

By measuring the frequencies f_{up} and f_{down} during the corresponding triangle ramps T_{up} and T_{down}, the Doppler- and beat-frequencies can be calculated. Recalling Equation (2.46), the distance to the target can be calculated from the laser wavelength λ, the triangle parameters and the drive current corresponding wavelength modulation constant $d\lambda/dt$. In addition, the target velocity can be calculated by plugging in Equation (2.48) into Equation (2.43) if the angle of incidence ψ is known.

2.4.5 LFI Sensor

As described in the previous sections, the optical power P_f of the laser under feedback is modulated based on an external target, which back reflects laser light into the laser cavity as described by Equation (2.31). Therefore it is required to measure the varying optical power inside the resonator of the laser cavity. A suitable solution is to integrate an intra-cavity photodiode into the lower DBR of the cavity as shown in Figure 2.16 [6]. A detailed description of the integration of a so-called monitoring diode is given by Grabherr et al. [2].
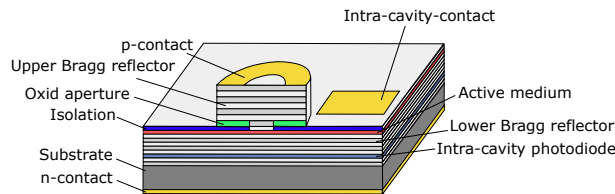
Figure 2.16: Integration of a photodiode into the lower DBR of the laser cavity [2].

Besides the sensor component itself, additional electrical components like a laser driver circuit, a receiver circuit, and a digital circuit are required to build up an LFI sensor to measure distance and velocity as described in Section 2.4.4. Figure 2.17 shows a block diagram of an LFI sensor, including the sensing component as well as the circuitry to operate an LFI sensor.
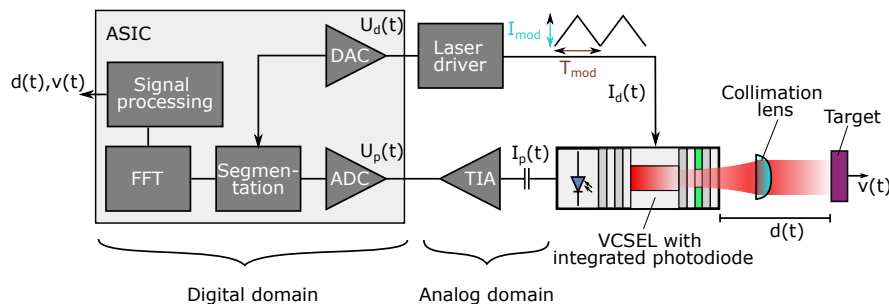
Figure 2.17: Block diagram of an LFI sensor with modulation source (DAC and laser driver) and analog frontend (TIA) for to capture the photodiode signal. Finally the measurement quantities (distance, velocity) are extracted in the digital domain [6].

The circuitry can be divided into analog and digital domains. The digital domain, which

could be realized by an Application Specific Integrated Circuit (ASIC), generates a triangular modulation signal $U_d(t)$, e.g., by an DAC. This signal is fed into a laser driver circuit, which converts the voltage signal into a current modulation signal $I_d(t)$. The VCSEL with integrated photodiode emits frequency-modulated light, which is collimated using, e.g., a collimation lens. The light travels along the optical axis of the laser until it hits the target at a distance $d(t)$, which moves with a velocity $v(t)$. A portion of the light is back-reflected and back injected into the cavity where both the local optical wave and the back-reflected optical wave interfere. The photodiode, integrated into the back DBR, measures the power modulation, which follows the inference signal according to Equation (2.31). The resulting current signal $I_p(t)$ is highpass filtered to suppress the DC power of the laser and amplified and converted into a voltage $U_p(t)$ by a TIA. This voltage signal is digitized by an Analog Digital Converter (ADC) and segmented into voltage segments $U_{p,up}(t)$ and $U_{p,down}(t)$ which either belong to the up-ramp or the down-ramp of the current modulation triangle. From both segments the peak frequencies f_{up} and f_{down} are calculated. With these frequencies f_d and f_b are calculated using Equation (2.48) and Equation (2.47). Afterwards, the target velocity $v(t)$ in beam axis and the target distance $d(t)$ are calculated by applying Equation (2.43) and Equation (2.46).

2.5 Conclusion

Mobile eye-tracking sensors are a key sensing technology for AR glasses, especially for retinal projection glasses, as discussed in Section 2.1. Moreover, they are the enabling technology for a wide range of applications from HCI applications over display enhancement applications to medical and well-being applications as introduced in Section 2.2, adding significant value to consumer AR glasses. The integration of eye-tracking sensors into the domain of consumer graded AR glasses sets high requirements in the area of gaze accuracy, robustness, sensor integration, power consumption as well as sensor update rate, which are discussed in detail within Section 2.2.5.

A detailed analysis of established state-of-the-art VOG eye-tracking sensor technology shows that VOG systems fulfill requirements regarding accuracy but are limited with regard to pupil detection and ambient light robustness, sensor integration, power consumption, and sensor update rate as discussed in Figure 2.7. Therefore, emerging eye-tracking sensor technologies are analyzed within section 2.3. The summary of different eye-tracking sensor approaches in fig. 2.9 shows that scanned laser eye-tracking sensors have the best fulfillment of all requirements.

Therefore, contributions within the first part of the thesis focus on closing the gap between the high gaze accuracy of established VOG sensors and the advantages of scanned laser eye-tracking sensor technology with respect to sensor integration and power consumption. In addition, known technological limitations like ambient light robustness, glasses slippage, and user calibration are addressed within the contributions in this part of the thesis.

As contributions in the first part of the thesis focus on integrating scanned laser eye-tracking

sensors into retinal projection AR glasses, the second and third part of the thesis focus on an alternative novel static laser approach, which enables mobile eye-tracking independent of the display technology. As potential technology LFI is identified due to the small size of the sensors and the capability to measure distance and rotational velocity with an outstanding update rate. Therefore this capture closes with a detailed description of this sensing technology in Section 2.4 to ease understanding of the contributions within the second and third part of the thesis.

The major contributions this thesis makes to enable energy-efficient mobile eye-tracking for AR glasses through optical sensor technology are discussed in the next chapter.

3 Major Contributions
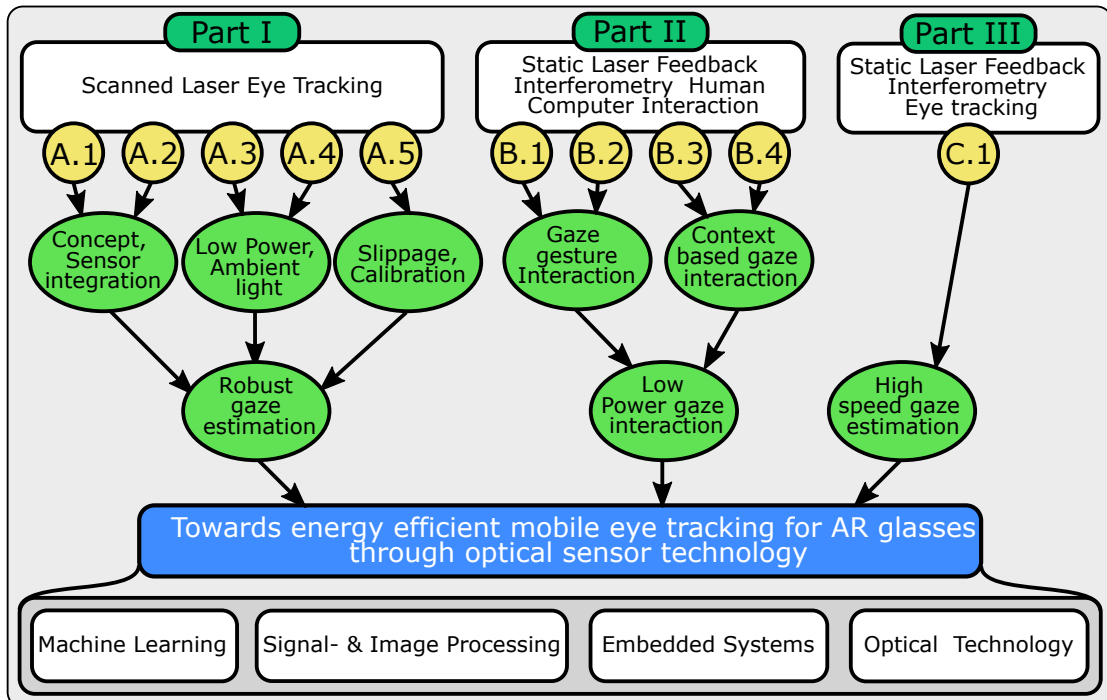
Figure 3.1: Overview and relationship of contributions in the thesis.

Figure 3.1 summarizes the contributions of this thesis. The major contributions are split into the three parts *Scanned Laser Eye Tracking, Static LFI for Human Computer Interaction for AR glasses,* and *Static LFI Eye Tracking.*

The first part of the thesis focuses on replacing VOG systems for retinal projection AR glasses through a scanned laser eye-tracking approach.

Contribution in the second part *Static LFI Human-Computer Interaction for AR Glasses* of the thesis focuses on low power gaze HCI through static LFI sensors.

To enable low-power gaze interaction with AR glasses, the publications B.1 and B.2 introduce static LFI sensors operating with the Doppler principle to measure rotational eye velocities. Furthermore, based on the measured eye velocities, a gaze gesture recognition algorithm is derived to enable a gaze gesture-based control of the glass's user interface.

Contribution within the third part *Static LFI Eye Tracking* focuses on fusing multiple static LFI sensors to develop a high-speed eye-tracking senor.

This thesis is based on a rich set of fundamentals from machine learning and signal- and image processing over embedded systems to optical technology, covering various aspects from computer science and electrical engineering.

The following sections summarize each part and its contributions regarding motivation, methods, and the achieved results. At the end of the chapter, a summary of the individual contribution is drawn.

3.1   Scanned Laser Eye Tracking

The contributions [127, 128, 129, 130, 131] focuses on scanned laser eye-tracking sensor approaches for retinal projection AR glasses in order to replace VOG systems and overcome the technological challenges of VOG systems as outlined in Section 2.3.6. The first two publications focus on the sensor integration concept, the second and third publication on ambient light robustness and minimization of power consumption, and the final publication on robustness against glasses slippage and reduction of calibration effort.

3.1.1   Scanned Laser Eye Tracking for Retinal Projection AR Glasses

This subsection is based on the publications A.1 *Low Power Scanned Laser Eye Tracking for Retinal Projection AR Glasses* and A.2 *A novel camera-free eye-tracking sensor for augmented reality based on laser scanning* in Appendix A.

Motivation

Retinal projection AR glasses are emerging as potential near-eye display technology with several advantages compared to other near-eye display technologies like high contrast and low power consumption. However, the main drawback of the retinal projection display technology is the single eye box due to the tiny exit pupil, limiting the effective eye box. Therefore, an eye box expansion mechanism like exit pupil steering is necessary to increase the eye box size. Furthermore, introducing an eye box expansion mechanism requires robust eye-tracking; thus, eye-tracking sensor technology is mandatory for the success of retinal projection AR glasses.

Integrating state-of-the-art VOG eye-tracking sensors into lightweight AR glasses is challenging

due to the large dimensions of camera optics and the required optical axis of the camera sensor w.r.t the visual axis of the eye. Furthermore, the power consumption of camera sensors limits the operation time of battery-constrained AR glasses. Therefore, the possibility of integrating a scanned laser eye-tracking system into a retinal projection AR glass as a replacement of VOG sensors is investigated in these publications to solve these limitations.

Methods

For this purpose, an IR laser was integrated into the RGB laser projection module of a retinal projection AR glasses system. The IR laser beam shares the same beam path as the RGB laser beams, and thus the existing MEMS micro mirror scanner used for image projection can be reused to scan the IR laser beam over the surface of a holographic free space combiner. By adding a additional optical function for the IR wavelength to the holographic beam combiner, the IR light is deflected towards the eye such that a 2D area on the surface of the eye is illuminated.

An external single-pixel photodetector was integrated into the glasses frame temple to generate a grey scale image from the illuminated area on the eye. The photodetector captured the intensity of the back scattered IR light for each scan position (α, β). Combining the single-pixel detector and the free space combiner yields a virtual single-pixel grey scale camera. As the laser beams are off-axis w.r.t the photodetector, dark pupil images are captured by the single-pixel camera. This allows the application of state-of-the-art VOG algorithms for pupil tracking and gaze estimation.

Results

Reusing the existing MEMS scanner leads to the comparable low power consumption of the proposed sensor of roughly 11 mW mainly driven by the photodetector circuitry and the IR laser. In addition, the holographic free-space combiner of the single-pixel camera enables virtual observation of the eye from a centered perspective through the glasses lens while not obscuring the user's FOV. This is a superior perspective for VOG algorithms as the pupil can be observed in a wide range of pupil positions. Finally, the system's output allows reusing state-of-the-art VOG algorithms. Therefore, the proposed system achieves similar accuracy as state-of-the-art VOG systems and immediately benefits from advancements in VOG algorithm developments. The achieved gaze accuracy of the system was $1.31\,°$, and the precision $0.01\,°$. The theoretical achievable single-pixel gaze angle resolution estimated from the geometric system analysis was $0.28\,°$. The deviation between this theoretical gaze accuracy and the achieved gaze accuracy is related to the optical geometry of the laboratory setup, which leads to a large pixel size compared to the glasses geometry and a defocusing of the laser beam and thus to a reduction of the image contrast.

3.1.2 Scanned LFI Eye Tracking for Retinal Projection AR Glasses

This subsection is based on the publications A.3 *A novel-eye-tracking sensor for AR glasses based on laser self-mixing showing exceptional robustness against illumination* and A.4 *A Highly Integrated Ambient Light Robust Eye-Tracking Sensor for Retinal Projection AR Glasses Based on Laser Feedback Interferometry* in Appendix A.

Motivation

Aside from power consumption and sensor integration, VOG systems further suffer from a signal loss in the presence of ambient light and a high-power consumption through the required image processing steps for pupil segmentation. In addition, other artifacts such as partly occluded pupils by the lashes or the eyelid and wearing mascara reduces the accuracy of VOG systems.

As AR glasses are everyday devices, robust operation and thus robust eye-tracking under various lighting conditions are mandatory. In addition, a similar robust gaze estimation for various pupil occlusions and mascara is required to achieve a great user experience for retinal projection systems with pupil steering.

Methods

A similar optical architecture as proposed by the publications discussed in Section 3.1.1 is chosen. Only the IR laser sensor inside the RGB laser projection module is replaced with an LFI sensor. The advantage of the LFI sensor is its coherent measurement principle, which is robust against external lighting. This aspect is investigated in the paper A.3 by exposing the sensor to various ambient lighting sources.

In A.4 the LFI sensor is integrated into an AR glasses prototype system in a laboratory setup to show the miniaturization potential as well as to investigate the system performance. In addition, a highly transparent IR HOE as free space combiner was fabricated to redirect the IR light of the LFI sensor towards the eye.

Compared to the external off-axis single-pixel photodetector used in the previous approaches the LFI sensor consists of an IR laser and a on-axis photodetector integrated into the back DBR of the laser. Therefore the photodetector is perfectly aligned with the illumination source, emphasizing the detection of the pupil through the bright pupil effect. Furthermore, the bright pupil response reduces the complexity of pupil segmentation algorithms as the pupil contour is immediately derived from the detected signals of the LFI sensor. As the bright pupil signal is corrupted by speckling, a multivariate Gaussian fitting approach was proposed to derive the pupil contour.

Results

The introduction of the LFI sensing modality led to a robust pupil detection method that is immune to ambient light at eye-safe exposure limits and provides robust pupil signals even for partly occluded pupils. Furthermore, as only a bright pupil signal is detected, the approach is robust against lashes and mascara. The robust signals from the pupil further reduce the computational complexity of pupil segmentation algorithms and thus reduce the overall system power consumption. A laboratory experiment with 15 participants unveiled a gaze accuracy of 1.674 ° and a precision of 0.945 ° of the proposed system.

3.1.3   A Holographic Single-Pixel Stereo Camera Eye Tracking Sensor for Retinal Projection AR Glasses

This subsection summarizes the publication A.5 *A holographic single-pixel stereo camera eye-tracking sensor for calibration-free eye-tracking in retinal projection AR glasses* in Appendix A.

Motivation

A significant challenge for VOG eye-tracking sensors in everyday devices such as AR glasses is the degradation of gaze accuracy due to glasses slippage [93]. The reduction in gaze accuracy when the glasses slip is mainly because gaze estimation algorithms assume a known static mapping function f between glasses- and the eye coordinate system, which is stationary over time. This mapping function f is derived from an initial system calibration performed by the user. If the glasses slip, f is no longer valid, and a recalibration is required [92]. Frequent recalibration by the user leads to a poor user experience. A possible solution to obtain f without a calibration step is to use a stereo camera setup, e.g., used in the Tobii Pro glasses [97]. However, this requires the integration of two camera sensors per eye, which drastically increases the power consumption of VOG systems and makes them impractical for lightweight AR glasses.

Methods

In this publication, a virtual single-pixel holographic stereo camera is porposed to achieve calibration-free, slippage-robust eye-tracking for the previously proposed scanned laser eye-tracking approaches, while keeping the advantgeous sensor integration and low power consumption. The main idea to create a stereoscopic perspective of the user's eye in a near-eye display relies on the spatial multiplexing of the HOE, which is used as a free space combiner in retinal projection systems.

By scanning over the spatial multiplexed HOE, two images of the eye from two perspectives generated by two optical functions embedded in the spatial multiplexed HOE are captured sequentially. By applying a cone reconstruction algorithm, the pupil disk can be reconstructed

from the two images in space with respect to the origin of the virtual camera. Therefore, the mapping f is derived from each pair of images. As the reconstruction algorithm relies only on a set of algebraic equations and omits demanding computational operations, a power-efficient implementation on AR glasses is reasonable. A spatial multiplexed HOE was fabricated using a holographic wavefront printer to assess the gaze accuracy of the proposed system experimentally.

Results

With the fabricated spatial multiplexed HOE the robustness of the proposed approach was investigated using a laboratory setup. As a result, a gaze accuracy of 1.35° and a precision of 0.02° over a FOV of 30° was achieved. The resolution of the individual camera sensors was 110 px × 240 px with a frame rate of 60 Hz. In addition, the reconstruction of the pupil disc in space using an artificial eye model was shown.

3.2 Static LFI Human-Computer Interaction for AR Glasses

The last part of the thesis focused on eye-tracking methods tailored to retinal projection AR glasses. However, aside from retinal projection display technology, other near-eye-display technologies like wave-guide displays or micro-LED displays exist. In addition, audio glasses such as the Amazon Echo frames [132] are emerging to the market without a near-eye-display. Therefore, gaze-based interaction concepts based on static LFI sensors are investigated through the contributions [4, 5, 133, 134] in this part of the thesis to enable gaze-based HCI concepts for AR glasses independent of the display technology.

3.2.1 A Novel Gaze Gesture Sensor for Smart Glasses Based on Laser Feedback Interferometry

This subsection is based on the publications B.1 *A Novel Gaze Gesture Sensor for Smart Glasses Based on Laser Self-Mixing* and B.2 *A compact low-power gaze gesture sensor based on laser feedback interferometry for smart glasses* in B.

Motivation

Various HCI input methods such as push buttons or capacitive sliders have been adapted from existing smart wearable technologies such as mobile phones to the field of smart glasses. The main drawback is that they do not allow hands-free interaction with the glasses, which reduces the immersion of AR glasses and limits the user experience, especially during cycling or driving. Aside from device interaction purposes, low-power always-on input methods like a push-button on a smartphone are used to control the display to save power and thus extend battery life. Since the display in AR glasses like the smartphone accounts for the tremendous

power demand, implementing a low power always-on input modality is also beneficial from a power-saving perspective.

Methods

Static LFI sensors are investigated to achieve low power always-on gaze gesture control for AR glasses. Static LFI sensors are capable of measuring the eye's rotational velocity as well as the distance between the sensor and the eye through a triangular modulation of laser wavelength, leading to a near range FMCW LIDAR. Due to the small size, high robustness against ambient light, and low power consumption, the sensor perfectly fits the requirements of AR glasses. From measured raw eye velocities, eye movement directions into the four quadrants up, down, left, and right were derived using a decision tree to classify four gaze symbols. In addition, blinks were classified from the distance measurement signal yielding a fifth gaze symbol. Finally, aside from classifying individual gaze symbols, an Finite State Machine (FSM) was proposed to model gaze gestures as a sequence of individual gaze symbols.

Results

A laboratory setup was built up to evaluate the proposed static LFI gaze gesture sensor. Velocity and distance features from two participants performing in a total of 162 gaze symbols were captured together with a camera sensor used to derive ground truth labels on a single sample level. As a result, the proposed gaze symbol classifier achieved a macro-F1-score of 93.44 % on a single sample scale. Furthermore, as the static LFI sensors are capable of measuring distance and velocity features at a sampling rate of 1 kHz a gaze gesture was recognized before the user finished the eye movement belonging to the gaze gesture, leading to a negative latency of about 100 ms. This allowed the user interface to start rendering interaction-dependent content even before the user finished the gaze gesture input.

3.2.2 A CNN-based Human Activity Recognition System for context-aware Smart Glasses

This subsection is based on the publications B.3 *A CNN-based Human Activity Recognition System Combining a Laser Feedback Interferometry Eye Movement Sensor and an IMU for Context-aware Smart Glasses* and B.4 *U-HAR: A Convolutional Approach to Human Activity Recognition Combining Head and Eye Movements for Context-Aware Smart Glasses* in Appendix B.

Motivation

State-of-the-art AR glasses rely on interaction concepts derived from existing wearable technology like push buttons, capacitive sliders, hand gesture control, or voice control. However,

these interaction methods rely on active interaction with the glasses, distracting the user and limiting the immersion of the AR glasses. To resolve this issue and enable true user immersion for AR glasses adding context awareness through recognition of human activities is proposed to control the user interface without active interaction. While the recognition of physical activities through a body-worn motion sensor or the recognition of cognitive activities through an eye-tracking sensor, are well established in research, the combination of both sensor modalities is less well explored. The combination of both sensors allows recognition of a rich set of human activities to infer a rich set of contextual information during everyday activities.

Methods

The input of the HAR system used to derive contextual information is eye- and head movement trajectories over time. The head movement trajectories are detected by an IMU sensor attached to the glasses frame. In contrast, the trajectories of the eyes in B.3 are detected by static LFI sensors, as also used in the previous section. In B.4, the static LFI sensors are replaced by a commercially available VOG sensor to highlight the operation of the HAR system independent of the eye movement sensor, thus enabling context awareness through HAR e.g., VR glasses with already integrated camera-based eye-tracking sensors.

To recognize human activities from the captured eye- and head movement trajectories, a windowing approach was chosen to split raw data into slices of 30 s from which features as input of a HAR classifier are derived. Furthermore, a CNN model was proposed to automate feature extraction. A CNN model allows the extraction of features on different time scales and reduces the input data size and thus the required system memory. Compared to other network architectures, CNN models further require a comparatively small number of parameters, which is beneficial primarily for integrating the HAR system into memory-constrained embedded systems. In B.4, the number of model parameters was further optimized by adopting a U-Net like CNN model for the HAR task.

To personalize the HAR classifier to each user and thus reduce false classification, transfer learning was applied to adapt the decision boundaries of a pretrained CNN model.

Finally, an experimental setup was the build-up to collect head- and eye-movement data from a set of participants, which performed the seven activities cycling, walking, talking, reading, typing, solving, and watching media. The activities were chosen to represent activities spanning from cognitive activities like reading to varied activities like talking involving both head and eye movements, as well as physical activities like cycling.

Results

By using state-of-the-art CNN models for activity classification and the adaption of the decision boundaries through transfer learning an macro F1-score of 88.15 % and 86.59 % was achieved for B.3 and B.4 respectively. The evaluation was carried out by using leave-one-

participant-out cross-validation Leave one Participant out Cross Validation (LOPOCV). Both HAR systems outperformed existing works on a more challenging set of activities. They showed generalization across users by evaluating the HAR systems of both works on a set of 15 and 20 participants, respectively. In addition, the importance of head- and eye movements for different activities was highlighted through an ablation study in both publications. The study supports the initial idea that the combination of both sensor modalities enhances HAR across various activities and thus enables context awareness through HAR for AR glasses.

3.3 Static LFI Eye Tracking

The second part of this thesis introduced static LFI sensors for HCI e.g., to recognize gaze gestures from measured relative eye movements. In this part of the thesis, the initial approach of static LFI sensors is extended to achieve full eye-tracking with static LFI sensors. The main advantages of the approach are the high update rate, the ambient light robustness, and a, compared to low power consumption in a VOG system. In addition, the proposed system operates independent of the display technology and thus can be used in AR and VR glasses.

This part is based on the publication C.1 *Static Laser Feedback Interferometry Gaze estimation* in Appendix C.

3.3.1 Motivation

VOG systems are limited in their update rate due to limitations of camera technology as well as the exponential increase of power consumption with respect to the update rate due to the camera sensor itself as well as the image processing required to infer gaze information from images In addition, VOG systems suffer from obtrusive sensor integration, loss of gaze estimation accuracy in dynamic lightning conditions and degradation of gaze accuracy in the presence of glasses slippage To overcome these limitations, this publication extends the gaze gesture approach. It introduces a model-based eye-tracking approach optimized for static LFI sensors.

Methods

Like VOG model-based eye-tracking approaches, a geometric model of the human eye is used to link LFI measurements to the pose of the eye. For this purpose, a measurement setup was built up, and the human eye of a participant was scanned with a static LFI sensor to derive the geometric eye model from LFI distance measurements. The experiment showed that most light of the IR laser beam is backscattered from the sclera, iris, and retina. Thus, the transparent optical components like the lens or the cornea can be omitted in the geometric eye model.

A static LFI simulation tool was introduced to generate measurement trajectories of a multi-

static LFI configuration and avoid human error. LFI sensors were modeled as point sources, and the human eye model was derived from a 3D scan of a human eye. The derived 3D model is rotated in a 3D Computer Aided Design (CAD) program based on an input eye movement trajectory (θ, ϕ). At each trajectory point, the measured distance and velocity for each modeled LFI sensor are calculated from the 3D model. In addition, a sensor noise model is added to the simulation tool to consider sensor noise.

An algorithm consisting of four stages was proposed to estimate the gaze vector from a set of static LFI distance and velocity measurements. In the first stage, the part of the eye (none, sclera, iris, retina) hit by an individual IR laser beam of an LFI sensors were classified by using the distance measure and the distance difference as input features for an Hidden Markov Model (HMM) classifier. In the second stage, the eyeball center is estimated via trilateration from the distance measures of at least three LFI sensors. The use of trilateration allows for continuous reestimation of the eyeball center. This reestimation adds slippage robustness to the proposed method. In the third stage, the gaze angle of the eye is estimated through continuous integration of measured rotational velocities. To compensate for errors introduced by sensor noise and an unknown initial pose of the eye frequent absolute eye pose estimation is added as a final stage to the gaze estimation algorithm. Distance measurements are fitted to the derived geometric eye model during this step.

3.3.2   Results

A laboratory setup was built up to characterize distance and velocity in a range of 20 mm - 30 mm and 0 $°/s$ - 500 $°/s$ to set up the sensor noise model required for simulation of multi LFI laser sensor distance and velocity measure trajectories. A distance noise of 68.66 μm, and a velocity noise of 2.5 $°$/s was measured. In addition, the classification accuracy of the proposed HMM model to classify the part of the eye hit by the laser was evaluated, and a macro F1-score of 93.33 % was reported. Finally, an actual eye trajectory, including different eye movements from fixations to saccades, was captured by an VOG system and fed into the simulation tool to generate distance and velocity measurement trajectories for six static LFI sensors. Based on the generated trajectories the gaze estimation algorithm was evaluated. A gaze accuracy of 1.79 $°$ was achieved.

3.4   Summary

To summarize the contributions made through this thesis, the radar chart from Section 2.3.9 is advanced by the eye-tracking sensor approaches contributed by this thesis. The resulting radar chart is shown in Figure 3.2.
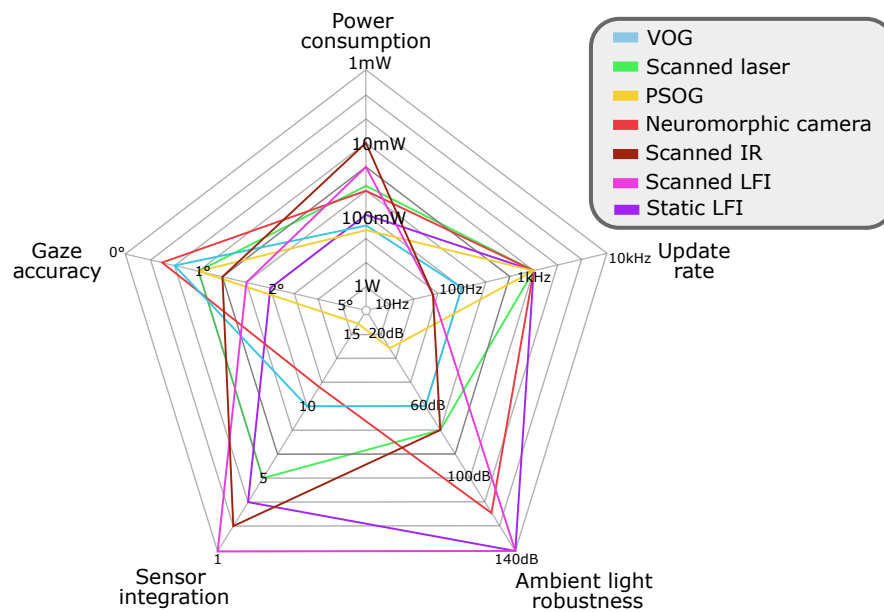
Figure 3.2: Comparison of the presented three developed eye-tracking sensor approaches compared to the state-of-the-art eye-tracking sensor technologies.

The following contributions within this thesis are summarized for each part in the next three subsections.

3.4.1 Scanned Laser Eye Tracking

The first part of the thesis contributes a scanned laser eye-tracking sensor technology for integration into retinal projection AR glasses. The approach closes the gap between established VOG systems by capturing 2D images of the eye with a highly integrated single-pixel camera setup.

Within the **publications A.1 and A.2** the *Scanned IR* approach is presented, which rely on a 2D scanner and an off-axis photodetector to capture dark pupil images. The main advantages are the high sensor integration and the redirection of the light via an HOE to create a superior perspective of the virtual camera compared to state-of-the-art VOG sensors. Therefore, these publications address the technological challenges of **camera sensor integration** as well as **power consumption**.

Within the **publications A.3 and A.4** the *Scanned LFI* approach is presented, which extends the previous publications of this part of the thesis by using an LFI sensor and its integrated in-axis photodiode to capture bright pupil images. This approach's main advantage is exploiting the red-eye effect to capture the bright pupil directly. Therefore, the sensor segments the pupil in the captured image and shows a robust pupil signal independent of the iris color or pupil occlusion through lids or lashes, which further reduces the algorithm power consumption

compared to VOG systems. In addition, the sensor operates coherently and thus is immune to ambient light. Within these two publications, the technological challenges of **ambient light robustness** and **power consumption** are addressed.

Within the final **publication A.5** of this part of the thesis, a methodology is contributed to achieve slippage robustness and minimize the calibration effort of the previously introduced approaches *Scanned IR* and *Scanned LFI* by spatial multiplexing of the HOE to capture the eye from two perspectives. This results in a single-pixel stereo camera system, which allows reconstruction of the pupil disk in 3D from a single pair of images. Therefore this publication addresses the technological challenges of **glasses slippage** and **calibration**.

Figure 3.2 summarizes the *Scanned IR* and *Scanned LFI* approaches with respect to the requirements derived in Section 2.2.5. Especially the *Scanned LFI* method shows superiority w.r.t sensor integration and ambient light robustness compared to all investigated technologies. In addition, the scanned laser sensor approaches achieve a comparable low power consumption. The theoretical system analysis in A.2 further shows that from a geometric point of view, a maximum gaze accuracy of $0.28°$ is achievable in a near-eye setting. Therefore, improvements regarding gaze accuracy are expected in the future.

3.4.2 Static LFI Human-Computer Interaction for AR Glasses

In the second part of the thesis, static LFI sensor technology for gaze interaction applications is contributed. The sensor modality allows deriving the unique features *distance* towards the eye and the eye's rotational *velocity* independent of the display technology with an outstanding sample rate of 1 kHz.

Within the **publications B.1 and B.2** gaze gesture interaction with this sensor modality is investigated. As the sensors consume only a fraction of the power of VOG sensors, they can be used as always on gaze gesture interaction sensor, e.g., for low power system wake up my gaze. Due to the high sample rate, it is further possible to classify a gaze gesture before the user finishes its execution leading to a negative sensor latency, which relaxes the system rendering constraints.

Within the **publications B.3 and B.4** the combination of gaze and head movement interaction are investigated to achieve context-awareness for AR glasses. Context-awareness is derived through HAR based on eye and head movements. For the collection of eye movements, the LFI sensor is used as a power-efficient alternative to VOG sensors. During an experimental study, data for seven activities is collected, and a classification approach to fuse head- and eye movement data with an CNN model is contributed. The same experiment is also made with an VOG sensor to show the generalization of the approach to existing AR glasses with VOG sensors integrated. The collected second dataset is published to emphasize research within context-aware AR glasses through HAR as an additional contribution to this part of the thesis.

3.4.3   Static LFI Eye Tracking

The main contribution of the **publication C.1** in this part of the thesis is a novel high-speed eye-tracking sensor approach based on highly integrated, ambient light robust static LFI sensors. Compared to VOG systems, the LFI sensors did not resolve the eye's surface in 2D as they have no spatial perception. This reduces the amount of captured data, which allows a higher update rate while reducing power consumption. With the introduction of the multi LFI sensor fusion algorithm based on a geometric model and spatial perception of the sensor is achieved by integration of observations over time. This allows obtaining absolute gaze estimation for static LFI sensors, including slippage robustness. Therefore within this publication the technological challenges of **power consumption**, **glasses slippage**, **update rate** and **ambient light robustness** are addressed.

In Figure 3.2 the *Static LFI* eye-tracking sensor technology developed within this part of the thesis is summarized with respect to the derived requirements for eye-tracking sensor technology for AR glasses within Section 2.2.5. The sensor technology achieves comparable update rates to emerging neuromorphic camera eye-tracking approaches while outperforming other sensor technologies regarding ambient light robustness.

4 Discussion

While the introduction of mobile eye-tracking for everyday devices enables new applications and opportunities to improve the immersion of AR glasses, e.g., through gaze-controlled interaction, it poses some risks to the user. The main risks are the medical impact of long-term IR exposure to the human eye and the impact on the user's privacy due to everyday eye-tracking. These aspects will be discussed in the following sections.

4.1 Health Impact of long-term IR Radiation on the Human Eye

The human eye is exposed to IR illumination regularly as the heat transfer from the sun to the earth occurs in the IR regime. While the human organism can handle this natural IR exposure. Recently, a rising number of artificial IR light sources like LIDAR sensors on cars or IR flooding illumination and IR dot projectors used for face recognition on smartphones are introduced. Finally, VOG eye-tracking systems rely on IR LEDs, e.g., to illuminate the eye region or produce glints on the cornea. While LIDAR sensors and IR illumination used for face recognition on smartphones emit artificial IR light only for a relatively short duration, state-of-the-art VOG systems with flooding IR illumination for AR glasses expose the eye with artificial IR illumination on a daily basis over a long period of time [135].

To address the potential risk of IR illumination of VOG systems, the research network Communication by Gaze Interaction (COGAIN) [135] initiated a study to explore safety issues in eye-tracking by IR illumination. The study identified three potential optical radiation hazards sources through IR illumination. The first source of the potential hazard to the eye is a thermal hazard to the cornea. This hazard mainly applies to IR illumination in the IR-B band ranging from 1400 nm - 1 mm wavelength as the tissue of the cornea has high transparency for IR wavelength in the IR-A band ranging from 780 nm - 1400 nm not absorbing much IR energy. The second potential hazard source is a retina tissue thermal burn. According to Vos et al. [136] the risk of a thermal burn of the retina exists primarily in the range from 400 nm - 1400 nm and thus including the IR-A band at which VOG IR illumination sources are emitting. Finally, it is known from workers dealing with hot materials, e.g., molten glass over a long period, that

there is a higher risk of developing cataract due to the intense IR radiation, especially from the IR-A band [137].

As a result of the study by COGAIN the International Commission on Illumination (CIE) issued in 2021 a technical report on the optical safety of eye trackers applied for extended duration [138]. The authors analyzed the different sources of potential radiation hazards for IR illumination. They pointed out that the most limiting exposure criterion is the IR exposure limit, leading to cataracts in long-term exposure conditions. The study further analyzed different state-of-the-art VOG systems and showed that the emitted dose of IR radiation for all tested VOG systems was below the IR exposure limit which increases the risk to develop cataract [138].

The scanned laser systems, as well as the static LFI systems introduced in this thesis, rely on coherent IR illumination from laser diodes with a compared to LEDs slight divergence angle. Therefore the authors in [138] pointed out that the long-term exposure limits for laser illumination sources need to be derive from the International Electrotechnical Commission (IEC) 60825-1 norm [139]. To protect participants' health during the experiments with the proposed static and scanned systems for all experimental setups, a risk assessment according to IEC 60825-1 was made. For the experimental setups in this thesis, the most restricting exposure limits, which assume a daily dose of IR radiation with a duration of >8h, were used to rate the eye safety.

The exposure limit for the scanned systems with geometry as described in appendix A.2 is according to [139] 9.25 mW for a wavelength of 820 nm and exposure duration of >8h. Compared to the exposure limit of 677 μW for a static laser of the same wavelength is because according to [139] scanning systems are rated as pulsed systems. The scanned IR system introduced in A.1 and A.2 used a IR laser with an optical power of 150 μW on the eye surface and the scanned LFI systems introduced in A.3 and A.4 used an IR laser with an optical power of 142 μW on the eye surface. Therefore, both scanned laser systems introduced in this thesis stay below the exposure limit of a static IR laser, ensuring eye safety even in fault conditions, e.g., when the MEMS micro scanner stops scanning the laser beam. The scanned LFI approach introduced in A.3 and A.4 further shows a robust signal in the presence of low-level illumination. While other VOG systems tend to increase the optical power of their IR flooding illumination in low lighting conditions, the scanned LFI system did not require to increase in the optical power.

The exposure limit for the static LFI approaches discussed in the publications B and C also stay at their operating wavelength of 850 nm below the exposure limit of 778 μW for each sensor. In addition, the mechanical design of the demonstrators and laboratory setups built during this thesis ensured that the combination of multiple LFI sensors did not violate exposure limits. Therefore, the static LFI approaches introduced in this thesis did not pose any hazard to the participants.

From a product development perspective, it cannot be ensured that for different head geome-

tries or in the case of glasses slippage, the laser beams of multiple LFI sensors did not hit the same spot and thus violate the exposure limits. Therefore, time multiplexing of the individual sensors is required to ensure eye safety for a product. Time multiplexing does not harm the performance of the sensors as the current update rate of 1 kHz can be preserved by slight adjustments on the triangular modulation rate of the individual static LFI sensors.

To finally rate the eye safety of an AR glasses, it is further required to ensure that the exposure limit of all visible and invisible light and laser sources combined stays below the exposure limits. Therefore, carefully considering eye safety on a system level is mandatory.

4.2 Privacy Concerns of Mobile Eye Tracking

Privacy concerns in the domain of mobile eye-tracking can be divided into concerns regarding user authentication and biometric identification as well as information obtained through the linkage between gaze- and environment information.

4.2.1 User Authentication and Biometric Identification

The advancement of sensor technology for the acquisition of gaze information inevitably leads to the extraction of features from the eye usable for user authentication and occular-based biometric identification [29].

Ocular-based biometric identification relies mainly on high-quality images of the user's eye to derive structural features of the iris or on high-quality scans of the retina to derive structural features such as the retina's blood vessels for retina recognition [140]. These methods allow the identification of a single user and are therefore reliable identification methods for security-critical applications [140].

The scanned laser eye-tracking method developed in this thesis can capture images of the eye, and the scanned LFI method can capture images of the retina.

According to [141] an image resolution of at least 200 pixels across the iris and a Signal to noise ratio (SNR) of at least 40 dB is necessary to gather enough information from the iris for biometric identification. Both aspects are not fulfilled by the scanned laser method as introduced in A.1 and A.2.

Retinal recognition systems rely on high-quality retina images to derive individual user features from the blood vessels [140]. However, the scanned LFI method introduced in A.3 and A.4 is not capable of resolving individual blood vessels of the retina as the interference signal of the LFI sensors are superimposed by speckle patterns.

In contrast, user authentication approaches perform recognition of a user in a closed set of N users, which is compared to biometric identification as less complex [142]. Therefore,

user authentication using eye-tracking features like number and duration of fixations, eye movement velocities, and amplitudes or blink frequency can be used to identify users with a certain accuracy [143].

The high update rate of 1 kHz together with the capability to direct measure eye velocities, makes the static LFI method introduced in B.1 a beneficial sensing modality to support gaze-based user authentication for AR glasses [144]. Therefore this aspect must be considered during development in the future.

4.2.2   Linkage between Gaze- and Environment Information

The energy-efficient eye-tracking methods presented in this thesis enable ubiquitous eye-tracking in everyday scenarios. However, the gaze information that is thus available in a previously impossible quantity poses an additional risk to the user's privacy, especially if gaze information is linked to environmental information, e.g., derived from a world camera. The linkage between the environment and gaze information enables applications like target advertisement [145] as it can be used to derive private information like user's interest, attention, or other sensitive information like gender, age, or race [146, 147].

Steil et al. [148] presented a potential solution to decouple environment and gaze information. The authors used a shutter in front of the world camera sensor, which is closed in privacy concerning situations like talking or reading messages on the smartphone. To recognize privacy concerning situations, they used the eye movement patterns captured by an VOG system. The publications B.3 and B.4 of this thesis can be used to improve the recognition of privacy concerning situations by adding context awareness to allow the user to define privacy concerning situations in which the world camera is disabled.

In addition, the system design can restrict the use of gaze information. For example, retina projection systems require gaze information for display enhancement methods like exit pupil steering or eye box switching. However, these applications are realized within a closed control loop inside the projection system. Therefore, it is not required to output gaze information outside the projection system.

Finally, the publications B.1 and B.2 of this thesis show that relative eye velocities are sufficient to interact with the AR glasses interface without complete gaze information.

4.2.3   Gaze based User State Estimation

Besides biometry and the linkage between gaze- and environment information, the user's gaze reveals additional user-sensitive information, which can be exploited on a large scale in an everyday eye-tracking setting. This information can be summarized as the user's state, containing the user's emotion [149], cognitive load [150, 32], user's attention [151] as well as medical features [37].

This information can be used either in a user-beneficial way, e.g., through early detection of medical issues, or in a non-beneficial way, e.g., through target advertising based on the user's emotion or attention.

Features from which the user's state is derived are pupil diameter variations, the user's scan path, the duration and number of fixations, (micro) saccades, and blink information. Although the scanned laser methods measure the pupil diameter variation with a low update rate and thus emotion recognition is rather complicated, the static LFI methods can resolve the eye movement-related features with a high update rate. Thus, the static LFI approaches pose a potential risk for gaze-based user state estimation. A potential miss-use of the sensors needs to be prevented, e.g., by adding a differential privacy approach [152].

5 Outlook

Energy-efficient and robust eye-tracking is a key technology required for the success of AR glasses. Furthermore, eye-tracking is crucial for novel hands-free interaction concepts with the glasses and furthermore enables advanced near-eye display technologies like retinal projection. The contributions of this thesis address the limitations of state-of-the-art eye-tracking sensor technology for AR glasses, and different approaches are proposed to solve individual short comes of existing eye-tracking sensors. Furthermore, the optical sensing technology approaches for eye-tracking provided in this thesis can be advanced in several ways, which will be discussed in the following.

5.1 Scanned Laser Eye Tracking

The main difference between the scanned laser eye-tracking sensor approach described in A.1 and A.2 and the scanned LFI eye-tracking approach described in A.3 and A.4 is the position of the photodetector. By combining both sensing methods, the bright as well as the dark pupil effect, can be captured in the same coordinate space as shown in Figure 5.1. This approach yields a dual bright- and dark pupil eye tracker. The pupil response signal from both sensors can be combined to increase the signal strength and improve pupil detection accuracy. In
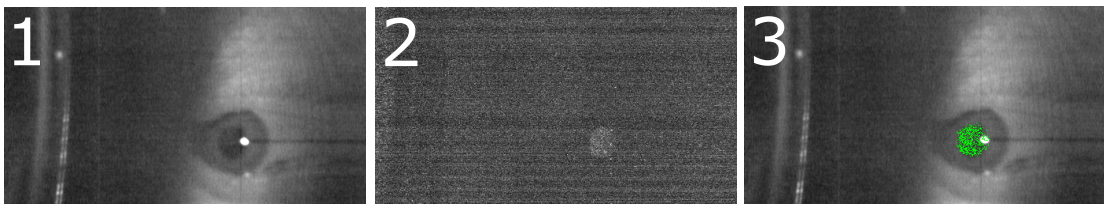
Figure 5.1: 1) Scanned laser image of the eye taken by external off-axis photodetector, 2) Scanned LFI image taken by the laser cavity integrated on-axis photodetector, 3) Scanned laser image with scanned LFI image overlaid in green
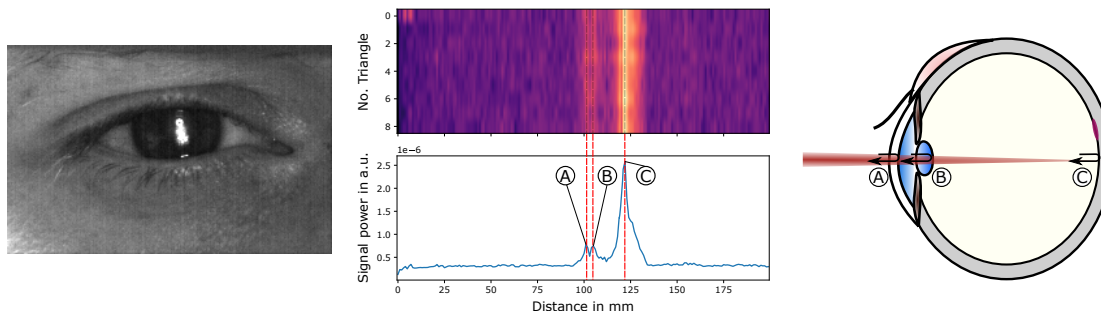
Figure 5.2: Left: Single static LFI sensor IR laser spot on the pupil of a participant captured with an IR camera, Center: Two subplots of the raw distance spectra (top) for a sequence of triangular modulation patterns and the averaged distance spectra (bottom) with three signal peaks A-C in the spectra. The frequencies on the x-axis are converted into corresponding distance measures. Right: Section of the human eye with the LFI laser beam penetrating the cornea and the lens and finally hitting the retina. At each change of the refraction index, a part of the emitted light is back-reflected (A, B), and the remaining light is back-reflected from the retina (C).

addition, the scanned laser image can be used to track landmarks on the eye to compensate for glasses slippage.

An alternative to compensate for slippage is shown in A.5 where a single-pixel stereo camera approach based on laser scanning is introduced. This approach works for both the scanned laser and the scanned LFI approaches. The two perspectives of the stereo camera are created by spatial multiplexing of the HOE in the horizontal dimension. Furthermore, by spatial multiplexing of the HOE in the vertical dimension as well, a multi-camera setup can be realized. This can be used to increase further the covered FOV of the scanned laser eye-tracking sensors or to increase the robustness and accuracy of gaze estimation.

5.2 Static LFI for Human-Computer Interaction for AR Glasses

The distance and velocity features of the static LFI sensors used for gaze gesture recognition in B.1 and B.2 and HAR in B.3 are derived from interference frequencies measured by the photodetector integrated into the laser cavity. The peak frequencies are derived from a distance spectrum obtained by an Fast Fourier Transform (FFT) of the captured interference signal recorded during a rising and falling triangular ramp. Figure 5.2 shows in the center the distance spectra measured by a static LFI sensor hitting the pupil of a participant perpendicularly (white spot on the pupil in the left image). The upper subplot of the figure shows that frequencies of multiple interference's appear in the spectra. The lower subplot shows the signal strength of the spectra calculated for each frequency bin averaged over ten consecutive distance spectra obtained from 10 triangular modulation slopes. From the frequencies and the known LFI sensor parameters, the corresponding distance can be calculated, and the individual peaks

in the spectra (A-C) can be mapped to different parts (cornea, lens, retina) of the eye. From this observation, a further research direction is the applicability of the static LFI sensor for medical or well-being applications in everyday devices. Especially if the sensor is integrated into AR glasses, this might allow early detection of biomarkers, which hint at eye disease like a cataract.

5.3   Static LFI Eye Tracking

The gaze estimation accuracy of the static LFI eye-tracking method is particularly limited by the cases where the LFI laser beams hit the sclera, as in this case, the gaze position is estimated by integrating velocity measurements, and an integration error builds up until measurement data can eliminate the integration error from the iris. One possible solution to increase the accuracy of the static LFI eye-tracking approach is to combine it with a camera sensor operating at a low frame rate. This combines the high update rate and ambient light robustness of LFI sensors with the high accuracy of camera-based eye-tracking systems while reducing the overall system's power consumption. Research towards hybrid eye-tracking systems through multi-sensor fusion offers the possibility to overcome the limitations of VOG systems such as limited sensor update rate and high-power consumption and is therefore of great importance for mobile eye-tracking for AR glasses.

A Scanned Laser Eye Tracking

This chapter includes the publications [127, 128, 129, 130, 131]:

1. **Johannes Meyer**, Thomas Schlebusch, Thomas Kuebler, Enkelejda Kasneci. "Low Power Scanned Laser Eye Tracking for Retinal Projection AR Glasses". In ACM Symposium on Eye Tracking Research and Applications (2020), https://doi.org/10.1145/3379157.3391995

2. **Johannes Meyer**, Thomas Schlebusch, Wolfgang Fuhl, Enkelejda Kasneci. "A novel camera-free eye tracking sensor for augmented reality based on laser scanning". In IEEE Sensors Journal (2020), https://doi.org/10.1109/JSEN.2020.3011985

3. **Johannes Meyer**, Thomas Schlebusch, Hans Spruit, Jochen Hellmig, Enkelejda Kasneci. "A novel-eye-tracking sensor for ar glasses based on laser self-mixing showing exceptional robustness against illumination". In ACM Symposium on Eye Tracking Research and Applications (2020), https://doi.org/10.1145/3379156.3391352

4. **Johannes Meyer**, Thomas Schlebusch, Enkelejda Kasneci. "A Highly Integrated Ambient Light Robust Eye-Tracking Sensor for Retinal Projection AR Glasses Based on Laser Feedback Interferometry". Proc. ACM Hum.-Comput. Interact. 6 (2022), https://doi.org/10.1145/3530881

5. **Johannes Meyer**, Tobias Wilm, Reinhold Fiess, Thomas Schlebusch, Wilhelm Stork, Enkelejda Kasneci. "A holographic single-pixel stereo camera eye-tracking sensor for

calibration-free eye-tracking in retinal projection AR glasses". In 2022 Symposium on Eye Tracking Research and Applications (2022), https://doi.org/10.1145/3517031. 3529616 🏆 Best Short Paper

A.1 Low Power Scanned Laser Eye Tracking for Retinal Projection AR Glasses

A.1.1 Abstract

Next generation AR glasses require a highly integrated, high-resolution near-eye display technique such as focus-free retinal projection to enhance usability. Combined with low-power eye-tracking, such glasses enable better user experience and performance. This research work focuses on low power eye tracking sensor technology for integration into retinal projection systems. In our approach, a MEMS micro mirror scans an IR laser beam over the eye region and the scattered light is received by a photodiode. The advantages of our approach over typical VOG systems are its high integration capability and low-power consumption, which qualify our approach for next generation AR glasses.

A.1.2 Research objectives

The number of commercially available augmented reality (AR) glasses has largely increased in the last years. A new shift in the commercial domain is the integration of eye-tracking sensors into AR glasses to increase user experience by adding new functions and introducing new ways of interaction with systems around us [153]. Examples for interactions with systems around us are driver assistance systems [151], human computer interaction (HCI) [154] or smart home control [155].

Additionally, the performance and resolution of AR applications can be increased by foveated imaging, where high resolution content is only projected sharply on the projection area corresponding to the fovea and lower-resolution content is used in peripheral regions. It is obvious that this technique is dependent on eye-tracking and can significantly reduce the required processing power for image rendering compared to a full-resolution rendering in the full field of view (FOV) [39]. A recent implementation of foveated imaging for VR applications was presented by Tobii [156].

In this research work we focus on the integration of low power eye tracking sensors in AR-Glasses utilizing retinal projection as near-eye display technique. Retinal projection displays are based on direct illumination of the eye's retina using an ultra low-power eye-safe laser projector. Unique aspects of this display technology are a high degree of integratability into a frame temple, providing outstanding design opportunities, as well as near focus-free image projection, providing a sharp projection independent on the user's accommodation state [157].
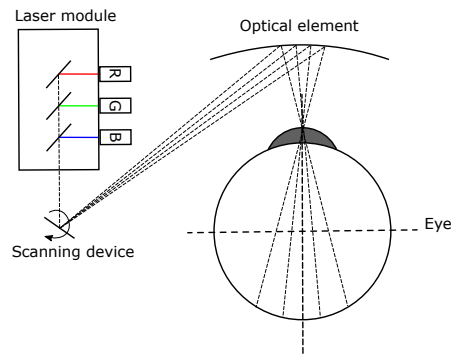
Figure A.1: Principle structure and optical path of a retinal projection system.

Figure A.1 illustrates the basic structure and the main components of a retinal projection system. A laser module combines the laser beams of diode laser emitters for red, green and blue wavelengths. The combined beam is led to a scanning device which deflects the laser beam two-dimensionally onto a holographical optical element (HOE). This element redirects the beam towards the retina of the eye, where it forms a projected image. For most retinal projection systems, eye-tracking is crucial as the scanned beam has to follow the pupil's position to enter the eye [158].

A.1.3 Hypothesis and problem statement

Current eye-tracking sensors for low cost, commercially available systems, e.g. Pupil Labs [53], use video-based oculography (VOG). These eye-tracking sensors rely on a set of infrared (IR) emitters to illuminate the eye and an IR camera to capture eye images. The gaze direction is then determined by image processing algorithms. While camera-based VOG systems are well-established and perform at high accuracy, there is only little potential for significant reduction of power consumption. Further, feasible orientation of the camera modules limits integration into a AR glasses frame and often interferes with the user's view [54, 112].

To solve the mentioned issues of current eye tracking technology new eye tracking sensor concepts with focus on a low power consumption and a high integratability are required. Furthermore, these sensors should be unobtrusive to ensure the user's comfort.

A.1.4 Approach and Methods

The concept we investigated is the integration of an eye tracking sensor into a retinal projection AR glasses by exploiting the existing laser projection unit and the HOE required for the display of the glasses.

This concept extends previous scanning laser eye tracking sensor approaches by [159] and [102]. The main difference is the use of a laser scan path over the full region of the eye. This

allows us to capture a full image of the user's eye and the use of robust slippage invariant state of the art VOG algorithms to extract the pupil position. The integration of the eye tracking sensor into a retinal projection display is shown in Figure A.1.

Figure A.2: Principle structure and optical path of a retinal projector with integrated eye-tracking capabilities. An IR laser is integrated into the RGB laser module and an optical receiver is placed near the eye.

Beside the components of the retinal projector the added components are an IR laser, an optical receiver and an additional laser beam deflection function of the optical element specific to the IR wavelength of the eye-tracking laser. The integration of the new components into a retinal projection system are shown in Figure A.2. The laser projection unit consists of a laser module and a scanning device. The laser projection module projects IR laser light (purple dotted line) onto the optical element. The optical element redirects the laser beam with a defined optical function (e.g. parabolic mirror) onto the surface of the eye. Based on the varying IR reflectivity of the different eye regions, the laser beam is scattered with a different intensity from the surface of the eye. The reflected light is measured by an optical receiver.

A.1.5   Preliminary results
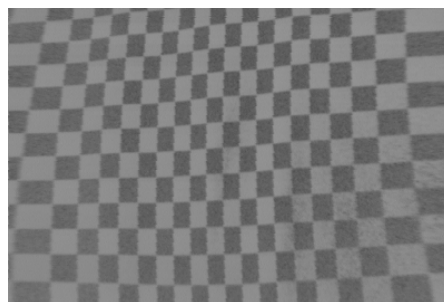
Figure A.3: The image of a 5 mm chess pattern in the eye-tracking region of the laboratory setup.
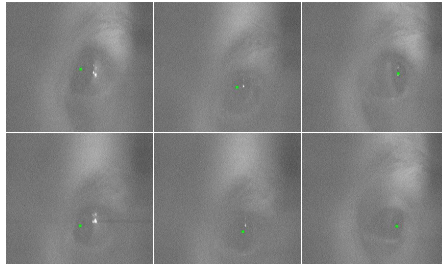
Figure A.4: Sample images captured with the proposed eye-tracking approach. The pupil position is detected by a state of the art CBF pupil detection algorithm [3].

To evaluate the feasibility of such a eye tracking sensor concept we created a laboratory setup by using similar components as for the retinal projection glasses demonstrator shown in Figure A.1. We integrated an IR laser into the projection module and used a semi transparent mirror to redirect the scanned laser beam towards the eye region. To measure the backscattered light from the eye region we used a separate photodiode and place it close to the eye region.

We ensured that the laboratory setup is a class 1 laser system according to IEC 60825-1 [139] and therefore does not pose any medical hazard to the eye. The emitted IR laser power towards the eye is less than $150\,\mu$W.

The power consumption of the proposed eye-tracking sensor is estimated roughly at $11\,$mW using off-the-shelf components. This estimation does not include the power budget of the MEMS micro mirror, which is already included in the power budget of the retinal display. [102] report that state of the art VOG eye-tracking sensors consume more than $150\,$mW of power, which is significantly higher than our sensor approach. Compared to the scanned laser approach by [102], a similar power consumption is achieved.

To evaluate the eye tracking sensor concept we carried out two experiments. One experiment evaluates the spatial resolution of the proposed eye tracking sensor. In this experiment we placed a 5mm chess pattern in the scan region of the eye tracking sensor. Figure A.3 shows the result of the experiment. The image is constructed by synchronized measurements of the backscattered light captured by the photodiode and measurements of the deflection angles of the micro mirror.

In the second experiment we evaluated the feasibility of capturing a human eye with the proposed eye tracking sensor and extracted pupil positions from the captured images using state of the art VOG algorithms. Figure A.4 shows the application of an VOG algorithm for pupil position extraction. To extract the pupil position we used the circular binary features (CBF) pupil detection algorithm by [3]. The estimated pupil center positions are marked by a green dot in Figure A.4. The result is very promising and proves the feasibility to apply standard VOG algorithms to our low-power eye tracking data. This enables robust slippage invariant pupil extraction while minimizing development effort.

69

A.1.6   Broader Impact

The field of low power eye tracking sensors for AR glasses is an emerging area of research [112]. With the increase of AR glasses eye tracking sensors will become state of the art wearable sensors. This impacts additional research fields in the psychological or medical domain e.g. long term supervision of the user's behaviour to detect early signs of medical diseases or mental disorders.

The demand for a low power consumption also impacts the research field of eye tracking algorithm design. New algorithm concepts will focus on low power consumption and high robustness outside the laboratory while keeping a sufficient gaze angle resolution to fulfill the requirements of AR glasses applications.

A.1.7   Future work

We presented a novel eye-tracking sensor for integration into our existing retinal projection AR glasses prototype. Currently, the viability of the approach has been demonstrated under laboratory conditions only. The head-worn demonstrator of the retinal projection system currently contains only red, green and blue projection lasers, as shown in Figure A.5. The next step of our work is focused on integrating the eye-tracking sensor into this demonstrator and evaluation of the performance under real-world conditions, e.g. in the presence of various illumination conditions. For this purpose, some improvements of the photodiode circuit are considered. In addition we want to carry out additional experiments on the laboratory setup to estimate the achievable gaze angle resolution for the AR glasses.

Another challenging task is the integration of pupil tracking algorithms into the head-mounted demonstrator under the constraint of limited computational resources. In particular, a compromise between the required gaze angle resolution for AR glasses applications and the required computational effort to achieve this resolution needs to be examined in more detail.

A.2 A Novel Camera-Free Eye Tracking Sensor for Augmented Reality based on Laser Scanning

A.2.1 Abstract

Next generation AR glasses require a highly integrated, high-resolution near-eye display technique such as focus-free retinal projection to enhance usability. Combined with low-power eye-tracking, such glasses enable better user experience and performance. We propose a novel and robust low-power eye-tracking sensor for integration into retinal projection systems. In our approach, a MEMS micro mirror scans an IR laser beam over the eye region and the scattered light is received by a photodiode. The advantages of our approach over typical VOG systems are its high integration capability and low-power consumption, which qualify our approach for next generation AR glasses. In here along with the technical components, we present a mathematical framework to estimate the achievable gaze angle resolution of our approach. We further show the viability of the proposed eye-tracking sensor based on a laboratory setup and discuss power consumption and gaze angle resolution compared to typical eye-tracking techniques.

A.2.2 Introduction

The number of commercially available augmented reality (AR) glasses has largely increased in the last years. A new shift in the commercial domain is the integration of eye-tracking sensors into AR glasses to increase user experience by adding new functions and introducing new ways of interaction with systems around us [153]. Examples for interactions with systems around us are driver assistance systems [151, 160], human computer interaction (HCI) [161, 154] or smart home control [155, 162].

Additionally, the performance and resolution of AR applications can be increased by foveated imaging, where high resolution content is only projected sharply on the projection area corresponding to the fovea and lower-resolution content is used in peripheral regions. It is obvious that this technique is dependent on eye-tracking and can significantly reduce the required processing power for image rendering compared to a full-resolution rendering in the full field of view (FOV) [39, 38]. A recent implementation of foveated imaging for VR applications was presented by Tobii [156].

Current eye-tracking sensors for low cost, commercially available systems, e.g. Pupil Labs [53], use video-based oculography (VOG). These eye-tracking sensors rely on a set of infrared (IR) emitters to illuminate the eye and an IR camera to capture eye images. The gaze direction is then determined by image processing algorithms. While camera-based VOG systems are well-established and perform at high accuracy, there is only little potential for significant reduction of power consumption. Further, feasible orientation of the camera modules limits integration into a smart glasses frame and often interferes with the user's view [54, 112].

71

These drawbacks limit the use of VOG based eye-tracking sensors for battery powered AR glasses in consumer applications. The requirements of AR glasses in this segment are low-power consumption and a high degree of integration into the frame temple [112]. Additionally, high stability and reliability are required for everyday use in the wild. A main problem that arises in the outside world is however reduced sensor performance due to artificial or natural light sources [56].

To meet these requirements, we propose a novel camera-free eye-tracking sensor based on retinal projection AR glasses. Our main contribution is a new miniaturized low-power eye-tracking sensor approach using state of the art VOG algorithms for integration into retinal projection AR glasses.

Retinal projection is a very promising near-eye display technique for AR glasses. It is based on direct illumination of the eye's retina using an ultra low-power eye-safe laser projector. Unique aspects of this display technology are a high degree of integratability into a frame temple, providing outstanding design opportunities, as well as near focus-free image projection, providing a sharp projection independent on the user's accommodation state [157].
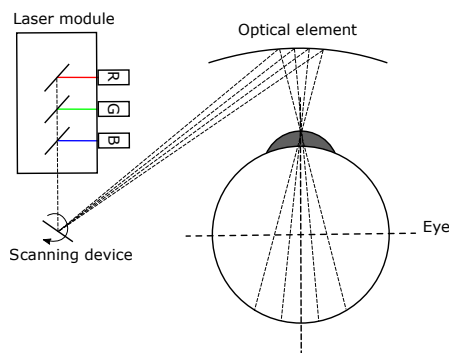
Figure A.5: Principle structure and optical path of a retinal projection system.

Figure A.5 illustrates the basic structure and the main components of a retinal projection system. A laser module combines the laser beams of diode laser emitters for red, green and blue wavelengths. The combined beam is led to a scanning device which deflects the laser beam two-dimensionally onto a holographical optical element (HOE). This element redirects the beam towards the retina of the eye, where it forms a projected image [14]. For most retinal projection systems, eye-tracking is crucial as the scanned beam has to follow the pupil's position to enter the eye [158].

We propose an extension of these retinal projection systems to fully integrate eye-tracking. More specifically, we place an additional IR laser diode in the laser module and a tiny photo-diode close to the joint in the frame temple. This resulting setup can be compared to photo sensor oculography (PSOG). PSOG exploits the varying IR reflectivity of different regions of the eye such as sclera, iris, pupil etc. [110].

Exploiting the scanning device and HOE already available for the projection system, this enables us to significantly simplify the PSOG setup by using only one IR diode laser and one photodiode for the whole tracking region. At the same time, spatial resolution is significantly increased. Moreover, system output are greyscale images suitable for processing using state of the art VOG algorithms.

The remaining of the paper is organized as follows. The next section discusses the state-of-the-art with regard to eye-tracking based on laser scanning. Section 3 presents our eye-tracking technique and its integration into retinal projection AR glasses as well as a mathematical framework to analyse the theoretical spatial and gaze angle resolution achievable by our technique. The evaluation in Section 4 shows the expected viability and resolution of the proposed eye-tracking sensor by an laboratory setup. Based on this setup the resolution for an head worn sensor is estimated utilizing the mathematical framework derived in Section 3. Afterwards a comparison between our sensor approach and VOG eye-tracking sensors is performed. Section 5 concludes this work and gives an outlook to our future research activities.

A.2.3 Related Work

First scanned lasers for retina imaging appeared in the clinical section. Webb et al. introduced a scanning laser opthalmoscope (SLO), which works according to a process of scanning a laser beam over the retina surface to capture an image of the retina [163]. The captured image was used for medical diagnosis of eye diseases. Eye motions during the scan process lead to distortions of the retina images. To resolve this issue, the SLO method was later extended by an eye-tracking system, as for example in [164, 165] who presented a binocular tracking scanning laser ophthalmoscope (TSLO). The beam of a super luminescent diode was collimated and deflected by a 15 kHz resonant scanner horizontally and by a 30 Hz mirror galvanometer scanner vertically onto the eye. A photomultiplier tube and a beam splitter were used to detect the reflections of the eye. The photomultiplier tube was sampled based on the position of the mirrors to extract an image. Afterwards, the eye position was calculated offline on a host device. For this techniques, [165] reported a possible tracking speed of 366 Hz at a resolution of $0,003°$. Such medical scanned laser eye-tracking approaches are characterized by high image resolution, however only during fixation phases of the eye. Furthermore, the optical setups are expensive, large and only suitable for laboratory use. To reduce the size of the optical setups [166] and [167] used micro-electro-mechanical systems (MEMS) micro mirrors as scanning devices.

A different scanned laser approach was described by [159]. They integrated an IR laser into an RGB laser module to track the eye position. The IR laser beam was deflected onto a polygon mirror for horizontal deflection and a galvanometer scanner for vertical deflection. The scanned laser beam was then guided by a prism onto the surface of the eye. To capture the reflections of the eye, a semi-transparent mirror was placed in front of the user's eye region.

The mirror directs the reflected light from the surface of the eye onto a collective lens on top of a photodiode. Based on the position of both scanning mirrors, the photodiode was sampled to capture an image of the eye. This setup was however rather large and was based on multiple mirrors which interfere with the user's view. Therefore, such eye-tracking setups can hardly be integrated into AR glasses.

To reduce the power consumption and size of scanned laser eye-tracking sensors, [101] introduced MEMS micro mirrors to scan a laser beam vertically and horizontally across the eye-tracking region. Based on this improvement, [102] integrated the technology into AR glasses. In their setup, a two dimensional resonant MEMS micro mirror and an IR laser were placed on the inside of the frame temple. The MEMS micro mirror deflected the light towards the surface of the eye. To receive reflections from the eye's surface, a photodiode was placed on the nose bridge of the glasses. To track the horizontal eye position, a linear trajectory was scanned by the MEMS micro mirror over the surface of the eye. The photodiode received the reflected IR light from the surface of the cornea. The maximum light intensity detected by the photodiode was then used to estimate the horizontal eye position. Afterwards, the vertical position of the eye was estimated by a hill climbing algorithm based on the amplitude shifts of the photodiode output in horizontal scanning direction. This approach achieves an angular resolution of approximately $1°$ at a temporal resolution of $3300\,Hz$ and a power consumption of less than $15\,mW$ [102].

The main draw back of this approach is the vulnerability to shifts of the glasses. Even small movements of the glasses on the user face lead to significant drifts and therefore inaccuracy of the eye position estimation [102].

Our approach combines the advantages of robust camera-based eye-tracking systems and the low-power consumption and small size of MEMS scanned laser eye-tracking sensors. Additionally, we integrate an optical element into the spectacle lens to obtain an eye-tracking sensor invisible to the user and therefore do not disturb the users view.

A.2.4   Scanned laser eye-tracking
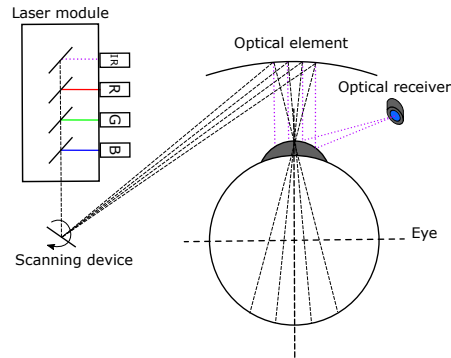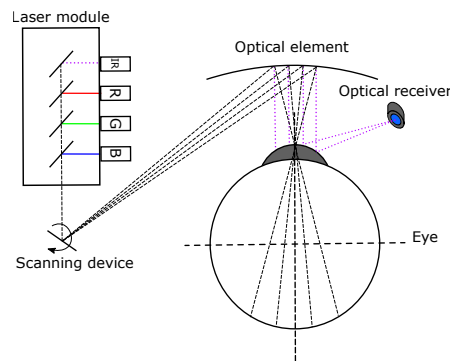
Figure A.6: Principle structure and optical path of a retinal projector with integrated eye-tracking capabilities. An IR laser is integrated into the RGB laser module and an optical receiver is placed near the eye.

The proposed eye-tracking sensor is based on a retina projector for AR glasses as described in [14]. The main components of the system are shown in Figure A.5. Additional components are an IR laser, an optical receiver and an additional laser beam deflection function of the optical element specific to the IR wavelength of the eye-tracking laser. The integration of the new components into a retinal projection system are shown in Figure A.6. The laser projection unit consists of a laser module and a scanning device. The laser projection module projects IR laser light (purple doted line) onto the optical element. The optical element redirects the laser beam with a defined optical function (e.g. parabolic mirror) onto the surface of the eye. Based on the varying IR reflectivity of the different eye regions, the laser beam is scattered with a different intensity from the surface of the eye. The reflected light is measured by an optical receiver.

The components of the sensor are described in more detail in the following.

Laser projection module

Figure A.7 shows a block diagram of a laser projection module.
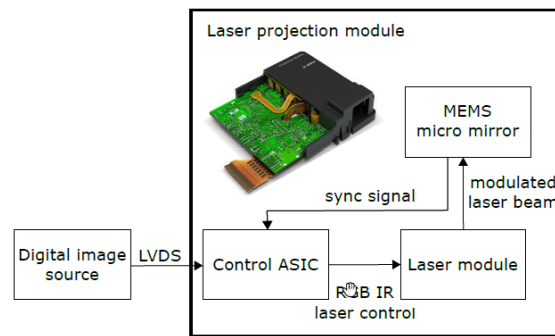
Figure A.7: Block diagram of the laser projection module. Based on incoming frames, the integrated laser module and the MEMS micro mirrors are controlled by a control ASIC to project the frames onto a surface.

The main component of the projection module is the control application specific integrated circuit (ASIC). The ASIC is fed with a digital image via a low voltage differential signaling (LVDS) interface.

Based on the incoming image stream, the lasers are modulated by the RGB values of each image pixel. The IR laser remains in an active state as long as eye-tracking remains active. The modulated and collimated RGB laser beam and the IR laser beam are directed onto the MEMS micro mirrors.

The micro mirror module consists of a fast axis MEMS micro mirror for the horizontal scan direction and a slow axis MEMS micro mirror for the vertical scan direction. The horizontal scan mirror is driven at resonance frequency and oscillates sinusoidal. The vertical mirror is actuated non resonantly by an external magnetic force. The start of a new line scan is reported by sync signals to the control ASIC.

Holographic optical element

An HOE is integrated into the AR spectacle lens to deflect the incoming scanned IR laser light field of the laser projection module towards the surface of the eye as shown in Figure A.8.
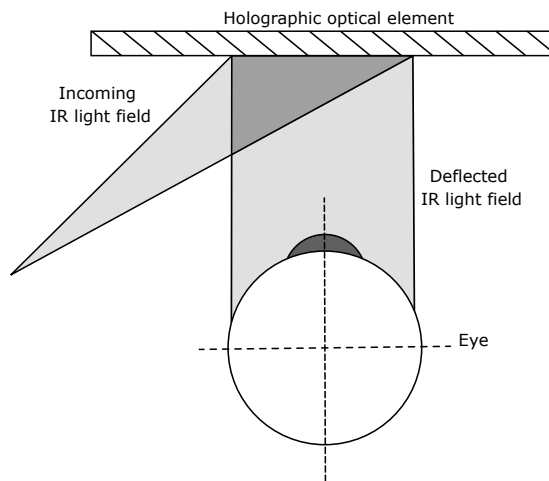
Figure A.8: Description of the function of an holographic optical element (HOE). The Bragg structure inside the HOE diffracts the incoming IR light along parallel lines towards the eye to form a rectangular region in the surface of the eye.

The HOE is realized by recording a mirror function into a photo polymer material. For a more detailed description of the recording process and integration of optical functions into HOEs using Bragg structures, we refer to [168].

The main advantage of HOEs over other optical elements like semi-transparent mirrors is the high selectivity and optical transparency. Ideally, the optical function is only active for the recording wavelength [158].

The HOE function controls the size and contour of the region which is scanned by the IR laser on the eye. For the scanned laser eye-tracking sensor, this scanned surface is referred to as eye-tracking region.

Receiving photodiode circuit

To capture a reflection map of the eye-tracking region a similar receiving photodiode circuit as known from previous approaches like PSOG or the scanned laser approach by [102] is used.Figure A.9 shows the system block diagram of the receiving photodiode circuit.
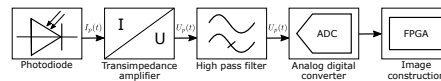
Figure A.9: System block diagram of the receiving photodiode circuit.

The main component is a photodiode which is sensitive to IR light. To increase the resolution and convert the photodiode current $I_p(t)$ into a measurable voltage $U_p(t)$, a transimpedance amplifier (TIA) circuit is used. The voltage $U_p(t)$ is AC coupled via a high pass filter to re-

move low frequency disturbances like ambient light. The amplified and filtered signal is then converted into the digital domain by an analogue to digital converter (ADC). A Field Programmable Gate Array (FPGA) converts the digitized samples into an image utilizing the sync signals of the laser projection module. To capture an image of the eye, the photodiode current is sampled at fixed time intervals by the ADC.
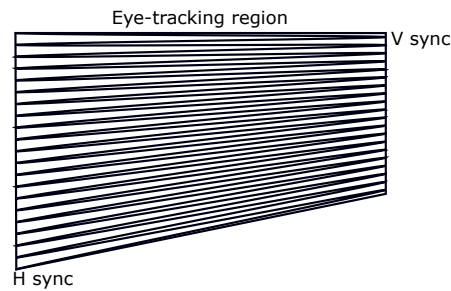
Figure A.10: Simplified sinusoidal scan path of the laser beam over the eye-tracking region. The sync signals are used to reconstruct an image based on the samples captured by the ADC.

The start of a new frame is indicated by the vertical synchronization signal V sync. With the rise of this signal, the ADC starts to sample the photodiode signal. The start of a new row of the frame is indicated by the horizontal synchronisation signal H sync. With these two signals, a reflectivity image of the surface of the eye is reconstructed by the FPGA.

Figure A.10 shows a simplified scan path (much less lines) of the laser beam in the eye-tracking region. The characteristic bounding box results from the geometric setup as depicted in Figure A.11.

The resolution of the image, and therefore the pixel size, is dependent on the bandwidth of the AFE, the sample rate of the ADC, the time-dependent sinusoidal angular scan speed of the laser projection system and the geometric setup.

Spatial resolution

State of the art camera-based eye-tracking approaches utilize pupil edge detection methods to extract the pupil location and calculate the gaze direction [70]. For this purpose, a sufficient spatial resolution in the captured eye region is required. The spatial resolution in VOG eye-tracking approaches is limited by the camera resolution and image distortions due to lens misalignments.

In the scanned laser eye-tracking sensor approach, pixel size and image distortion are related to the scan frequency of the MEMS micro mirrors, the optical path length and the performance of the photodiode circuit [169].

Horizontal distortions occur due to a sinusoidal-like velocity pattern of the laser spot along the scan path shown in Figure A.10. Combined with equitemporal sampling of the photodiode

circuit, this leads to barrel distortions [170].

The spatial resolution is characterized by analysing the image data of a high contrast linear or square edged test pattern with known properties in the eye-tracking region captured by the photodiode circuit. The known properties of the test pattern and the resulting number of pixels in the image are used to calculate pixel sizes at various positions in the image. With this, image distortions due to varying optical path length and the sinusoidal angular velocity pattern are quantified.

Gaze angle resolution

A gaze angle is described by the pupil center position and eye rotation angles. With rotation of the eye, the pupil center, and therefore, the pupil edges are rotated as well. To assess the gaze angle resolution of the proposed scanned laser eye-tracking sensor, a mapping from spatial resolution to angular resolution is required. It is dependent on the geometrical system design, the laser projection unit, the sample rate of the photodiode circuit, the HOE function, and the rotation angle of the eye. For the calculations, a parallel deflection function of the HOE as described in Figure A.8 is assumed.

Figure A.11 shows the geometrical dimensions of the AR glasses setup.
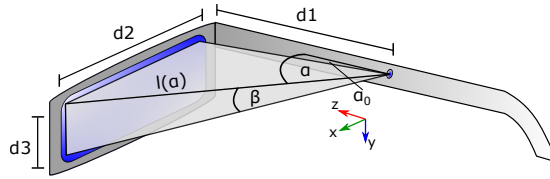
Figure A.11: Integration of the proposed scanned laser eye-tracking sensor into AR glasses. The HOE is integrated into the spectacle lens and the laser projection module is integrated into the frame temple.

The laser projection module is integrated into the right frame temple with a distance d_1 to the hinge between frame temple and spectacle lens. The angles α and β describe the horizontal and vertical scan angles of the laser projection module. The distance d_2 denotes the width of the HOE inside the spectacle lens. The distance d_3 describes the height of the HOE.

As mentioned earlier, pixel sizes vary throughout the eye tracking region. They are dependent on the oscillation frequencies f_h and f_v as well as the maximum scan angles α_{max} and β_{max} of the horizontal and vertical micro mirrors, respectively. In addition, the sample speed f_s of the ADC and the distance d_1 are key parameters.

The x-coordinate of a point on the HOE can be expressed by d_1 and α:

$$x(\alpha) = d_1 \cdot \tan(\alpha). \tag{A.1}$$

The angle α changes sinusoidal in time. For the geometry shown in Figure A.11, this is

$$\alpha(t) = \frac{\alpha_{max}}{2} \cdot \sin(\omega_h \cdot t - \frac{\pi}{2})) + \frac{\alpha_{max}}{2} + \alpha_0. \tag{A.2}$$

α_0 denotes the horizontal offset angle between frame temple and the right edge of the HOE integrated in the spectacle glasses. The pixel size s_x along the horizontal direction can be expressed as the difference between a first horizontal position at angle α and a second position at an angle incremented by $\Delta\alpha$

$$s_x = x(\alpha + \Delta\alpha) - x(\alpha). \tag{A.3}$$

$\Delta\alpha$ denotes the angle increment between two consecutive samples of the ADC and is dependent on the sample speed f_s of the ADC. Therefore,

$$\Delta\alpha = \alpha(t + \frac{1}{f_s}) - \alpha(t). \tag{A.4}$$

This can be used to eliminate t in Equation (A.2):

$$\Delta\alpha = \alpha - \frac{\alpha_{max}\cos\left(\frac{2\pi f_h}{f_s} + \mathrm{acos}\left(\frac{2\alpha_0 - 2\alpha + \alpha_{max}}{\alpha_{max}} + 1\right)\right)}{2} + \frac{\alpha_{max}}{2}. \tag{A.5}$$

Now, the horizontal pixel size s_x from Equation (A.3) can be expressed by the position on the HOE given by α and known system constants only:

$$s_x(\alpha) = -d_1 \cdot \left(\tan(\alpha) + \tan\left(\alpha_0 + \frac{\alpha_{max}}{2} - \frac{\alpha_{max}\cdot\left(\cos\left(\frac{\omega_h}{f_s} + \mathrm{acos}\left(\frac{2\alpha_0 - 2\alpha - \alpha_{max}}{\alpha_{max}}\right)\right) - 1\right)}{2}\right)\right). \tag{A.6}$$

The vertical axis is operated in a triangular-like fashion with a linearly increasing $\beta(t)$ until β_{max} and a linear flyback to $\beta = 0$. As apparent from Figure A.11, the optical path length increases with increasing deflection of the micro mirrors. With this, the vertical position is dependent on the length of the projection of the optical path between micro mirror and HOE onto the xz-plane. This can be expressed as

$$l(\alpha) = \frac{d_1}{\cos(\alpha)}. \tag{A.7}$$

The resulting y-coordinate is then calculated using β:

$$y(\alpha, \beta) = l(\alpha) \cdot \tan(\beta). \tag{A.8}$$

Analogy to the horizontal case, the pixel size s_y along the vertical direction is expressed as the difference between a first vertical position defined by α and β and a second position at an angle incremented by $\Delta\beta$. Here, $\Delta\beta$ is the linear angle increment between two consecutive

horizontal scan lines and can be expressed as

$$\Delta\beta = \frac{\beta_{max}}{N_h} \tag{A.9}$$

with the number of horizontal lines N_h scanned over the eye-tracking region. N_h is calculated by

$$N_h = \frac{f_h}{f_v \cdot k_v}. \tag{A.10}$$

The constant $k_v < 1$ describes the fraction of forward scan to flyback time which is required to drive the horizontal mirror back to the starting position.

With this, the vertical pixel size s_y is

$$s_y(\alpha, \beta) = \frac{y(\alpha, \beta + \Delta\beta) - y(\alpha, \beta)}{\cos(\beta)}. \tag{A.11}$$

The division by $\cos(\beta)$ results in a projection of the pixel size vector perpendicular to the optical path onto the plane of the HOE.

Now, pixel dimensions can be expressed dependent on mirror positions only (given by α and β). To calculate gaze angle resolution, this spatial resolution has to be mapped to eye rotational resolution.
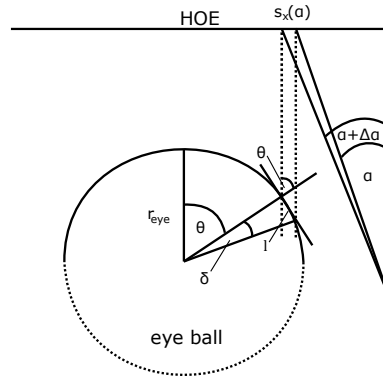
Figure A.12: Mapping of horizontal eye rotation to spatial resolution on the HOE. A parallel deflection function of the HOE is assumed.

For calculation, we assume a simplified eye ball model as shown in Figure A.12 with radius r_{eye}. It is rotated by θ in the horizontal plane and ϕ in vertical direction around its center. This rotation causes the pupil edge to propagate through different pixels along the reconstructed image. Neglecting methodologies like subpixel interpolation, eye rotation is only noticeable by the eye-tracking system if the pupil edge propagates from one pixel to an adjacent one. Thus, we define an infinitesimal angular increment δ of eye rotation that keeps the pupil edge in the current pixel. In this view, δ is the position-dependent angular resolution of our eye-tracking system.

Based on Figure A.12, δ_h for horizontal gaze angle resolution is calculated by

$$\delta_h(\theta) = \operatorname{atan}\left(\frac{s_x(\alpha(\theta))}{r_{eye} \cdot \cos(\theta)}\right) \tag{A.12}$$

with $\alpha(\theta)$ as mapping function to map an eye rotation angle θ to a horizontal angle α. With the assumption that the eye-tracking region, and therefore the HOE, is centered over the eye, the mapping function $\alpha(\theta)$ is

$$\alpha(\theta) = \operatorname{atan}\left(\frac{\frac{d_2}{2} - r_{eye} \cdot \sin(\theta)}{d1}\right). \tag{A.13}$$

The vertical gaze angle resolution δ_v in analogy to Equation (A.12) is

$$\delta_v(\phi) = \operatorname{atan}\left(\frac{s_y(\alpha(\theta), \beta(\phi))}{r_{eye} \cdot \cos(\phi)}\right) \tag{A.14}$$

with

$$\beta(\phi) = \operatorname{atan}\left(\frac{\frac{d_3}{2} - r_{eye} \cdot \sin(\phi)}{d1}\right). \tag{A.15}$$

The overall gaze angle resolution is calculated by the euclidean distance of the horizontal and vertical gaze angle resolution

$$\delta(\theta, \phi) = \sqrt{\delta_v(\theta)^2 + \delta_h(\phi)^2}. \tag{A.16}$$

A.2.5   Evaluation

To determine the gaze angle resolutions of the proposed head worn eye-tracking sensor, a laboratory setup is used. It is shown in Figure A.13.
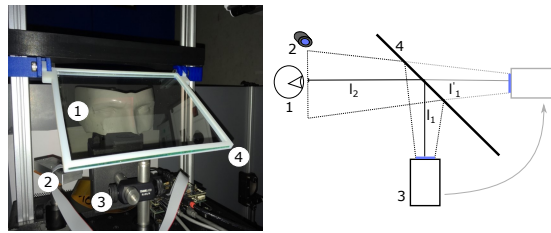
Figure A.13: Laboratory setup to evaluate the accuracy and precision of the proposed eye tracking sensor.

The subject ① is placed in a distance of $l_2 = 90\,mm$ in front of an semitransparent mirror ④ to align the eye tracking region with the subjects eye. The semitransparent mirror mimics the HOE with a parallel beam deflection function as shown in Figure A.8. The mirror is used to deflect the scanned IR laser beam of the laser projection module ③. The laser module

is placed in a distance of $l_1 = 30mm$ towards the semi-transparent mirror. The photodiode circuit ① is directly orientated towards the subjects eye to receive as much backscattered IR light as possible and thus improve sensitivity. The mirror allows the laser projection module to be virtually rotated so that it points directly at the eye tracking region without disturbing the user's view. In addition, artefacts of the captured images caused by eyelids and eyelashes are reduced, which improves the robustness of the sensors.

Table A.1 shows the geometrical and electrical properties of the the laboratory setup. The overall distance d_1 is the sum of l_1 and l_2.

Table A.1: Electrical and geometrical properties of the laboratory setup.

α_0	$\pm\alpha_{max}$	$\pm\beta_{max}$	f_s	f_h	f_v	k_v	d_1	d_2	d_3
0°	15°	9°	22 MHz	21kHz	60Hz	0.83	120mm	69mm	39mm

The laboratory setup is a class 1 laser system according to IEC 60825-1 [139] and therefore does not pose any medical hazard to the eye. The emitted IR laser power towards the eye is less than $150\,\mu$W. Laser class 1 would allow $670\,\mu$W for a steady beam, even significantly more in scanned operation. Laser safety is therefore ensured for all single error cases as required by IEC 60825-1 as well.

The power consumption of the proposed eye-tracking sensor is estimated roughly at $11\,$mW using off-the-shelf components. The main components that affect power consumption are the TIA and the ADC. With a higher degree of integration, e.g. by a custom ASIC, further power reductions are expected.

[102] report that state of the art VOG eye-tracking sensors consume more than $150\,$mW of power, which is significantly higher than our sensor approach. Compared to the scanned laser approach by [102], a similar power consumption is achieved.

Spatial resolution

To prove the mathematical framework in Appendix A.2.5, the spatial resolution as described in Appendix A.2.4 is calculated by placing a chess pattern of defined size in the eye-tracking region and compared with the theoretical spatial resolution. The experimental spatial resolution is determined based on the known properties of the chess pattern image. The number of pixels per chess field is counted and divided by the known length of a chess field to determine the pixel size.

Figure A.14 and Figure A.15 show the resulting theoretical and experimental spatial resolutions, as color surface and annotated numbers respectively.
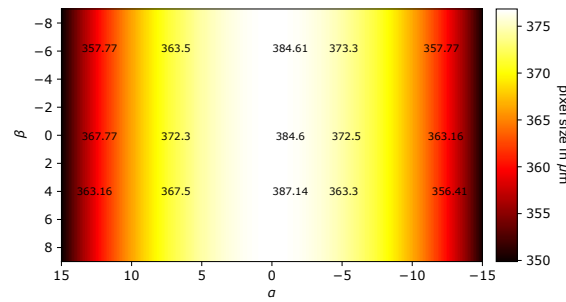
Figure A.14: Calculated pixel size for the horizontal direction based on Equation (A.6) with the parameters of Table A.1. In addition, the measured horizontal pixel sizes from the chess pattern are annotated.

The spatial resolution in horizontal direction shown in Figure A.14 is dominated by the non-linear horizontal scan speed of the MEMS micro mirror described by Equation (A.2). The sinusoidal change of velocity results in smaller pixels in the left and right edge regions of the eye tracking region. In the centre of the eye tracking region, the peak velocity is reached and thus increasing pixel size. In consequence, the sensor resolution is higher towards the left and right sides of the eye tracking region.
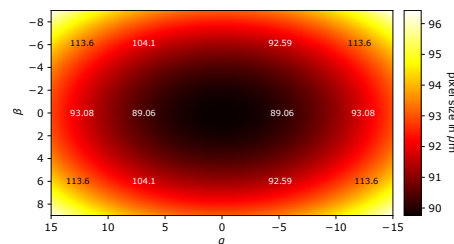
Figure A.15: Calculated pixel size for the vertical direction based on Equation (A.11) with the parameters of Table A.1. In addition, the measured vertical pixel sizes from the chess pattern are annotated.

For vertical direction the pixel sizes increases for increasing angles α and β, as shown in Figure A.15. This effect is mainly caused by increasing optical path length of the laser beam for increasing angles α and β. This effect is superimposed by the effect of the sinusoidal change of velocity, which leads to additional distortions in the horizontal direction.

Experimental gaze angle resolution

To estimate the achievable gaze angle resolution a experiment with the laboratory setup is performed. A subject sits approximately 0.5m away from a chart with visual markers M and fixate the markers. For each marker a set of N images is taken. The markers are placed on the chart to cause eye rotation angles θ and ϕ in the range of $\pm 20°$ and of $\pm 10°$.

Figure A.16 shows a subject fixating different markers. The images are captured with the

laboratory setup.
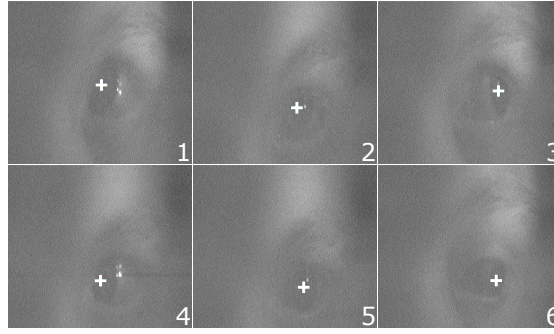
Figure A.16: Images of a subject fixating different markers on a chart. The images are captured with the proposed eye-tracking approach. The pupil position is detected by a state of the art VOG pupil detection algorithm [3]. The numbers indicate which marker of the chart is fixated by the subject.

To extract the pupil positions we use the state of the art VOG circular binary features (CBF) pupil detection algorithm by [3]. The estimated pupil center positions are marked with a green dot in Figure A.16. The result is very promising and proves the feasibility to apply state of the art VOG algorithms to our low-power eye tracking data. This enables slippage robust pupil extraction for the scanned laser eye tracking sensor with minimal algorithm design effort. In addition the sensor benefits directly from advances in VOG algorithms.

To calculate the resolution and precision of the sensor we perform the standard 9 point chart marker based calibration method similar to [53]. For each calibration marker M_c N images are captured and the corresponding pupil coordinates P_c are calculated using the CBF algorithm. With the known calibration coordinates and the related pupil coordinates a second order polynomial function

$$x_c = a_0 + a_1 x_c + a_2 y_c + a_3 x_c y_c + a_4 x_c^2 + a_5 y_c^2 \tag{A.17}$$

for x coordinates and

$$y_c = b_0 + b_1 x_c + b_2 y_c + b_3 x_c y_c + b_4 x_c^2 + b_5 y_c^2 \tag{A.18}$$

for y coordinates is fitted using a least mean square optimizer to estimate the coefficients a and b.

In addition to the calibration markers M_c test markers M_t are placed on the calibration chart. These markers are used to estimate the spatial accuracy and precision of the proposed eye tracking sensor. Figure A.17 shows the results of the experiments in gaze angle coordinate space.
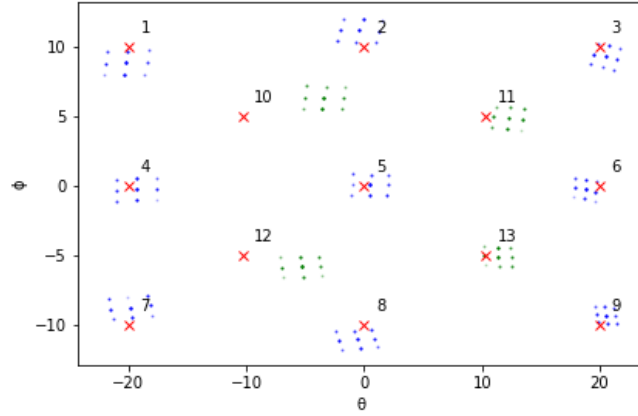
Figure A.17: Result of the accuracy and precision experiment with calibration markers M_c (1-9) and test markers M_t (11-13) and the corresponding pupil coordinates P. The arrow indicates the error between corresponding marker coordinates and estimated pupil coordinates.

To estimate the resolution based on the experiment the average angular distance between the position of the test markers M_t and a set of corresponding locations of fixations P_t is calculated for every test marker

$$Res(M_t, P_t) = \frac{1}{N} \sum_{e=0}^{N} \sqrt{\left(x_{M_t} - x_{P_t}\right)^2 + \left(y_{M_t} - y_{P_t}\right)^2}. \tag{A.19}$$

In addition the precision is estimated by the Root Mean Square (RMS) of successive samples of fixations P_e for a given test marker M_t. It is calculated by

$$Pre(P_t) = \sqrt{\frac{1}{N} \sum_{e=0}^{N-1} \left(x_{P_t} - x_{P_t+1}\right)^2 + \left(y_{P_t} - y_{P_t+1}\right)^2} \tag{A.20}$$

for every test marker. The resulting mean resolution of the test markers is $1.31°$ and the precision is $0.01°$ for this setup.

A possible source of error, which reduces the achieved gaze accuracy, is that the head of the test person is not fixed during the experiment. In addition, a large part of the captured images covers the face around the eye. The pupil information is therefore only contained in a small subset of pixels in the center of the image.

Theoretical Gaze angle resolution

In addition to the experimental estimation of the gaze resolution of the proposed eye tracking sensor the the theoretical gaze angle resolution is estimated by Equation (A.16). The distance r_{eye} in Equation (A.16) is derived from the Emsley's reduced eye model. Based on this model, the distance r_{eye} between the iris and the center point of the eye ball is $9.77\,\text{mm}$. The result is

shown in Figure A.18.
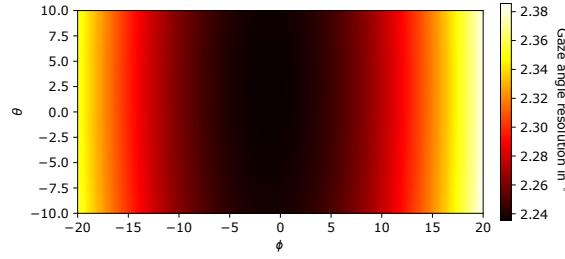
Figure A.18: Calculated gaze angle resolution of the proposed eye tracking sensor based on the laboratory setup based on Equation (A.16) using the geometry of the Emsley's reduced eye model.

Recalling Figure A.12, δ is much smaller for given $s_x(\alpha)$ close to $\theta \approx 0, \phi \approx 0$. To some extent, this compensates the lower spatial resolution in the center of the eye tracking region as shown in Figure A.14. The theoretical mean single-pixel gaze angle accuracy of the laboratory setups is around $2.3\,°$, which is significant lower as the experimental estimated gaze angle resolution. Thus, our simplified mathematical approach neglecting the effect of pupil tracking algorithms can be understood as an upper boundary estimation for gaze angle resolution.

Based on this assumption the theoretical gaze angle resolution of the head worn eye tracking sensor is calculated using the electrical and geometric parameters shown in Table A.2. The main differences between the laboratory setup and the proposed head worn demonstrator are the reduced distance d_1 towards the HOE, the offset angle α_0 and the maximum angles for α_{max} and β_{max}. Due to the glasses geometry α is in a range between 0 and α_{max} an β is in a range between 0 and β_{max}. The theoretical gaze angle resolution of the proposed head worn

Table A.2: Electrical and geometrical properties of the AR glasses setup.

α_0	α_{max}	β_{max}	f_s	f_h	f_v	k_v	d_1	d_2	d_3
15°	30°	18°	22 MHz	21kHz	60Hz	0.83	22.5mm	22.5mm	7.3mm

eye tracking sensor is shown in Figure A.19.
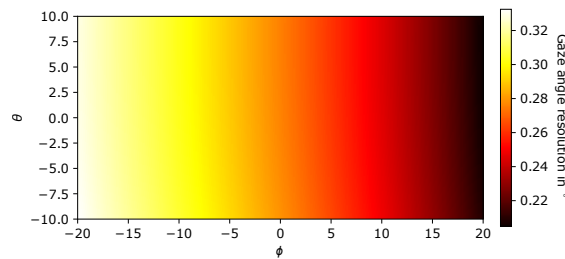
Figure A.19: Calculated gaze angle resolution of the proposed eye tracking sensor for the glasses geometry based on Equation (A.16) using the geometry of the Emsley's reduced eye model.

The offset angle α_0 adds an increasing optical path length with increasing angles α which leads to an decrease in resolution for extreme eye rotational angles especially towards the lower left edge, as apparent from Figure A.19. The theoretical mean single-pixel gaze angle accuracy for this setup is around $0.28\,°$ and is therefore a upper boundary estimation of the gaze angle resolution. A further increase in gaze angle resolution by the use of an pupil tracking algorithms is expected.

Compared with the scanned laser approach of [102], the calculated gaze angle resolution of the proposed laser based eye-tracking sensor for the glasses geometry is higher especially for relevant eye rotation angles around the center. Compared to state of the art VOG eye-tracking sensors like [53], our approach is capable to achieves higher gaze angle resolution. Based on this results a less complex pupil tracking algorithm can be used to reduce the computational effort and therefore power consumption, while keep comparable gaze angle resolutions to VOG sensors.

Limitations

The temporal resolution of the eye-tracking sensor is limited by the frame rate of the retinal projection system, which currently is $60\,\text{Hz}$. To overcome this limitation, a faster micro mirror could be used. However, as a collimated laser beam requires sufficient micro mirror aperture, miniaturization of the micro mirror is restricted. Therefore, frequencies up to about $120\,\text{Hz}$ are technically feasible.

The low-power consumption is achieved because components already contained in the projection system such as the micro-mirror module are not considered in the eye tracking power budget. This is valid as long as the sensor is a replacement for an VOG sensor in a retinal projection system, for example. For stand-alone application of the eye tracker, contribution of the micro-mirror module to the power budget has to be considered.

A.2.6   Conclusion

We presented a novel eye-tracking sensor for integration into our existing retinal projection AR glasses prototype. Compared to VOG eye-tracking sensors, our technology achieves a significant reduction in power consumption at comparable gaze angle resolution and full integration into a frame temple. Besides the photodiode in the frame temple, the sensor is completely invisible to the user. Furthermore, our approach is used in combination with state of the art robust VOG eye-tracking algorithms.

The achievable gaze angle resolution is evaluated experimentally based on a laboratory setup. It shows comparable gaze angle resolution in comparison with VOG sensors. In addition the theoretical gaze angle resolution of the proposed head worn eye tracking sensor is calculated, which leads to an even better gaze angle resolution as the laboratory setup.

Based on these results the next step is to integration of the proposed eye tracking sensor into our AR glasses demonstrator to perform experiments and evaluate the sensor under real conditions, e.g. in the presence of various illumination conditions.

Additionally, the eye tracking algorithms used must be transferred to an embedded platform and integrated into the demonstrator.

A.3 A Novel -Eye-Tracking Sensor for AR Glasses Based on Laser Self-Mixing Showing Exceptional Robustness Against Illumination

A.3.1 Abstract

The integration of eye-tracking sensors in next-generation AR glasses will increase usability and enable new interaction concepts. Consumer AR glasses emphasize however additional requirements to eye-tracking sensors, such as high integratability and robustness to ambient illumination. We propose a novel eye-tracking sensor based on the self-mixing interference (SMI) effect of lasers. In consequence, our sensor as small as a grain of sand shows exceptional robustness against ambient radiation compared to conventional camera-based eye trackers. In this paper, we evaluate ambient light robustness under different illumination conditions for video-based oculography, conventional scanned laser eye tracking as well as the SMI-based sensor.

A.3.2 Introduction

Next generation AR glasses are already integrating eye-tracking sensors to enable new gaze-based interaction concepts with the glasses and the surroundings [171, 172, 173]. In addition, new rendering schemes to reduce the power consumption and increase the image quality during image projection like foveated imaging are introduced [39, 174]. These rely mainly on eye-tracking for pupil position estimation [158].

The state of the art eye-tracking sensor technology is video oculography (VOG). VOG sensors use infrared (IR) illumination and an IR camera to capture images of the eye surface and determine the pupil position utilizing image processing techniques. Such technology is meanwhile well established and provides high accuracies [54].

To perform best, VOG systems require high-contrast, disturbance-free eye images. In real world conditions, variations in illumination affect image quality and in the end lead to a low pupil detection ratio [56]. To improve the detection ratio, recent algorithms using deep neural networks are applied which introduce additional computational effort and high power consumption for real-time applications [175].

To overcome the susceptibility of VOG sensors to ambient light, we propose a novel eye-tracking sensor for AR glasses based on lasers featuring the self-mixing interference (SMI) effect [122]. Our main contribution is the successful application of the SMI effect for an exceptionally robust eye-tracking sensor as well as systematic comparison to competing technologies. We use two micro-electro-mechanical system (MEMS) micro mirrors to scan the IR laser beam of a vertical cavity surface emitting laser (VCSEL) with a cavity-integrated photodiode across the eye region. Strong pupil signals are based on the high reflectivity of the retina in the infrared wavelength range.

The remaining of the paper is organized as follows: Section 2 discusses previous work in the area of MEMS scanned laser eye-tracking sensors. In Section 3, we explain the underlying SMI VCSEL technology and introduce an index for disturbance robustness. The evaluation in Section 4 compares the performance of the selected three technologies under different artificial illumination conditions up to direct exposure to an intense camera flash. Section 5 concludes this work and gives a brief outlook to future work.

A.3.3 Related Work

MEMS scanned laser eye-tracking sensor concepts categorize into two classes that differ in the way of scanning the eye surface as well as in the algorithm for pupil position estimation.

Sarkar et al. [101] introduced small MEMS micro mirrors to scan a laser beam vertically and horizontally across the eye's surface. The micro mirror and the IR laser were attached to the frame temple of the glasses and a photodiode was integrated close to the nose pads [102]. The laser beam was scanned along a straight line over the eye. The micro mirror angle at which the photodiode observed a corneal glint was found to be linked to the horizontal pupil position. For the vertical pupil position, a hill climbing algorithm based on the amplitude shifts of the photodiode output in horizontal scanning direction was used. An optical bandpass filter in front of the photodiode reduced susceptibility to ambient radiation. Sakar et al. [102] reported an angular gaze resolution of approximately $1°$ with a temporal resolution of $3300\,Hz$ at a power consumption of less than $15\,mW$. Recently, a further implementation of this sensor concept was presented by [176].

One drawback of this approach is a high susceptibility to movements of the glasses movements of the glasses while being worn. As mirror scan angles at glint occurrence are directly linked to pupil positions, recalibration is required each time the glasses move [102].

To approach shift-invariance, [159] used a different scan path of the IR laser beam. Their setup was based on two micro mirrors for horizontal and vertical deflection to scan the IR laser beam across the entire eye region. An external photodiode captured the reflectivity of different regions of the eye that could be used to reconstruct an image of the eye region. Then, state of the art VOG image processing algorithms were used to extract pupil positions. Again, an optical long-pass filter reduced susceptibility to ambient radiation. However, only radiation of shorter wavelength can be blocked by such a filter and interference with infrared sources is possible.

Both prior works used an external photodiode to record laser beam reflectivity in the eye region. Our approach exploits the SMI effect to record such a reflectivity map at outstanding robustness to ambient radiation as the laser cavity itself is a very narrow bandpass and the SMI effect relies on the laser's coherence length.

A.3.4 Methodology
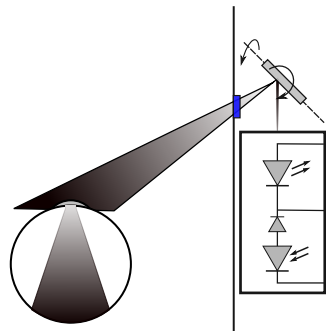
Figure A.20: Setup of the proposed eye-tracking sensor integrated in a frame temple

Figure A.20 shows the IR VCSEL with integrated photodiode and a MEMS micro mirror integrated in a frame temple. The MEMS micro mirror deflects the laser beam of the IR VCSEL towards the eye. By using a two-axes mirror, a rectangular region covering the eye is scanned.

At a specific mirror scan angle, the IR laser beam enters the eye through the pupil and is retro-reflected towards the laser cavity. The retina acts as retro-reflector similar to VOG eye-tracking methods based on the bright eye effect [62]. This results in a strong interference inside the laser cavity, which is observable at the integrated photodiode output. Aligning this photodiode output as a grey-value pixel with the mirror scan angles, a two dimensional reflectivity image of the scan region is formed. Then, pupil extraction can be done in a similar way to the VOG sensors. Configuring the laser power in such a way that strong SMI interference is only maintained for retinal reflections. This significantly enhances the image contrast and much simpler image processing algorithms can be applied.

Self-mixing interference laser sensor

The proposed eye-tracking sensor relies on a VCSEL semiconductor laser. The VCSEL's technology has several advantages over edge emitting lasers (EEL). They are characterized by a very low threshold current and thus a low power consumption. Furthermore, they have small dimensions and can easy be manufactured and tested, enabling low cost devices [117].

Due to the small size of the aperture, a circular emission profile and low beam divergence is achieved. This allows the use of small optical components, which results in a high integratability in the glasses frame temple [117]. The size of the sensor without external optical components is shown in Figure A.21.

The SMI effect relies on backscattering of a fraction of the coherent laser light into the laser cavity. Inside the cavity, it interferes with the lasing field and modulates the amplitude and frequency of the lasing field. The cavity serves as an optical mixer and amplifier as well as an optical filter due to the narrow optical bandwidth [117].

To obtain information about the reflectivity of the structure that causes backscattering, the optical power of the modulated laser field is recorded by a photodiode. In our sensor, the photodiode is directly integrated into the bottom reflector of the cavity yielding a tiny sensor module [2].
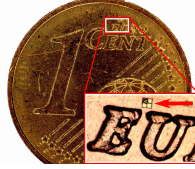
Figure A.21: Size of the sensor in comparison to a Euro cent coin. The size of the sensor is approximately 180x180 μm.

Image quality metric

VOG pupil detection algorithms rely on the recognition of features such as the pupil's edges. To achieve a high pupil detection rate however, a high contrast and a noise-free image with suitable resolution is required [56]. To take these requirements into account, the structural similarity index (SSIM) is used as an image quality metric to describe the robustness of an eye-tracking sensor against disturbance by ambient light.

The SSIM expresses image quality degradation as a combination of structural, illuminance and contrast distortions of a disturbed image F relative to a reference image R [177]. It is calculated as

$$SSIM(R,F) = \frac{(2\mu_R\mu_F + C_1)(2\sigma_{RF} + C_2)}{(\mu_R^2 + \mu_F^2 + C_1)(\sigma_R\sigma_F + C_2)}. \tag{A.21}$$

Here, μ and σ describe the mean illuminance and standard deviation of illuminance, respectively, which can be interpreted as an estimation of the image contrast. The cross correlation σ_{RF} between R and F expresses structural distortions and the constants C_1 and C_2 are required to avoid the denominator becoming zero.

The index is calculated using a local $N \times N$ sliding window over the whole frame to consider local distortions. Then, the arithmetic mean of all m window positions represents an overall quality metric, the mean structural similarity index (MSSIM):

$$MSSIM(R,F) = \frac{1}{m} \sum_{j=1}^{j} SSIM(R_j, F_j). \tag{A.22}$$

A.3.5 Evaluation
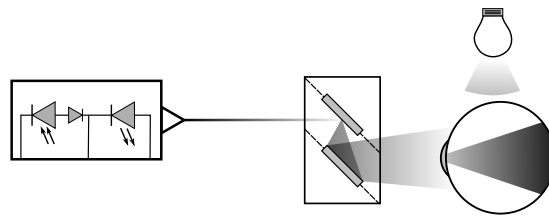
Figure A.22: Laboratory setup to evaluate the disturbance immunity of the proposed eye-tracking sensor against disturbances induced by artificial light sources.

Figure A.22 shows the laboratory setup of the proposed eye-tracking sensor. The laser beam of the VCSEL is directed to the micro mirror module, which consists of two MEMS micro mirrors. Each micro mirror deflects the laser beam in one orthogonal direction resulting in a two-dimensional scan region covering the eye and its surroundings. Synchronization signals indicate the start of a new frame as well as the start of a new horizontal line. These signals are temporal synchronized with the output of the VCSEL's integrated photodiode and used to reconstruct a spatial mapping of the retro-reflectivity sensed by the SMI effect.

Ambient light sources

To evaluate the robustness against ambient radiation of the proposed eye-tracking sensor, it is exposed to different light sources. To evaluate even extreme illumination scenarios, high illuminance, modulated light sources and light sources with broad wavelength spectra are used. Figure A.23 shows the optical [1] and frequency spectra [2] of the light sources considered in the experiment.

---

[1]Measured with Ocean Optics HR4000
[2]Measured with Ultra fast photodiode UDP-200-SP
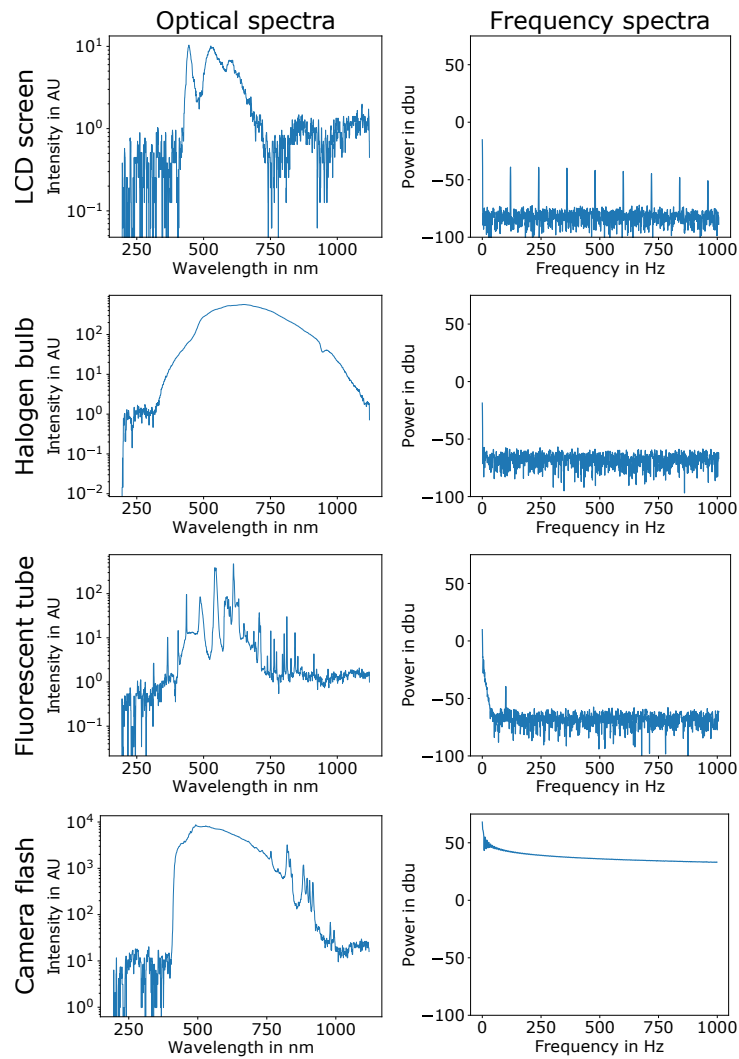
Figure A.23: Optical and temporal characterization of the light sources used to disturb the three eye-tracking sensors.

To assess ambient sources relevant to data glasses, we considered a liquid crystal display (LCD) (watching TV or using PC), a fluorescent tube as well as a halogen bulb (indoor illumination) and a photoflash [3] as very strong disturbance in terms of transient and spectral bandwidth. Measurements in direct sunlight are not carried out, as no uniform conditions can be guaranteed for all measurements. For this reason the halogen lamp is used, which imitates the wavelength spectrum and intensity of the sun on a bright day.

---

[3] Canon Speedlite 580EX II

Disturbance immunity

For each light source, we evaluate three different eye-tracking approaches: a system based on a scanned IR laser comparable to the system of [159], the proposed eye-tracking sensor based on the SMI effect and a commercial VOG sensor by Pupil Labs.

For each system, test cards optimized to the underlying technology are used to obtain a high resolution and contrast-rich image. Therefore, a chess pattern is used for the scanned IR laser sensor and the VOG sensor. For the SMI sensor a black coloured test chart with circular retro-reflectors of different diameters is used, mimicking pupils of different diameters.
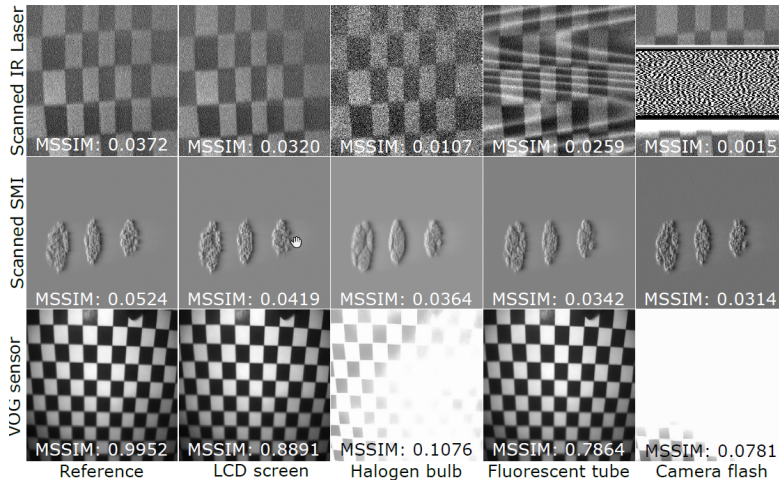
Figure A.24: Results of the disturbance evaluation for the three eye-tracking sensors.

First, the eye-tracking sensors are placed in front of their respective test chart to capture a high quality disturbance free reference image R without any external illumination. Afterwards, the measurement is repeated for each of the four chosen light source to capture a disturbed image F. Figure A.24 shows the resulting images F for each eye-tracking sensor and illumination condition.

Afterwards, the MSSIM is calculated by Figure A.21 and Figure A.22 with $C_1 = C_2 = 0.26$ and $N = 9$ similar to [178]. For differentiation of intra-sensor image quality variations to ambient light disturbance, the MSSIM of two consecutive reference images (one as R and one as F) is calculated as a representation of sensor noise. The MSSIM is included as annotations for every experiment in Figure A.24 with a low MSSIM indicates a low repeatability within one row.The scanned IR laser sensor shows an overall low MSSIM due to Gaussian noise of the photodiode and a reduced dynamic range due to currently low amplification in the analog frontend. With rising illumination, noise increases and affects the MSSIM. Further, periodic flickering of the fluorescent tube and the high illumination change by the photoflash produce strong interference in the images, thus leading to poor MSSIM.

The scanned SMI setup shows a low reference MSSIM due to noise induced by speckling on

the targets surface as well. Image noise is considered a minor impact for eye-tracking using SMI signals, however directly affects the MSSIM. However, exceptional robustness to ambient radiation is obvious from the images, as other disturbances like periodic flickering and high illumination changes (photo flash) are almost invisible in the recorded images.

The VOG sensor as our eye-tracking Gold Standard reference is characterized by an overall high MSSMI and very good image quality. Even periodic flickering of fluorescent lights is effectively suppressed in the sensor. Only the strong illumination change of the photo flash strongly affects the image quality and leads to an low MSSIM. To compare these three very different eye-tracking sensor principles, the MSSIM is normalized by dividing the respective MSSIM by the MSSIM of the corresponding reference frame. This way, repeatability issues are suppressed to emphasize on ambient light robustness. Figure A.25 shows the resulting normalized MSSIM (MSSIM-n) with high values indicating high image quality.
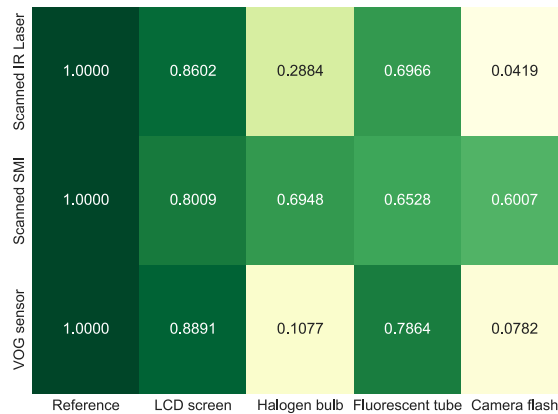
Figure A.25: Visualization of the MSSIM-n to compare the eye-tracking sensors for the different illumination conditions.

Concluding, the proposed scanned SMI sensor is very robust in terms of ambient light suppression seen in overall high MSSMI-n indices. Although overall imaging quality is poor compared to recent VOG systems, the technology is well suited for eye-tracking. Exceptional image contrast is not required as the sensor principle exploits the retro-reflectivity of the retina to separate the pupil from surrounding eye regions. This way, edge- or template-based pupil detectors are not required. Based on this feature, the pupil position is detectable with less computational effort. In addition rotations of the eye in relation to the sensor may result in partial shadowing of the retina by the sclera. This leads to distortions of the image of captured retro-reflections, which can be used as additional features to estimate the gaze angle. For extreme eye rotations, the optical path of the laser beam can be deflected via a parabolic mirror to reduce large angles of incidence and cover the whole eye region.

A.3.6   Conclusion

We presented a novel eye-tracking sensor based on MEMS micro mirrors and a laser featuring the SMI effect. Compared to VOG sensors and scanned laser eye-tracking with separate photodiode, exceptional robustness to ambient radiation is shown. Paired with the sensor's low power consumption and its tiny size, it is a promising technology for fully integrated always-on ubiquitous eye-tracking in next generation AR glasses.

An important next step is to repeat the experiment with the human eye instead of test charts, especially as there exists only little data about the SMI effect from human retinal reflections. Furthermore, the pupil detection accuracy and precision of the proposed eye-tracking sensor will be evaluated. For outdoor use, robustness to sunlight should be analysed, although no relevant degradation is expected following to the halogen bulb and photo flash experiments.

One promising increase the pupil-iris contrast further sensor is the modulation of the VC-SEL wavelength by current modulation. Similarly to frequency modulated continuous wave (FMCW) radar systems, this introduces depth resolution of the sensor. However, for enhanced performance, modifications of pupil detection algorithms are required to exploit these new sensor capabilities.

A.4   A Highly Integrated Ambient Light Robust Eye-Tracking Sensor for Retinal Projection AR Glasses Based on Laser Feedback Interferometry

A.4.1   Abstract

Robust and highly integrated eye-tracking is a key technology to improve resolution of near-eye-display technologies for augmented reality (AR) glasses such as focus-free retinal projection as it enables display enhancements like foveated rendering. Furthermore, eye-tracking sensors enables novel ways to interact with user interfaces of AR glasses, improving thus the user experience compared to other wearables. In this work, we present a novel approach to track the user's eye by scanned laser feedback interferometry sensing. The main advantages over modern video-oculography (VOG) systems are the seamless integration of the eye-tracking sensor and the excellent robustness to ambient light with significantly lower power consumption. We further present an algorithm to track the bright pupil signal captured by our sensor with a significantly lower computational effort compared to VOG systems. We evaluate a prototype to prove the high robustness against ambient light and achieve a gaze accuracy of 1.62 °, which is comparable to other state-of-the-art scanned laser eye-tracking sensors. The outstanding robustness and high integrability of the proposed sensor will pave the way for everyday eye-tracking in consumer AR glasses.

A.4.2   Introduction

Robust and highly integrated eye-tracking sensors are a key technology to improve resolution of display technologies like focus-free retinal projection for augmented reality (AR) glasses e.g. by enabling display enhancement methods like foveated rendering [39, 40, 38]. Furthermore eye-tracking allows to steer the exit pupil increasing the display's field of view (FOV) [39, 105] of AR glasses. In addition to display enhancement techniques, eye-tracking sensors enable novel ways to seamlessly interact with the user interface of AR glasses [179, 5, 4], improving thus the user experience.

Video oculography (VOG)camera sensors are the state-of-the-art in mobile eye-tracking, tracking either the pupil in the 2D image and estimate gaze using a geometric 3D eye model [9], or track the pupil and corneal reflections from additional infrared (IR) LEDs and use a regression-based approach to determine gaze direction, as shown by [63]. In both cases, the key to robust eye tracking is robust detection and tracking of the pupil under a variety of conditions, which is, as shown in Figure A.26, not always the case with current VOG systems.

A well-known issue with state-of-the-art VOG sensors is the limited dynamic range of camera sensors, leading to a loss of the pupil signal in presence of varying ambient light or in bright sun light [85, 56], as illustrated in Figure A.26 a).

Furthermore the high off-axis integration of camera sensors in current VOG systems, as
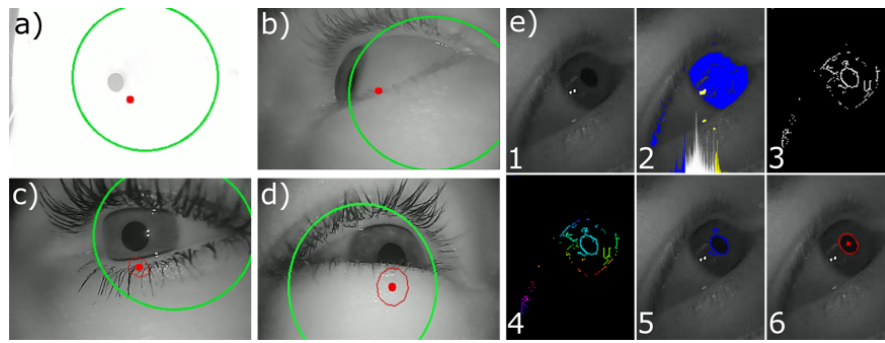
Figure A.26: Key challenges of VOG based eye-tracking sensors is a robust detection of the pupil, which is limited due to a) limited dynamic range of camera sensors to operate under a wide range of ambient illumination settings e.g. in bright sun light, b) detection of the pupil over the whole field of view due to the high off-axis integration of camera sensors, c) false pupil detection e.g. due to mascara, other disturbances, dirt on the lens and d) false detection due to partly occluded pupils by lashes or eye lids. Furthermore the pupil detection is rather computational complex as several image processing steps are required to extract the pupil as shown in e)

illustrated in Figure A.26 b) [100, 53], leads to a loss of pupil detection especially if the gaze vector points away from the cameras optical axis, which allows robust eye-tracking only in a part of the user's FOV [89]. This problem is solved by adding more camera sensors to cover a larger field of view, such as proposed by [97] or [90]. However, this leads to more complex sensor integration as well as higher power consumption stem from additional sensors and higher complexity eye-tracking algorithms.

Additional cases which leads to a false pupil detection are due to the wearing of mascara [104], as false edges are considered as pupil edges by the pupil detection algorithm, illustrated in Figure A.26 c). A similar case is shown in Figure A.26 d) where a part of the pupil is occluded by the eyelid, which also leads to a false pupil detection [180, 56]. This issues is addressed by more advanced pupil detection algorithms e.g. by using convolutional neural networks like PupilNet [74], RITNet [76] or the Deep VOG approach by [67]. The main drawback of these advanced algorithms is increased demand of processing power which increases the power consumption of VOG eye-tracking systems.

Finally, VOG algorithms require several steps of image processing to extract pupil features from camera images, as shown in Figure A.26 e), which illustrates the processing steps of the VOG algorithm proposed by [53]. There are several variants of the algorithm with optimization of individual steps of the pupil detection pipeline to improve detection accuracy and robustness, e.g. ELSE [69], PURE [70] or PUREST [71]. The increased robustness is accompanied by higher computational requirements. In recent years this issue is addressed by pupil detection algorithms which are optimized with respect to computational requirements and latency e.g. by [84] or [181].

A.4 A Highly Integrated Ambient Light Robust Eye-Tracking Sensor for Retinal Projection AR Glasses Based on Laser Feedback Interferometry

The presented disadvantages of VOG sensors and the corresponding power consuming eye-tracking algorithms indicate that the sensing technology itself puts some hurdles for eye-tracking sensor integration into AR glasses. To enable robust eye-tracking and overcome these limitations, we introduce a low power eye-tracking sensor approach using laser feedback interferometry (LFI) sensing technology to integrate eye-tracking capability into retinal projection AR glasses.

The LFI sensor is composed of a tiny vertical cavity surface emitting laser (VCSEL), operating at the infrared (IR) regime. In addition, a photodetector is integrated into the laser cavity using semiconductor processes. The small sensor size enables high integration into the frame temple of AR glasses. Integration of the photodetector enables the LFI sensing method, a coherent sensing method, leading to a high robustness against ambient light as most light stemming not from the lasers own radiation is suppressed [117]. Thus the sensor is capable to robustly operate in presence of ambient light [128].

To solve the sensor integration problem and the high-off-axis integration of VOG systems, we further propose to integrate the LFI IR laser sensor into a retinal projection AR glasses system which consists of a micro-electro-mechanical system (MEMS) micro mirror based laser scanner and a holographic optical element (HOE) to steer the laser beam towards the eye.

Furthermore, we exploit the unique sensing modality of the LFI sensor and propose a low-power pupil detection and tracking algorithm by exploiting the characteristic bright pupil signal.

Our contribution is three fold:

(i) We propose an highly integrated eye-tracking sensor approach for retinal projection AR glasses based on an ambient light robust LFI sensor. By combining the LFI sensor with a highly transparent IR HOE and a MEMS micro mirror we further solve the highly-off-axis sensor integration. In addition, the eye tracker is invisible to the user, as it is fully integrated into the frame temple.

(ii) We propose an algorithm optimized to detect and track the pupil based on the characteristically bright pupil signal captured by the LFI sensor and

(iii) We evaluate the resulting gaze accuracy of the proposed algorithm and the ambient light robustness of the proposed sensor experimentally in a prototype setup.

Compared to the work of [129] we switch from an IR laser with external photodiode to the LFI sensor with integrated photodiode and further show the high integratability into a glasses frame. In addition we manufacture the high transparent IR HOE, which is mandatory for the system.

Compared to the work of [128] we apply the LFI sensor to human eyes and proof the proposed

bright pupil effect. We further evaluate gaze accuracy in a human study with 20 participants and further propose an power saving algorithm for pupil detection algorithm.

In the upcoming Section the state-of-the-art w.r.t. scanned IR laser eye-tracking sensors is discussed. Afterwards, in Appendix A.4.4, we introduce the proposed eye-tracking sensor and describe briefly the system components. Furthermore, we describe the underlying sensing principle of the LFI sensor technology as well as the origin of the observed bright pupil pattern. In addition, we describe our algorithm to detect and track the pupil. In Appendix A.4.5, we describe our setup used to evaluate the gaze accuracy and compare it to a VOG system. Further, we show the robustness against artificial light. Finally, we compare our results with other state-of-the-art scanned IR laser eye-tracking approaches, discuss the applicability for AR glasses w.r.t. power consumption, sensor integration, glasses slippage and system latency, and finally draw a conclusion from our work.

A.4.3   Related Work

One of the first works which address scanned IR laser eye-tracking technology for AR glasses was introduced by [101]. The authors used a 2D MEMS mirror to scan the beam of an laser operating in the IR regime in a 2D pattern over the eye's surface. The photodiode, which receives backscattered light, was integrated close to the nosepad while the scan unit consisting of the IR laser and the 2D MEMS mirror were integrated into the glasses frame temple. The photodiode detects corneal reflections originating from the eye's surface [102]. To obtain the horizontal gaze angle, the MEMS mirror scan angle under which a corneal reflection was detected by the photodiode is captured. To further obtain the vertical gaze angle [102] proposed a hill climbing algorithm using the the photodiode amplitude variation as feature. To address ambient light robustness, an optical bandpass filter was applied to the front of the photodiode. The authors reported a gaze resolution of $\approx 1°$ with an update rate of $3300\,\mathrm{Hz}$ while their system consumes less than $15\,\mathrm{mW}$ power.

A major drawback of their method is the vulnerability to glasses slippage. As gaze angles are directly linked to the MEMS mirror scan angles via calibration, the system requires calibration after occurrence of slippage of the glasses [102].

To achieve slippage robustness, the authors most recently released MindLink [103], which incorporates five photodiodes attached around the spectacle frame and a 2D MEMS micro mirror placed in the nose pad of the glasses. With this improved setup, the authors reported a gaze accuracy of <1° over a FOV of 40° x 25° and achieved an update rate of 500 Hz.

[129] approach slippage robustness by scanning an IR laser beam with a 2D scan path over the surface of the eye. The scan path was formed using two 1D MEMS mirrors for vertical and horizontal deflection. Backscattered light from the eye is measured by a photodiode, which is placed in the frame temple. The measured intensity variation over both the horizontal and the vertical scan angles is used to construct a gray scale image of the eye's surface. By applying

a state-of-the-art VOG algorithm [3], they achieved a gaze accuracy of 1.31° with an 60 Hz update rate. The authors further reported a power consumption of 11 mW and estimated a theoretical resolution of 0.28° with an improved optical design. To increase robustness against ambient light they propose to use optical filters in front of their photodiode circuitry, similar to [102].

Most recently, EyeWay Vision [105] released a scanned IR laser based eye-tracking sensor to steer the exit pupil for their retinal projection AR glasses. In a previous evaluation of the prototype system by [104] a gaze accuracy of 1.72° at a sampling rate of the corneal reflection signals of 4000 Hz was reported. For absolute eye-tracking accuracy and to compensate translation movements of the eye with respect to the glasses e.g. due to slippage, a stereo camera with a sample rate of 120 Hz was added to the laboratory setup.

All above-mentioned related approaches used a photodiode to capture back reflected light of an IR laser, which was scanned in a 2D pattern over the surface of the eye. Sarkar et. al.[102, 103] and EyeWay Vision [104] focus on glint features from the cornea, the limbus or the retina to estimate the gaze direction while [129] reconstruct a gray scale image and extract the dark pupil from the image by applying state-of-the-art VOG algorithms.

All methods have drawbacks with respect to ambient light robustness, which are addressed by protecting the photodiode with optical filters from ambient light. Furthermore, the glint feature based approaches by Sarkar et. al. and EyeWay Vision tend to have issues with slippage. To address this issue they either add an reference sensors or additional photodiodes, which adds to the overall power budget of these systems. Furthermore the approaches of [129] and [104] are validated only in a laboratory setup and the sensor integration is not fully solved.

In our approach, we address the issue of ambient light robustness and sensor integration by using the LFI sensor technology. We further follow the path of [129] and use a 2D scan pattern to reconstruct gray scale images to extract the bright pupil feature. With this approach, we address the issue of a high power consumption by exploiting the bright pupil effect to directly detect the pupil in an image to reduce computational complexity.

A.4.4 Scanned laser feedback interferometry

Figure A.27 illustrates the integration of the LFI sensor into the retinal projection AR glasses to form a scanned LFI eye-tracking sensor. The LFI sensor component is added to the RGB module, integrated in the glasses frame temple. The IR laser of the VCSEL is coupled via a prism into the beam path of the visible light of the RGB lasers and the combined beam is scanned via a MEMS mirror module over the HOE surface. The HOE acts as a wavelength selective mirror which parallelised the incoming beam pattern and redirects it towards the eye region.

The HOE is recorded into a photopolymer (Bayfol HX TP photopolymer) by constructing a reference wavefront and an imaging wavefront and expose the photopolymer with both
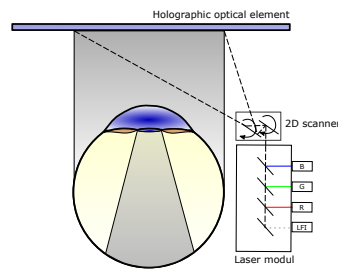
Figure A.27: The LFI sensor added to the RGB module and shares the same optical path as the visible light. The holographic optical element acts as a wavelength selective mirror and redirects the scanned laser pattern to the eye's surface.

wavefronts. As our photopolymer is only active for visible light, we recorded the HOE at a wavelength of 650 nm with an angular offset such that if the HOE is played back at 850 nm under an different angle the desired wavefront is reconstructed. [182] and [183] gave a detailed description of the recording HOEs using photopolymer with an angular offset.

The MEMS mirror module contains two 1D MEMS mirrors to scan in a 2D pattern over the HOE. The horizontal mirror scans in a sinusoidal pattern, while the vertical MEMS mirror is non resonantly actuated using an electrodynamic driver to steer the sinusoidal pattern vertically over the HOE. With the known geometry and the mirror deflection angles $\alpha_h(t)$ and $\beta_v(t)$, the corresponding intersection point of the laser beam with the HOE can be calculated. For a detailed description of the geometry and the image generation we refer to [129]. The scan pattern is illustrated in Figure A.28. HOEs are characterized by a high wavelength selectivity
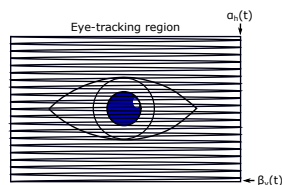
Figure A.28: Scan pattern of the laser beam over the eye's surface.

and optical transparency allowing integrating them invisible to the user into the glasses lenses [182].

Laser feedback interferometry

The key element in our scanned eye-tracking approach is the LFI sensor. LFI is a widely applied interferometry sensing method [6], which is used e.g. to measure displacement or velocity of solid targets. Recently, LFI sensors have also been applied to AR glasses e.g. for gaze gesture recognition [4, 5] and human activity recognition [133]. This works address the applicability of static LFI sensors for a near-eye setting.
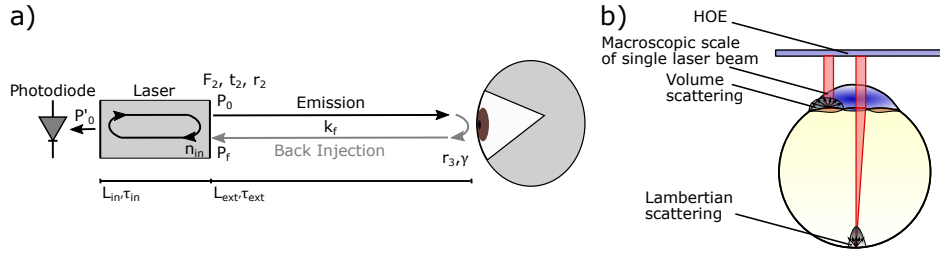
Figure A.29: a) LFI sensing scheme modeled by the well known Coupled-cavity model. Emitted light from the laser is backscattered from the eye's surface and backinjected into the cavity. The photodiode integrated into the back mirror monitors the optical power inside the cavity, which changes based on variation of the feedback path [4, 5].) Macroscopic scale of the laser beam hitting the outer surface (sclera, iris) of the laser or the retina of the eye.

To describe the basic sensing method of LFI sensors, the coupled-cavity model as shown in Figure A.29 a) is used. The laser with its cavity length L_{int} and laser round trip time τ_{int} emits coherent light with an optical power P_0 towards the eye's surface. The laser hits the eye ball with an incident angle γ and dependent on the reflectivity and absorption, summarized by r_3, and the scattering behavior of the tissue, a fraction of the emitted power P_f is backinjected in the cavity of the laser. τ_{ext} describes the round-trip time of the laser to cross the distance L_{ext}. τ_{ext} is given by the speed of light c_0 and the refractive index n_{ext} inside the external cavity [6].

The backinjected wave interferes inside the cavity with the locally oscillating wave, which results in a optical power modulation of the laser

$$P_0' = P_0 \left(1 + m \cdot \cos\left(\phi_{fb}\right)\right). \tag{A.23}$$

The feedback power P_0' relies on the laser's optical power P_0, the modulation depth m and variations of the feedback phase ϕ_{fb}. The photodiode inside the Bragg reflector measures a tiny fraction of the optical modulated feedback power P_0' [6].

While scanning the surface of the eye region, two effects influence the modulated feedback power P_0'. The first effect is an amplitude modulation due to varying reflectivity r_3 and scattering behavior of the different parts of the eye, which influence the modulation depth m. According to [1] m is given in a second order approximation by

$$m = 2 \cdot k_f \cdot \tau_p \cdot \left(\frac{\eta_i}{\eta_d} - 1\right) + k_f \cdot \tau_{int} \cdot (1 - F_2) \left(\frac{1 + r_2^2}{t_2^2}\right). \tag{A.24}$$

The feedback rate k_f describes the normalized reflected field injection rate, τ_p the photon lifetime and the fraction of η_i and η_d the differential efficiency between pump efficiency and quantum efficiency of the cavity. F_2 describes the fraction of total power which is coupled out of the front mirror of the laser cavity. The mirror is further described by its transmitivity t_2 and

r_2. The feedback rate can be rewritten with respect to the three-mirror model by

$$k_f = \frac{t_2^2 \sqrt{\frac{P_0}{P_f}}}{r_2} / \tau_{int} \qquad with \qquad \sqrt{\frac{P_0}{P_f}} \propto \sqrt{\frac{r_3}{r_2}}. \qquad (A.25)$$

Considering a constant transmitivity t_2 and reflectivity r_2 of the front mirror and a constant output of laser power P_0, the coupling factor is mainly affected by a variation of the power of backscattered light P_f due to an increase of reflectivity r_3 of the target and scattering behavior as shown in Figure A.29 b).

Figure A.29 b) shows the macroscopic scale of a single laser beam reflected by the HOE for two deflection angles of the micro mirror. The left beam position describes the beam hitting the outer tissue of the eye (sclera, iris) where volume scattering effects dominate the overall scattering and thus a rather low portion of light is backinjected into the laser cavity. The right beam position describes the beam hitting the retina. In this case the lens of the eye focuses the laser beam onto the retina and as the retina surface is dominated by Lambertian scattering [184], a large portion of light is back scattered. This effect is also referred as red eye effect or bright pupil effect, which is varying in severity across different human eyes [185].

The second effect which influences the modulated feedback power P_0' is given by the modulation of the feedback phase ϕ_{fb} due to speckling effects. Speckling describes the additive superposition of several backscattered signal components with random amplitude and phase. The sum of this components leads to a random modulation of the phase ϕ_{fb} of the back injected light and in particular whether constructive or destructive interference dominates the signal [186]. With respect to the eye this effect is well known from optical coherence tomography (OCT) imaging, where the signal from the retina is characterized by dark an bright speckle patterns [187].

Bright pupil detection
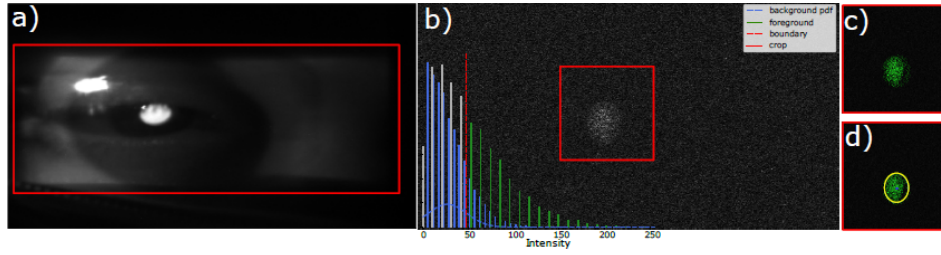
Figure A.30: a) Image of the scan area (bright area inside red box) on a person's eye taken with an IR camera looking directly through the HOE from the outside. b) Background: Modulated feedback power P'_0 measured by the integrated photodiode of the LFI sensor over the scan area. Foreground: Histogram of the retinal area pixel intensity distribution (green and grey) and the non-retinal area distribution (blue). c) Segmented bright retinal area pupil pixels from b) using the intensity boundary (red dashed line). d) Multivariate Gaussian fit of the retinal area pixels in c) with pupil center in blue and pupil contour in yellow.

Figure A.30 a) shows the region of the eye, scanned by the LFI sensor. An IR camera looking directly through the HOE from the outside towards the eye. The pupil appears bright when the IR laser beam hits the retina during the 2D scan, also known as the bright pupil effect. This supports our assumption that the reflectivity r_3 as well as the scattering behavior changes and therefore changes the coupling factor k_f in presence of the retina, resulting in amplitude modulation according to Equation (A.24). In particular, by integrating the photodiode into the back reflector of the laser cavity, IR illumination and sensing element are perfectly aligned on axis to support the bright pupil effect. In addition, the effect of speckling is clearly visible, leading to a normal distributed pattern of bright and dark speckles.

To detect the location of the pupil and therefore track the eye for each full 2D scan the following three steps are applied to each recording.

Image reconstruction: The photodiode signal of the LFI sensor is sampled in equidistant time steps t to capture the modulated feedback power $P'_0(t)$, while the MEMS mirror scans the laser beam over the surface of the eye. To generate an image of the eye region, the modulated feedback power $P'_0(t)$ and the mirror deflection angles $\alpha_h(t)$ and $\beta_v(t)$ are sampled in the same equidistant time steps. A series of samples $(P'_0(t), \alpha_h(t), \beta_v(t))$ are used to construct an image using the mirror deflection angles as pixel coordinates $(I_x \approx \alpha_h, I_y \approx \beta_v)$ on the HOE and the modulated feedback power P'_0 as intensity value $(I(x, y) = P'_0)$ of the pixel. In Figure A.30 b) in the background, a reconstructed image is shown. The pupil appears in the center of the image as a bright pattern, marked by the red box.

Pupil segmentation: Similar to VOG-based eye-tracking sensors, segmentation of the pupil is required to determine the pupil contour and center. To separate the retinal area pixels from the non-retinal area pixels of the image, a histogram-based approach is used. In Figure A.30 b) the normalized histogram of the image is shown in gray and green, containing information about both the retinal area and the non-retinal area. This histogram is overlaid by a second

normalized histogram (in blue), which includes only the first ten lines of the image and represents the non-retinal area area probability density distribution (PDF) $\mathcal{N}_b(\mu_b, \sigma_b)$, since there is no pupil in the first ten lines of the image. To extract the retinal area and thus separate the pupil from the non-retinal area , the intensity limit I_b (red dashed line in Figure A.30 b)) is calculated by $I_b = \mu_b + \sigma_b$ based on the non-retinal area PDF. Using this limit, the normalized histogram of the image is divided into non-retinal area intensity values (gray) and retinal area intensity values (green).

Figure A.30 c) shows a cropped area around the bright pupil pattern for illustration. The remaining retinal area pixels are highlighted in green.

Pupil ellipse fitting: The segmented bright pupil pattern is given as a set of p_i tuples, containing the pixel coordinates as well as the pixel intensity $p_i = (x_i, y_i, P'_{0_i})$. To obtain the pupil ellipse from this set of tuples, a multivariate Gaussian distribution $\mathcal{N}_p(\mu_p, \Sigma_p)$ is fitted using least squares optimization. μ_p hereby represents the pupil center and the main components of the covariance matrix Σ_p represent the horizontal and vertical axis of the ellipse representing the pupil contour. In Figure A.30 d), the resulting ellipse contour is annotated in yellow as well as the center of the ellipse as a blue dot. A pupil ellipse is therefore given by $\mathcal{E}_i = (\mu_{p0}, \mu_{p1}, 3 \cdot \Sigma_{p00}, 3 \cdot \Sigma_{p11})$.

A.4.5 Evaluation
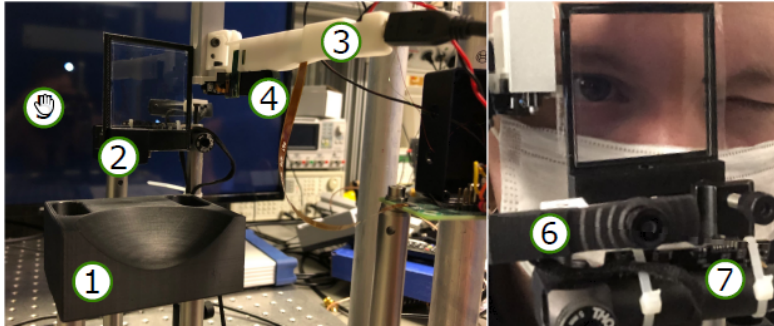
Figure A.31: Laboratory setup to evaluate the proposed scanned LFI eye-tracking sensor. The left image shows the laboratory setup from the perspective of a participant and the right image shows a participant inside the setup.

Figure A.31 shows the laboratory setup used to evaluate the scanned LFI eye-tracking sensor. The LFI sensor component itself is based on an research prototype adapted from an optical communication application where IR VCSELs with monitoring photodiodes in the back DBR are common. A detailed description of the sensor component is given by Grabherr et. al. [2].

The glasses frame temple ③ with the integrated laser module and the MEMS micro mirror module ④ is based on a modified BML500P [188], an optical microsystem developed for AR glasses. The MEMS mirrors are used to scan the IR laser across the surface of the HOE, which

is integrated into a flat glasses lens ②. The high transparency of the HOE allows a participant to sit in front of the laboratory setup and look through the glasses lens towards a display ⑤ on which stimuli markers are displayed. The participant's head is fixed in front of the HOE and the glasses frame temple by a chin rest ① to minimize head movements that could lead to erroneous measurements. In addition a Pupil Core V1 [53] is added to the setup. The world camera ⑥ monitors the stimuli markers on the display and an eye camera ⑦ observes the participant's eye from a bottom-up perspective through the HOE. The mirror signals $\alpha_h(t)$ and $\beta_v(t)$ as well as the interference signal $P'_0(t)$ are captured in the setup by an oscilloscope.

The Lab setup complies according to IEC 60825-1 [139] regularization to a class 1 laser system and therefore does not pose any risks to the eye. The optical power of the IR laser beam surface was limited to an optical power of 142 μW on the eye's surface, whereas the IEC 60825-1 allows a maximum optical power of 778 μW for an 8-hour continuous emission to the retina.

The low required optical power is favorable to minimizes power consumption of our scanned LFI eye-tracking sensor. Using off-the-shelf components, the power consumption of our system is estimated roughly at 30 mW. The main components contributing to the overall systems power consumption are the transimpedance amplifier (TIA) (THS4567   10 mW), which is used to amplify the interference signal $P'_0(t)$ measured by the integrated photodiode, and the analog digital converter (ADC) (MAX19191 with   15.3 mW). The gain of the TIA was set to 940 during the experiments. With further integration, additional power reduction is expected. The estimated power consumption is comparable to reported power consumption of other scanned IR laser eye-tracking sensors. In example, [102] reported a power consumption of 15 mW for their system. A major advantage of our approach is that we reuse the existing MEMS micro mirror of the RGB projection similar to [105], and therefore, did not require an additional scanner which would increase the power consumption.

Gaze accuracy

To evaluate the performance of the scanned LFI eye-tracking sensor and prove the robustness of our approach, we conducted a study with 20 participants (4 female, 16 male, mean age 34 SD(10.83)). The participant's eye colors ranged from dark brown to blue-gray. Half of the participants required vision correction ranging from +1 dpt to -2.75 dpt. Except for participant P14 who wore contact lenses, participants did not wear vision correction during the study. None of the participants were of Asian ethnicity, so the robustness of the effect of reduced bright pupil response in Asian populations as reported by [185] was not tested. All participants gave their written consent after being informed about the nature of the study.

During the study, participants sat approximately 0.6 m away from a 36" display and positioned their head on the chin rest. To set the calibration and test marker coordinates, participants were first asked to look straight through the HOE towards the display. Then, the center marker describing the resting position of the eye at $\theta = 0$ and $\phi = 0$ was adjusted to align with straight gaze. After setting the calibration and test marker coordinates, participants were asked to
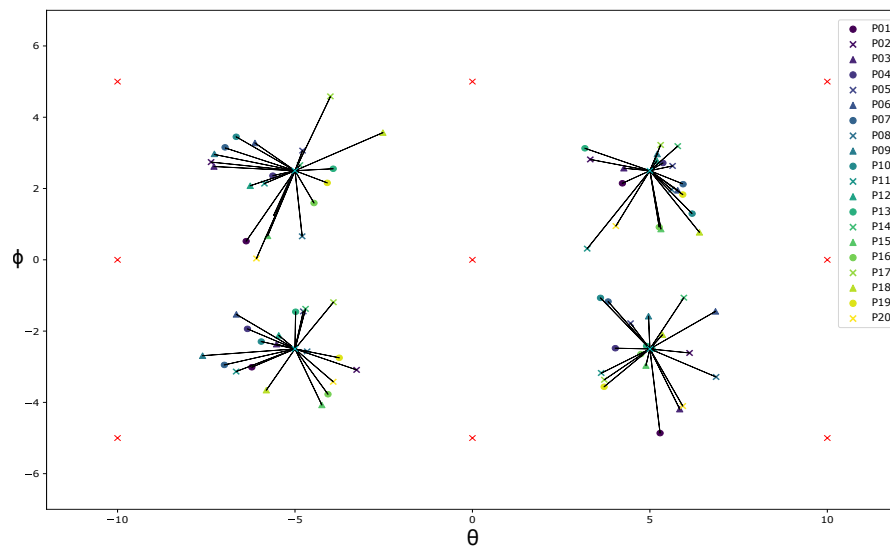
Figure A.32: Results of the gaze accuracy experiment. Participants were asked to fixate the calibration markers (red crosses) and the test markers (cyan crosses). The calculated mean gaze position per test marker and participant is added as a colored marker. In addition, an arrow shows the correlation between the calculated gaze position and the test marker.

follow and fixate on the stimuli markers on the monitor. In a sequence 9 reference markers (red crosses in Figure A.32) and 4 test markers (cyan crosses in Figure A.32) are presented for approximately 5 seconds each with three repetitions resulting in a total of 39 stimuli markers presented per participant. During the experiment, scanned LFI data and images from the Pupil Core eye camera were recorded for each marker location. For each point, the first and last second of recorded data were discarded to ensure that the participant had time to fixate on the next stimuli marker. In addition, scanned LFI sensor images were discarded if no pupil was detected due to blinking. In the Pupil Core data, detected pupil positions with a confidence < 0.8 are discarded in order to eliminate errors due to blinking as well. The pupil core camera was placed 8 cm away from the eye, which is rather large. To compensate the larger distance, the camera focus was tuned to receive sharp images at that distance. To compensate accuracy losses due to the increased distance we reduced the camera angle w.r.t. eye compared to a head worn configuration.

After data acquisition and cleaning of the raw data, the standard 9-point polynomial regression algorithm is used to map the data from pupil position space to gaze angle space. The regression algorithm was trained for each user individually and for both the scanned LFI sensor and the Pupil Core VOG sensor separately. Figure A.32 shows the mapped gaze points for each participant and the 4 test points for the LFI eye-tracking sensor.

To evaluate the scanned LFI eye-tracking sensor based on the captured data, we use the accuracy as evaluation metric, which is defined as the average angular offset between estimated fixation location and the corresponding marker position. In addition, we evaluate the

Table A.3: Accuracy and precision of our approach and the Pupil Core eye tracker over all participants

	Scanned LFI		Pupil Core V1	
	Precision °	**Accuracy °**	**Precision °**	**Accuracy °**
P1	1.991	2.591	0.127	1.802
P2	0.438	1.976	0.172	1.935
P3	0.718	1.239	0.644	1.597
P4	0.587	0.982	0.778	1.585
P5	1.001	1.122	0.746	0.988
P6	1.661	1.792	0.925	1.542
P7	0.512	1.408	0.830	1.820
P8	0.743	1.623	0.974	1.990
P9	1.039	2.877	0.523	1.339
P10	0.905	2.408	0.411	1.616
P11	0.455	1.888	0.830	0.820
P12	1.011	1.082	0.775	0.875
P13	0.743	1.597	0.158	0.966
P14	0.960	1.211	0.892	1.026
P15	0.569	1.077	0.058	0.847
P16	1.733	1.812	0.106	1.126
P17	1.106	1.655	0.477	1.442
P18	0.685	1.951	0.459	1.371
P19	1.138	1.750	0.291	1.181
P20	0.914	1.446	0.791	2.427
Mean	0.945	1.674	0.548	1.415
Std	0.4162	0.5052	0.3014	0.4305

precision, which is defined according to [53] as the root mean squared (RMS) error between successive samples. Table A.3 summarizes precision and accuracy results of the study for the scanned LFI eye-tracking sensor and the Pupil Core.

Our scanned LFI eye-tracking sensor achieves a mean gaze accuracy of 1.674°, which is comparable to the accuracy reported by other scanned laser eye-tracking approaches e.g. the 1.72° reported by [104]. The accuracy of the Pupil Core is 0.232° lower compared to our approach. In our experiments, we did not achieve the stated precision and accuracy of the Pupil Core, which is to some extent due to our laboratory setup as the scanned IR pattern appears as a varying IR illumination, which distorts the dark pupil tracking of the Pupil Core. The results of the study show that the scanned LFI eye-tracking sensor is capable to track the bright pupil with a reasonable accuracy.

Ambient light robustness

A further requirement to eye-tracking sensors for consumer AR glasses is a robust operation under variation of ambient light. To evaluate the ambient light robustness, our scanned LFI

eye-tracking sensor as well as the Pupil Core VOG sensor are exposed to different illumination sources, while a participant was looking straight through the HOE such that the HOE and thus the parallel laser rays were perpendicular to the eye. Figure A.33 summarizes the results of this study. The first row shows images taken with the Pupil Core V1 eye-tracking sensor
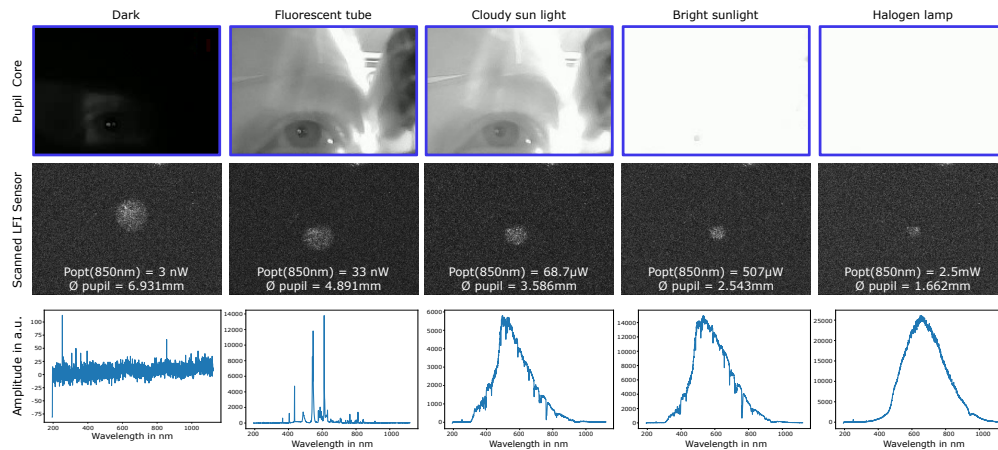
Figure A.33: Comparison of the Pupil Core V1 eye-tracking sensor and our approach with respect to ambient light robustness. The first row shows images captured by the Pupil Core. The second row shows images captured by our approach with annotated optical power of the light source at 850 nm and the estimated pupil diameter from the pupil contour. The last rows shows spectrograms of the different light sources.

using the pupil capture software (V1.17.71) with default settings while the second row shows images captured with the scanned LFI eye-tracking sensor. The last row shows a spectra of each illumination source captured by an OceanOptics4000 optical spectrometer. In addition, we measured the optical power at the wavelength of 850 nm on the eye's surface as both the the Pupil Core eye-tracking sensor and our scanned LFI eye-tracking sensor operates at 850 nm. The results are annotated in the second row of the image.

The first lighting situation we investigated was a completely dark laboratory with no external light sources. Under this condition, both sensors track the pupil as expected. The second lighting situation we investigated is office lighting. Under this very controlled lighting condition, both sensors also worked perfectly. Also under cloudy sunlight this condition, both sensors show stable operation. In bright sunlight ($P_{opt}(850\,\text{nm}) = 507\,\mu\text{W}$) the dark pupil appears only as a tiny dark spot in the camera image, which is no longer robustly detected. While the VOG camera sensor saturates, the scanned LFI eye-tracking sensor still is capable to robustly detect the bright pupil. As already a improved version of the VOG system (Pupil Core V2) is available, which we did not used for the experiment, the image quality for the bright sun light condition might improve.

As final lightning condition, we used a halogen lamp, which is a broadband thermal radiator with characteristically high intensity in the IR wavelength region. With a measured optical power of 2.5 mW at 850 nm the eye region was exposed by a five times higher intensity com-

pared to bright sun light. Even under this condition, the scanned LFI sensor is capable to detect the bright pupil reliably, leading to an outstanding dynamic range of the scanned LFI eye-tracking sensor. The observed high robustness to ambient light is in line with earlier work by [128].

A.4.6 Discussion

To assess the quality of our scanned LFI eye-tracking approach with respect to the state of the art of scanned IR laser eye tracking approaches and discuss the results and potential limitations, we compare our approach with other scanned IR laser eye tracking approaches in Table A.4.

Table A.4: Comparison between different scanned IR eye tracking approaches and our approach

	[102]	[129]	[104]	**Ours**
Tracking method	Corneal reflection	Dark Pupil tracking on rasterized 2D image	Corneal reflection & Stereo image	Bright Pupil tracking on rasterized 2D image
IR Scanner	2D MEMS mirror	2x 1D MEMS mirrors	2D MEMS mirror	2x 1D MEMS mirrors
Accuracy	>1°	1.31°	1.72°	1.67°
Precision	-	0.01°	0.0091°	0.945°
Diag. FOV	35.35°	44.72°	16.97°	22.36°
Power	15 mW	11mW	-	30mW
Sample rate	3300 Hz	60 Hz	4000 Hz	60 Hz

The works of Sarkar et. al. and Holmqvist et. al. differ from the work of Meyer et. al. and our approach mainly with regard to the chosen tracking method. They track corneal reflections with a rather high sampling rate while the work of Meyer et. al. and our work rely on a rasterized 2D image and tracking of either a dark or a bright pupil. All approaches are in the same range of absolute gaze accuracy and power consumption. Furthermore, they are evaluated on a comparable diagonal FOV. The main improvements in our work compared to the state of the art is the *robustness of pupil detection*, which is extremely important for consumer AR glasses. By using the presented LFI measurement method, the sensor is almost immune to ambient light. Due to the signal characteristics of the bright pupil and the proposed algorithm, our approach overcomes several limitations of VOG eye-tracking systems, as it is robust against eyelashes that interfere with the pupil, mascara that causes false pupil detection and eyelids that partially occlude the pupil. In addition, the sensor works independently of eye color and iris structure.
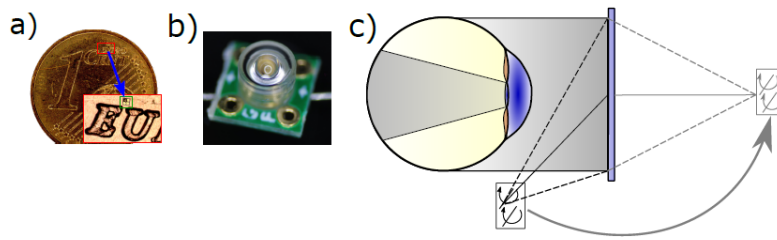
Figure A.34: Sensor integration of the scanned LFI eye-tracking sensor. a) shows a microscope image of the 160 μm x 180 μm LFI sensing element on a coin for scale (blue arrow). b) Encapsulated optical module of the research prototype composing of the LFI sensor as well as the beam shaping lens. The lens diameter is roughly 2 mm. c) virtual rotation of the MEMS scanner to the center of the FOV by the HOE to solve the off-axis integration issue of VOG sensors

Sensor integration

In addition to robust pupil detection, our approach can be fully integrated into AR glasses with retinal projection, as the optical path of the IR laser uses the same optical path as visible light. Moreover, the VCSEL as the optical transmitter and the photodiode as the optical receiver of the LFI sensor element are highly integrated in a single chip, as shown in Figure A.34 a). In combination with the beam shaping optics a diameter of the optical module of below 2 mm is possible (Figure A.34 b)), allowing thus direct integration into the RGB laser module. Compared to other scanning laser approaches, such as e.g. shown by [102], our setup does not require any components to be mounted outside the spectacle temple or even in the spectacle frame. The use of an HOE allows us to virtually rotate our MEMS scanner to the front of the glasses lens as shown in Figure A.34 c). Images taken with the scanned LFI eye-tracking sensor therefore appear as if taken from the perspective of a camera viewing the eye centrally from the outside through the lens. Compared to VOG systems, this effect is possible without any camera arms interfere with the users FOV. In addition, this perspective allows covering the whole eye region and it is possible to track the pupil over a large FOV.

Power consumption

[102] estimated the power consumption of VOG camera sensors at 150 mW, while our sensor consumes only about 30 mW, which is a significant improvement and allows real-time operation in lightweight consumer AR glasses. In addition, our proposed pupil detection algorithm requires less computationally intensive image processing steps to extract the pupil contour compared e.g. to the Pupil Core algorithm [53]. As the output of our pupil detection algorithm is an ellipse contour $\mathcal{E}$, the power consumption to derive an absolute gaze vector e.g. by using a geometrical 3D model approach as proposed by [68] is comparable to VOG eye-tracking systems.

Glasses slippage

A major problem that causes eye-tracking sensor accuracy to degrade is the effect of glasses slippage [93]. This issue also affects our sensor performance as we are working on image data. The impact of slippage to our sensor might however be less significantly affecting our results as the camera axis in our approach is close to the optical axis of the eye. Compared to the work of [129] we only capture the bright pupil and do not gather further any information from the eye region. Thus slippage compensation by tracking landmarks like the eye corners as introduced by [95] is not feasible. A possible solution to achieve slippage robustness for our approach is to adopt the approach of [92] to derive slippage robust features from a geometric 3D eye model, which we will consider and evaluate as part of our future research.

Update rate, latency and motion blur

Due to the tight coupling of the optical path of the RGB projection and the IR path the update rate is limited to 60 Hz, which is compared to [102] and [104] rather low. A faster scanning MEMS mirror would improve the update rate to some extend. However, since the diameter of the laser beam determines the minimum required mirror diameter, mirror miniaturization is limited, resulting in a maximum technically feasible scan frequency of 120 Hz. Compared to VOG systems latency is rather low as in our approach we capture images pixel by pixel, and therefore, the foreground background segmentation can be performed in parallel to image capturing, leading to a latency of 0.0166 s to calculate the pupil ellipse $\mathscr{E}$. A camera sensor captures all pixels in parallel while in contrast our system captures images sequential. Thus a fast saccadic movements of the pupil during image acquisition can lead to elliptical distortions of the captured ellipse $\mathscr{E}$.

Gaze angle dependency of bright pupil effect

The bright pupil effect appears only if both the illumination axis and the sensor axis are close to each other. The scanned LFI system moves both the light source and the detector in parallel, leading to a perfect alignment of both axes. However, a disadvantage of our system is the collimated nature of the laser beam compared to a diverging IR light source. If the laser beam is not approximately perpendicular to the retina due to the Lambertian scattering less light is back injected into the laser cavity leading to a reduced bright pupil response. This effect is independent from external illumination. A possible solution is to use a parabolic mirror function for the IR HOE which follows the curvature of the eye.

A.4.7 Conclusion

In this work, we present a novel scanned LFI eye-tracking sensor approach, which is able to track the pupil with high robustness. Compared to VOG sensors and other scanned laser approaches, we highlight the outstanding robustness to ambient light and the high integrate-

ability of our sensor approach into retinal projection AR glasses. We introduced the sensing technology, derived the physical foundations to describe the signal occurrence and propose a pupil extraction algorithm, which is optimized for the bright pupil signal characteristics measured by our sensor approach. To validate the accuracy of our scanned laser eye-tracking sensor we build a prototype using a modified retinal projection AR glasses setup based on the BML500P, a retinal projection system.

Our eye-tracking sensor achieves a mean accuracy of 1.674°, which is comparable to scanning laser eye-tracking approaches e.g. by [104]. We further solve typical problems of VOG eye-tracking sensors, e.g. the highly off axis integration of camera sensors by using an IR HOE to virtually place the laser scanner in front of the participants eye.

With the advancements especially in ambient light robustness and by the nearly invisible integration of the eye-tracking sensor we pave the way for eye-tracking sensors to become standard sensors for upcoming AR glasses, which will enable new application areas of eye-tracking e.g. long-term gaze monitoring for early detection of mental disorders.

Acknowledgements

A.5 A holographic single-pixel stereo camera eye-tracking sensor for calibration-free eye-tracking in retinal projection AR glasses

A.5.1 Abstract

Eye-tracking is a key technology for future retinal projection based AR glasses as it enables techniques such as foveated rendering or gaze-driven exit pupil steering, which both increases the system's overall performance. However, two of the major challenges video oculography systems face are robust gaze estimation in the presence of glasses slippage, paired with the necessity of frequent sensor calibration. To overcome these challenges, we propose a novel, calibration-free eye-tracking sensor for AR glasses based on a highly transparent holographic optical element (HOE) and a laser scanner. We fabricate a segmented HOE generating two stereo images of the eye-region. A single-pixel detector in combination with our stereo reconstruction algorithm is used to precisely calculate the gaze position.
In our laboratory setup we demonstrate a calibration-free accuracy of 1.35° achieved by our eye-tracking sensor; highlighting the sensor's suitability for consumer AR glasses.

A.5.2 Introduction

Eye-tracking is a key technology to improve current near-eye displays such as holographic retinal projection displays [174]. It can enable foveated rendering to increase the system's perceived resolution while minimizing the system's power consumption, or increase the overall field of view (FOV) by using exit pupil steering [158, 38, 39]. Thus, robust eye-tracking sensors are required for future near-eye displays such as augmented reality (AR) glasses.

Current state of the art video-oculography (VOG) meets the requirements for gaze accuracy but is limited in sensor integration and robustness against slippage of glasses. In addition, VOG systems often require at least a single marker calibration to achieve high gaze accuracy; thus limiting the usability of VOG sensors in everyday AR glasses.

In order to achieve calibration-free eye-tracking [96] have introduced a stereo camera approach for 3D reconstruction of the pupil using a closed-form stereo reconstruction algorithm. However, this approach implements multiple camera sensors to cover a sufficiently large FOV; resulting in the integration of additional components with a high power consumption.

To address drawbacks of current VOG stereo approaches for AR glasses, we introduce a single-pixel holographic stereo camera sensor approach, which is capable of capturing a stereo perspective view of the eye-region. We embedded a holographic optical element (HOE) in the eyeglass lens, which we used to redirect a scanned laser beam to the eye-region; followed by the capturing of backscattered light using a single-pixel detector. The HOE is designed to achieve a stereo perspective vision by performing two holographic functions; transforming a single scanned beam into two separate beams which propagate towards the eye. HOEs based on photopolymer technology provide excellent transparency and low noise which, combined

with their high integrability, makes them particularly interesting for future AR glasses. In addition to that, the presented stereo camera VOG sensor consumes only a fraction of the power of current VOG stereo systems.

Our main contributions are i) we solved the sensor integration problem of headword stereo camera VOG systems by using an HOE and a single-pixel detector in a scanned laser system to generate stereo images and ii) we fabricated a suitable HOE and demonstrated the calibration-free 3D reconstruction of the eye's optical axis in our experimental setup by applying the closed-form stereo reconstruction algorithm proposed by [189]. We further evaluated the gaze accuracy and precision of the overall system.

Our main contribution compared to previous work by [128],[129] is the proposed and fabricated segmented HOE to create a stereo perspective of the eye to enable stereo reconstruction algorithms e.g. by [189] to reconstruct the optical axis of the eye without calibration.

In the next section, we present related work focusing on stereo camera eye-tracking for mobile applications. Sec. A.5.4 describes our single-pixel stereoscopic holographic camera sensor system design as well as the reconstruction algorithm to derive the 3D gaze vector from a pair of images. Afterwards, in Sec. A.5.6 and Sec. A.5.7 our approach is evaluated in a laboratory setup to derive gaze accuracy and precision. Finally, in Sec. A.5.8, we conclude our work and discuss limitations and further research directions.

A.5.3   Related Work

One of the first works introducing a stereo camera approach for calibration-free eye-tracking was published by [96]. The authors used a pair of infrared (IR) cameras to capture a set of images to create a stereo perspective of the left eye. The pair of cameras were then integrated into the side of the glasses' frame and an image of the eye was relayed onto the camera sensors via a semi-transparent mirror (hence resulting in a rather bulky design). By applying a closed-form stereo reconstruction algorithm [189], a 3D perspective of the pupil plane was reconstructed by extracting pairs of pupil ellipses. The system achieved an accuracy of 2.2° over a FOV of 30°.

[190] combined a stereo camera approach with additional glint features and reported a gaze accuracy of 1.6°. To solve the camera integration issue they integrated the camera pair below the glasses by adding a mechanical arm, leading thus to a highly off-axis camera integration. A similar approach is proposed by [191], which combined corneal images captured by an RGB camera as well as pupil ellipse features extracted from images captured by a second IR camera in a mobile stereo camera setup. The cameras were mounted on a mechanical arm to integrate them into a head-worn setup, leading to similar highly off-axis integration. By fusion of corneal images and IR images, the authors achieved a mean accuracy of 2.19° over a FOV of 30°. Additionally, this approach showed robustness to calibration drift.

Currently, the Tobii Pro glasses [97] are to the best of our knowledge the only commercial

mobile eye tracker, working with a pair of cameras. They have embedded the camera sensors and IR illumination directly into the glass lenses inside the user's FOV. After a single point calibration, a gaze accuracy of 0.6° is reported combining stereo reconstruction methods as well as corneal reflection methods. The system shows to be robust to slippage [93].

To summarize the state of the art, all current approaches use IR camera sensors that are either directly integrated into the lens of the eyeglasses or mechanically integrated by camera brackets, which results in an impairment of the user's vision. Only the work of [96] tend to reduce the impairment by using a semi-transparent mirror. However, their approach still requires two power consuming camera sensors and additional imaging optics per eye.

In summary, to address challenges related to camera integration and reduce the power consumption of the overall system, we present the combination of an HOE with a single-pixel detector. With this, we demonstrate an eye-tracking system creating a stereo holographic image with two virtual cameras.

A.5.4 Holographic Single-Pixel Stereo Camera

The holographic single-pixel stereo camera eye-tracking sensor, as shown in Figure A.35 a), consists of three main components: a 2D laser scanner with a MEMS (micro-electromechanical system) micro mirror as scanning unit, an HOE embedded into the glasses' lens and a single-pixel detector. The MEMS scanner with its scan angles α and β deflects the laser beam towards the glasses' lens to scan along the surface of the embedded HOE. A review on MEMS scanner technology is given by [192]. The laser component is integrated into an RGB laser module as described in our previous work [129]. The HOE is deflecting the laser beam towards the user's eye, forming the total image plane I_i at the eye's surface.

The global holographic function of the HOE is defined by considering the whole scanning area as a diverging point source, originating from the scanning point of the MEMS scanner. The HOE splits up the incoming beam into two diverging point sources, which propagate towards the user's eye under different angles. The origin points of both wave fronts generated by the HOE lay in the space behind the HOE's surface, stretching out two diverging cones towards the eye's surface as outlined in Fig. A.35 a).

a)

b)
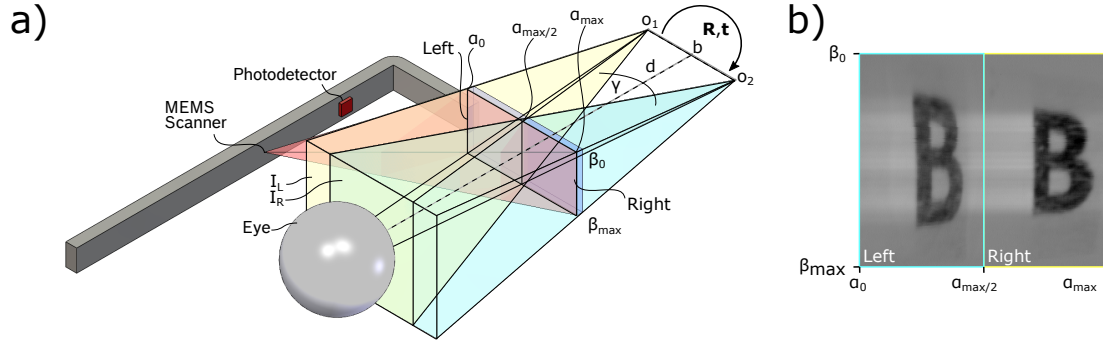
Figure A.35: a) Schematic integration of the proposed eye-tracking sensor into a glasses frame. The MEMS laser scanner and the single-pixel detector are integrated into the frame temple while the HOE with its two sub-HOEs (Left and Right) with the corresponding wave fronts (yellow and blue) are integrated into one glasses lens. b) Image of a capital B letter printed on paper. The paper is placed in the overlapping image plane (I_L, I_R), captured by the holographic single-pixel stereo camera system

Each sub-HOE can be regarded as a virtual laser scanner, which in combination with the single-pixel detector can be described as a virtual camera sensor with its own coordinate space c_i, a camera matrix $\mathbf{K}_i$ and an optical center of the camera o_i as the origin of c_i, with $i \in \{1, 2\}$ for each virtual camera sensor. The coordinate spaces of both virtual cameras are linked to each other by the rotation matrix $\mathbf{R}$ in conjunction with the translation vector $\mathbf{t}$.

To reconstruct an image from the scanned laser beam and the HOE, the backscattered light from the eye's surface is captured by the single-pixel detector consisting of a photosensitive diode, a transimpedance amplifier and an analog-to-digital converter (ADC), which samples the photodiode current in equidistant time steps. Therefore, the sample rate of the ADC determines the pixel clock rate of the outlined system, which consequently determines the resolution of the virtual camera sensors. The position of an individual pixel on the sensor plane of the virtual camera relative to the image plane I_i is given by the deflection angles α (horizontal) and β (vertical) of the MEMS scanner, whereas the intensity of an individual pixel is given by the scattered light measured by the single-pixel detector at each scanning position.

By splitting the HOE into two non-overlapping sub-HOEs with overlapping image planes, the same scanning point in the global image plane results in two diffracted light signals, originating from the two virtual laser scanners with different propagation angles. Thus, one full scan from $\alpha_0 - \alpha_{max}$ and $\beta_0 - \beta_{max}$ of the MEMS scanner generates two image frames of the same object as shown in Figure A.35 b), which can be reconstructed into two images from the object under two different perspectives. The frame rate of the proposed stereo system therefore depends on the number of full scans the laser scanner performs per second.

HOE fabrication

The HOE is fabricated by means of holographic wave front printing. Details on the recording setup employed as an extended immersion-based holographic wave front printer, are outlined in [182]. The HOE is made up of individual sub-holograms, so called Hogels, which are aligned in an array-based structure. Each Hogel is recorded by sequentially relaying two coherent recording wave fronts onto a photopolymer-based holographic film. The resulting interference pattern leads to a photopolymerization-based modulation of the local refractive index in the volume of the holographic film, which results in a manifestation of the recording wave front's characteristics in the form of a 3D diffraction grating.

The HOE is recorded in development grade Bayfol® HX TP* photopolymer [193] by Covestro, with a photopolymer thickness of $16\,\mu$m and a protective polyamide layer of $60\,\mu$m. Individual Hogels are recorded via two monochromatic wave fronts, modulated by means of two phase-only spatial light modulators. Both wave fronts originate from a common single-mode laser source with $\lambda = 639\,$nm. For demonstration purposes and ease of experimentation a wavelength in the visible spectrum has been chosen; however, future HOE-based systems are planned to operate at non-visible IR wavelengths as outlined in [183]. Each Hogel performs an individual optical transformation, which contributes to the HOE's global holographic function. The HOE realizes a combiner functionality, whereby each of the two sub-HOEs performs a point-source-to-point-source transformation from a large off-axis to a close to on-axis configuration. Both sub-HOEs have the same off-axis recording point source placed at the MEMS scanner position. Under reconstruction of the HOE each sub-HOE deflects a diverging wave front propagating with an angle of $\pm\,7.1°$ relative to the sub-HOE's respective surface normal.

Stereo calibration

To reconstruct the pupil-normal-vector in space, the parameters $\mathbf{K}_1, \mathbf{K}_2, \mathbf{R}$, and $\mathbf{t}$ of the proposed single-pixel stereo holographic eye-tracking camera sensor must be known. These parameters can be determined by a camera calibration procedure using a chess pattern. In a first step, images of a checkerboard pattern are captured with both virtual cameras, varying the orientation and position of the checkerboard. Afterwards, the method of [194] is applied to determine the camera matrix $\mathbf{K}_i$ of each virtual camera, as well as the relative position $\mathbf{t}$ and orientation $\mathbf{R}$ of both cameras to each other. The stereo calibration process is required only once after the HOE is embedded in the eyeglass lens and the laser projector is integrated into the eyeglass temple.
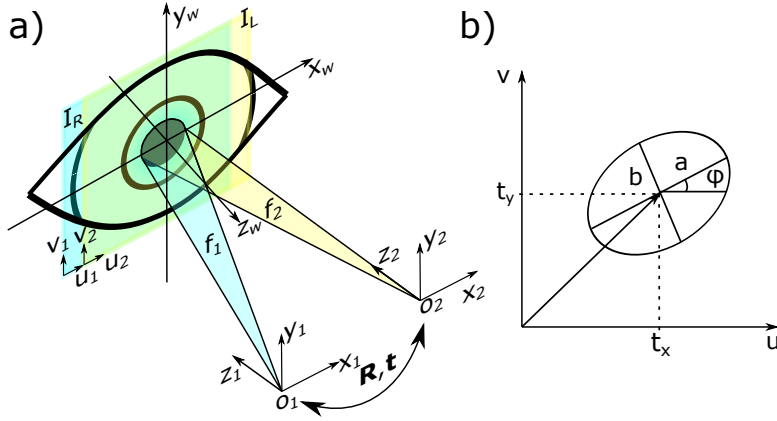
Figure A.36: a) Correspondence between the world coordinate systems c_w and the camera coordinate systems c_i. the origin of each camera coordinate system o_i is the origin of one of the two cones (f_1, f_2), which are constructed from the pupil ellipse in the 2D image space. b) Parameterized description of the pupil ellipse $\mathscr{E}$ in the image plane

Stereo correspondence

The correspondence of the pupil in the stereo system can be described by its appearance as ellipse $\mathscr{E}_i$ in each camera frame I_i. With the given pupil ellipse $\mathscr{E}_i$ and the known optical center of the camera, a cone, e.g. f_1, can be constructed to re-project the pupil ellipse from a 2D image plane to 3D, as shown in Figure A.36 a). From a single cone, the 3D position and orientation of the pupil cannot be derived due to the fact that many valid pupil projections exist, which yield the same cone [9]. To solve this problem, we make use of the correspondence of the pupil and construct a second cone f_2. As both cones are constructed from the same pupil in the 3D space, the pupil position $\mathbf{t}_1$ and orientation $\mathbf{R}_1$ with respect to c_1 can be derived from the intersect of both cones. Aside from $\mathbf{t}_1$ and $\mathbf{R}_1$ the pupil size a and b is derived. To reconstruct the pupil ellipse, the closed form solution as proposed by [189] is used.

Each virtual-holographic single-pixel stereo camera operates in its own coordinate space c_i with its camera center o_i and an image plane I_i in which the normalized pupil ellipse $\hat{\mathscr{E}}_i$ is defined. Since the contour of the pupil in the image plane is assumed to be an ellipse or circle and both cameras observe the same pupil, the intersection of the two cones f_1 and f_2 again results in a conic intersection. Therefore, the projected ellipse are defined by two conics $\mathbf{A}$ given by

$$x^T \mathbf{A}_1 x = 0 \qquad x^T \mathbf{A}_2 x = 0. \tag{A.26}$$

An ellipse $\mathscr{E}_i$ is described in the image plane by its principal axis a and b, its position in the image given by t_x and t_y as well as the orientation ϕ with respect to the u-axis, as shown in Figure A.36 b). With this information the conic matrices A_1 and A_2 can be derived by applying

an affine transformation $\mathbf{S}$

$$
\mathbf{S}_i = \begin{pmatrix} \cos(\phi_i) & -\sin(\phi_i) & -t_{x_i}\cos(\phi_i) + t_{y_i}\sin(\phi_i) \\ \sin(\phi_i) & \cos(\phi_i) & -t_{x_i}\sin(\phi_i) - t_{y_i}\cos(\phi_i) \\ 0 & 0 & 1 \end{pmatrix} \tag{A.27}
$$

to the ellipse matrix $\mathbf{H}_i$ for virtual camera i in normal form in the pupil plane

$$
\mathbf{H}_i = \begin{pmatrix} \frac{1}{a_i^2} & 0 & 0 \\ 0 & \frac{1}{b_i^2} & 0 \\ 0 & 0 & -1 \end{pmatrix} \tag{A.28}
$$

which results in $\mathbf{A}_i = \mathbf{S}_i^T \mathbf{H}_i \mathbf{S}_i$. As $\mathbf{H}_i$ is defined in the pupil plane, we can without loss of generality assume a focal distance $f = 1$ and therefore rescale and translate the ellipse parameters in the image plane I_i by

$$
\hat{t}_{x_i} = \frac{1}{f_{x_i}}(t_{x_i} - c_{x_i}), \quad \hat{t}_{y_i} = \frac{1}{f_{y_i}}(t_{y_i} - c_{y_i}), \quad \hat{a}_i = \frac{a_i}{f_{x_i}}, \quad \hat{b}_i = \frac{b_i}{f_{y_i}} \tag{A.29}
$$

where f_{x_i} and f_{y_i} are the the focal distances and c_{x_i} and c_{y_i} are the camera centers from $\mathbf{K}_i$. After normalization of the ellipse $\mathscr{E}_i$ towards the pupil plane, the relation between the camera coordinate system c_i and the world coordinate system c_w is given by $x_i = \mathbf{R}_i x_w + \mathbf{t}_i$. For points in the pupil plane I_i, this can be rewritten as $x_i = \mathbf{G}_i \mathbf{u}_w$ where $\mathbf{u}_w = (x_w, y_w, 1)$ are homogeneous coordinates in the pupil plane and $\mathbf{G}_i$ is a 3×3 matrix consisting of the first two rows of $\mathbf{R}_i$ and the last row contains the corresponding translation vector $\mathbf{t}_i$

$$
\mathbf{G}_i = (\mathbf{r}_{i1} \mathbf{r}_{i2} \mathbf{t}_i) \quad i = 1, 2. \tag{A.30}
$$

Considering the pinhole camera model $u_i = \frac{x_i}{z_i}$ and $v_i = \frac{y_i}{z_i}$, the pupil plane is linked to the world coordinate system c_w:

$$
z_i \mathbf{u}_i = \mathbf{G}_i \mathbf{u}_w \qquad i = 1, 2. \tag{A.31}
$$

With this correspondence between c_w and c_i, a pupil ellipse $\mathscr{E}$ in the pupil plane is defined according to [189] by

$$
\mathbf{u}_w^T \mathbf{H} u_w = 0 \tag{A.32}
$$

and its projection with respect to Equation (A.26) by

$$
\mathbf{u}_i^T \mathbf{A}_i \mathbf{u}_i = 0 \qquad i = 1, 2. \tag{A.33}
$$

Inserting Equation (A.31) into Equation (A.33) yields

$$\mathbf{u}_w^T \mathbf{G}_i^T \mathbf{A}_i \mathbf{G}_i \mathbf{u}_w = 0 \qquad i = 1, 2. \tag{A.34}$$

As both equations Equation (A.32) and Equation (A.34) describe the same cone, $\mathbf{H}$ can be written in a generalized form

$$\mathbf{G}_i^T \mathbf{A}_i \mathbf{G}_i = k_i \mathbf{H} \tag{A.35}$$

with k_i describing an unknown scaling factor of the cone [189].

A.5.5 Stereo reconstruction

For a set of two virtual cameras, Equation (A.35) provides 12 constraints as it contains two real valued symmetric 3×3 matrices with six parameters each. As we have only 10 unknown parameters, three from $\mathbf{R}_1$ and $\mathbf{t}_1$, as well as the scalars k_1, k_2, a and b, the system is overdetermined and we can solve it e.g. for $\mathbf{R}_1$ and $\mathbf{t}_1$ independently as shown by [189].

In a first step, Equation (A.35) is reduced to

$$\left(\mathbf{R}_1^T \mathbf{A}_1 \mathbf{R}_1\right)^{2\times 2} = k_1 \mathbf{H}^{2\times 2}, \left(\mathbf{R}_2^T \mathbf{A}_2 \mathbf{R}_2\right)^{2\times 2} = k_2 \mathbf{H}^{2\times 2} \tag{A.36}$$

where 2×2 denotes the upper left submatrix of the corresponding 3×3 matrix. By substituting the known stereo correspondence $\mathbf{R}_2 = \mathbf{R}\mathbf{R}_1$ and $\mathbf{t}_2 = \mathbf{R}\mathbf{t}_1 + \mathbf{t}$ obtained from stereo calibration as described in Appendix A.5.4 and elimination of $\mathbf{H}^{2\times 2}$, Equation (A.36) can be rewritten to

$$\left[\mathbf{R}_1 \left(\mathbf{A}_1 - k\mathbf{R}^T \mathbf{A}_2 \mathbf{R}\right) \mathbf{R}_1\right]^{2\times 2} = \mathscr{O}^{2\times 2} \quad with \quad k = \frac{k_1}{k_2}. \tag{A.37}$$

To achieve a (2×2) zero matrix $\mathscr{O}$, as stated by Equation (A.37), the determinant of the corresponding (3×3) matrix must be equal to zero. Therefore, Equation (A.37) yields

$$det\left(\mathbf{A}_1 - k\mathbf{R}^T \mathbf{A}_2 \mathbf{R}\right) = 0 \tag{A.38}$$

which shows that k is an eigenvalue of the matrix $\mathbf{R}^T \mathbf{A}_2 \mathbf{A}_1$. By solving for k and denoting the left side of Equation (A.38) by $det\,(\mathbf{C})$ Equation (A.37) can be rewritten as follows:

$$\left(\mathbf{R}_1 \mathbf{C} \mathbf{R}^T\right)^{2\times 2} = \mathscr{O}^{2\times 2} \tag{A.39}$$

As we already used $det\,(\mathbf{C}) = 0$, Equation (A.39) only provides two independent solutions. With this constraint, $\mathbf{R}_1$ can be found. In a first step the two non-zero eigenvalues λ_1 and λ_2 and the corresponding eigenvectors $\mathbf{s}_1$ and $\mathbf{s}_2$ of $\mathbf{C}$ are calculated. Afterwards, the third column $\mathbf{r}_{31}$ of $\mathbf{R}_1$ is given by

$$\mathbf{r}_{31} = \pm norm\left(\sqrt{|\lambda_1|}\mathbf{s}_1 \pm \sqrt{|\lambda_2|}\mathbf{s}_2\right). \tag{A.40}$$

Equation (A.40) yields four different possible solutions of $\mathbf{r}_{31}$, which is due to the fact that the two intersecting cones f_1 and f_2 have two ellipses in common, as [189] shows. Considering the geometric conditions, that both virtual cameras are positioned on the same side of the ellipse and the gaze vector is directed away from the pupil plane towards the virtual camera, only one possible solution remains [96]. To obtain the correct $\mathbf{r}_{31}$, we must ensure that both z-components of $\mathbf{r}_{31}$ and $\mathbf{r}_{32} = \mathbf{R}\mathbf{r}_{31}$ are positive. Using this criterion, in a second step the remaining vector $\mathbf{r}_{31}$ is selected and the other columns of $\mathbf{R}_1$, $\mathbf{r}_{21}$ and $\mathbf{r}_{11}$, are given by the corresponding eigenvectors of $\mathbf{A}_1 - \mathbf{r}_{31}\mathbf{r}_{31}^T\mathbf{A}_1$, which finally leads to a solution for $\mathbf{R}_1$. Afterwards, the remaining parameters can be resolved in the following order

$$\mathbf{t}_1 = \begin{pmatrix} \mathbf{r}_{11}^T\mathbf{A}_1^T \\ \mathbf{r}_{12}^T\mathbf{A}_1^T \\ \mathbf{r}_{21}^T\mathbf{A}_2^T\mathbf{R} \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ -\mathbf{r}_{21}^T\mathbf{A}_2^T\mathbf{t} \end{pmatrix}, \quad k_1 = -\mathbf{t}_1^T\mathbf{A}_1\mathbf{t}_1, \quad a^2 = \frac{k_1}{\mathbf{r}_{11}^T\mathbf{A}_1\mathbf{r}_{11}}, \quad b^2 = \frac{k_1}{\mathbf{r}_{12}^T\mathbf{A}_1\mathbf{r}_{12}} \tag{A.41}$$

to obtain the length of the principle axis a and b of the pupil ellipse $\mathscr{E}$ as well as the center of the pupil with respect to the camera coordinate system c_1. From $\mathbf{t}_1$ and $\mathbf{R}_1$ the pupil-normal-vector can be derived to calculate the gaze angles.
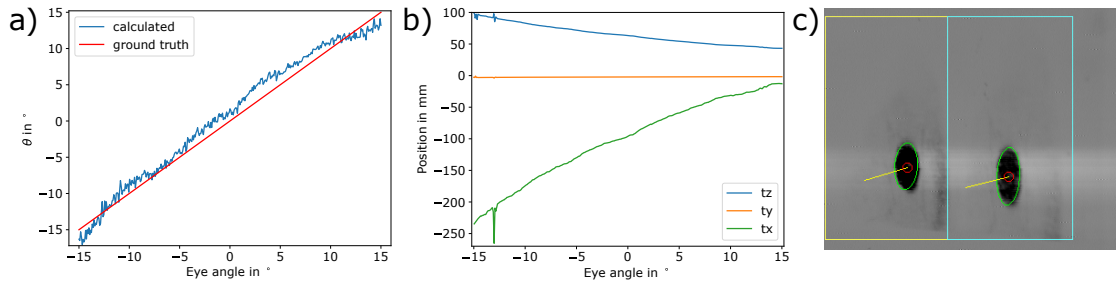
Figure A.37: a) Calculated gaze angle θ (blue) based on stereo reconstruction and ground truth from rotation stage (red), b) Position of pupil center w.r.t. virtual camera 1 $\mathbf{t}_1$, c) A captured stereo image with annotated ellipse contour, center, and reconstructed gaze vector

A.5.6 Evaluation
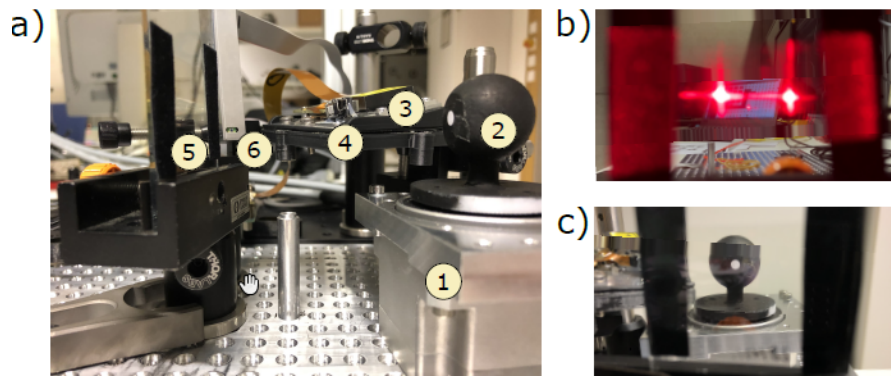
Figure A.38: a) Laboratory setup consisting of the fabricated HOE (5), the MEMS laser scanner (3,4), a single-pixel detector (6) and an artificial eye (2), mounted on a precision, motorized rotation stage (1), b) Origin of the two virtual cameras captured by a camera placed in the pupil plane, c) Image of the artificial eye through the glasses' lens demonstrating the excellent transparency of the HOE.

To evaluate the accuracy and precision of our single-pixel holographic stereo camera eye-tracking sensor and to demonstrate the working principle, we implemented a laboratory setup as outlined in Figure A.38.

An artificial eye model (2) is attached to a precision motorized rotation stage (1) [1], which allows the artificial eye to be rotated with a resolution of 0.0089° and a precision of ±0.00083°. Furthermore a laser scanner[2] consisting of a laser module (3) and a MEMS micro mirror (4) is used to scan a 2D field across the fabricated HOE (5), which generates the stereo perspective. To capture the reflected light for image reconstruction, a photodiode (6) is used.

The HOE is fabricated to operate in accordance with the red laser wavelength ($\lambda = 639\,\mathrm{nm}$) of the laser scanner. The two optical center points of the two virtual cameras o_1 and o_2 are shown in Figure A.38 b) captured from the pupil plane. Finally, Figure A.38 shows an image of the artificial eye model taken from the outside of the lens to highlight the high transparency of the HOE. The artificial eye model uses a white pupil to emphasize the specular reflection of red light at the pupil, causing the photodetector to perceive a dark pupil, similar to the dark pupil shown by [129] on a real eye using an IR laser scanner.

Tab. A.5 summarizes the geometrical and electrical parameters of the prototype system used in the laboratory to evaluate our approach.

To determine the camera parameters ($\mathbf{K}_1$, $\mathbf{K}_2$, $\mathbf{R}$, and $\mathbf{t}$) of the system, the stereo system is first calibrated as described in Appendix A.5.4 using a 5×3 checkerboard with a square size of 3 mm. Afterwards, the artificial eye is rotated in 0.5° increments from -15° to 15° to cover a total FOV of 30 °. For each position, 100 image pairs are acquired with the single-pixel-detector

---

[1] Rxhq 50-12T0.3 Jenny Science
[2] BML050 [195]

Table A.5: Summary of design parameters of the stereo setup in the laboratory setup according to Figure A.35.

Baseline b	Distance d	Camera angle γ	Camera resolution (w×h)	Size of HOE (w×h)	Pixel clock	Frame rate
20 mm	80 mm	14.2°	110 px × 240 px	20 mm × 10 mm	22 MHz	60 Hz

as described in Appendix A.5.4. Then, the two sub-images per image are isolated and the contour of the pupil ellipse is extracted for each sub-image. Based on the extracted contours, the ellipse $\mathscr{E}_i$ is fitted using a least squares method and the ellipse parameters (a_i, b_i, t_{xi}, t_{yi} and ϕ_i) are extracted. After rescaling the ellipse parameters by applying Equation (A.29) and computing A_1 and A_2, the reconstruction algorithm, described in Appendix A.5.5, is used to determine the gaze vector as well as the position of the ellipse with respect to c_1. Figure A.37 summarizes the results of the experiment.

Figure A.37 a) shows the gaze angle resulting from the reconstructed gaze vector for the artificial eye. The mean gaze accuracy is 1.35° with a mean precision of 0.02°. The gaze accuracy is mainly limited by the pupil segmentation and the ellipse fitting accuracy. The accuracy can be improved by using more advanced ellipse fitting algorithms e.g. DeepVOG by [67]. Figure A.37 b) shows the pupil center $\mathbf{t}_1$ relative to $\mathbf{c}_1$. While the y-axis is constant, the x-axis varies as the rotation of θ around y_w leads to a shift of the pupil center along x_w. Furthermore, the distance d between camera center c_1 and the pupil plane is close to the camera distance of 80 mm described in Tab. A.5.

A.5.7   Discussion

Overall, the experiment shows that stereo reconstruction leads to a comparably high gaze accuracy, even with the low resolution of the two virtual cameras. In terms of sensor integration, our approach has several advantages over VOG stereo systems, as the use of an HOE allows to capture image information from multiple angles.

[129] reported a power consumption of 11 mW of a similar setup, which is already significant lower compared to VOG systems. Therefore, our approach reduces power consumption significantly further compared to a stereo VOG setup, since we generate two virtual cameras from a single laser scanner and only have to sacrifice image resolution.

A limitation of our approach is the high dependence on robust detection of the pupil ellipse $\mathscr{E}$, since the reconstruction fails if the cones f_1 and f_2 do not intersect. [128] addressed this general issue by replacing the IR laser and the single-pixel-detector by an laser feedback interferometry (LFI) sensor to increase the robustness of pupil detection, an approach that could be applied also in this working context.

Finally the stereo reconstruction algorithm only estimates the normal vector of the pupil and thus the optical axis of the eye omitting the offset angle κ between optical and visual axis [63].

To determine the user dependent κ at least a one time single marker calibration is required. Otherwise κ needs to derived from average population.

A.5.8   Conclusion

In this paper, we presented a novel approach to mobile stereo camera eye-tracking for AR glasses, which combines a MEMS laser scanner with a segmented HOE. We fabricated the HOE and demonstrated its functionality for the proposed use-case. We applied a stereo reconstruction algorithm to the stereo images captured by our holographic single-pixel virtual stereo camera, which achieved an accuracy of 1.35°.

To increase the robustness of the stereo reconstruction as well as the FOV covered by the virtual cameras, we plan to increase the number of virtual cameras and optimize the orientation of the cameras w.r.t. to the eye. In addition, we plan to integrate our system into a head-worn demonstrator based on the BML500P [3] to further investigate precision and accuracy also taking human error into account.

Acknowledgements

---

[3] [188]

B Static LFI HCI for AR Glassess

This chapter includes the publications [5, 4, 133, 134]:

1. **Johannes Meyer**, Thomas Schlebusch, Hans Spruit, Jochen Hellmig, Enkelejda Kasneci. "A Novel Gaze Gesture Sensor for Smart Glasses Based on Laser Self-Mixing". In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (2021), https://doi.org/10.1145/3411763.3451621

2. **Johannes Meyer**, Thomas Schlebusch, Hans Spruit, Jochen Hellmig, Enkelejda Kasneci. "A compact low-power gaze gesture sensor based on laser feedback interferometry for smart glasses". Proc. SPIE 11788, Digital Optical Technologies 2021 (2021), https://doi.org/10.1117/12.2593772

3. **Johannes Meyer**, Adrian Frank, Thomas Schlebusch, Enkeljeda Kasneci. "A CNN-based Human Activity Recognition System Combining a Laser Feedback Interferometry Eye Movement Sensor and an IMU for Context-aware Smart Glasses". Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5 (2021), https://doi.org/10.1145/3494998

4. **Johannes Meyer**, Adrian Frank, Thomas Schlebusch, Enkelejda Kasneci. "U-HAR: A Convolutional Approach to Human Activity Recognition Combining Head and Eye Movements for Context-Aware Smart Glasses". Proc. ACM Hum.-Comput. Interact. 6 (2022), https://doi.org/10.1145/3530884

B.1 A Novel Gaze Gesture Sensor for Smart Glasses Based on Laser Self-Mixing

B.1.1 Abstract

The integration of gaze gesture sensors in next-generation smart glasses will improve usability and enable new interaction concepts. However, consumer smart glasses place additional requirements to gaze gesture sensors, such as a low power consumption, high integration capability and robustness to ambient illumination. We propose a novel gaze gesture sensor based on laser feedback interferometry (LFI), which is capable to measure the rotational velocity of the eye as well as the sensor's distance towards the eye. This sensor delivers a unique and novel set of features with an outstanding sample rate allowing to not only predict a gaze gesture but also to anticipate it. To take full advantage of the unique sensor features and the high sampling rate, we propose a novel gaze symbol classification algorithm based on single sample. At a mean F1-score of 93.44 %, our algorithms shows exceptional classification performance.

B.1.2 Introduction

In recent years, various smart glasses have been released into the market. As a successor of virtual reality (VR) glasses, they follow a more natural design while integrating similar sensing technology into the glasses. In general, they consist of a light projection engine and a set of sensors to capture user inputs and monitor the user's state.

To enable human computer interaction (HCI) with smart glasses, a variety of interaction concepts are possible. These concepts can be classified into touch based interactions e.g. via a track pad on the glasse's temple introduced by the Google Glass [196], an external on-body device e.g. a controller like as in the Magic Leap One [197] or via touchless interactions such as spoken commands via voice recognition as used by the Echo Frames [132]. Also, hand gestures captured by a camera sensor have been shown by the HoloLens [17].

An additional touchless interaction concept is making use of the user's gaze by tracking their eye movements by means of eye-tracking sensors. Gaze-based interaction allows a fast and natural input, leading to an intuitive and unobtrusive way of interaction with smart glasses while maintaining social acceptance and user's privacy [17].

To the best of our knowledge, no commercially available smart glasses solution so far utilizes gaze gestures as input modality, mainly due to the challenges arising from integration and due to limited available power and space constraints [17]. Furthermore, a high robustness of the sensor against variable external illumination is required to allow for operation in uncontrolled outdoor environments e.g. in bright sun light [56].

To overcome these limitations and enable gaze gesture based interaction for the next genera-

tion of smart glasses, we present a novel low power gaze gesture sensing approach based on laser feedback interferometry (LFI). This multimodal sensor is capable of measuring distance towards the eye as well as eye rotational velocity with a high sample rate of up to 1 kHz. The sensor is based on a small vertical cavity surface emitting laser (VCSEL) in the near infrared (IR) spectrum, which allows for a space constrained integration into the glasse's frames. Additionally, the LFI sensing principle is due to the coherent sensing scheme only sensitive to its own radiation, allowing a robust operation in presence of ambient radiation, as shown in [128].

In the next section, we give an overview of the state of the art regarding gaze gesture sensor concepts for AR glasses and discuss the limitations of existing sensor concepts. In Appendix B.1.4, we introduce the sensing principle and the sensor concept. In addition, we provide an overview of the measurement features produced by the LFI sensor on the human eye. Based on these features, we introduce an optimized and robust gaze symbol classification algorithm in Appendix B.1.5. Afterwards, we evaluate the proposed sensor concept in Appendix B.1.6 using a laboratory setup. In the last section, we conclude our findings and discuss further steps.

B.1.3   Related Work

Several eye-tracking sensor concepts have been investigated in the past. The successors of these different concepts, which are widely used in research and commercial applications, are video-oculography (VOG) and electro-oculography (EOG) [15]. In addition, novel low power eye-tracking sensor approaches for AR glasses based on microscanners and infrared (IR) lasers have emerged in recent years [101, 129].

Drews et al. used a VOG eye-tracking sensor with a 60 Hz sample rate to capture the absolute gaze position and extract gaze gestures from the input data stream [16]. They described a gaze gesture as a sequence of atomic eye movements. Atomic eye movements, also refereed to as strokes of the eye, are single unidirectional eye movements, e.g. left or upwards and can be interpreted as gaze symbols. A set of gaze symbols forms a gaze gesture protocol. Drews et al. used gaze symbols to control the user interface by linking a sequence of gaze symbols to a gaze gesture.The main advantage of their approach over dwell based interaction approaches, e.g. by [198], is that an absolute calibration of the eye-tracking sensor is not required because only relative eye positions are tracked.

The main disadvantage of VOG sensor based approaches is the high power consumption required by the sensors and the image processing, as well as the limited sample rate of the sensors [17]. Furthermore, they are sensitive to ambient light which disturbs the captured images leading to a low detection rate and therefore limiting the interaction capabilities in the wild [56].

To overcome this limitations Bulling et al. integrated EOG sensors into a wearable glasses demonstrator to measure relative movements of the eye and extract gaze gestures for mobile

human computer interactions [20]. They placed two electrode pairs for horizontal and vertical eye movements around the eye and used microcontroller to sample the four EOG channels and extract the relative eye movements. In addition, they used a light sensor and an accelerometer to compensate external artefacts in the sensor data. They reported a sample rate of 250 Hz with an overall power consumption of 769 mW.

They use a gaze gesture protocol with 16 symbols to encode the gaze gestures. The overall correct classification rate for different gaze gestures of varying complexity was given as 87 %.

Compared to VOG sensor concepts, EOG sensors benefit from computational light weight signal processing. In addition, they are more robust against ambient illumination. The major drawback limiting the use of EOG sensor for gaze-based interaction in recent smart glasses is the use of electrodes on the skin [15].

Our gaze gesture sensor concept based on LFI combines the benefits of non-intrusive integration of VOG sensors and the calibration free measurement of relative eye movements by EOG sensors. Furthermore, our sensor principle is robust against external illumination and is therefore capable of operating in the wild.

B.1.4 Gaze gesture sensor based on laser feedback interferometry
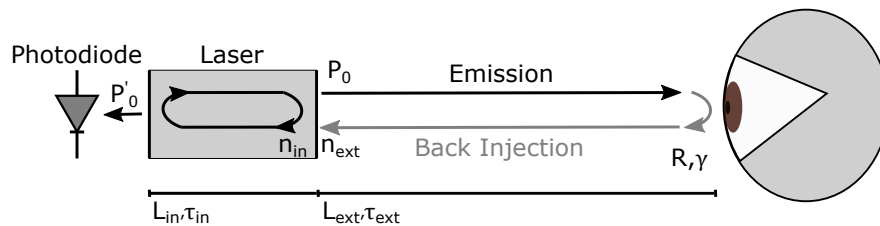
Figure B.1: Coupled cavity model of a laser feedback interferometry sensor. The laser emits light which is scattered by the eye and back injected into the laser cavity. The photodiode monitors the laser power, which varies with changes in the feedback path.

Laser feedback interferometry or self-mixing interferometry (SMI) is a widely known interferometry measurement method. It is used in the industry as well as in laboratory environments to measure displacement and velocity of solid targets, as well as fluids and distance. Due to the high distance and velocity resolution, it is also widely used in vibrometry applications [121].

Sensing principle

Figure B.1 shows the coupled cavity model to introduce the basic sensing principle of LFI sensors. A laser with an optical output power P_0 emits a coherent laser beam towards the surface of the eye. The laser beam hits the eye under an angle of incidence γ, is attenuated by volume scattering effects and absorption described by a reflectivity R and is back injected

into the laser. τ_{ext} denotes the time the laser beam requires to travel over the distance L_{ext} towards the eye. τ_{ext} is dependent on the speed of light c_0 and the external refraction index n_{ext} of the external medium [6].

The back injected light interferes with the local oscillating field, which is often referred to as self-mixing interference, resulting in a modulation of the optical power

$$P_0' = P_0 \left(1 + m \cdot \cos\left(\phi_{fb}\right)\right). \tag{B.1}$$

The varying feedback power P_0' is dependent on the optical power P_0, the modulation depth m and a varying phase ϕ_{fb} of the backscattered light field. A small fraction of the varying feedback power P_0' is measured by a photodiode, which is integrated into the back mirror of the laser cavity [6].

A more in-depth description of the coupled cavity model is given by the rate equations introduced by [119]. A solution of the rate equation is the excess phase equation

$$\phi_{fb} - \phi_s + C\sin\left(\phi_{fb} + \arctan\left(\alpha\right)\right) = 0. \tag{B.2}$$

The feedback phase is expressed as a function of the signal phase ϕ_s, Acket's feedback parameter C and Henry's line width enhancement factor α. Considering operation of the LFI sensor in the weak feedback regime ($C < 1$) and a constant α, Equation (B.2) has only a single solution and ϕ_{fb} is, therefore, only dependent on ϕ_s, leading to

$$\phi_s = \frac{4\pi n_{ext} L_{ext}}{\lambda} \tag{B.3}$$

with λ describing the wavelength of the laser. Considering $n_{ext} \approx 1$ is constant due to the operation of the sensor in free space, only changes in λ and L_{ext} lead to a varying phase ϕ_s and , consequently, to a varying phase ϕ_{fb}. This results, with respect to Equation (B.1), in a modulation of the optical power which is measured by the photodiode. Changes in the wavelength λ occur by a modulation of the laser drive current. That leads to a periodic heating and cooling of the resonator and, thus, to a periodic change of the cavity length. The variation of the cavity length leads to a periodic modulation of the wavelength, which allows for continuous measurement of the distance according to Equation (B.4).

To distinguish between both effects, we compute the partial derivative of Equation (B.3) with respect to time, which leads to

$$f_0 = \frac{2L_{ext}}{\lambda^2} \frac{d\lambda}{dI} \frac{dI}{dt}\bigg|_{L_{ext}=const.} \tag{B.4}$$

and

$$f_d = \frac{2v_{ext}\cos\left(\gamma\right)}{\lambda}\bigg|_{\lambda=const.} \tag{B.5}$$

133

Considering a known $\frac{d\lambda}{dI}$, which is a static process parameter of the laser, and a controlled current modulation $\frac{dI}{dt}$, the distance to the eye can be calculated by extracting the so called beat frequency f_0 by applying a fast Fourier transform (FFT) to the measured varying optical power and rearranging Equation (B.4) with respect to L_{ext}.

Movements of the eye ($\frac{dL_{ext}}{dt} = v_{ext}$) lead to a shift of the beat frequency f_0 by the so called Doppler frequency f_d. With a known angle of incidence γ and a measured Doppler frequency, Equation (B.5) can be rearranged with respect to v_{ext} to obtain the surface velocity of the eye.

In order to separate f_0 and f_d and, thus, simultaneously measure the distance and velocity of the eye, a triangular modulation similar to frequency modulated continuous wave (FMCW) radar is applied to the drive current of the laser [122]. By separating the up- and down ramp signals into two segments and applying an FFT on each segment, an f_{up} and an f_{down} frequency is captured. f_0 and f_d are obtained from these measurements by

$$f_0 = \frac{f_{up} + f_{down}}{2} \tag{B.6}$$

and

$$f_d = \frac{f_{up} - f_{down}}{2} \tag{B.7}$$

respectively.
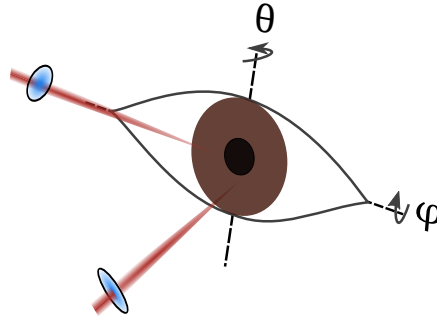
Figure B.2: Positioning of the LFI sensors with respect to the rotational axes of the human eye.

The LFI sensor measures surface velocity as a dot product between the laser normal and the surface velocity vector of the moving eye at the intersection point between laser and eyeball. If the eye's rotational axis and the laser beam are aligned, a simplified description by the angle of incidence γ as described in Equation (B.5) is possible. Therefore, two LFI sensors are required to measure both, the horizontal movement around the θ-axis and vertical movement around the ϕ-axis. Figure B.2 shows the positioning and the laser beam directions of two LFI sensors to comply with these requirements.

From an integration point of view, the sensor for vertical movements can be integrated into the AR glasse's frame below the spectacle and the sensor for horizontal movements can be integrated into the AR glasses frame temple. The size of a single LFI sensor is mainly

determined by the diameter of the lens in front of the laser chip which is roughly 4 mm. This results in a size of the final sensor that is comparable to the twin-eye sensor shown in [6].

Sensor features on the eye
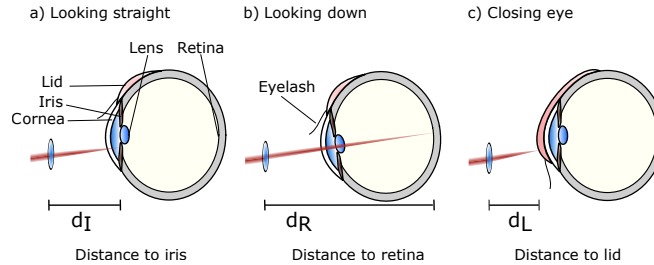
Figure B.3: Different movement related position changes of the eye and the lid resulting in a change in distance measurement due to the geometry and scaffold of the eye.

Based on Equation (B.4) and Equation (B.5) and the positioning of the LFI sensors shown in Figure B.2, four features are measured by the sensors. For each rotational axis of the eye, the surface velocity and the distance are measured resulting in v_θ and v_ϕ as velocity related features and d_θ and d_ϕ as position related features.

Figure B.3 shows a sectional view of the eye and a fixed LFI sensor position for three different positions of the eye and the lid. In Figure B.3 a), the gaze is directed straight ahead. The laser beam of the LFI sensor for vertical rotations penetrates the cornea and backscattering occurs at the iris. For this setup, the LFI sensor measures the distance between sensor and iris d_I.

In Figure B.3 b), the eye is slightly rotated along the vertical direction downwards and the laser beam penetrates the cornea and the lens and is backscattered from the retina. The measured distance in this case is the distance between sensor and retina d_R.

In Figure B.3 c), the eye is directed straight forward but the lid is closed by a blink. In this case, the distance d_L between sensor and lid is measured. By subtracting d_I and d_L, the thickness of the lid can be calculated, which is around 4 mm, and by subtracting d_R and d_L the diameter of the eye is approximately calculated, which is around 24 mm [61].

B.1.5 Gaze gesture algorithm

Similar to related works by [16], [20] and [199], the gaze gesture algorithm is based on a gaze gesture protocol which encodes atomic movements of the eye into gaze gestures denoted by symbols.

We define a gaze symbol set consists of four basic atomic eye movements which are annotated by small letters. For the horizontal axis, movements to the left l and to the right r around θ are possible. In addition, eye movements in the vertical axis around ϕ are denoted by u and

d for up and down movements of the eye, respectively. In addition to movements of the eye, blinks (*b*) are considered an additional atomic movement. In order to cover all types of eye movements, fixations of the eye, as well as slow movements, are described by an additional symbol *n* for non-movements.

The entire set of symbols *S* thus consists of six unique symbols $S = \{l, r, u, d, b, n\}$.

Feature extraction

The novel sensor concept described in Appendix B.1.4 allows for the extraction of a unique set of features from the human eye with a high sampling rate. A major advantage over the state of the art is the use of the distance measurement between the glasse's frame and the eye as an additional feature, as discussed in Appendix B.1.4.

To obtain unique features from the measured velocities, v_θ and v_ϕ are treated as velocity vector components. This velocity vector can be represented in polar coordinate space by an angle

$$\epsilon = \arctan(\frac{v_\theta}{v_\phi}) \tag{B.8}$$

and a vector length

$$v = \sqrt{v_\theta^2 + v_\phi^2} \tag{B.9}$$

which lead us to an two dimensional feature vector *F* with ϵ and v as gaze symbol features. Figure B.4 shows the feature space, which is spanned by ϵ and v, in polar coordinate space.
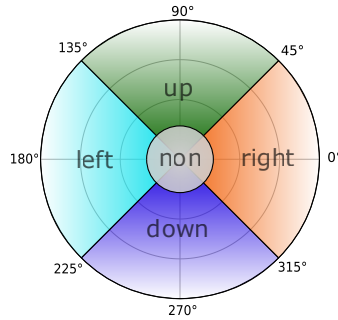
Figure B.4: Gaze gesture velocity feature space shown in a polar coordinate system.

The grey area in the centre covers the area of fixations and slow eye movements e.g. drift or tremor. The size is determined by a certain absolute velocity threshold v_n for v. The sensitivity and robustness of the gaze symbol algorithm can be controlled by varying v_n. The main advantage of this representation is the wide range of allowed angles of 90° per direction, leading to a robust gaze symbol detection and an easy execution of gaze gestures for the user.

To also be able to measure blinks d_θ is added as a third feature to *F*. The feature vector describes a point in cylindrical coordinate denoted by $F = \{\epsilon, v, d_\theta\}$.

Gaze symbol classification

The first step of the classification process is to map a feature vector F measured by the LFI sensors to a gaze symbol from the gaze symbol set S. For this purpose, a decision tree, which describes the feature space shown in Figure B.4 and the additional distance information d_θ, is used. The main advantage of this single sample classification approach over other classification approaches, e.g. by [20], is the invariance with respect to time. This allows for a robust classification, which is insensitive to sensor drift, as is the case with EOG sensors [15]. Furthermore, this classification approach does not require the detection of a movement sequence in the input sensor signal stream to extract significant movements which belong to a gaze symbol as proposed by [199].

B.1.6 Evaluation
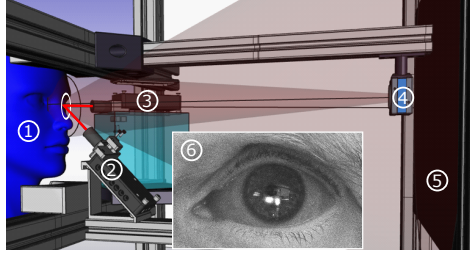
Figure B.5: CAD sketch of the laboratory setup to validate the proposed LFI gaze gesture sensor approach.

To validate the proposed features of the human eye measured by the LFI sensors and evaluate the gaze gesture recognition algorithm, a laboratory setup is used, as shown in Figure B.5.

The subject ① is placed in front of a monitor ⑤ which is used to create stimuli for eye movements. Its head is fixated by a head- and chin rest to suppress unintended head movements. The LFI sensor at position ② is responsible for measurements of v_ϕ and d_ϕ while the LFI sensor at position ③ is responsible for measurements of v_θ and d_θ. An IR camera ④ captures a video of the subject during the experiments. A sub Figure ⑥ shows a frame captured by the camera showing the focused laser spots of the two IR lasers on the iris of a subject as white spots.

Table B.1: Properties of the LFI sensors of the laboratory setup.

LFI Sensor	$d\lambda/dI$	λ	γ	P_0
②	$0.406\,\text{nm/mA}$	$848\,\text{nm}$	$45°$	$390\,\mu W$
③	$0.396\,\text{nm/mA}$	$856\,\text{nm}$	$45°$	$390\,\mu W$

With the low optical power of the IR VCSELs, the mechanical setup of chin and head rest and the positioning of the lasers in the laboratory setup, a class 1 laser system according to IEC 60825-1 (class 1 optical power limit $780\,\mu$W) is achieved and, therefore, the experiments do

not pose any medical hazard to the subject's eye.

The power consumption of the proposed gaze gesture sensor is roughly 140 mW, which is estimated based on an STM32G473 microcontroller including all required peripherals to capture and process the data from the LFI sensors. With optimized logic blocks and an subsampling scheme enabled by an higher integration using a custom application specific integrated circuit (ASIC) a further power reductions to 30 mW is expected.

A comparable power consumption of 150 mW is reported by [101] for state of the art VOG eye tracking sensors excluding image processing. Compared to [20] a much lower power consumption can be achieved.

Feature validation

To validate the proposed features of the human eye from Appendix B.1.4, the laboratory setup is used. Figure B.6 shows the measured data of the two LFI sensors from a subject performing the gaze symbol sequence $[l, r, u, d, b]$.
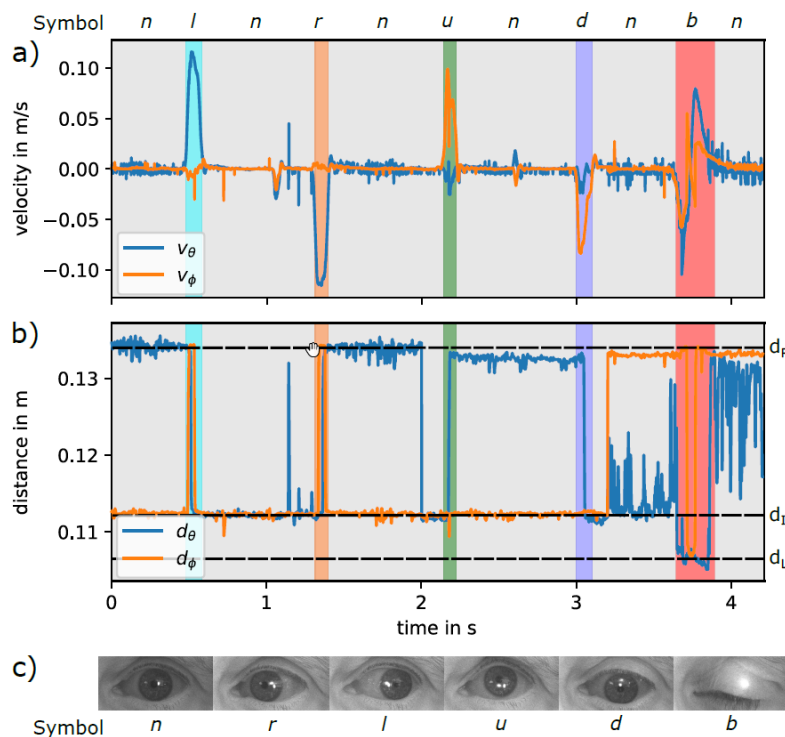
Figure B.6: Data captured with the laboratory setup to evaluate the proposed features of the eye. a) shows the measured velocity, b) shows the measured distance and c) shows eye movements captured with the IR camera and the corresponding gaze symbol. The background colour refers to the corresponding atomic movement of the eye.

The LFI sensors are modulated to achieve an update rate of 1 kHz for distance and velocity

138

measurement. Figure B.6 a) shows the measured velocity and Figure B.6 b) shows the measured distance. The movement threshold v_n is set to 0.02 m/s to distinguish between atomic eye movements and non-relevant movements of the eye. In the distance plot we added three distance lines corresponding to the distances described by Figure B.3 a) - c). The measured distance between d_R (retina) and d_I (iris) is 23,36 mm and the measured distance between d_I (iris) and d_L (lid) is 5,01 mm, which both correspond to the known anatomy of the human eye.

Gaze gesture symbol classification

To validate the proposed gaze gesture symbol classifier described in Appendix B.1.5, a set of gestures is recorded using the laboratory setup as described in Appendix B.1.6. Two male subjects with blue and brown eyes were instructed to perform 18 eye movement gestures and 9 blinks each. To capture natural trajectories of the eye, during the execution of the gaze gestures no visual stimuli were used to guide the gaze of the subjects.

In a first step, the ground truth is manually annotated utilizing the IR camera images to the measured data with the corresponding gaze symbols, similar to related work by [200]. To distinguish between movements and non-movements, the velocity threshold $v_n = 0.02\,ms$ is applied. Afterwards, we extract the features from the measured data according to Appendix B.1.5 and use the proposed gaze symbol classification algorithm to classify atomic eye movements. This allows us to evaluate the classification approach on the sample-level. In addition, we treat the multiclass classification problem for evaluation purpose as a binary one-vs-all classification problem. This evaluation approach is commonly used in the literature [201, 202] and allows us to compute the F1-score as evaluation metric. Table B.2 shows the evaluation results.

Table B.2: Properties of the LFI sensors of the laboratory setup.

Gaze symbol	Gestures	Symbols	Precision	Recall	F1-Score
l	36	3783	0.906	0.995	0.949
r	36	3505	0.916	0.997	0.955
u	36	3229	0.820	0.976	0.891
d	36	3624	0.873	0.945	0.908
b	18	3063	0.971	0.967	0.969

B.1.7 Conclusion

We present a novel low power gaze gesture sensor concept based on IR VCSELs and the LFI effect, which allows for seamless integration of gaze gesture sensors into next generation smart glasses. This sensor enables new gaze based interaction concepts like true hands-free interaction. We also introduced a computational lightweight gaze symbol classification algorithm with a sufficient classification accuracy and precision.

Based on the promising results obtained with the laboratory setup, our next goal is to integrate the sensors into our head-mounted demonstrator. This will allow us to evaluate the gaze

gesture sensor during everyday activities to investigate the robustness against glasses slippage.

Furthermore we want to develop a gaze gesture algorithm which recognizes a user interaction with the glasses by detecting a sequence of gaze symbols and investigate the robustness against unwanted interactions during everyday activities.

B.2 A compact low-power gaze gesture sensor based on laser feedback interferometry for smart glasses (invited)

B.2.1 Abstract

The integration of gaze gesture sensors in next-generation smart glasses will improve usability and enable new interaction concepts. However, consumer smart glasses place additional requirements to gaze gesture sensors, such as a low power consumption, high integration capability and robustness to ambient illumination. We propose a novel gaze gesture sensor based on laser feedback interferometry (LFI), which is capable to measure the rotational velocity of the eye as well as the sensor's distance towards the eye. This sensor delivers a unique and novel set of features with an outstanding sample rate allowing to not only predict a gaze gesture but also to anticipate it. To take full advantage of the unique sensor features and the high sampling rate, we propose additionally a novel gaze gesture classification algorithm based on single sample. At a mean F1-score of 93.44 %, our algorithms shows exceptional classification performance at a negative latency between gaze gesture input and command execution.

B.2.2 Introduction

Consumer smart glasses evolve their full potential, if the user has intuitive means to interact with and control the device. In the past a variety of interaction concepts are brought into the market, which can be classified into touch based interactions e.g. via a track pad on the glasse's temple introduced by the Google Glass [196], an external on-body device e.g. a controller like as in the Magic Leap One [197] or via touchless interactions such as spoken commands via voice recognition as used by the Echo Frames [132]. Also, the use of hand and finger gestures captured by a camera sensor have been shown by the HoloLens [17].

An additional touchless interaction concept is making use of the user's gaze by tracking their eye movements by means of eye-tracking sensors. Gaze-based interaction allows a fast and natural input, leading to an intuitive and unobtrusive way of interaction with smart glasses while maintaining social acceptance and user's privacy [17].

Gaze-based interaction allow further to adaptive control and activate the image projection system based on the users interaction, which allows to extend the battery life of lightweight augmented reality (AR) glasses to all-day operation.

To cover both aspects, user interaction and system control, low-power always-on gaze gesture sensors are well suited.

To the best of our knowledge, no commercially available smart glasses solution so far utilizes gaze gestures as input modality, mainly due to the challenges arising from integration and due to limited available power and space constraints [17]. Furthermore, a exceptional robustness

of the sensor against variable ambient illumination is required to allow for operation in uncontrolled outdoor environments e.g. in bright sun light [56].

To fulfill these requirements and enable gaze gesture based interaction for the next generation of smart glasses, we present a novel low power gaze gesture sensor based on laser feedback interferometry (LFI) and extend our previous work [5] by an gaze gesture detection algorithm. This multimodal sensor is capable of measuring distance towards the eye as well as eye rotational velocity with an outstanding sample rate of up to 1 kHz. The sensing component consits of a tiny infra red (IR) vertical cavity surface emitting laser (VCSEL) with an integrated photodiode. Due to its coherent sensing scheme, the sensor is only sensitive to its own emitted radiation, allowing a robust operation in presence of ambient radiation, as shown by Meyer et.al. [128]. Another advantage of this sensing technology is the omission of an imaging process for gaze sensing. This enables unobtrusive, socially acceptable sensor integration and reduces privacy concerns as no images of the eyes are captured.

In the next Section, we give an overview of the state of the art regarding gaze gesture sensors as well as gaze interaction concepts for AR glasses and discuss the limitations of existing concepts. In Appendix B.2.4, we introduce the sensing principle and the sensor concept of the proposed LFI sensor. In addition, we provide an overview of the measurement features captured by the LFI sensor on the human eye. Based on these features, we introduce an optimized and robust gaze gesture recognition algorithm in Appendix B.2.4. Afterwards, we evaluate the proposed sensor concept in Appendix B.2.5 using a laboratory setup. In the last section, we conclude our findings and discuss further steps.

B.2.3   Related Work

Several eye-tracking sensor concepts have been investigated in the past. The successors of these different concepts, which are widely used in research and commercial applications, are video-oculography (VOG) and electro-oculography (EOG) [15]. In addition, novel low power eye-tracking sensor approaches for AR glasses based on microscanners and infrared (IR) lasers have emerged in recent years [101, 129].

Most gaze interaction concepts relie on VOG. VOG concepts for smart glasses rely on video cameras, which are mainly integrated into the glasses frame. They record the movements of the eye and track the position of the pupil and extract the gaze vector by using computer vision algorithms.

Bednarik et.al. introduce a gaze interaction method by interpreting the gaze vector as mouse courser to interact with UI elements [198]. By fixating and dwelling on an UI element, the UI element is activated and an interaction is performed. The dwell time is required to differentiate between explicit interactions and random eye movements and thus solve the so-called Midas touch problem. The main disadvantage of this approach is that, due to the required dwell time, only a few interactions can be performed in a defined period of time. Furthermore, an

absolute calibration of the eye tracker is necessary to obtain the absolute gaze position.

Drews et al. used a VOG eye-tracking sensor to capture the absolute gaze position and extract gaze gestures from the input data stream [16]. They described a gaze gesture as a sequence of atomic eye movements. Atomic eye movements, also refereed to as strokes of the eye, are single unidirectional eye movements, e.g. left or upwards and can be interpreted as gaze symbols. A set of gaze symbols forms a gaze gesture protocol. Drews et al. used gaze symbols to control the user interface by linking a sequence of gaze symbols to a gaze gesture. In addition, they used a timeout of 1000 ms to distinguish between gesture inputs and natural eye movements. The main advantage of their approach over dwell based interaction approaches is that an absolute calibration of the eye-tracking sensor is not required because only relative eye positions are tracked.

In recent years gaze based pursuit interactions are investigated [18]. Pursuit interaction exploit pursuit eye movements by displaying moving objects, which are pursued by the human gaze. [18] combined the gaze vector and optical flow of a series of images to obtain pursuit informations for targets shown on a display. Similar to the gaze gesture approach, they only require relative eye movements, but in order to obtain the velocities for detection of pursuit eye movement , the optical flow needs to be calculated from a series of incoming images, resulting in high computational complexity.

The main disadvantage of VOG sensor based approaches is the high power consumption required by the sensors and the image processing, as well as the limited sample rate of the sensors [17]. Furthermore, they are sensitive to ambient light which disturbs the captured images leading to a low detection rate and therefore limiting the interaction capabilities in the wild [56].

Our LFI-based gaze gesture sensor concept combines the advantages of non-intrusive integration of VOG systems with the advantage of directly capturing the relative velocity of the eye with an outstanding sampling rate and a low power consumption. Due to the fact that the LFI sensor captures relative eye movements, it is ideally suited for a gaze gesture-based interaction concept similar to Drews et.al.. Furthermore, our sensor principle is robust against external illumination and is therefore capable of operating in the wild.

B.2.4   Gaze gesture sensor based on laser feedback interferometry
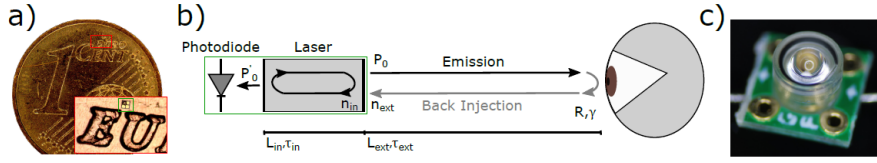
Figure B.7: a) 160 μm x 180 μm VCSEL with integrated photodiode on a cent coin b) Coupled cavity model of a laser feedback interferometry sensor. The laser emits light which is scattered by the eye and back injected into the laser cavity. The photodiode monitors the laser power, which varies with changes in the feedback path. c) Encapsulated optical module including the LFI sensor and the beam shaping optics.

Laser feedback interferometry or self-mixing interferometry (SMI) is a widely known interferometry measurement method. It is used in the industry as well as in laboratory environments to measure displacement and velocity of solid targets, as well as fluids and distance. Due to the high distance and velocity resolution, it is also widely used in vibrometry applications [121].

Sensing principle

Figure B.7 b) shows the coupled cavity model to introduce the basic sensing principle of LFI sensors. A laser with an optical output power P_0 emits a coherent laser beam towards the surface of the eye. The laser beam hits the eye under an angle of incidence γ, is attenuated by volume scattering effects and absorption described by a reflectivity R and is back injected into the laser. τ_{ext} denotes the time the laser beam requires to travel over the distance L_{ext} towards the eye. τ_{ext} is dependent on the speed of light c_0 and the external refraction index n_{ext} of the external medium [6].

The back injected light interferes with the local oscillating field, which is often referred to as self-mixing interference, resulting in a modulation of the optical power

$$P_0' = P_0 \left(1 + m \cdot \cos\left(\phi_{fb}\right)\right). \tag{B.10}$$

The varying feedback power P_0' is dependent on the optical power P_0, the modulation depth m and a varying phase ϕ_{fb} of the backscattered light field. A small fraction of the varying feedback power P_0' is measured by a photodiode, which is integrated into the distributed Bragg reflector (DBR) of the laser cavity itself [6].

To understand the link of the phase ϕ_{fb} to our observation goal (distance and velocity), we consider the rate equations introduced by [119]. A solution of the rate equation is the excess phase equation

$$\phi_{fb} - \phi_s + C \sin\left(\phi_{fb} + \arctan\left(\alpha\right)\right) = 0. \tag{B.11}$$

The feedback phase is expressed as a function of the signal phase ϕ_s, Acket's feedback parameter C and Henry's line width enhancement factor α. Considering operation of the LFI

sensor in the weak feedback regime ($C < 1$) and a constant line width enhancement factor α, Equation (B.11) lead to a single solution [6] and ϕ_{fb} is, therefore, only dependent on ϕ_s, leading to

$$\phi_s = \frac{4\pi n_{ext} L_{ext}}{\lambda} \tag{B.12}$$

with λ describing the wavelength of the laser. Considering $n_{ext} \approx 1$ is constant due to the operation of the sensor in free space, only changes in λ and L_{ext} lead to a varying phase ϕ_s and , consequently, to a varying phase ϕ_{fb}. This results, with respect to Equation (B.10), in a modulation of the optical power which is measured by the photodiode. Changes in the wavelength λ occur by a modulation of the laser drive current. That leads to a periodic heating and cooling of the resonator and, thus, to a periodic change of the cavity length. The variation of the cavity length leads to a periodic modulation of the wavelength, which allows for continuous measurement of the distance according to Equation (B.13).

To distinguish between both effects, we compute the partial derivative of Equation (B.12) with respect to time, which leads to

$$f_0 = \frac{2L_{ext}}{\lambda^2} \frac{d\lambda}{dI} \frac{dI}{dt}\bigg|_{L_{ext}=const.} \tag{B.13}$$

and

$$f_d = \frac{2v_{ext} cos(\gamma)}{\lambda}\bigg|_{\lambda=const.} \tag{B.14}$$

Considering a known $\frac{d\lambda}{dI}$, which is a static process parameter of the laser, and a controlled current modulation slope $\frac{dI}{dt}$, the distance to the eye can be calculated by extracting the so called beat frequency f_0 by applying a fast Fourier transform (FFT) to the measured varying optical power and rearranging Equation (B.13) with respect to L_{ext}.

Movements of the eye ($\frac{dL_{ext}}{dt} = v_{ext}$) lead to a shift of the beat frequency f_0 by the so called Doppler frequency f_d. With a known angle of incidence γ and a measured Doppler frequency, Equation (B.14) can be rearranged with respect to v_{ext} to obtain the surface velocity of the eye.

In order to separate f_0 and f_d and, thus, simultaneously measure the distance and velocity of the eye, a triangular modulation similar to frequency modulated continuous wave (FMCW) radar is applied to the drive current of the laser [122]. By separating the up- and down ramp signals into two segments and applying an FFT on each segment, an f_{up} and an f_{down} frequency is captured. f_0 and f_d are obtained from these measurements by

$$f_0 = \frac{f_{up} + f_{down}}{2} \tag{B.15}$$

and

$$f_d = \frac{f_{up} - f_{down}}{2} \tag{B.16}$$

respectively. Recalling Equation (B.16), the triangle modulation allows to extract the direction

of the velocity as well. The modulation frequency of the triangle signal therefore corresponds to the sample frequency of the LFI sensor. Under this consideration, the upper sample frequency of the LFI sensor is determined by thermal time constants of the laser cavity, which is for spacial confined semiconductor VCSELs in the range of 100 kHz [117].

Sensor concept

Figure B.8 shows a system diagram of the whole LFI sensor. The output of the photodiode is AC coupled to an transimpedance amplifier (TIA) to reject the DC component of the varying optical output power and only measure the cosine part of Equation (B.10).

Inside the digital logic the output of the TIA is sampled by an analogue digital converter (ADC). The sampled data is synchronized with the up- and down ramp of the triangle signal synthesized by an internal digital to analogue converter (DAC) and fed to the laser driver. Afterwards, an FFT is applied to the two signal segments to obtain f_{up} and f_{down} and use Equation (B.15) and Equation (B.16) to calculate f_0 and f_d, respectively.

The LFI sensor measures surface velocity as a dot product between the laser normal and the surface velocity vector of the moving eye at the intersection point between laser and eyeball. If the eye's rotational axis and the laser beam are aligned, a simplified description by the angle of incidence γ as described in Equation (B.14) is possible. Therefore, two LFI sensors are required to measure both, the horizontal movement around the θ-axis and vertical movement around the ϕ-axis. Figure B.9 shows the positioning and the laser beam directions of two LFI sensors to comply with these requirements.
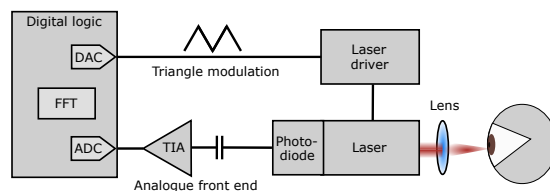
Figure B.8: System diagram of the whole LFI sensor including the sensing element, the modulation circuit and the data processing element [6].
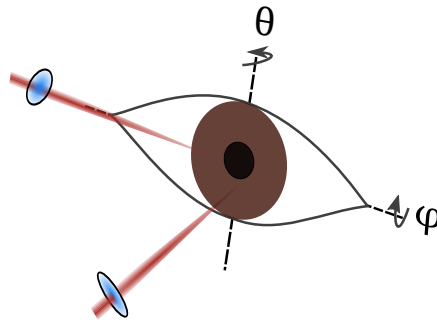
Figure B.9: Positioning of the LFI sensors with respect to the rotational axes of the human eye.

From an integration point of view, the sensor for vertical movements can be integrated into the AR glasse's frame below the spectacle and the sensor for horizontal movements can be integrated into the AR glasses frame temple. The size of a single LFI sensor is mainly determined by the diameter of the lens in front of the laser chip, shown in Figure B.7 c) is in the range of 1 mm - 2 mm.
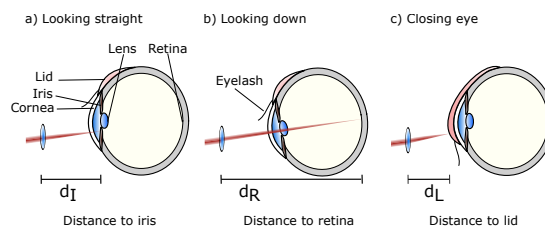
Figure B.10: Different movement related position changes of the eye and the lid resulting in a change in distance measurement due to the geometry and scaffold of the eye.

Sensor features on the eye

Based on Equation (B.13) and Equation (B.14) and the positioning of the LFI sensors shown in Figure B.9, four features are measured by the sensors. For each rotational axis of the eye, the surface velocity and the distance are measured resulting in v_θ and v_ϕ as velocity related features and d_θ and d_ϕ as position related features.

Figure B.10 shows a sectional view of the eye and a fixed LFI sensor position for three different positions of the eye and the lid. In Figure B.10 a), the gaze is directed straight ahead. The laser beam of the LFI sensor for vertical rotations penetrates the cornea and backscattering occurs at the iris. For this setup, the LFI sensor measures the distance between sensor and iris d_I.

In Figure B.10 b), the eye is slightly rotated along the vertical direction downwards and the laser beam penetrates the cornea and the lens and is backscattered from the retina. The measured distance in this case is the distance between sensor and retina d_R.

In Figure B.10 c), the eye is directed straight forward but the lid is closed by a blink. In this case, the distance d_L between sensor and lid is measured. By subtracting d_I and d_L, the thickness of the lid can be calculated, which is around 4 mm, and by subtracting d_R and d_L the diameter of the eye is approximately calculated, which is around 24 mm [61].

Gaze gesture algorithm

Similar to related works by Drews et. al.[16] Bulling et. al. [20] and Findlin et.al. [199], the gaze gesture algorithm is based on a gaze gesture protocol which encodes atomic movements of the eye into gaze gestures denoted by symbols.

Figure B.11 a) shows a graphical description of our gaze gesture protocol. We define four basic atomic eye movements which are annotated by small letters. For the horizontal axis, movements to the left l and to the right r around θ are possible. In addition, eye movements in the vertical axis around ϕ are denoted by u and d for up and down movements of the eye, respectively. In addition to movements of the eye, blinks (b) are considered an additional atomic movement. In order to cover all types of eye movements, fixations of the eye, as well as slow movements, are described by an additional symbol n for non-movements. The entire set of symbols S thus consists of six unique symbols $S = \{l, r, u, d, b, n\}$.

Feature extraction

The novel sensor concept described in Appendix B.2.4 allows for the extraction of a unique set of features from the human eye with a high sampling rate. A major advantage over the state of the art is the use of the distance measurement between the glasse's frame and the eye as an additional feature, as discussed in Appendix B.2.4.
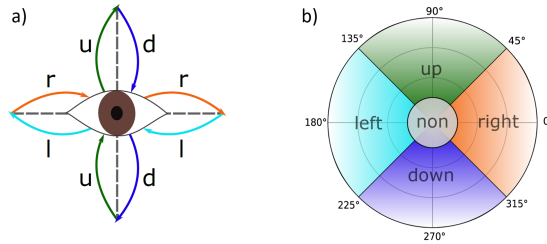
Figure B.11: a) Graphical description of the proposed gaze gesture alphabet. The small letters denote atomic unidirectional movements of the eye. b) Gaze gesture velocity feature space shown in a polar coordinate system.

To obtain unique features from the measured velocities, v_θ and v_ϕ are treated as velocity vector components. This velocity vector can be represented in polar coordinate space by an angle

$$\epsilon = \arctan(\frac{v_\theta}{v_\phi}) \tag{B.17}$$

and a vector length

$$v = \sqrt{v_\theta^2 + v_\phi^2} \tag{B.18}$$

which lead us to two dimensional feature vector F with ϵ and v as gaze symbol features. Figure B.11 b) shows the feature space, which is spanned by ϵ and v, in polar coordinate space.

The grey area in the centre covers the area of fixations and slow eye movements e.g. drift or tremor. The size of this area is independent on the angle ϵ and is, therefore, only determined by a certain absolute velocity threshold v_n for v. The sensitivity and robustness of the gaze gesture algorithm can be controlled by varying v_n. The other areas in the feature space are each associated with exactly one symbol of the symbol set S.

The main advantage of this representation is the wide range of allowed angles of 90° per direction, leading to a robust gaze symbol detection and an easy execution of gaze gestures for the user. In addition, the use of speed as a feature for extracting gaze symbols allows atomic eye movements to be performed individually without exceeding a certain angle of eye's rotation.

Taking Equation (B.14) into account, it appears that the LFI sensors are capable of measuring any kind of velocity at the eye. This includes blinking, which would be misclassified using the approach previously described as an opening or closing movement of the eye due to the up and down movement introduced by the moving eyelid. To avoid this misclassification, the distance measurement d_θ is evaluated for differentiation, as already described in Appendix B.2.4. By adding the distance measurement d_θ as a third feature to F, the feature vector describes a point in cylindrical coordinate denoted by $F = \{\epsilon, v, d_\theta\}$.

Gaze symbol classification

The first step of the classification process is to map a feature vector F measured by the LFI sensors to a gaze symbol from the gaze symbol set S. For this purpose, a decision tree, which describes the feature space shown in Figure B.11 b) and the additional distance information d_θ, is used. The main advantage of this single sample classification approach over other classification approaches, e.g. by Bulling et. al. [20], is the invariance with respect to time. This allows for a robust classification, which is insensitive to sensor drift, as is the case with EOG sensors [15]. Furthermore, this classification approach does not require the detection of a movement sequence in the input sensor signal stream to extract significant movements which belong to a gaze symbol as proposed by Findling et. al. [199].

Gaze symbol preprocessing

One drawback of the single sample classification approach is that even slight movements of the eye in the wrong direction lead to a different gaze symbol. To overcome this limitation and increase robustness of the gaze gesture recognition, the gaze symbol stream is preprocessed.

Figure B.12 shows the preprocessing of the gaze symbol stream.
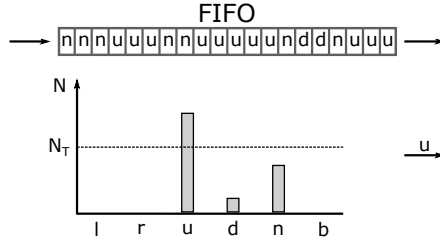
Figure B.12: Gaze gesture symbols accumulated in a FIFO memory.

A first in first out (FIFO) buffer with length N is filled with incoming gaze symbols. In addition, for every symbol in S, a corresponding bin is created. These bins contain the number of associated symbols from the FIFO. If the number of symbols in a bin exceeds a certain threshold N_T, the corresponding gaze gesture symbol is set as an active symbol until N_T is undershot. The FIFO is initialized at beginning of the algorithm with n as default symbol to always output an active symbol.

Gaze gesture recognition
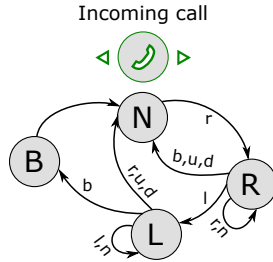
Figure B.13: Description of a gaze gesture using a FSM to accept a incoming call displayed in the AR glasses user interface.

Similar to Drews. et. al. [16], we define a gaze gesture as a sequence of atomic movements of the eye described by the corresponding gaze gesture symbols. To describe a gaze gesture, a finite state machine (FSM) is used. This allows for a flexible design of custom gaze gestures. To manipulate the state of the FSM, the preprocessed gaze symbol stream is used.

Figure B.13 shows an FSM with a gaze symbol sequence $[r, l, b]$ as gaze gesture to accept a incoming call. When the state R is reached, a timer is started with a given time t_0 to reset the FSM to its initial state N if a timeout is reported before reaching the final state B. This increases the robustness and reduces the sensitivity to unintended interactions. Furthermore, a higher robustness of gaze gesture recognition can be achieved by increasing the length of the gaze symbol sequence of a gaze gesture.

The transition to the final state B is issued as soon as the blink gaze symbol b is detected in the

preprocessed gaze symbol stream, which occurs as soon as enough b symbols have entered the FIFO and N_T is exceeded. This allows for the execution of the gaze gesture command before the blink is even finished, resulting in a negative latency between finishing the gesture and execution of the command.

Due to the fact that no sliding window or time and power consuming transformation of time series data, as proposed by Bulling et. al. [19], is used to recognize the gaze gesture, a low power gaze gesture recognition for smart glasses is achieved.

B.2.5   Evaluation
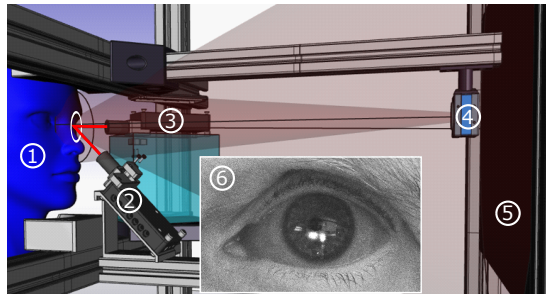
Figure B.14: CAD sketch of the laboratory setup to validate the proposed LFI gaze gesture sensor approach. The grey scale image shows the eye of a subject with the two laser spots on the iris captured by the reference camera.

To validate the proposed features of the human eye measured by the LFI sensors and evaluate the gaze gesture recognition algorithm, a laboratory setup is used, as shown in Figure B.14.

The subject ① is placed in front of a monitor ⑤ which is used to create stimuli for eye movements. Its head is fixated by a head- and chin rest to suppress head movements, which would lead to a faulty velocity measurement. The LFI sensor at position ② is responsible for measurements of v_ϕ and d_ϕ while the LFI sensor at position ③ is responsible for measurements of v_θ and d_θ. An IR camera ④ captures a video of the subject during the experiments. A sub Figure ⑥ shows a frame captured by the camera showing the focused laser spots of the two IR lasers on the iris of a subject as white spots. The lasers are not aligned to hit the eye at the same position, as already described in Appendix B.2.4.

Table B.3: Properties of the LFI sensors of the laboratory setup.

LFI Sensor	$d\lambda/dI$	λ	γ	P_0
②	0.406 nm/mA	848 nm	45°	390 μW
③	0.396 nm/mA	856 nm	45°	390 μW

Considering the sensor system diagram shown in Figure B.8, the laser together with the laser driver and the lens are integrated into the black housing shown in Figure B.14. The triangular modulation signals are synthesized by an external waveform generator and the photodiode

signal of the sensors is sampled by an oscilloscope together with the synthesized triangular modulation signal to extract f_{up} and f_{down}, which are required to calculate f_d and f_0 by Equation (B.15) and Equation (B.16), respectively. Afterwards the distance- and the velocity are calculated by Equation (B.13) and Equation (B.14) with the known laser parameters shown in Table B.3.

With the low optical power of the IR VCSELs, the mechanical setup of chin and head rest and the positioning of the lasers in the laboratory setup, a class 1 laser system according to IEC 60825-1 is achieved and, therefore, the experiments do not pose any medical hazard to the subject's eye.

The power consumption of the proposed gaze gesture sensor is roughly 140 mW using of the shelf components. With optimized logic blocks and a subsampling scheme enabled by a higher integration using a custom application specific integrated circuit (ASIC), a further power reductions to 30 mW is expected. A comparable power consumption of 150 mW is reported by Sarkar et.al. [101] for state of the art VOG eye tracking sensors, however excluding image processing. Compared to the EOG approach of Bulling et al. [20], a much lower power consumption can be achieved.

Feature validation

To validate the proposed features of the human eye from Appendix B.2.4, the laboratory setup is used. Figure B.15 shows the measured data of the two LFI sensors from a subject performing the gaze symbol sequence $[l, r, u, d, b]$.
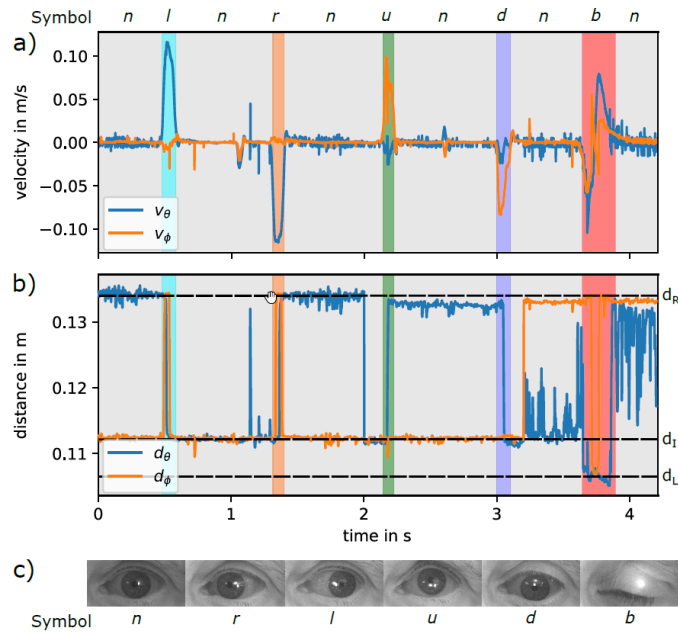
Figure B.15: Data captured with the laboratory setup to evaluate the proposed features of the eye. a) shows the measured velocity, b) shows the measured distance and c) shows eye movements captured with the IR camera and the corresponding gaze symbol. The background colour refers to the corresponding atomic movement of the eye.

The LFI sensors are modulated to achieve an update rate of 1 kHz for distance and velocity measurement. Figure B.15 a) shows the measured velocity and Figure B.15 b) shows the measured distance. The movement threshold v_n is set to 0.02 m/s to distinguish between atomic eye movements and non-relevant movements of the eye. In the distance plot we added three distance lines corresponding to the distances described by Figure B.10 a) - c). The measured distance between d_R (retina) and d_I (iris) is 23,36 mm and the measured distance between d_I (iris) and d_L (lid) is 5,01 mm, which both correspond to the known anatomy of the human eye. The distance measurement shows that the iris of the eye scatters IR light and the cornea and the lens of the eye let IR light pass through without significant reflection and backscattering only occurs on the retina.

In Figure B.15 c), the images captured by the IR reference camera show the laser spots on the subject's eye for different atomic eye movements described by the corresponding gaze symbol. Similar, the background colours of the plots in a) and b) refer to the corresponding atomic eye movements.

Gaze gesture symbol classification

To validate the proposed gaze gesture symbol classifier described in Appendix B.2.4, a set of gestures is recorded using the laboratory setup as described in Appendix B.2.5. Two male subjects with blue and brown eyes were instructed to perform 18 eye movement gestures

and 9 blinks each. To capture natural trajectories of the eye, during the execution of the gaze gestures no visual stimuli were used to guide the gaze of the subjects.

In a first step, the ground truth is manually annotated utilizing the IR camera images to the measured data with the corresponding gaze symbols, similar to related work by Santini et. al. [200]. To distinguish between movements and non-movements, the velocity threshold $v_n = 0.02\,ms$ is applied. Considering Emsley's reduced eye model (reye = 11.11mm) this equals to a velocity threshold of 103.14 °/s, close to the suggested velocity threshold used in the velocity threshold Identification algorithm (I-VT). Afterwards, we extract the features from the measured data according to Appendix B.2.4 and use the proposed gaze symbol classification algorithm to classify atomic eye movements. This allows us to evaluate the classification approach on the sample-level. In addition, we treat the multiclass classification problem for evaluation purpose as a binary one-vs-all classification problem. This evaluation approach is commonly used in the literature e.g. Startsev et.al. [201] or, Hoppe et. al. [202] and allows us to compute the F1-score as evaluation metric. Table B.4 shows the evaluation results.

Table B.4: Results of the gaze gesture symbol classification algorithm.

Gaze symbol	Gestures	Samples	Mean duration	F1-Score
l	36	3783	105.08 ms	0.949
r	36	3505	97.36 ms	0.955
u	36	3229	89.69 ms	0.891
d	36	3624	100.67 ms	0.908
b	18	3063	170.17 ms	0.969

The overall high F1-scores, shows high precision and recall capability of the proposed algorithm and underlines the robustness of the features acquired with our novel LFI gaze gesture sensor.
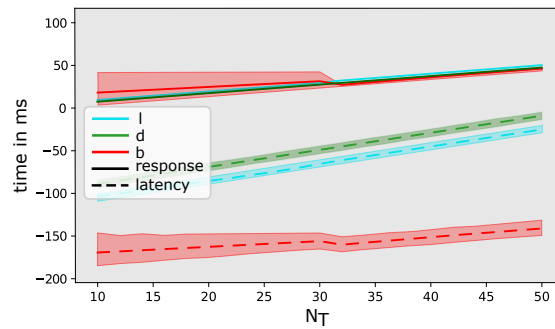
Figure B.16: Response time t_{res} (solid line) and latency t_{lat} (dashed line) for the left-, up- and blink atomic eye movements evaluated for different thresholds N_T. The light shaded regions around the lines represent the standard deviation of all atomic eye movements captured during the experimental validation.

Gaze symbol preprocessing validation

To increase the robustness of the recognition algorithm, the gaze symbols are preprocessed according to the proposed method in Appendix B.2.4. Based on the results shown in Table B.4, a FIFO size N of 100 elements is used, which represents an 100 ms window. This preprocessing smooths the incoming gaze symbol stream to increase the robustness, but also introduces a response and latency time to the gaze gesture symbol output, which is used to control the FSM. We define the response time t_{res} as time between the start of a gesture (v exceeds v_n) and the first indication of a gaze gesture symbol in the output symbol stream (N_T is exceeded). In addition the latency, t_{lat} is defined as the time between the end of a gesture (v falls below v_n) and the first indication of a gaze gesture symbol in the output symbol stream (N_T is exceeded). Figure B.16 shows the latency and response time for different thresholds N_T. A part of the gestures described in Table B.4 is used for the evaluation. The solid and dashed lines represent the mean response time and latency, respectively, while the light shaded regions around the lines represent the standard deviation.

Figure B.16 shows another significant advantage of the proposed gaze gesture algorithm over state of the art approaches. Beside the low response time, we achieve a negative latency, meaning that the proposed algorithm is able to recognize a single gaze gesture before the user has finished its execution. This allows for a seamless interaction with the smart glasses UI because the integrated control unit reacts before the user finishes the gesture input.

Gaze gesture recognition validation

Based on the example in Appendix B.2.4, the proposed FSM based gaze gesture recognition is validated. A subject is reading an text until a incoming call appears. When the incoming call appears, the subject performs the gaze gesture as described in Figure B.13. Figure B.17 shows the measured velocities v_θ and v_ϕ as solid and dashed lines respectively. Every data point is classified, preprocessed and fed into the FSM. The corresponding state of the FSM is annotated to the velocity trajectories. The black dashed vertical line indicates the time at which the gesture is successfully recognized. Similar to Drews et.al. [16], we set the timeout t_0 to one second.

This example shows the robustness of the detection and preprocessing algorithm. Furthermore, the negative latency between gesture recognition and the completion of the gesture is illustrated. The blink gesture and, thus, the execution of the user interaction already starts while the eyelid is closing, which is described by the negative speed of v_θ and v_ϕ. Especially during reading activities, the eye primarily makes small movements to the right when scanning over a line of text and larger movements to the left when jumping back to the beginning of a new line. In this experiment, the FSM reaches the state L at the 2 seconds mark on the time axis. If a blink occurs at this position, a false gesture input is executed.

To address this issue and make this approach more robust against unintended activations, an
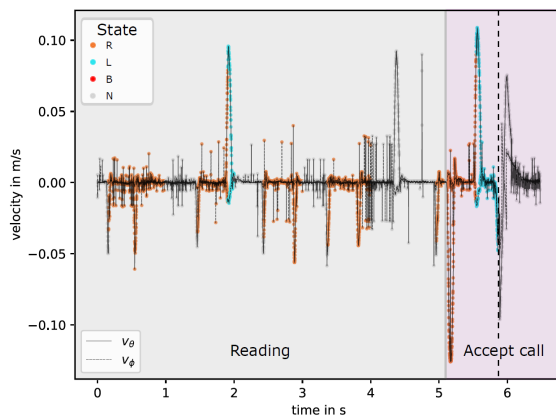
Figure B.17: Application of the proposed gaze gesture recognition approach to an example user interaction. The colourized dots on the velocity trajectory denotes the corresponding state of the FSM described in Figure B.13.

incoming call can be classified as an event-based interaction and the FSM for user interaction detection is armed only when the event occurs. Furthermore, the time-out t_0 can be decreased to reset the FSM earlier or a more complex gesture consisting of a longer sequence of atomic eye movements can be used.

B.2.6   Conclusion

We present a novel low power gaze gesture sensor concept based on IR VCSELs and the LFI effect, which allows for seamless integration of gaze gesture sensors into next generation smart glasses. This sensor enables new gaze based interaction concepts like true hands-free interaction. In addition, the high sampling rate of the sensor, combined with our gaze gesture recognition algorithm, results in negative latency and ,thus ,ensures a fast response time and a high user experience.

Based on the promising results obtained with the laboratory setup, our next goal is to integrate the sensors into our head-mounted demonstrator. This will allow us to evaluate the gaze gesture sensor during everyday activities to minimize unwanted activations and investigate the robustness against glasses slippage.

In addition the proposed LFI sensor opens options for further applications like fatigue detection based on blink speed and duration, recognition of user's emotions, classification ofthe user's mental health as well as early detection of eye diseases.

B.3 A CNN-based Human Activity Recognition System Combining a Laser Feedback Interferometry Eye Movement Sensor and an IMU for Context-aware Smart Glasses

B.3.1 Abstract

Smart glasses are considered the next breakthrough in wearables. As the successor of smart watches and smart ear wear, they promise to extend reality by immersive embedding of content in the user's field of view. While advancements in display technology seems to fulfill this promises, interaction concepts are derived from established wearable concepts like touch interaction or voice interaction, preventing full immersion as they require the user to frequently interact with the glasses. To minimize interactions, we propose to add context-awareness to smart glasses through human activity recognition (HAR) by combining head- and eye movement features to recognize a wide range of activities. To measure eye movements in unobtrusive way, we propose laser feedback interferometry (LFI) sensors. These tiny low power sensors are highly robust to ambient light. We combine LFI sensors and an IMU to collect eye and head movement features from 15 participants performing 7 cognitive and physical activities, leading to a unique data set. To recognize activities we propose a 1D-CNN model and apply transfer learning to personalize the classification, leading to an outstanding macro-F1 score of 88.15 % which outperforms state of the art methods. Finally, we discuss the applicability of the proposed system in a smart glasses setup.

B.3.2 Introduction

Followed by smart watches and smart ear wear smart glasses are considered as the next break-through in the smart wearable domain because they are able to enhance human perception and embedding user interfaces seamlessly into the user's field of view (FOV). In certain situations, e.g. when driving, this highly integrated display causes distraction to the user as information e.g. notifications appear in the FOV. State of the art smart glasses try to adapt established user interface control concepts from the smartphone e.g. by adding tactile input modalities like capacitive sliders, buttons or touch sensors to the glasses frame [203], [204] or voice based control [132]. Furthermore, human gaze is investigated as a natural input modality for controlling user interfaces with smart glasses [5],[16].

All of these concepts require active interaction with the glasses. Furthermore, for voice based interaction the user's privacy is not covered, what might reduce the social acceptance of smart glasses [17]. To address this drawback, we propose a pervasive computing approach to control and adapt the user interface based on the user's activities. This leads to a reduction of required user interaction and enables unconscious non-intrusive control of the user interface of smart glasses.

A major advantage for HAR with smart glasses is that the glasses provide access to the human

eyes and thus the human gaze to derive features for activity recognition. Combined with head positions or head movements, the result is a unique set of features that represents a wide range of human perception. This allows recognizing a rich set of human activities ranging from mainly cognitive activities like reading e.g. shown by Islam et. al. [205] to activities of physical nature like walking [206].

Besides of these advantages, smart glasses also introduce challenges arising from the sensor integration due to space constraints as well as the limited available power for a continuous operation in a HAR task. While head movements can be easily captured by accelerometers and gyroscope sensors, which are available as highly integrated microelectromechanical systems (MEMS) with low power consumption and small sensor size, challenges arise for eye movement sensors.

Eye movement sensors, also referred to as eye tracking sensors, are dominated by video oculography (VOG) sensors and electro oculography (EOG) sensors [15]. VOG sensors, e.g. the Pupil Invisible [90] rely on video cameras, which are integrated into the glasses frame or the frame temple. They capture video frames from the eye and use computer vision algorithms to extract the pupil contour and estimate the user's gaze. The main disadvantages of VOG sensors are the size of the camera sensors and the required optics as well as the required angle under which the sensor needs to be positioned to capture images of the eye. In addition, the computer vision algorithms add significantly to the overall power budget for VOG systems. Camera sensors itself are furthermore sensitive to varying external illumination by limited dynamic range which occurs during every day activities, reducing the pupil detection accuracy [56]. EOG sensors measure the electrical potential of the electrical dipole between the retina and the cornea with electrodes placed on the user's skin to estimate the gaze vector. While EOG sensors are insensitive to external illumination and consume less power compared to VOG sensors, they cannot be integrated unobtrusively and contactless into glasses as they require attaching electrodes to the skin [15].

To overcome these limitations and to be able to reliably measure eye movements to enable HAR for smart glasses, we introduce laser feedback interferometry (LFI) sensors. The LFI sensors have three main advantages compared to VOG and EOG sensors.

i) **Sensor integration** The sensor is based on a small vertical cavity surface emitting laser (VCSEL) in the near infrared (IR) spectrum with a cavity-integrated photodiode, which allows for a space constrained integration into the glasses frame. Due to the invisible IR laser beam, the sensor is unobtrusive and contactless for the user.

ii) **Sensed features** The multimodal sensor is capable of measuring distance towards the eye as well as eye rotational velocity with an outstanding sample rate of up to 1 kHz. From the distance information, characteristic blink patterns can be extracted, while the rotational velocity is directly related to eye activity without any further signal processing.

iii) **Ambient light robustness** The VCSEL with cavity-integrated photodiode enables to apply

158

the LFI sensing principle, a self-coherent sensing scheme. This leads to a high sensitivity to its own radiation while suppressing most other radiation entering the laser cavity. This allows a robust operation in the presence of ambient radiation, as shown by Meyer et. al. [128].

To summarize the advantages of the LFI sensor compared to camera-based VOG systems, we like to highlight the low power consumption of the sensors and the ability to measure the rotational velocity without any further signal processing. In addition, the coherent sensing scheme of the LFI sensors allows for high robustness to ambient light, enabling robust operation in a wide range of lighting conditions from darkness to bright sunlight, as shown by Meyer et. al. [128].

In this work, we propose a novel way to capture eye movements for HAR utilizing LFI sensors to enable a pervasive control of the user interface of upcoming smart glasses. We show that the combination of eye and head movements lead to a high activity recognition rate for a rich activity set which consists of seven different activity classes from physical as well as cognitive domains. To prove this, we conduct experiments and collect simultaneously LFI sensor data as well as data from an inertial measurement unit (IMU) consisting of an accelerometer as well as a gyroscope sensor from 15 participants performing seven different activities. We further design a one dimesional convolutional neural network (CNN) based classifier to recognize the activities. We investigate the effect of different sensing modalities as well as the effect of transfer learning on the classification accuracy and discuss the proposed sensors as well as the proposed classifier with respect to sensor integration and power consumption.

We highlight the contribution of this work as follows. First, we integrate a novel LFI based sensor modality to smart glasses to capture eye movements and combine them with an IMU to obtain a unique HAR data set. To the best of our knowledge, this is the first work of HAR systems, which combines LFI sensor modalities and IMU sensor modalities for smart glasses. Secondly, we implemented the activity recognition system into a smart glasses demonstrator. Experiments with 15 participants were conducted to evaluate the performance of the system. Using state of the art leave one participant out cross validation (LOPOCV) we obtain a macro F1-score of 88.15 %, which is outstanding for the chosen diverse activity set.

In the next section, we discuss the related work in body-worn HAR, eye-tracking based HAR and related work that combines head and eye tracking to perform HAR in a head-worn setup. In Section B.3.4 we provide background knowledge of the LFI sensing principle as well as the features the LFI sensor is capable to capture from the human eye. Afterwards, in Section B.3.5 we introduce our data recording setup and describe the experimental design to generate our data set. In addition, we give an overview of the data processing and the obtained data set. In Section B.3.6 we introduce our 1D-CNN model to classify the human activities and compare the classification performance with a baseline model. Furthermore, we show the effect of transfer learning as well as the influence of both sensor modalities on the classification accuracy for different activities. We conclude this work with a discussion on the applicability and the main challenges of the proposed system.

B.3.3 Human activity recognition

HAR is a well established field of research which deals with concepts of inferring the current goals and behavior of one or multiple human individuals based on a series of observations [207]. The main motivation for building and implementing a precise HAR system is its wide range of applicability. Domains such as active and assisted living (AAL), healthcare monitoring or surveillance can derive a clear benefit from employing HAR algorithms. What is more, within the last decade HAR found its way into consumer electronics with products such as Microsoft Kinect [208, 209] or Nintendo Switch [210] and also tapped into the consumer sports sector with smart footwear providing feedback during running or golfing [211, 212, 213].

At present, there are two main approaches of how human activity data is collected: video-based systems and sensor-based systems [214]. In video-based systems cameras are installed to gain insights on human behavior from images and videos, typically employed for surveillance or recognizing activities of a group of people [215]. On the other hand, sensor-based systems rely on observations stemming from sensors either attached to an individual's body [216] or from ambient sensing devices such as RFID-tags [217] installed within a participant's environment.

In this work, we present a sensor-based HAR system and therefore limit our literature review to sensor-based HAR approaches.

HAR Based on Body Worn Sensors

Modern smart wearables from smart phones to smart ear wear are equipped with MEMS IMU sensors capable of measuring body motion in a ubiquitous and unobtrusive fashion. This intensified research has led to a multitude of publications in the area of sensor-based HAR. One of the most renown works was published by Kwapisz et al. [218]. They collected an activity data set with the six physical activities *walking, jogging, ascending stairs, descending stairs, sitting* and *standing* from 29 participants. They sampled the smart phone's accelerometer with 20 Hz, which the participants wore in their front trouser pocket. The raw data stream was split into windows of 10 s duration with each window containing 200 samples. For each window they derived the statistical features *mean, standard deviation* and *time between peaks* and used them for training a decision tree classifier. They achieve an overall accuracy of 91 %. This high accuracies only drops for *climbing stairs* as it frequently was misclassified as *walk* [218]. This effect is also known as inter-class similarity where the sensor features of two activities are not distinguishable by the classifier [206].

To increase classification performance, Wahl et. al. [219] integrated an IMU sensor, a light sensor as well as an heart rate sensor into a glasses frame and recorded data of nine participants during a full day. They derive 25 statistical features to classify nine activities where two mainly cognitive activities where present in their activity set. They reported a overall accuracy of 77 % utilizing a LOPOCV validation scheme.

In a more recent work by Hayashi et al. [220] audio data was added as further sensor modality

and captured together with smart phone's accelerometer data at a sample rate of 16 kHz and 200 Hz, respectively.

The participants were asked to pursue everyday occupation in a dedicated apartment. Data was recorded for 19 different participants and labelled to match 22 activity classes such as *cooking, eating, reading, sleeping*, etc. leading to a diverse activity set, with 12 physical activities, 4 cognitive activities and 6 mixed activities. Similarly to Kwapisz et al. [218], Hayashi et al. [220] made use of statistical features (*mean, variance, entropy, correlation*) which were extracted from non-overlapping sliding windows with a duration of one second each. These features were then used to train a deep neural network for activity classification. For training and predicting on data originating from the same participant, F1-scores of 80 % are reported. However, classification performance drops as soon as the classifier is reviewed using a LOPOCV scheme as different participants perform the same activity in different ways, leading to variation in statistical characteristics. This effect is also called intra-class heterogeneity [206].

To counteract intra-class heterogeneity, Hayashi et al. [220] followed the idea of transfer learning and retraining a subset of the deep neural network's parameters with a few samples from the left-out participant. This adaptation method reduces the accuracy degradation in the LOPOCV validation scheme and the F1-Score of 80 % is obtained.

HAR Based on Eye Tracking

One of the first works utilizing eye-tracking for HAR was published by Bulling et al. in 2008 [221]. They classify whether a participant was *reading*, solely based on the participant's eye movements recorded with an EOG sensor. Within the collected data set of eight participants they achieved an overall recognition rate of 80.2 % by using a hidden Markov model as classifier.

In a follow-up study, they extended the activity set by office activities like *reading, copying, handwriting, watching a video* and *web browsing*. For classification, they extracted statistical features from the raw EOG readings of eight participants such as *mean, variance* or *maximum peak*. Applying a support vector machine (SVM), lead to a precision of 76 % across multiple participants [19].

In 2013, Bulling et al. [222] stepped away from a discrete activity set and abstracted distinct activities with high level cues which group a set of activities. They choose *spatial, physical, social* and *cognitive* cues as high-level abstraction. Eye movements of four participants were recorded using an EOG sensor. Detected saccades were encoded into a string representation based on the saccade's direction. For each context and participant, a separate binary classifier (string kernel SVM) was trained. The reported results show a high F1 score of 91.21 % for the recognition of social interactions but less accurate performance when predicting a physical activity context (F1 score of 74.78 %).

Steil et al. [223] approached the challenge of HAR in an unsupervised way. 10 participants wore

a VOG head-mounted eye-tracking sensor [53] throughout a full day of their ordinary life and were not restricted in terms of the activities to be performed. Saccadic direction, fixation duration and blink rate were encoded into strings and altogether form a bag-of-words representing an individual's visual behavior. This bag-of-words was passed to a Latent-Dirichlet-Allocation (LDA) topic model, which tried to find reoccurring topics, i.e. activities, within the visual data. With the help of an exemplary ground truth annotation of activity classes, detected topics could be assigned to an activity class. Steil et al. [223] reported the highest average F1-score for the activity of reading (74.75 %) but mentioned that good recognition performance is highly dependent on the individual participant, the activity duration and the number of topics to be discovered by the LDA model. Nevertheless, unsupervised approaches have the advantage of being able to deal with an arbitrary set of activities.

Braunagel et al. [21] investigated eye tracking based HAR in the context of conditionally autonomous driving. They recored the eye movements of 84 participants using a VOG sensor while they performed secondary tasks (*reading, video, idle*) in a driving simulator. For classification, Braunagel et al. [21] used visual scanpaths, i.e. they mapped the recorded gaze data to a string of symbols, with each symbol representing a certain region within the gaze space. The advantage of visual scanpaths compared to statistical feature extraction is that scanpaths preserve the temporal order of the gaze signal. The maximum reported average F1-score of 84 % supports this argument.

A very recent publication from Lan et al. [224] stepped away from using hand-crafted statistical features for HAR and leveraged the automatic feature learning capabilities of deep learning models. An individual's visual behavior was encoded into a spatial-temporal graph. Each node in the graph represents a gaze vector and has connecting edges to temporally adjacent gaze vectors. Edge weights were computed based on differences in angular orientation and euclidean distance of the connected nodes. The resulting weight matrix was passed to a CNN which learns meaningful features such that recognition performance is maximized. The gaze vectors of eight participants where captured by an VOG sensor from Pupil Labs while the participants perform six mainly cognitive activities (*Browsing, playing online games, reading, searching answers in a list, watching video, typing*). The reported F1-scores were most promising, 96 % on average over data of all participants, when considering gaze data within windows of 30 seconds in duration.

HAR Combining Eye-Tracking and Body Worn Sensors

Bulling et. al. [225] were the first to combine information from eye and head movements to classify *reading* activities in an everyday environment, using an EOG sensor to extract eye movement features and a body-worn IMU sensor to extract body movements. Using a combined evaluation of body and eye movement features, they achieved 87.8 % performance with an SVM across 8 participants, while performance degraded by 24 % when head movement features were discarded.

Ishimaru et al. [226] advances the initial study of Bulling et. al. [225] by a larger set of in total 5 activities from cognitive as well as physical domain. For data acquisition, eight participants wore the commercially available Google Glass and performed the activities *talking, watching, solving* a mathematical problem, *reading* and *sawing*. Head motion was recorded using Google Glass' built-in IMU. Actual eye movement recording is not possible with Google Glass which is why Ishimaru et al. [226] resorted to the data provided by the glasses' proximity sensor for blink detection. The recorded accelerometer data was converted into a motion feature by using the averaged variance across all axes. The mean blink frequency as well as the center of distribution of all recorded blink frequencies were used as eye motion features. Ishimaru et al. [226] evaluated the performance for both sensor modalities and their combination by training a person-dependent decision tree. Results indicate that the combination of head motion and blink features yields and average F1-score of 82 % whereas restricting the features space to head or eye features resulted in F1-scores of 63 % and 67 % respectively.

In a later study Ishimaru et. al. replaced the Google Glass with commercial EOG glasses [227] leading to EOG signals as well as head acceleration signals as input modalities for their k-nearest neighbor (KNN) classifier. They collected an activity data set of four activities (*typing, eating, reading* and *talking*) from two participants and reported and overall accuracy of 70 %.

The analysis of related work shows that the area of HAR with body worn sensors and the area of HAR based on eye tracking are well established fields with a lot of previous work. Furthermore, it is shown that HAR with body worn sensors most likely emphazise the features of physical activities and HAR with eye tracking more likely emphazise the features of cognitive activities. Therefore, high recognition accuracies are present in both subfields. The combination of both sensor modalities in a head worn setup is only investigated by a few works of Ishimaru et al. where he was also mainly focusing on cognitive tasks. To capture eye movements he used an EOG sensor in his latest study, which has disadvantages in terms of product integration for future smart glasses.

In our work, we replace the EOG sensor with a promising LFI sensor and furthermore use state of the art method to improve the HAR accuracy in a balanced activity data set and show the classification robustness over a larger set of participants by applying the state of the art LOPOCV method.

B.3.4   Laser Feedback interferometry
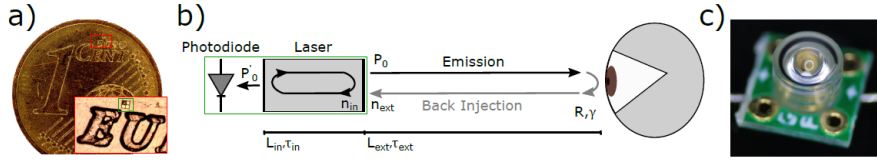
Figure B.18: a) 160 μm x 180 μm VCSEL with integrated photodiode as sensing element on a cent coin b) Coupled cavity model of a laser feedback interferometry sensor. The laser emits light which is scattered by the eye and injected back into the laser cavity. The photodiode monitors the laser power, which varies with changes in the feedback path. c) Encapsulated optical module including the LFI sensor and the beam shaping optics used in the experiments [5].

Laser feedback interferometry is a widely known interferometry measurement method [6]. It is used in the industry as well as in laboratory environments to measure displacement and velocity of solid targets, as well as fluids and distance. Due to the high distance and velocity resolution, it is also widely used in vibrometry applications [121]. Recent works by Meyer et. al. [4],[5] apply LFI sensors in a smart glasses setting to recognize gaze gestures and show the basic functionality of this sensing modality in a near eye setup.

Sensing Principle

Figure B.18 b) shows the coupled cavity model to introduce the basic sensing principle of LFI sensors. A laser with an unattenuated optical baseline output power P_0 emits a coherent laser beam towards the surface of the eye. The laser beam hits the eye under an angle of incidence γ, is attenuated by volume scattering effects and absorption described by a reflectivity R and is injected back into the laser. τ_{ext} denotes the time the laser beam requires to travel over the distance L_{ext} towards the eye. τ_{ext} is dependent on the speed of light c_0 and the external refraction index n_{ext} of the external medium [6].

The back injected light interferes with the local oscillating field, which is often referred to as self-mixing interference, resulting in a modulation of the optical power

$$P_0' = P_0 \left(1 + m \cdot \cos\left(\phi_{fb}\right)\right). \tag{B.19}$$

The modulated feedback power P_0' is dependent on the optical power P_0, the modulation depth m and a varying phase ϕ_{fb} of the backscattered light field. A small fraction of the modulated feedback power P_0' is measured by a photodiode, which is integrated into the distributed Bragg reflector (DBR) of the laser cavity itself [6].

To understand the link of the phase ϕ_{fb} to our observation goal (distance and velocity), we consider the rate equations introduced by Lang and Kobayashi [119]. A solution of the rate equation is the excess phase equation

$$\phi_{fb} - \phi_s + C \sin\left(\phi_{fb} + \arctan\left(\alpha\right)\right) = 0. \tag{B.20}$$

The feedback phase is expressed as a function of the signal phase ϕ_s, Acket's feedback parameter C, describing the coupling strength between target (eye) and laser cavity, and Henry's line width enhancement factor α. As we operate the LFI sensor in the weak feedback regime ($C < 1$) and a constant line width enhancement factor α, Equation B.20 leads to a single solution [6] and ϕ_{fb} is, therefore, only dependent on ϕ_s, leading to

$$\phi_s = \frac{4\pi n_{ext} L_{ext}}{\lambda} \tag{B.21}$$

with λ describing the wavelength of the laser. Considering $n_{ext} \approx 1$ is constant due to the operation of the sensor in free space, only changes in λ and L_{ext} lead to a varying phase ϕ_s and consequently, to a varying phase ϕ_{fb}. This results, with respect to Equation B.19, in a modulation of the optical power which is measured by the photodiode. Changes in the wavelength λ occur by a modulation of the laser drive current. That leads to a periodic heating and cooling of the resonator and, thus, to a periodic change of the cavity length. The variation of the cavity length leads to a periodic modulation of the wavelength, which allows for continuous measurement of the distance according to Equation B.22.

To distinguish between both effects, we compute the partial derivative of Equation B.21 with respect to time, which leads to

$$f_0 = \frac{2L_{ext}}{\lambda^2} \frac{d\lambda}{dI} \frac{dI}{dt}\bigg|_{L_{ext}=const.} \tag{B.22}$$

and

$$f_d = \frac{2v_{ext} cos(\gamma)}{\lambda}\bigg|_{\lambda=const.}. \tag{B.23}$$

Considering a known $\frac{d\lambda}{dI}$, which is a static process parameter of the laser, and a controlled current modulation slope $\frac{dI}{dt}$, the distance to the eye can be calculated by extracting the so called beat frequency f_0 by applying a Fast Fourier Transform (FFT) to the measured varying optical power and rearranging Equation B.22 with respect to L_{ext}.

Movements of the eye ($\frac{dL_{ext}}{dt} = v_{ext}$) lead to a shift of the beat frequency f_0 by the so called Doppler frequency f_d. With a known angle of incidence γ and a measured Doppler frequency, Equation B.23 can be rearranged with respect to v_{ext} to obtain the surface velocity of the eye.

In order to separate f_0 and f_d and, thus, simultaneously measure the distance and velocity of the eye, a triangular modulation similar to frequency modulated continuous wave (FMCW) radar is applied to the drive current of the laser [122]. By separating the up- and down ramp signals into two segments and applying an FFT on each segment, an f_{up} and an f_{down} frequency is captured. f_0 and f_d are obtained from these measurements by

$$f_0 = \frac{f_{up} + f_{down}}{2} \tag{B.24}$$

and

$$f_d = \frac{f_{up} - f_{down}}{2} \tag{B.25}$$

respectively. Recalling Equation B.25, the triangle modulation allows extracting the direction of the velocity as well. The modulation frequency of the triangle signal therefore limits the update rate of the LFI sensor.

The main advantage of the LFI sensor compared to Time-of-Flight (ToF) sensors is its FMCW operating method, which allows distance and velocity to be measured simultaneously, whereas ToF sensors only measure absolute distance [126]. A rotation of the sclera surface in a fixed distance is therefore not measurable by TOF sensors as the absolute distance does not change.

Sensor Features on the Eye
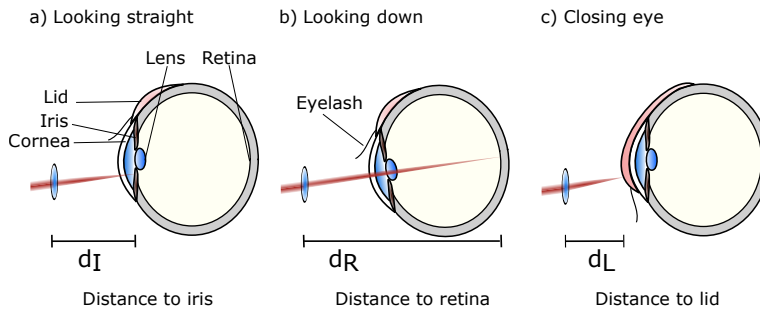
Figure B.19: Different movement related position changes of the eye and the lid resulting in a change in distance measurement due to the geometry and scaffold of the eye [5].

Based on Equation B.22 and Equation B.23 the LFI sensor is capable to measure the surface velocity in laser beam axis v_{eye} and the distance between the sensor and the eye d_{eye}. Both contain valuable features for HAR. Figure B.19 describes the distance features measurable on the eye. A sectional view of the eye and a fixed LFI sensor position is shown for three different positions of the eye and the eyelid. In Figure B.19 a), the eye is directed straight ahead. The laser beam of the LFI sensor penetrates the cornea and backscattering occurs at the iris. For this arrangement of sensor and eye, the LFI distance measurement gives the distance between sensor and iris d_I as feature.

In Figure B.19 b), the eye is slightly rotated downwards, the laser beam penetrates cornea and lens and the main backscattering occurs from the retina, leading to a second distance feature d_R describing the distance between the sensor and the retina. In a temporal context, this feature occur if the pupil crosses the laser beam during eye activities.

Lastly in Figure B.19 c), the lid is closed by for blinking. During a blink the distance d_L between sensor and lid is measured. This lead to a blink feature and in a temporal context it contains information about blink duration and blink frequency, which are used e.g. by the works of Ishimaru et al. [226] or Steil et al. [223].

B.3.5 Recording head and eye movement data set

Our data set was captured from 15 voluntary participants (age (years) = 25.9 (SD 4.5); 10 male; 5 female). Before taking part in the experiment, all participants provided written consent for using their data for research purposes. Two participants wore contact lenses during the experiment. The data set was anonymized to remove any personally identifiable information.

Apparatus

As LFI sensors are not available in commercial smart glasses the data set was captured with a custom head worn research apparatus shown in Figure B.20.
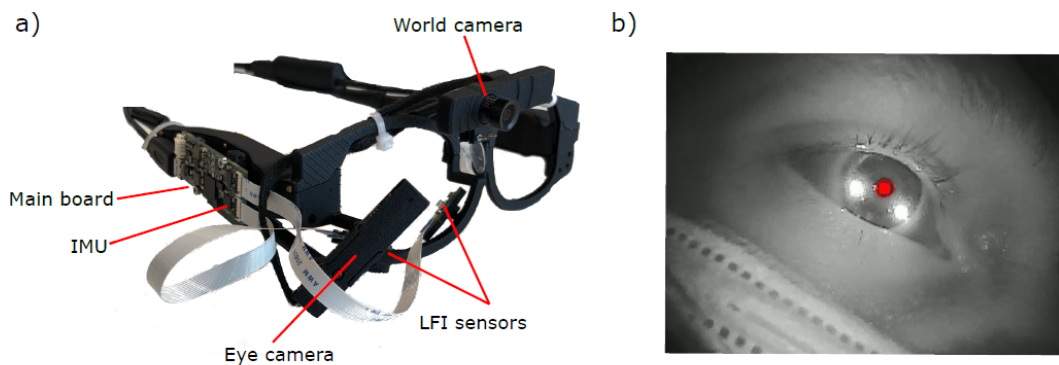
Figure B.20: a) Research apparatus used to capture eye movement data by LFI sensors as well as head movement data by an IMU sensor. The colored arrows denote the coordinate space of the accelerometer. The coordinate space of the gyroscope is aligned to it. b) Image of the eye of an participant captured with the eye camera showing the detected pupil (red mark) by the pupil labs algorithm as well as the two LFI laser spots on the eye (bright spots).

The apparatus consists of a main board which reads out and supervises the two LFI sensors (shown in Figure B.18 c)) attached to the lower glasses frame. The laser beams of the LFI sensors are directed towards the eye ball to measure the eye's rotational velocity as well as the distance towards the eye with a sample rate of 1 kHz. To ensure an eye safe operation of the lasers, the optical power of the lasers is limited to 360 μW. With this low optical power a class 1 laser system according to IEC 60825-1 (well below class 1 optical power limit of 770 μW) is achieved and, therefore, the experiments do not pose any medical hazard to the participant's eye.

Furthermore, an IMU sensor (BMI270) containing a MEMS gyroscope as well as a MEMS accelerometer is attached to the main board. The IMU is used to capture head movements with a sampling rate of 860 Hz. Both data streams are synchronized by the main board and streamed via USB to a laptop, which is used during the experiment to record the data.

In addition to the IMU and LFI sensors we attach a Pupil Labs Core eye tracker [53]to the

apparatus to capture world video frames as well as eye video frames. To record the video data the Pupil Capture (v1.17.71) software was used. The eye video camera is IR sensitive and we used it to align the laser spots of the LFI sensors to the surface of the eye as shown in Figure B.20 b). This step is required to adjust the alignment of the LFI sensors to different head shapes before the actual measurement, as our rigid glasses frame of our research apparatus is adaptable to any interpupillary distance (IPD). The world video frames were used to annotate the captured head and eye movements with the corresponding activity label.

Experiment Design

Our activity set consists of seven activities (*talk, read, video, walk, type, solve* and *cycle*) combining cognitive as well as physical activities. The experiment itself was split into two parts, a stationary part on the laptop and an outdoor part where the laptop was carried in a bag pack. After a short introduction into the functionality of the setup and an alignment of the LFI sensors towards the participant's eye, the experimenter initiated the recording session and starts a casual conversation with the participant. After the *talk* part of the experiment the experimenter left the room and the participant followed for the remaining stationary part of the experiment the instructions on the laptop, where a website guides the participant through the *solve*, *read*, *video* and *type* parts of the experiment. During the *solve* activity the participant was asked to solves a logic test. Afterwards, a text about smart glasses was presented to the participant to *read* and a video about smart glasses to watch. Based on the video and the text read, the participants were then asked questions, which they had to answer by *typing* their answer in text boxes on the screen.

With the stationary part completed, the experimenter returned and stowed the recording laptop in a backpack to be worn by the participant and asked the participant to *walk* around indoor and outdoor and afterwards to *cycle* around the research campus. Figure B.21 shows world camera frames of different activities captured during the experiment.
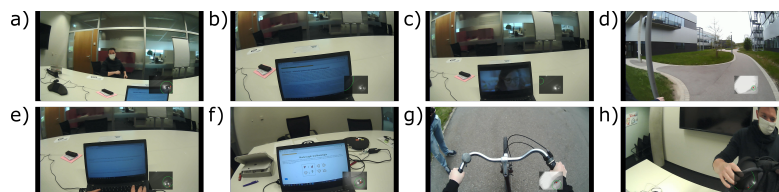
Figure B.21: a) - g) show world camera frames captured during the activities *talk, read, video, walk, type, solve* and *cycle*. In addition, h) shows a world camera frame captured between the indoor and the outdoor part of the experiment where the recording laptop was stowed into the backpack.

Data Processing

In order to classify activities from the head and eye movement data in a supervised fashion, it is required to associate each data sample to an activity label, which corresponds to an activity that the participant was performing at each sample's time of recording.

This activity labels are derived from the world video. In the first step, each participant's recording is loaded into the Pupil Player (v2.4). Next, the videos recorded with the eye and world cameras as well as the corresponding time stamps are exported and the time stamps of the LFI sensors as well as the IMU sensors are synchronized to the world video timestamps. Afterwards the exported world video is compressed and then loaded into a labeling software for annotation purpose.

In order to assign an activity label to the head and eye movement recordings, the labels saved in the time domain of the world video need to be transcribed to the time domain of the IMU and LFI sensors. Since human activity usually extend over several seconds at a rather low frequency compared to the sensor output rate, and in order to reduce the amount of data, both the LFI and IMU data streams were down sampled to a common data rate of 120 Hz. Due to the remaining difference in sampling frequency of head and eye movements with ($\approx$120 Hz) and world ($\approx$30 Hz) camera, multiple samples of the LFI sensors and the IMU correspond to a single frame in the world video. For label transcription, the start and end frame of each labeled activity block in the time domain of the world video is fetched. Next, all LFI and IMU samples that are associated to a world frame within the range of the given start and end frame receive the label of the corresponding activity block. This process results in labeled LFI and IMU features.

Data Exploration

Table B.5 gives an overview of the data set captured during the HAR experiment. A total of 18,5 h head and eye movements were recorded across 15 participants and 7 activities. The mean duration per class are balanced, while the distribution of duration across participants is skewed. Participants spend the most time *typing* (865 s), while *cycling*(534 s) and *reading*(506 s) were sampled the shortest. This skew occurs because the duration of each activity was not limited during the experiment to ensure that participants behave as naturally as possible without being pressured by an expiring clock. The only exception was the activity *video*, which was naturally limited by the playtime of the video. Furthermore, the experimenter paid attention that the activities *walk* and *cycle* do not take unnecessarily long to prevent the participants from becoming bored. To get better understanding of the recorded data and the underlying pattern of the extracted features, Figure B.22 shows a 30 s time span of all input features for all activities.
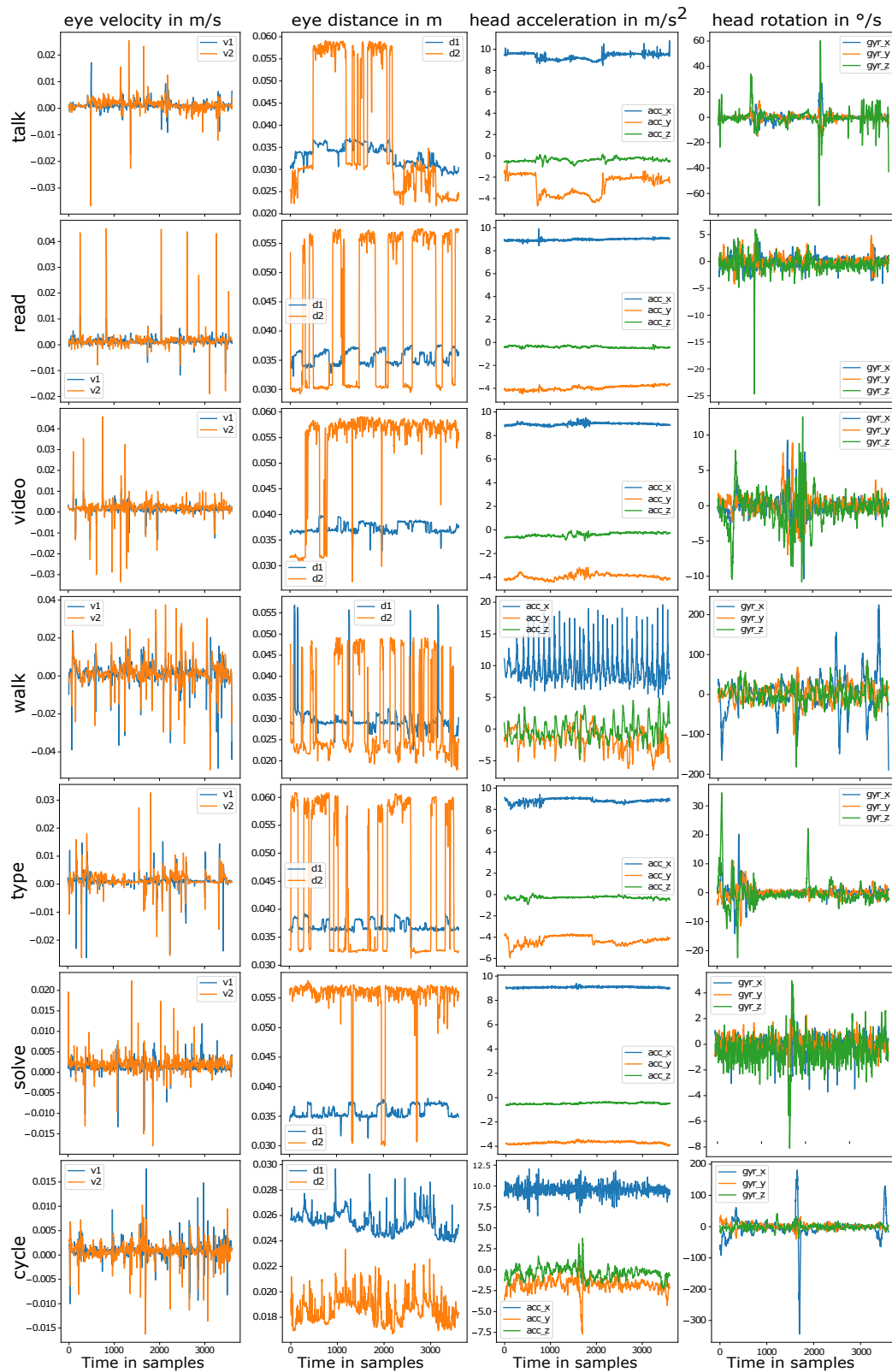
Figure B.22: 30 second windows of the raw head and eye movement features of participant P1 for the seven activities

Table B.5: Duration in seconds of each individual experiment itemized by activity.

	talk	read	video	walk	type	solve	cycle	total
P1	251	408	620	623	385	774	547	3608
P2	475	268	621	547	628	383	611	3533
P3	620	541	618	857	841	857	920	5254
P4	693	412	622	680	376	617	382	3782
P5	664	222	619	475	806	313	262	3361
P6	609	691	620	503	1296	743	426	4888
P7	408	454	618	371	943	483	439	3716
P8	269	605	617	605	1735	869	295	4995
P9	772	760	618	760	589	730	605	4834
P10	745	536	616	632	1080	769	454	4832
P11	867	528	618	846	875	799	780	5313
P12	714	466	621	742	580	447	496	4066
P13	722	448	620	785	1452	880	628	5325
P14	552	690	620	684	843	557	631	4577
P15	598	562	619	626	547	871	527	4350
∅	597	506	619	649	865	673	534	4443

The two left columns show the participants eye movement features eye velocity (v1, v2) as well as distance (d1, d2) towards the eye, captured by the LFI sensors. The two right columns show the participants head movement features captured by the accelerometer (accx, accy, accz) and the gyroscope (gyrx, gyry, gyrz).

The first apparent pattern one can observe is the difference in the variance of the head-mounted sensors amplitudes. Whilst the gyroscope and accelerometer measurements depicted in Figure B.22 show a quite steady line with few to no variation for the stationary activities, the corresponding sensor readings oscillate across all three axes for the physical activities *cycle* and *walk*.

Considering the detected eye velocities, the strongest re-occurring pattern can be observed for the activity *read*. Especially the velocity of the second sensor shows a well-defined pattern with sequentially high positive velocity peaks and small negative velocities in-between, which align well with the human reading pattern. Using small saccades, the eye slowly moves towards the end of a line of text before jumping back with a large saccade to the start of the next line. In addition, the distance signal shows a regular step pattern which occurs if the laser beam crosses the sclera and iris and drops into the pupil and gets back reflected from the retina of the eye, as described in Figure B.19.

B.3.6   Evaluation

Within this section, the synchronized eye and head movement data stream is used to build and train models for HAR. Prior to training, the data stream needs to be prepared to make classification feasible.

Data Preparation

Because different individuals may perform the same activity with varying intensity, the sensor readings may be offset or scaled differently across participants. This introduces undesired bias and inconsistency into the data. To avoid this, the distributions of all features within each participant's data stream are standardized to zero mean and unit variance.

Furthermore, the duration of human activity typically spans from several seconds up to numerous minutes. Thus, a single sensor sampled at sub-second resolution will most likely lack the necessary information to recognize its corresponding human activity. To overcome this issue, multiple consecutive sensor readings are combined using a sliding window of fixed duration. For each window, all containing sensor readings are extracted and considered as one sample and the corresponding label is determined by the most frequent label occurring during the window. In this way, a set of labeled windows is generated for each participant used for HAR.

Due to the varying duration of each individual activity, resampling on window scale is used to balance the data set by up sampling all under represented classes to match the most frequent class. As a result, all activity classes are represented equally often in the data set.

With this data preparation steps, the data set has been transformed into a three-dimensional tensor $D \in \mathbb{R}^{N \times T \times S}$, with N being the total number of extracted and resampled windows across all participants' data streams, T being the number of sensor readings per window and S being the number of sensor modalities considered. The sliding window size was fixed similar to previous work by Lan et al. [224] to 30 seconds with an overlap of 30 %. This results in a total number of $N = 4377$ sliding windows each containing $T = 3600$ sensor readings. As two triaxial head features (accelerometer and gyroscope) and four eye features (eye velocity and distance) from two sensors are considered, a total of $S = 10$ sensor modalities are used.

Classification Approach

The data set D is split into a train and a test set by applying the concept of LOOCV. To be precise, D is split into $D_{train} \in \mathbb{R}^{N_{train} \times T \times F}$ and $D_{test} \in \mathbb{R}^{N_{test} \times T \times F}$, with $N_{train} + N_{test} = N$, such that D_{test} consists of all windows assigned to the left-out participant's data stream and D_{train} contains the windows from all remaining participants. With 15 participants, this leads to 15 possible splits for D. Each classification model is trained and tested separately for each permutation of D_{train} and D_{test} and the model's overall performance is computed by averaging across the individual performances for each permutation, leading to a macro F1-score as classification performance indicator for each model. This method allows getting an understanding of how well a trained model generalizes to data stemming from an unknown participant. To ease comparison, each classifier's performance is visualized using the confusion matrices of the same three participants. These participants were chosen to represent test sets with overall low (P12), average (P6) and high (P11) macro F1-scores across the investigated

classifiers.

Baseline Model

As Ishairmaru et.al [226] reported a high F1-score of 82% with a person-dependent decision tree for HAR based on handcrafted head and eye features, we chose a random forest classifier (RFC) as our base line model. As the RFC is not capable to learn features automatically, manual feature engineering is necessary. As suggested by previously mentioned works [19, 221, 228, 229, 226], a statistical analysis of the computed windows yields the desired input feature vector for the RFC. To be specific, *mean, variance* and *L2-norm* are computed along the temporal axis for each sensor modality and window in D_{train} and D_{test}. The resulting feature values are stacked into a feature vector $v \in \mathbb{R}^{F*S}$ where S is the number of considered sensor modalities and F the number of statistical features to be computed (here $F = 3, S = 10$). Thus, the RFC is trained and tested on the matrices $V_{train} \in \mathbb{R}^{N_{train}\times 30}$ and $V_{test} \in \mathbb{R}^{N_{test}\times 30}$.
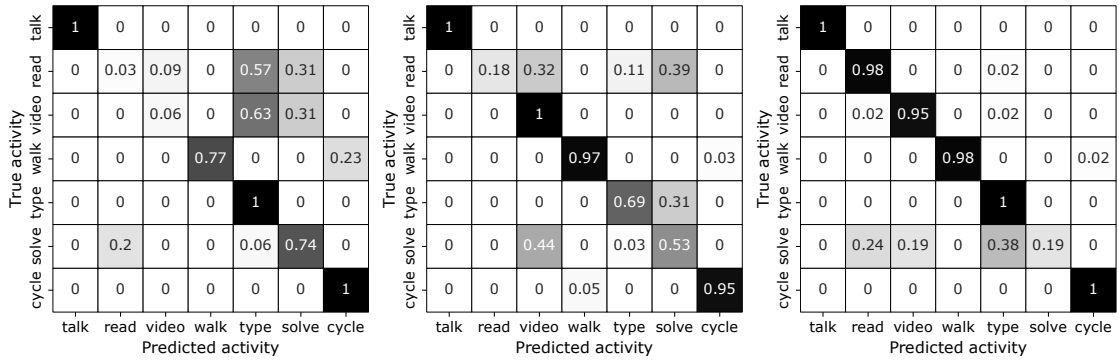
Figure B.23: Normalized confusion matrix for the RFC for participants P12 (F1-score 59.77 %), P6 (F1-score 73.36 %) and P11 (F1-score 88.31 %).

Figure B.23 shows the average macro F1-score after LOOCV has run for each participant. The RFC achieves an average accuracy of 74.32 % and an average macro F1-score of 71.64 %. From the confusion matrices depicted in Figure B.23 it can be seen that the classifier's confusion is only present within the set of stationary (*solve, read, talk, video* and *type*) and physical (*walk* and *cycle*) activity classes, respectively. Furthermore, it can be observed that within the stationary activities, the activities that frequently involve head and eye movements, namely *type* and *talk*, lead to a higher recognition rate than eye movement intensive activities such as *solve, video* and *read*. One possible reason for this observation could be that the statistical features do not take into account the time-dependent pattern of eye movements which could be helpful in distinguishing between *solve* and *read*, as shown in Figure B.22.

1D-CNN Model

To better represent time-dependent patterns of the head and especially the eye movement patterns and get rid of the hand crafted statistical features, we chose a CNN classification approach to improve our classification result on activities. CNNs use a set of convolving kernels that move across an input plane, in our case an TxF matrix. CNNs typicaly consists of several convolutional layers which are chained together to build up a classifier. The lower convolutional layers extract local salient patterns of the input signal, whereas the deeper layers operate on high-level, abstract patterns provided from the previous layers [230]. Using the concept of backpropagation, kernel weights are learned and updated such that a certain error metric is minimized. This is also referred to as automated feature learning and removes the necessity of manual feature selection. CNNs are well known for image classification. However, multiple works have managed to successfully apply CNNs for HAR tasks [230, 231, 232, 233]. Here, the CNN operates on a small subset of data by gradually moving a convolving kernel of fixed size across the window. To capture temporal relationships, one-dimensional convolutions along the time axis are used. Each sensor modality is processed by a separate set of kernels. Each kernel then aggregates multiple sequential data points of that modality.
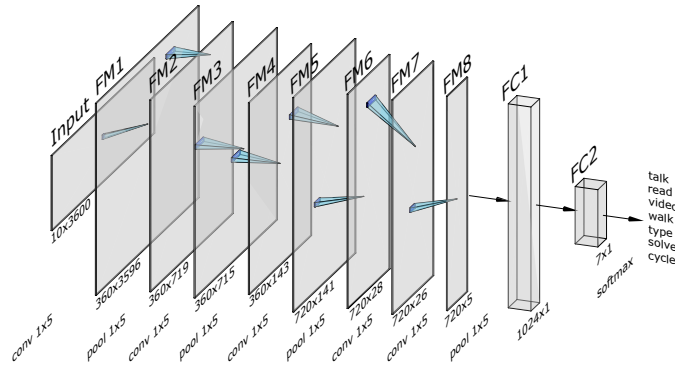
Figure B.24: Architecture of the 1D-CNN model. A batch normalization layer and a leaky ReLU follow each convolutional operation, which are omitted for visualization purpose. The feature maps (FM) height corresponds to the number of convolutional layers whereas their width represents the temporal axis.

Table B.6: Hyperparameters of the 1D-CNN model.

Parameter name	Value
Learning rate	1E-3
Epochs	9
Exponential learning rate decay	0.95
Weight decay	1E-4
Window size	30s
Window overlap	0.3

The 1D-CNN model architecture, shown in Figure B.24, is inspired by the work of Yang et al. [230] and was adjusted to fit for the HAR task. The CNN1D-model is trained and evaluated

using all windows in D_{train} and D_{test} respectively. To process a single window, the CNN1D-model utilizes four single strided convolutional layers (FM1, FM3, FM5, FM7), each followed by a maximum pooling layer (FM2, FM4, FM6, FM8), a leaky ReLU activation and a 1×5 max-pooling operation. The latter leads to the fact that the temporal resolution becomes coarser with increasing network depth, which allows extracting more coarse time-dependent features with increasing convolutional layers. This coarse-grained, high-dimensional time series features computed by the last pooling layer (FM8) is passed through two fully connected layers (FC1, FC2) with a dropout layer in-between to map the feature sequence to the desired amount of classes. In the final step, a softmax activation function is applied to transform the network's output into a probability distribution across class labels. The model is trained using cross entropy loss and the Adam optimizer [234] with the hyper parameters given in Table B.6.
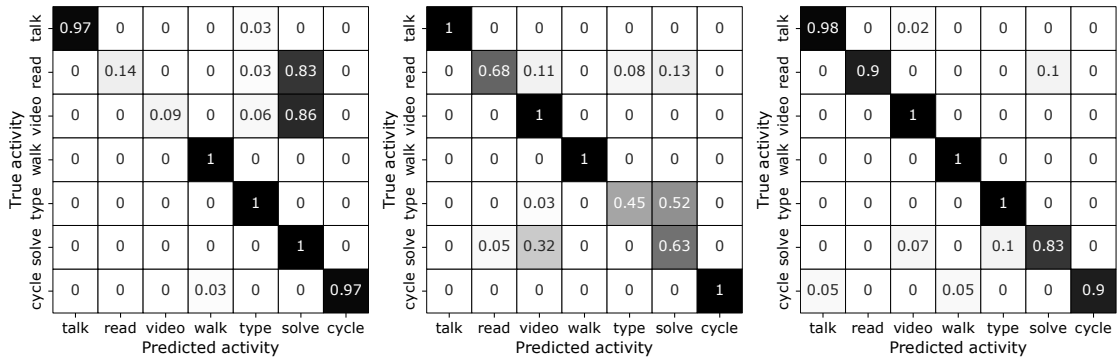
Figure B.25: Normalized confusion matrix for the 1D-CNN model for participants P12 (F1-score 73.87 %), P6 (F1-score 82.25 %) and P11 (F1-score 94.55 %).

Figure B.25 shows the confusion matrices of the selected participants. After LOOCV, the 1D-CNN model achieves an average accuracy of 82.13 % and a macro F1-score of 80.98 %. Compared to the RFC baseline model, the overall HAR rate increases for all participants. This is mainly caused by the improved accuracy for the reading class. The 1D-CNN seems to be able to extract patterns that are specific to the *read* class, which are not represented by the statistical features alone. In addition, the 1D-CNN model handles intra class similarities between the eye movement-intensive activities *read*, *solve* and *video* better.

1D-CNN with Transfer Learning
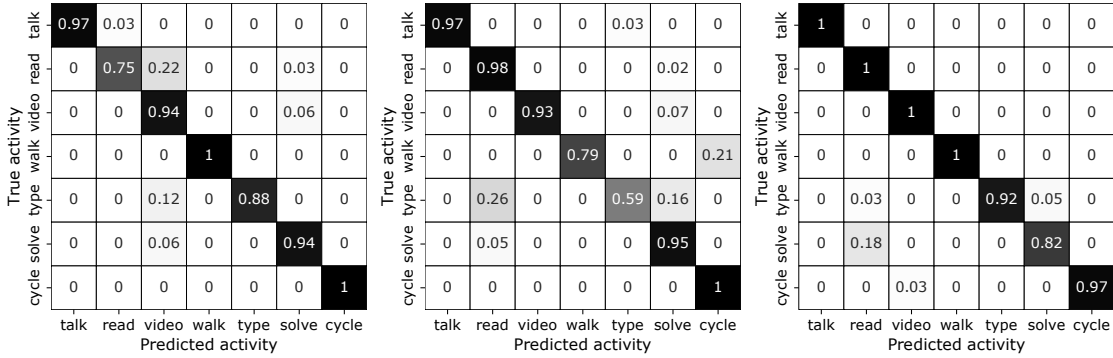
Figure B.26: Normalized confusion matrix for the 1D-CNN model after three transfer samples for participants P12 (F1-score 92.41 %), P6 (F1-score 88.34 %) and P11 (F1-score 95.97 %).

Human activity is highly diverse and versatile because different participants perform the same activity in a different manner. This leads to intra class heterogeneity [206] and a general model, which shows high recognition accuracy over all participants, is hard to build. To tackle this issue we apply transfer learning as proposed by Chikhaoui et al. [235] to further increase the recognition accuracy of the 1D-CNN model. The 1D-CNN model architecture is modified as follows. For every permutation of LOOCV, the model is trained as described in Section B.3.6. After training has finished, the weights in the convolutional layers and in the first linear layer (FC1) are frozen such that they will not be updated during further backpropagation. Furthermore, the dropout layer between the two linear layers (FC1, FC2) is removed. Next, only the last linear layer (FC2) is reinitialized and trained for a few epochs, using three 30 second windows per class from the left out participants data set.

To increase generalization the three transfer learning samples are chosen randomly within an activity class. With the chosen sample size of three transfer samples per activity only a small subset of all samples per activity are used to retrain the model to minimize overfitting.

The model still relies on the feature extractor built from all other participants, which is mainly in the convolutional layers but learns to interpret the extracted features in the context of the current test participant by personalizing their last linear layer to the participant-specific data. Figure B.26 shows the confusion matrices of the three selected participants after applying transfer learning. With this approach, the overall classification accuracy increases, leading to a macro F1-score of 88.15 %.

Especially for participant P12 the classification accuracy improves as the classifier now resolves the disambiguity between *solve,* and *read* in a better way. Also for participant P6 the classification accuracy increases slightly. Overall transfer learning leads to an improvement of ≈ 7 % with only three samples per class. The hyper parameters used during the transfer learning step are denoted in Table B.7.

176

Table B.7: Hyper parameters of the transfer learning part of the 1D-CNN model.

Parameter name	Value
Learning rate	1E-3
Epochs	10
Exponential learning rate decay	0.95
Weight decay	1E-4
Transfer samples per class	3

Model Performance Comparison

In Table B.8 the macro F1-scores of the three models (RFC, 1D-CNN and 1D-CNN Trans) are shown for each participant.

Table B.8: Macro F1-score of the three models, investigated for HAR, for each participant. The bold marked scores show the highest achieved macro F1-score over the three classifiers while the scores marked with an underline show the lowest macro F1-score.

	RFC	1D-CNN	1D-CNN Trans
P1	71.88	81.85	**87.48**
P2	62.26	80.84	**87.79**
P3	87.96	90.56	**95.81**
P4	60.05	69.17	**75.25**
P5	71.67	90.09	**88.61**
P6	73.97	81.97	**88.34**
P7	63.55	79.57	**95.63**
P8	72.91	**84.09**	82.78
P9	73.09	78.34	**83.94**
P10	65.08	76.43	**92.72**
P11	84.01	94.49	**95.97**
P12	59.36	69.33	**92.60**
P13	74.94	**81.78**	81.07
P14	78.47	69.65	**86.22**
P15	75.38	86.54	**88.04**
∅	71.64	80.98	**88.15**

The RFC shows in nearly all cases the lowest macro F1-score, which indicates that the extraction of statistical features from sampled windows of the sensors do not represent the patterns required to recognize activities well. With this approach, we achieve similar results as related work by Ishimaru et. al. [227].

As we introduce a representation of the sequence and extract features automatically by applying the 1D-CNN model, the overall macro F1-score increases to a similar region as reported by Ishimaru et. al. [226].

By further personalization of the classifier to the participant using transfer learning, on the 1D-CNN model a further increase in performance was observed, which outperforms related work by Ishimaru et. al. [226, 227]. It leads to a similar performance like related work focusing either on physical activities recognized with body worn sensors e.g. [218] or cognitive activities

recognized with eye-tracking sensors [21].

In Table B.9 the three classifiers are compared on an activity level.

The primary physical activities *cycle* and *walk* show an overall high recognition rate. In addition, *talk* as an activity consisting of head and eye movements shows a high recognition rate over all classifiers. Especially the *read* and *solve* activity take advantage of the extraction of sequential features by the 1D-CNN model. Transfer learning improves mainly the recognition rate for the *read* and *video* activities, which are heavily dependent on eye movements. This indicates that these activities are dominated by an intra class heterogeneity over all participants. Another observation can be made for the activity *type*. For this activity transfer learning leads to a degradation of the classification accuracy compared to the other classifiers. This shows that there is an inter class similarity (also present in Figure B.25 for P6) between *type* and *solve* and by personalization, the decision boundaries shifts either to *solve* or *type* activity and therefore one increases while the other decreases.

Impact of Sensor Modalities

To investigate the impact of the sensor modalities, especially the impact of head movement features versus the impact of eye movement features, feature importance is studied. During this study all three models are retrained with only a subset of all features, more precisely the features were separated into IMU features (triaxial accelerometer and gyroscope features) and LFI features (d_{eye} and v_{eye}) per sensor. Afterwards, we retrain and evaluate all models with only LFI features from both sensors (LFI1 & LFI2) and only IMU features and compared the achieved macro F1-score with the macro F1-score reached if all sensor modalities where considered. In addition, the effect on the macro F1-score when only a single LFI sensor is used is investigated to answer the question of how the change in angle of incidence γ of the laser beam affects the classification accuracy.

Figure B.27 shows the results of the influence of the sensor modalities on the macro F1-score.

Table B.9: Averaged macro F1-score for each activity over all participants of the three models, investigated for HAR. The bold marked scores show the highest achieved average macro F1-score over the three classifiers while the scores marked with an underline show the lowest average macro F1-score.

	RFC	1D-CNN	1D-CNN Trans
talk	94.59	<u>93.39</u>	**95.06**
read	<u>34.92</u>	68.01	**91.78**
video	<u>70.02</u>	74.88	**87.85**
walk	95.08	**96.12**	<u>95.03</u>
type	84.53	**85.83**	<u>77.17</u>
solve	<u>47.83</u>	60.14	**74.74**
cycle	<u>85.13</u>	94.92	**96.59**
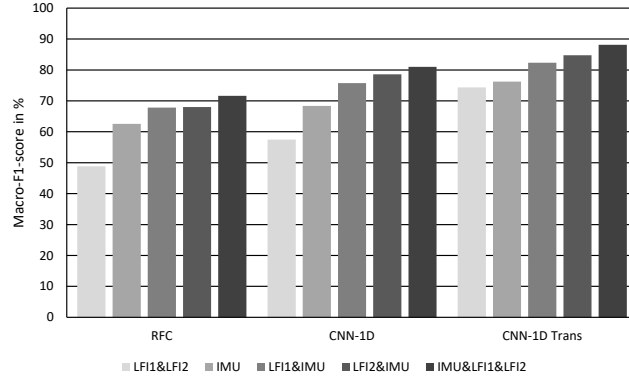
Figure B.27: Impact of the sensor modalities on the macro F1-score for the three models

Is it clearly visible that the combination of head and eye movement features are mandatory to reach high classification accuracy for all three models. For the 1D-CNN model and the RFC it is also shown that the impact of the IMU features is higher than the impact of the LFI features. A further interesting point to note is that the influence of the angle of incidence γ of the laser beam shows only small contribution to the overall performance of the models but it has to be noted that the sensor position of LFI2 leads to a slightly better performance.

In Table B.10 the effect of head and eye movement features on the macro F1-score are investigated per activity. As expected from related work, the physical activities *walk* and *cycle* are

Table B.10: Macro F1-scores per activity class averaged across all participants when solely considering eye movement features (LFI1 & LFI2) and head movement features (IMU) as well as their combination. For each classifier and sensor modality the best and worst values are emphasized in bold and underlined respectively.

	RFC			1D-CNN			1D-CNN Trans		
	LFI1& LFI2	IMU	both	LFI1& LFI2	IMU	both	LFI1& LFI2	IMU	both
talk	62.39	92.12	94.59	55.01	95.02	93.39	**81.74**	92.26	95.06
read	40.50	<u>23.61</u>	<u>34.92</u>	58.20	<u>31.74</u>	68.01	81.53	<u>49.25</u>	91.78
video	33.67	60.13	70.02	<u>46.99</u>	59.26	74.88	73.42	82.77	87.85
walk	60.60	**89.93**	**95.08**	**74.21**	**97.54**	**96.12**	76.27	93.25	95.03
type	70.31	70.21	84.53	67.96	81.05	85.53	62.36	78.63	77.17
solve	<u>33.54</u>	36.61	47.83	48.02	38.57	<u>60.14</u>	<u>60.45</u>	51.01	<u>74.74</u>
cycle	**63.45**	79.41	85.13	71.89	91.84	94.92	88.16	**93.87**	**96.59**

dominant in the IMU data, while the LFI data mainly cover the cognitive activities *video* and *reading*. To some extend the LFI features provide information about physical activities. This might result from glasses slippage during physical activities, which is encoded in the distance information d_{eye}.

Looking only at the features of the LFI sensor, the velocity feature is the most relevant feature for HAR, as it describes the eye movement pattern of the different activities. Especially for

the activity *reading*, which is dominated by eye movements, a clear velocity pattern can be observed, as shown in Figure B.22.

The distance feature mainly contains information about blinking and passing of the pupil through the laser beam. This information is only weakly correlated with activities, since an increased blink frequency per window correlates with higher cognitive load as well as drowsiness.

B.3.7   Discussion

Despite the overall high accuracy on the diverse activity set that we reach by combining LFI sensors and IMU sensor, there are still limitations, we will discuss below and point out a direction of how to improve them.

Power Consumption

Power consumption is an important factor since the glasses are battery powered and for continued HAR a steady stream of sensor data is mandatory. The power consumption per LFI sensor prototype is $\approx 70\,\text{mW}$ with a sample rate of 1 kHz using of-the-shelf components. This can be reduced to about 15 mW by using an application specific integrated circuit (ASIC) [4], while the power consumption of the IMU is $\approx 1.2mW$ at full operation mode [236]. As we down sample the data to 120 Hz a small reduction of power consumption is achievable by adjusting the sampling rate of the sensors. In addition, the power consumption can be further reduced by using just a single LFI sensor, as the study of impact on sensor modalities in Section B.3.6 shows that using a single LFI sensor already a reasonable classification performance is achieved.

However, the main driver of power consumption is the classification algorithm, which should run on a resource constrained embedded processor in the frame temple. For the RFC algorithm Kumar et. al. [237] present a resource and power-efficient implementation while keeping a high classification performance.

With respect to the 1D-CNN model, energy-efficient implementations for CNN-based models are established, e.g. in dedicated neural network accelerators or a tensor processing units [83], [82]. The remaining drawback of the 1D-CNN model for use on a resource constrained embedded processor is the memory footprint mainly driven by the fully connected layers (FC1, FC2) which hold 3686400 out of 3703816 model parameters. To reduce them a further network optimization could be achieved by model pruning or other optimization techniques [238]. Another method could be to rely on a pure CNN based model like proposed by Perslev et. al. [239].

Transfer Learning

To exceed the state of the art with respect to classification performance we apply transfer learning to improve the 1D-CNN model. This requires the user to perform each activity of the activity set once for 90 seconds to adapt the final layer of the 1D-CNN model to the user. This may degrade the user's experience even if it is a known procedure for smart wearables e.g. the initial setup of face recognition or fingerprint recognition in a smart phone [240].

In addition, the last layer (FC2) of the 1D-CNN model needs to be re-trained, which needs to be done on the device. While the inference on the device is a computationally simple process, the post-training to fit the model is computationally intensive and therefore takes some time. One solution to improve model adaptation time is to perform the post-training in the cloud.

Glasses Slippage

Glasses slippage is a well-known issue in the eye tracking domain [93] and it also affects the LFI sensors in some cases. Glasses slippage was to some extend present during our experiments, mainly during the physical activities *walk* and *cycle* as well as *talk*. Therefore, the effect of slippage is already reflected in our reported macro F1-scores. But in some cases we observed that the eye was covered by the eye lids of a participant and the laser spots of the LFI sensors hit the eye lid. Under these circumstances no robust velocity features were captured by the LFI sensors leading to a reduction in classification performance. To ensure that the laser beam of the LFI sensor hits the eye, the laser beam can be scanned over the surface of the eye [102], [129]. Another solution could be achieved by modification of the optics to split the laser beam of the sensor to several beams in a line as proposed by Riemensberger et. al. [241].

B.3.8 Conclusion

In this work we introduce the LFI sensor for eye movement detection. The LFI sensor is very advantageous as it consumes very low power, is fully unobtrusive as the IR light is invisible to the human eye, and is very robust to ambient light.

We attached the LFI sensors together with an IMU to a glasses demonstrator to capture eye and head movements and recorded a large dataset of 15 participants performing seven activities. With the proposed 1D-CNN model and by applying transfer learning, we outperform the state of the art and achieve a macro F1-score of 88.15 % on a diverse activity set consisting of both physical and cognitive as well as social activities. We further investigate which sensor modality leads to high classification performance on each activity level. Furthermore, we discuss our work critically with respect to power consumption, transfer learning and the impact of slippage. Based on our results, we are confident that this work will advance the path towards enabling always-on context awareness based on HAR for smart glasses.

B.4 U-HAR: A Convolutional Approach to Human Activity Recognition Combining Head and Eye Movements for Context-Aware Smart Glasses

B.4.1 Abstract

Smart glasses are considered the next breakthrough in wearables. Their advancement lies in the ubiquitous embedding of content in the user's field of view, promising an immersive extension of reality. While advances in display technology seem to fulfill this promise, interaction concepts are inherited from existing wearable solutions that require the user to actively interact with the glasses, limiting the user experience. One way to improve the user experience and drive immersive augmentation is to reduce user interactions to a necessary minimum by adding context awareness to smart glasses.

To achieve context awareness, we propose an approach based on human activity recognition which incorporates features derived from the wearer's eye and head movement. Towards this goal, we combine an eye-tracker and an IMU to collect eye and head movement features from 20 participants performing seven cognitive and physical activities to derive context information. From a methodological perspective, we introduce U-HAR, a convolutional network optimized for activity recognition on a device with constrained power and memory resources. By applying few shot learning, we achieve an outstanding macro-F1 score of 86.59%, allowing us to derive contextual information that will pave the wave for a more immersive user experience in future smart glasses.

B.4.2 Introduction

After the success of smartphones, smart earbuds and smartwatches, smartglasses are expected to be the next breakthrough in the domain of smart wearables and will replace the smartphone as personal smart wearable in long term as recently announced by Apple [1]. Their advantage over other wearables is the seamless embedding of visual content into the field of view (FOV) of a user, resulting in a true augmentation of the environment. A major drawback of smart glasses, preventing full immersive augmentation of the environment, arises from the state of the art interaction concepts experienced by the user. The user still has to actively interact with the glasses, e.g. through spoken commands as shown in the Echo frames [132], by hand gestures used in the Hololens [51], touch interaction concepts introduced by Google Glass [203] or gaze gestures [5].

One way to improve the user experience and drive immersive augmentation is to minimize user distraction from interacting with smart glasses. A possible solution to reduce user interaction and thus drive immersive augmentation is to add context awareness to smart glasses. A possible way to achieve context awareness in a smart glass environment is to derive contextual

---

[1]https://www.computerworld.com/article/3642649/analyst-apples-ar-glasses-will-run-mac-chips.html

information from the wearer's activities by using Human Activity Recognition (HAR).

In this work, we investigate the use of HAR to derive contextual information in a smart glasses setup. For this purpose, we built up an apparatus to simultaneously collect head movement data with an inertial measurement unit (IMU) and eye movement data with a commercial video oculography (VOG) eye-tracking sensor during our experiment. For evaluation purposes, 20 participants performed seven physical or cognitive tasks (i.e. talk, solve, read, watch a video, type, walk and cycle), leading thus to an unique dataset with a total duration of 1514 minutes. In addition to this novel dataset, we propose a convolutional neural network (CNN) model to temporal fuse head- and eye-movement information to recognize human activities, which we coined **U-HAR**. By applying few shot learning, the CNN model is further personalized to achieve an macro F1-score of 86.59 % in a leave-one-participant-out cross validation (LOPOCV) scheme. With this outstanding accuracy our model outperforms earlier works, e.g. by [227, 226] by a large margin. In addition to the above methodological novelty, we investigate the relevance of the sensor features for different human activities and discuss the proposed system in terms of its applicability for smart glasses. Our contribution is three fold:

(i) **Context aware smart glasses:** We propose to employ HAR to derive contextual information to minimize user interaction with smart glasses to increase user experience, to minimize power consumption through contextual aware display control and improve privacy as contextual information can be derived without the need of a world camera sensor.

(ii) **U-HAR:** We adapt a CNN (**U-HAR**) model architecture and add few shot learning to efficiently recognize a rich set of human activities with a high accuracy from eye- and head movement features.

(iii) **Dataset:** We collected a comparable large dataset of 20 participants with commercially available VOG and IMU sensors commonly used in AR glasses to demonstrate the performance of our system and publish the dataset for further research [2].

In the following Section, we provide an overview of related work with a focus on HAR combining both features from head- and eye-movements. In Section B.4.4, we describe the apparatus and the experiment used to record the dataset. In Section B.4.5 **U-HAR** is introduced, our proposed CNN model to classify activities from the gathered dataset and describe the network architecture. To assess the performance of our network, we compare in Section B.4.6 the proposed model with a baseline model and investigate the performance improvement that can be achieved by adapting the model with few shot learning. Furthermore, we investigate the impact of sensor modalities on HAR accuracy for the different activities of our set of activities. During the final Section, HAR for context aware smart glasses with an focus on the applicability of our model highlighting challenges w.r.t. sensor integration, power consumption and few shot adaption is discussed and a conclusion from our work is drawn.

---

[2]https://atreus.informatik.uni-tuebingen.de/seafile/d/978f6631b6b34f7c9139/

B.4.3   Related Work

A huge amount of smart wearables like smartphones, smart earbuds and smartwatches are equipped with an IMU sensor to measure body motion in an unobtrusive and ubiquitous fashion. This stimulates research on HAR using body motion features. One of the probably most renowned works in this context was the work of [218]. In this work, the authors sampled an IMU sensor of a smartphone with 20 Hz while 29 participants performed 6 mainly physical activities: *jogging, ascending stairs, walking, sitting, standing* and *descending stairs*. The time series data they gathered was split over the time axis into equidistant slices of 10 seconds duration. From each of the slices the authors derived the following statistical features: *mean, time between peaks* and *standard deviation*. The individual features from each windowed time series slice were used to train a decision-tree-classifier. [218] achieved with this method an accuracy of 91 % by applying 10-fold-cross-validation.

To expand the activity space compared to the physical activity dominated activity space by [218], [220] added a microphone as additional sensor modality to an IMU setup to gather audio data and derive audio features at a sampling frequency of 16 kHz along with IMU features at a sampling frequency of 20 Hz. The authors, recorded IMU and audio data for 19 participants in a rather unconstrained experiment environment and labelled 22 activity classes e.g. *reading, cooking, sleeping, eating*, etc., forming thus a diverse activity dataset composed of 4 cognitive activities, 12 physical activities and 6 mixed activities containing physical as well as cognitive elements. Similar to [218], [220] derived statistical features like (*variance, mean, correlation* as well as *entropy*) from non-overlapping windows spanning over 1 s of the total data stream. They achieved an F1-score of 80 % using a deep neural network (DNN) while train and predict on windows extracted from the same participant. However, this rather high accuracy drops significantly as soon as a LOPOCV validation scheme is applied. This effect also known as interclass heterogeneity[206], shows that human activities vary across participants as the same activities are performed in a slightly different ways by different participants, leading to a variation of the statistical features.

To counteract accuracy degradation due to interclass heterogeneity, [220] apply transfer learning and thus retrained a few of the final layers of the DNN model using a small amount of samples stemming from the left-out participant. With the use of transfer learning the authors achieved again an F1-score of 80 % even while using the LOPOCV validation method.

In contrast to the previous mentioned works of [218] and [220], [224] used an video based eye-tracking sensor (VOG) from Pupil Labs [56] to capture eye-movement data and extract gaze vector information as features during an experiment with eight participants. The participants perform during the experiment the following six mainly cognitive activities: *reading, playing online games, Browsing, typing, watching video* and *searching in a list*.

The authors reported an macro F1-scores of 96 % over all participants, when using sliding windows with 30 s duration to slice the data stream.

The work of [226] is one of the first works, which recognize human activities by combining eye- and head-movement features. According to the authors this combination of features allows to classify physical as well as cognitive activities. During the experiment eight participants wore the Google glasses, a commercial smart glasses to acquire data while the participants perform the following activities: *solving a mathematical puzzle, watching media, talking, reading* and *sawing*. While head motion was recorded with the builtin accelerometer of the google glasses, the recording of eye movements was not possible as the google glasses only provide a proximity sensor to detect blink events. Thus only blink information is considered as eye-movement feature.

From the captured accelerometer data the *averaged variance* over all axes was extracted as head-movement feature. As eye movement features [226] derived the *mean* blink frequency and the *center of distribution* from the detected blinks. To evaluate theire approach they trained an individual decision tree classifier per participant using features from both sensor modalities. The combined evaluation using both sensor modalities yielded an macro F1-score of 82 % while [226] reported a degradation of accuracy to 63 % and 67 % if they restrict the feature space to solely to head- or eye-movement features.

To gather a richer feature set from eye movements [227] used a commercial available glasses with integrated electro oculography (EOG) sensor instead of the Google Glasses. In addition the captured head movement data using an accelerometer. Data from both sensing modalities are combined and finally fed into a k-nearest neighbor (KNN) classifier to classify *talking, reading, eating* and *typing* gathered from two participants. They achieved an overall accuracy of 70 %.

Most recently [133] propose the combination of an IMU sensors with an optical laser feedback interferometry (LFI) sensor to capture head- and eye-movement data from 15 participants, which performed the seven activities *reading, watching media, solving, typing, talking, walking* and *cycling*. Instead of statistical features they extract the features automatically using a CNN model and achieve an overall F1-score of 88.15 % by personalizing their network using transfer learning.

Related work indicates that significant effort is made to classify physical activities using data from an body- or headworn IMU sensor and cognitive activities from an eye-tracking sensor. Thus, high recognition accuracy is reported isolated in both sub fields using either IMU sensors for physical activities or eye tracking sensors for cognitive activities. The few works dealing with both eye- and head movement data in a headworn setup rely either on special sensors e.g. the LFI sensor from [133] or EOG sensors which did not allow for a robust measurement of gaze signals in an everyday device, as the dry electrodes of an EOG sensor require constant good contact with the skin to obtain an optimal EOG signal [227]. Furthermore, related work restricted their studies to a small number of participants thus limiting evaluation with respect to generalization for a larger set of participants.

To overcome this limitations and provide context awareness for a broad spectra of smart glasses

with integrated VOG sensors, we used a commercial VOG sensor similar to the sensor used by [224] and add a commercial available IMU sensor for further collection of head movement data from a large number of participants; resulting in a unique dataset that includes data from physical and cognitive activity domains. Inspired from related work we further propose a state of the art CNN model together with the aspect of transfer learning to improve HAR accuracy in a balanced activity dataset containing physical, cognitive as well as mixed activities, and show the robustness of classification in a large number of participants by applying the LOPOCV method

B.4.4   Collection of head- and eye-movement dataset

The dataset was collected from 20 participants (10 male, 10 female) with a mean age of 27 (±4.3). All participants gave their written consent to use their data for research purpose before taking part in the experiment. Further, participants affected by visual impairment were asked to wear contact lenses to ensure a high data quality.

Apparatus

To collect the data, the participants wore a research apparatus during the experiment which is shown in Figure B.28 a). Similar to [224], we used the commercially available Pupil Labs Core
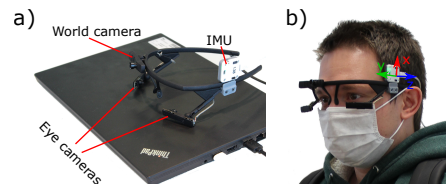
Figure B.28: Research apparatus used during the experiments to collect eye- and head-movement data.

[53] VOG eye-tracking system as the basis of our research apparatus, as shown in Figure B.28 a). The VOG sensor system includes two eye cameras pointed towards the eye, capturing images at 120 Hz of a participant's left and right eyes. In addition, the frame holds a third camera, which points at the world scene. To collect head movement data, a custom board consisting of a microcontroller and an IMU sensor (BMX055) is mechanically attached to the frame of the Pupil Labs Core. The microcontroller controls and reads the IMU's triaxial accelerometer, the triaxle gyroscope and the triaxial magnetometer at a sampling rate of 120 Hz. In Figure B.28 b), a participant wears the visualized apparatus and the axis orientation of the accelerometer and the magnetometer are drawn. The rotational axes of the gyroscope are perpendicular to the indicated acceleration axes. Both the Pupil Labs Core and the custom IMU board are connected to a laptop via USB to simultaneously collect eye and head movement data during the experiment.

Experiment design

The experiment itself is divided into two parts, a stationary part and a physical part. During the stationary part, participants sit in front of the recording laptop and perform the following activities in the following order *talk, read, solve, watching video* and *type on the keyboard*, which are mainly cognitive and mixed activities. In this part of the experiment, the execution time per activity was not limited in order to elicit a natural behavior of the participants during the execution of the activities. During the physical part of the experiment, participants go for a
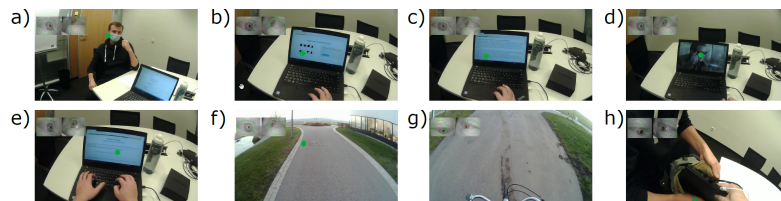
Figure B.29: a) - g) frames captured by the world camera during the 7 activities *talk, solve, read, video, type, walk* and *cycle* in this order. In addition, h) shows a captured frame during the switching between the stationary and the physical experiment part.

walk and ride a bike to *cycle* outdoors. Before the start of the experiment, the Pupil Labs Core eye tracker is calibrated using the built-in single marker calibration method [99] to provide accurate gaze estimation in the participants' FOV . Afterwards, the recording is started and eye movements are recorded utilizing the Pupil Capture (v1.15) software. At the beginning of the experiment, the experimenter initiates a conversation to *talk* with the participant. Afterwards the experimenter leaves the participant alone and the participant is guided by a website on the laptop through the stationary part of the experiment. For the *solve* activity, a logic test with a set of tasks containing a mathematical quiz and extension of a series of geometric patterns was presented to the participant. Afterwards, the participants *reads* a text about recent advances in the domain of smart glasses and watches a *video*. Afterwards the participant has to answer questions related to the shown text and the video, by *typing* answers on the keyboard of the laptop. After finishing the stationary part, the laptop is stashed in a backpack by the experimenter to start the physical part of the experiment which consists of the activities *walk* and *cycle*.

The activities were selected based on related work e.g. by [133] as well as to span a large variation between pure eye movement related activities (*read, video*), mixed activities (*type, solve*), social activities (*talk*) and pure head movement dominated activities (*walk, cycle*) .

Data processing

After a successful recording, the recorded world and eye videos are exported using the Pupil Player (v2.4). For each frame of the eye videos, the pupil center is extracted based on the provided 3D eye model pupil detection method [9].

Figure B.30 shows images of the extracted pupil center for four different cases. In some cases, e.g. during a blink (Figure B.30 c)), the pupil could not be detected, also indicated by a low pupil detection confidence value of less than 0.9. In these cases, we forward filled the current pupil position with the last valid pupil position (pupil detection confidence of more than 0.9) to obtain a continuous stream of features over time. Furthermore, we applied a moving median filter with kernel size 10 on the pupil position data stream to smooth out high saccades or jumps due to falsely detected pupils e.g. shown in Figure B.30 d).
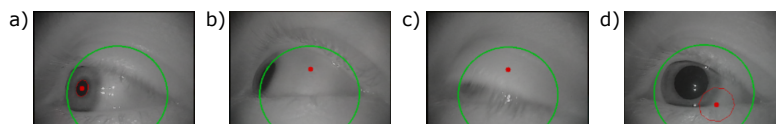
Figure B.30: Four different cases of pupil detection (red dot) affect data quality. a) Well detected pupil b) no detected pupil due to steep camera angle c) no detected pupil during blink d) false detected pupil due to eye lashes

In Figure B.29, images captured by the world camera are shown for each activity. The images are used to manually assign an activity label to each timestamp and thus label the record. Each world video frame is assigned with one of the seven activity labels or *none* if no activity is performed by the participant e.g. during the transition between the stationary part of the experiment and the physical part. As the world camera samples at a sampling rate of $\approx 30\,\text{Hz}$ and the eye video frames are sampled at $\approx 120\,\text{Hz}$, four eye video frames are assigned with the activity label extracted from one world video frame. The same labels are assigned to the IMU data stream, which is also sampled at $\approx 120\,\text{Hz}$ and synchronized with the pupil labs data stream. Figure B.36 shows 30 s windows of the sensor features captured from participant P1 for each performed activity.

To get insights into the collected dataset, Table B.11 summarizes total duration, mean duration and the standard deviation per activity. A more detailed version on individual participant level is given in Table B.15.

Table B.11: Total duration, mean duration and standard deviation in minutes of the individual activity over all participants.

	NULL	talk	read	video	walk	type	solve	cycle	total
total	98	137	165	209	237	284	266	119	1514
mean	1.96	6.85	8.25	10.45	11.85	14.2	13.3	5.95	-
std	2.06	2.58	2.36	0.42	2.32	5.15	5.88	1.02	-

In total 1514 minutes of raw data were recorded from 20 participants. The order of stationary and physical part were not fixed. Also experiments take place in different environments like in the office or at the participants home. Thus e.g. cycling was performed on different bikes and walking at different locations to reduce bias by experiment design. The large difference in mean and deviation between activities caused by the experiment design as participants were not restricted in duration per activity to ensure a naturally behavior during the experiment.

B.4.5   U-HAR Network architecture
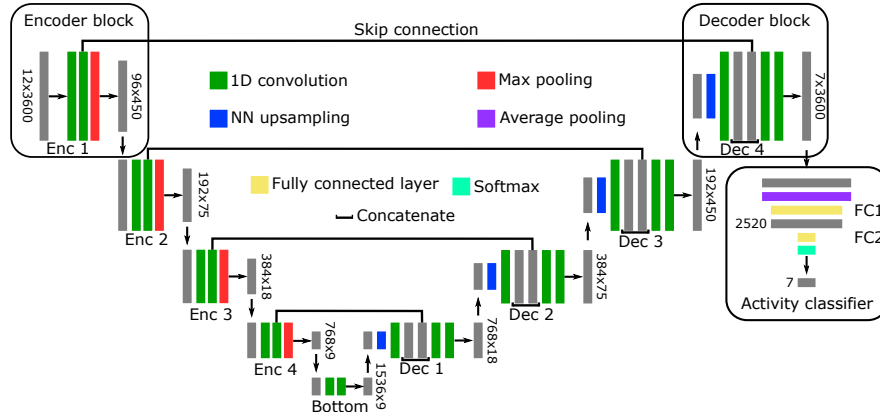
Figure B.31: Network architecture of the proposed **U-HAR** model consisting of a U-Time [7] like encoder and decoder structure and a activity classifier based on two fully connected layers

To classify the collected data according to the seven activities in the activity set, we propose a CNN structure followed by a fully connected classification structure as shown in Figure B.31. The basic network architecture of the so-called U-Nets was introduced by [242] for two-dimensional biomedical image segmentation and adapted by [7] for segmentation of one-dimensional time series data. Perslev et. al. shows a reasonable classification accuracy of sleep stages [239]. Other works, e.g. [243], adapt the network architecture to the domain of autonomous driving to classify lane changes. As HAR refers to the concept of inferring activities based on a time series of observations [206], which is similar to the applications of Perslev et. al. and Eslpas et. al., we decided to adapt the model architecture and apply it to HAR. The idea of the encoder and decoder structure is to transform a time series input with T samples from sensors providing F features into a time series output with T samples for C classes, resulting in a single sample classification. Since human activities normally ranges from several seconds to minute scale, classification on a higher time scale above a single sample is mandatory. To this end, an activity classifier was added to the network architecture, which allows activities to be classified over a longer time period. More details about the classifier are provided in Section B.4.5.

Data preparation

To avoid bias introduced by sensor offsets or the scaling of sensor data introduced by different participants performing activities with non-uniform intensity, all data streams of each participant are normalized to unit variance and zero mean. A sequence of normalized sensor readings are combined to create a window with a fixed number of sensor readings T by applying similar to related work a sliding window approach to the input data stream. All sensor readings (pupil position, accelerometer, gyroscope and magnetometer) of one window are extracted and handled as a single sample for classification. The class label C is derived

from the most frequent occurrence of a label within a single window. As shown in Table B.11, the duration of each individual activity varies across participants. Therefore, the dataset on window scale has to be rebalanced to match the most frequent class by upsampling all under-represented classes C. After applying the data preparation steps, the dataset is transformed into a 3D tensor $D \in \mathbb{R}^{N \times T \times F}$ with N as the total number of samples and F the number of features per sample. Similar to the work of [224] we choose a window length of 30 s and an overlap of 30 % between windows, which leads at a sample rate of 120 Hz to $T = 3600$. This leads to a total number N of 3318 samples. The IMU provides in total 9 features from the 3D accelerometer, gyroscope and magnetometer and the VOG sensor a 3D pupil position vector, leading in total to $F = 12$ features.

Encoder block

The encoder consists of four equally built up convolutional blocks. Similar to the initial architecture proposed by [7], each block contains two convolutional layers each with a five dimensional kernel and same padding to preserve the input dimensions. All convolutional layers are using batch normalization. Afterwards, the temporal resolution is decreased by applying max pooling with kernel sizes (8, 6, 4, 2) after each corresponding layer, while increasing the number of filters respectively. In this way, the temporal resolution is decreased from 8.33 ms to 400 ms allowing deeper layers to learn more abstract features of human activities. Through this aggressive down-sampling, the number of trainable parameters and therefore the memory footprint is reduced, desirable for operating HAR on embedded hardware inside the glasses frame temple.

Decoder block

Analogous to [7], the decoder consists of four transposed convolution blocks. Each of these blocks performs nearest neighbor (NN) up-sampling to increase the time resolution. Each up-sampling block is followed by a convolutional layer with kernel size of (2, 4, 6, 8) and batch normalization for Dec1 - Dec4, respectively. Afterwards, the feature maps of the up-sampled tensor and the corresponding encoder tensor are concatenated along the filter axis and fed into two convolutional layers, each followed by batch normalization as well to process the feature maps in each decoder block. After Dec4, a pointwise convolution is performed to map the feature map obtained by the encoder decoder structure to a TxC tensor, which assigns each sensor reading in the given window T a classes C confidence score.

Activity classifier

The activity classifier serves as a trainable link between the single sample level class confidence score and the label level assigned on the window size T.

Therefore, the activity classifier maps the class confidence scores of size TxC to a classification

output of size $1xC$, effectively removing the temporal axis from the intermediate representation. For this purpose, the activity classifier is added to the output of Dec4, consisting of an average pooling layer with kernel size of $S = T/K$ and two fully connected layers FC1 and FC2. The average pooling layer calculates the average along the temporal axis leading to an output with size SxC which is flattened and fed into the first fully connected layer FC1, followed by a dropout layer. FC2 maps the output of FC1 to the desired $1xC$ feature vector with a softmax activation.

Within the proposed structure, K serves as an additional hyper parameter, which can be used to adjust the count of trainable parameters of the activity classifier. This allows the optimization of the network in terms of recognition accuracy and power consumption as well as memory requirements, making it possible to adapt the network structure, to a certain extent, to the available embedded hardware.

Few shot adaption of the activity classifier

As shown by [220], a major challenge in human activity recognition is intraclass heterogeneity [206] leading to a degeneration of HAR accuracy. To counteract this degeneration, [220] proposed transfer learning to personalize the HAR classifier. [235] introduced few shot learning (FSL), a distinct form of transfer learning, to personalize a generalized HAR model.

We adapt this method to the **U-HAR** model by fetching a small fraction of samples per class from the left-out participant after training the **U-HAR** model using the LOPOCV scheme to personalize the decision boundaries of the activity classifier. During this model adaption phase, the weights of the convolutional layers of the encoder blocks and the decoder blocks and the first fully connected layer (FC1) are frozen. Furthermore, the dropout layer between FC1 and FC2 is removed and the weights of FC2 are reinitialized. Afterwards FC2 is retrained by a few epochs using a small set of samples per class stemming from the leave out participant. By freezing the weights of the encoder and decoder the model keeps its feature extractor build from all but the left out participant and learns a personalized activity classifier bound to the individual participant.

B.4.6 Evaluation

To evaluate the proposed **U-HAR** model as well as the effect of few shot learning (**U-HAR-FSL**), we split the dataset D into a test and train subset. to be precise, D is split into $D_{test} \in \mathbb{R}^{N_{test} \times T \times F}$ and $D_{train} \in \mathbb{R}^{N_{test} \times T \times F}$, with $N_{train} + N_{test} = N$. As we used LOPOCV D_{test} contain all samples of the left out participant. With 20 participants, this leads to 20 permutations of D_{train} and D_{test} for which the model is trained and tested separately. To rate the model's overall accuracy the macro F1-score is calculated across all permutations. The macro F1-score gives insights how well the model generalizes across participants and how well it classifies data from a unknown subject. Table B.12 gives an high level overview of the resulting F1-Score

Table B.12: Average macro F1-score over all participants for each activity for the investigated HAR models.

	SVM	U-HAR	U-HAR-FSL
talk	<u>73.03</u>	92.15	**97.20**
read	**91.97**	<u>77.74</u>	89.42
video	<u>66.89</u>	67.38	**78.60**
walk	<u>89.88</u>	95.59	**96.37**
type	<u>88.07</u>	**88.15**	86.89
solve	<u>20.16</u>	**66.57**	64.40
cycle	<u>77.99</u>	92.29	**95.87**
Overall	<u>66.86</u>	82.09	**86.59**

of all classifiers per activity as well as the overall macro F1-score.

For better comparison between the proposed **U-HAR** model and the state-of-the-art, we benchmark our dataset using a SVM as baseline model. Afterwards the **U-HAR** model without and with transfer learning **U-HAR-FSL** is used to classify activities from the dataset. To ease comparison between the three models, the performance of each classifier is visualized based on the confusion matrices in the upcoming sections stemming from three participants, selected to represent an overall low (P15), an average (P7) and a well (P13) performing participant.

Baseline model

[19] reported an F1-Score of 76% using a SVM model to predict human activities from statistical eye movement features extracted from EOG sensor readings. We extract similar to previous works [19, 221, 228, 229, 226], the following statistical features *L2-norm*, *variance* and *mean* from the temporal axis of each sensor reading to derive $3xF$ input feature vector from each sample. The SVM is trained on all but the the leave out participant and tested on the left out participant and classification is approached in a one-vs-all setting. As hyper parameters a RBF kernel function is used and we tune the regularization parameter to 1 to maximize classification performance of the SVM. Further optimization of the hyper parameters or the use of additional features derived from the time series data might lead to further improvements of the overall performance.
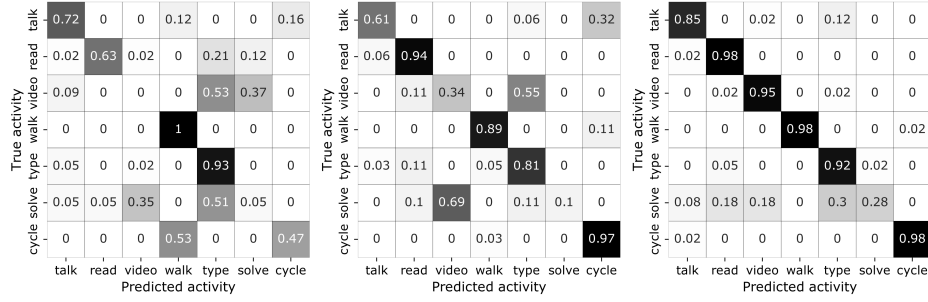
Figure B.32: Confusion matrices of P15 (macro F1-score 50.81 %), P7 (macro F1-score 59.53 %) and P13 (macro F1-score 82.13 %) obtained by the SVM classifier based on statistical features.

Figure B.32 shows three confusion matrices obtained from the selected participants to provide an overview of the classifier's performance with respect to the different activities. The macro F1-score of the SVM classifier is 66.86 % with an accuracy of 69.63 %, which is roughly 10 % below the reported performance from [19]. One possible reason could be that the activity set consists of cognitive as well as physical activities, and therefore, the activities are harder to recognize. Another reason could be that statistical features did not serve as an effective representation of the underlying patterns. Of particular note is that miss classification exists only in the subset of cognitive and physical activities for P15. More specifically, the activity *walk* is never miss classified as *read* while confusion exists for *walk* and *cycle*. The confusion matrices of participant P7 as well as the successfully performing participant P13 show confusion between *solve*, *read* and *type*, exhibiting that these classes share a common set of statistical features. This is a well-known problem in HAR and [206] denote it as interclass similarity.

U-HAR model

To get rid of statistical features and improve on the representation of time-dependent patterns, the dataset D is evaluated in the following using the proposed **U-HAR** model, introduced in Section B.4.5.
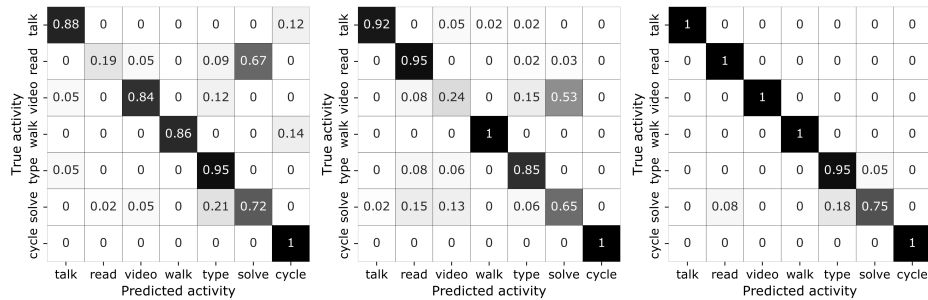
Figure B.33: Confusion matrices of P15 (macro F1-score 75.52 %), P7 (macro F1-score 75.52 %) and P13 (macro F1-score 95.59 %) obtained by the **U-HAR** model introduced in Section B.4.5 without few shot learning.

Similar to the SVM classification approach, the dataset is split into a training set consisting of all participants except the left-out participant, resulting in 20 permutations of training and test splits. For all 20 permutations, the model is trained over 6 epochs using cross-entropy-loss and Adam optimization with a learning rate of 4E-5 and a weight decay of 1E-6. Furthermore the learning rate was reduced by 1E-6 per epoch. In comparison with the SVM baseline model, the macro F1-score improves significantly by 15.23 %, reaching 82.09 % with a accuracy of 82.84 %. For the average pooling layer, we choose a kernel size of 10 ($K = 360$), effectively down sampling the segmented output to $360x7$, as a good trade off between network size and classification accuracy. Comparing the confusion matrices in Figure B.33 with the corresponding confusion matrices in Figure B.32, there is a clear improvement in recognition for the activity *talk* among the three participants. Furthermore, the physical activities *cycle* and *walk* are well recognized among all participants as well. As with the SVM model for participants P15 and P7, there still exists confusion between the cognitive activities *read*, *solve* and *video*, which hints to interclass similarity.

U-HAR model with few shot learning
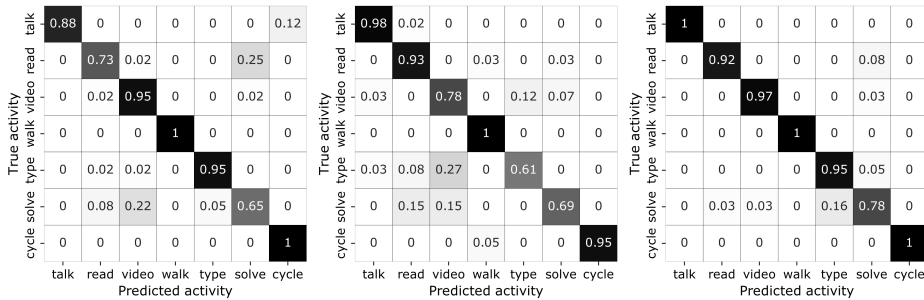
Figure B.34: Confusion matrices of P15 (macro F1-score 89.87 %), P7 (macro F1-score 84.74 %) and P13 (macro F1-score 94.57 %) obtained by the U-HAR model after adapting the decision boundaries of the activity classifier block by applying few shot learning.

To further improve the accuracy of activity recognition, we apply few shot learning as proposed in Section B.4.5 to the activity classifier. For this purpose, we randomly select three samples (windows of 30 seconds) for each activity class of the omitted participant and retrain the last fully connected layer over 10 epochs using a learning rate of 1E-4 and an exponential learning rate decay of 0.99. Figure B.34 shows the three confusion matrices derived from the chosen participants after few shot learning. The macro F1-score increased to 86.59 % and the mean accuracy increased to 86.99 %. Across all three participants, the recognition of the activities *read*, *solve* and *video* improves. As we personalize the activity classifier to some extent by adjusting the decision boundaries through few shot learning, the interclass similarity observed in Figure B.32 and Figure B.33 resolved.

Model comparison

To compare the performance of the classification approaches, we analyze the performance of the three models across all participants in the experiment. Table B.13 summarizes the model performances for each participant. In nearly all cases, the SVM classifier performs worst leading to the overall lowest macro F1-score, indicating that the used statistical features does not lead to a good representation of the underling temporal patterns to recognize activities. The SVM model reaches similar performance as related work by [227]. The **U-HAR** model improves the performance as it automatically extract features to represent the underling temporal patterns much better and due to the convolutional layers it is capable of to also gather information from the temporal sequence of incoming data. The macro F1-score increases to be comparable with related work e.g. [226]. The additional personalising of the activity classifier through transfer learning, as mentioned in Section B.4.5, an additional increase in activity recognition performance is observable. The proposed **U-HAR** model combined with few shot learning outperforms the related work e.g. by [226, 227]. It also reaches similar activity recognition performance when compared to related work that solely focus on activities from the physical domain recognized by IMUs e.g. by [218]. To analyze which activity profits from

Table B.13: Macro F1-score of the baseline model, the **U-HAR** as well as **U-HAR-FSL** model over all participants. The highest achieved macro F1-scores for each participant over all models are bold marked while the lowest F1-scores are underlined.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	∅
SVM	56.22	47.77	57.84	62.45	69.87	88.89	59.43	80.87	63.02	72.62	62.44	82.12	60.47	50.81	68.97	67.56	73.3	78.47	57.70	76.35	66.86
U-HAR	75.40	80.31	83.13	74.69	76.66	93.67	78.71	81.78	80.55	**79.99**	77.56	**95.59**	81.98	75.52	77.98	91.65	**83.36**	81.54	**79.12**	92:59	82.09
FSL	**91.59**	**88.28**	**87.11**	**78.56**	**89.47**	**95.76**	**84.74**	**87.96**	**87.96**	71.30	**77.67**	94.57	**86.20**	**89.87**	**86.03**	**95.20**	75.03	**93.91**	77.63	**95.04**	**86.59**

each model design we further calculate per activity and model the average macro F1-score over all participants. In Table B.12 the results are summarized. The physical dominated activities *walk* and *cycle* show together with the mixed activity *talk* the an overall high accuracy over all HAR models. Especially the *solve* activity improves by automatic extraction of features while taking into account the sequence data. The few shot adaption scheme improves the recognition of *read* and *video*, which are cognitive activities mainly described by eye movement features. The improvement through few shot adaption indicates that in particular for this classes intraclass heterogenity is dominating. Especially for reading intraclass heterogenity is well known as each individual has a slightly different scan pattern during reading mainly influenced how trained an individual is on this task. Furthermore a high interclass similarity exists for the activities *video* and *solve*, which leads to a high confusion between the activities and thus a low accuracy. To improve on this additional features are required. This can either be derived from the existing sensors e.g. pupil diameter variation or blinks as used by [226] or by adding additional senor modalities like an microphone as proposed by [220]. Another potential improvement is to step away from fine grained activities to classification on a higher contextual hierarchy like proposed by [222] e.g. classify activities into physical context or cognitive context to control the glasses UI based on this higher context levels. Finally the **U-HAR** model provides in the final stage of the decoder as intermediate representation a

sample wise activity classification. From this representation further information can be drawn e.g. to distinguish composed activities containing multiple sub activities. Furthermore inter class similarity could be investigated in at this stage.
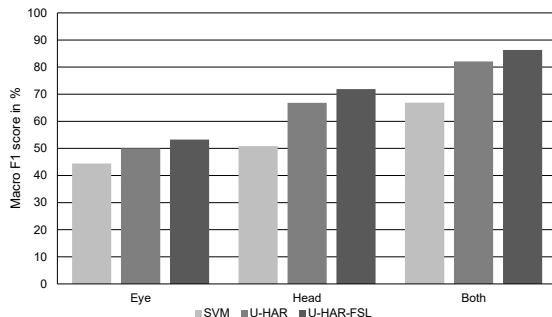
Figure B.35: Effect of eye related features (pupil position) and head related IMU features on the macro F1-score.

Impact of the sensing modalities

To analyze effects of both sensing modalities, we conducted a study on the importance of these sensor modalities. More specifically, the three models (i.e., SVM, **U-HAR** and **U-HAR-FSL**) are retrained with either the head movement related features (IMU) or the eye movement related features (VOG). Figure B.35 summarizes the resulting macro F1-scores after retraining the three models with only head or eye movement features. Our results in Table B.14 show that a high classification score is only obtained if both head and eye movement features are taken into account. At least for activity sets which contain physical and cognitive activities the use of both sensor modalities is mandatory. The head movement features have a higher impact on classification accuracy compared to the eye movement features for all three classifiers. Furthermore it is shown that the high classification accuracy of cognitive activities *read* and *video* is mainly driven by the eye movement features while the mainly physical activities *walk* and *cycle* are mainly classified by the head movement features.

Table B.14: Averaged resulting macro F1-scores per activity when solely considering head movement features or eye movement features and their combination. For each classifier and sensor modality, both the lowest and highest F1-scores are highlighted.

	SVM			U-HAR			U-HAR-FSL		
	Eye	Head	Both	Eye	Head	Both	Eye	Head	Both
talk	24.05	86.15	73.03	30.38	90.96	92.15	31.92	97.21	97.20
read	**92.84**	14.07	**91.79**	**80.59**	28.43	77.74	**88.45**	27.30	89.42
video	73.85	58.49	66.89	40.26	77.25	67.38	51.86	78.04	78.60
walk	68.14	**88.11**	89.88	65.36	**94.61**	**95.59**	58.15	96.54	96.37
type	72.14	51.18	88.07	76.94	71.35	88.15	68.78	84.43	86.89
solve	8.45	14.46	20.16	52.59	35.29	66.57	42.55	38.39	64.40
cycle	11.71	75.67	77.99	25.22	91.90	92.29	45.93	**95.87**	**95.37**

B.4.7   Discussion and limitations

As shown in Section B.4.6, the combination of eye and head motion features leads to higher
HAR accuracy. Therefore, the proposed approach of combining eye and head movements and
fusing them through the **U-HAR** network enables HAR-based context awareness for smart
glasses. In order to apply the proposed system in a real application, several aspects needs to
be considered, which are discussed below.

Sensor integration

While integrating IMU sensors into smart glasses is straightforward as they are already inte-
grated into every smart wearable, integrating eye-tracking sensors seems to be more challeng-
ing. The power consumption of VOG eye tracking sensors is magnitudes higher compared
to IMU sensors, and the size and the position requirements of the video cameras in order to
obtain robust pupil detection over the whole FOV of an participant are challenging, as shown
in Figure B.30. To overcome the limitations of power consumption and sensor integration,
[101, 103] and [129, 128] proposed new eye tracking sensor approaches to replace the camera
sensor and facilitate integration by using scanned infrared lasers to track a person's eye.

Power consumption

The power consumption of smart glasses is dominated by the power consumption of the
display, similar to smartphones. Therefore, a smart glass system directly benefits from a
reduction of the display's duty cycle. The contextual control of the display proposed in this
paper is a possible solution to this issue, as it allows the display to be activated contextually.
This implies an always-on HAR with continuous classification of activities by the **U-HAR**
model. Besides power memory is a further constrained resource on an embedded processor
on the glasses. However, as the **U-HAR** model mainly rely on convolutional layers, energy
efficient deployment of the model e.g. by using tensor processing units (TPUs) or dedicated
DNN accelerators is possible [83], [82]. The remaining constrain is the small available memory
on embedded devices. The memory footprint of the **U-HAR** model is mainly determined
by the parameters of FC1 and FC2. With an average pooling kernel size of 10 ($K = 360$), this
two layers hold 6.373.087 parameters, which are 98.61% of the total network's parameters. A
straightforward way to decrease this rather large number of parameters is the adoption of
the hyper parameter K. In fact ($K = 50$), results in a kernel size of 72 of the average pooling
layer, reduces the parameters by a factor of 24.6 while keeping a sufficient macro F1- score of
80.53 % . Further optimization of the model architecture e.g. by improving the 1D CNN layers
e.g. by adaption of bottleneck layers as proposed by [244] or by model pruning or other model
optimization techniques are still possible [238] to reduce model size and thus computational
complexity. Furthermore the IMU features can be fused into a 3D orientation vector of the
head to further reduce network parameters.

Few shot learning

As suggested by related work to improve classification performance and handle intraclass heterogeneity we used few shot learning to enhance our **U-HAR** model. This require to obtain labels from the user to adapt FC2 and thus personalize the activity classifier. The label collection may degrade user's experience although it is a known operation from other wearables e.g. the initial configuration of face-recognition in a smartphone [240].

Privacy

The proposed **U-HAR** system relies on privacy preserving sensor features as no direct link between the features and the individual exists. If inference of contextual information occurs on the edge, as proposed in the previous sections, no raw potential private data needs to be transmitted. Furthermore the proposed system did not require a world camera sensor to derive contextual information, which adds on the social acceptance of such a glasses system.

B.4.8 Conclusion

In this work, we present an approach which enables context-awareness through human activity recognition for smart glasses. We built a research apparatus and conducted experiments with 20 participants to record and published a unique activity dataset. Our dataset contains a distinctive combination of seven activities stemming from physical and cognitive domains. We recorded eye movement features using a VOG sensor and head movement features using an IMU sensor for all activities leading to a total recording duration of 1514 minutes. We further adapted the U-Time CNN model architecture proposed by [7], applied it to human activity recognition and proposed the **U-HAR** model structure. We advanced the network architecture to enable few shot adaptation to personalize the activity classifier to each individual and counteract accuracy losses due to intraclass heterogeneity. With the few shot adaptation approach, we achieved a macro F1-score of 86.59 % using LOPOCV. This result significantly outperforms related work. As part of our future work, we will focus on deployment of the classifier in a smart glasses prototype. In addition, we like to investigate how to gather labels from users in a every-day setting to apply the proposed few shot adaption to increase classification performance and thus minimize missclassification to optimize usability. Finally we like to investigate how a user can extend the classifier with its personal activities in a on device setting.

B.4.9   Supplement Material
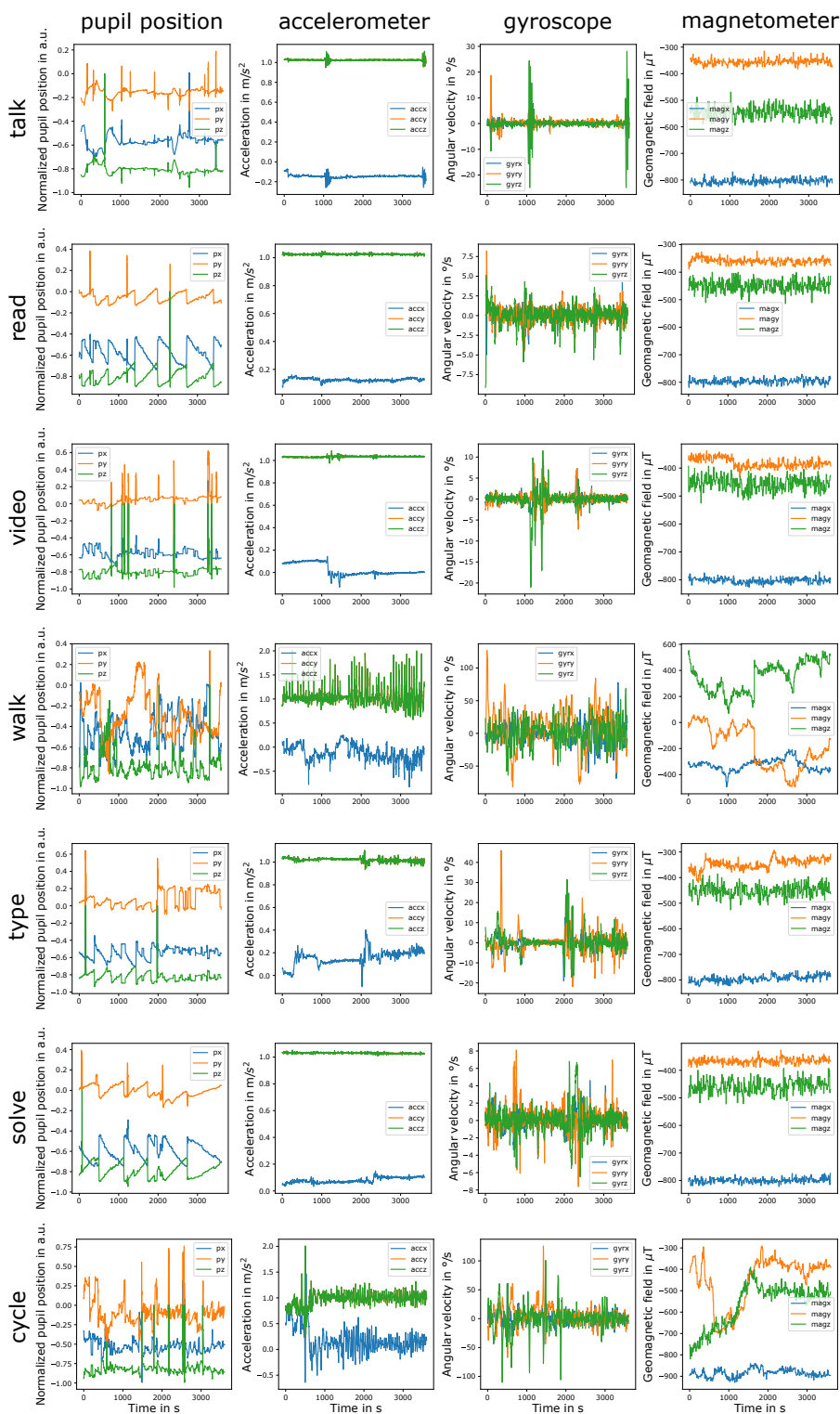
Figure B.36: 30 second windows of the raw pupil position and IMU features of participant P1 for the seven activities

Table B.15: Duration in minutes of each individual experiment itemized by activity.

	NULL	talk	read	video	walk	type	solve	cycle	total
P1	1	4	6	11	16	9	6	7	59
P2	6	8	7	11	14	18	12	5	79
P3	6	12	9	10	13	28	17	7	103
P4	11	6	13	11	10	13	35	5	104
P5	4	9	8	10	11	17	12	5	78
P6	4	5	7	10	14	13	7	6	67
P7	8	14	11	10	8	22	18	6	96
P8	5	6	7	10	10	9	7	7	60
P9	6	5	10	11	8	14	13	7	74
P10	4	7	5	10	11	14	12	5	68
P11	2	7	6	10	12	7	11	6	61
P12	3	9	5	10	14	6	14	6	68
P13	4	6	9	10	14	13	11	5	72
P14	6	4	12	11	10	15	12	5	75
P15	5	5	11	10	14	12	16	6	78
P16	4	4	8	10	13	13	12	6	71
P17	5	6	7	10	13	22	14	9	87
P18	6	8	8	10	10	11	13	5	72
P19	5	5	11	11	8	15	16	5	75
P20	4	5	5	10	14	13	9	6	67
total	98	137	165	209	237	284	266	119	1514

C Static Laser Feedback Interferometry Eye Tracking

This chapter contains the manuscript:

I. **Johannes Meyer**, Stefan Gehring, Enkelejda Kasneci. "Static Laser Feedback Interferometry based Gaze Estimation for Wearable Glasses". Submitted to IEEE Transactions on Systems, Man, and Cybernetics: Systems (2022)

C.1 Static Laser Feedback Interferometry based Gaze Estimation for Wearable Glasses

C.1.1 Abstract

Fast and robust gaze estimation is a key technology for wearable glasses, as it enables novel methods of user interaction as well as display enhancement applications, such as foveated rendering. State-of-the-art video-based systems lack a high update rate, integrateability, slippage robustness and low power consumption. To overcome these limitations, we propose a model-based fusion algorithm to estimate gaze from multiple static laser feedback interferometry (LFI) sensors, which are capable of measuring distance towards the eye and the eye's rotational velocity. The proposed system is ambient light robust and robust to glasses slippage. During evaluation a gaze accuracy of $1.79°$ at an outstanding update rate of $1\,kHz$ is achieved while the sensors consume only a fraction of the power compared to state-of-the-art video-based system.

C.1.2 Introduction

Eye tracking is a key sensing technology for wearable glasses required to enable applications ranging from gaze contingent interaction [161, 18, 5] to display enhancement methods like foveated rendering [38, 39, 40] or exit pupil steering[42, 43]. State of the art eye tracking systems rely on video oculography (VOG), which is a camera-based system to infer the users

gaze from a sequence of images. However, due to power constraints imposed by the wearability of these glasses, the update rate of VOG systems is limited by the high power consumption of the camera sensors and the image processing algorithms required to estimate user's gaze from captured images [84]. Especially see-through AR glasses further require an eye tracking system, which operates robustly in the presence of ambient light [56]. This further limits the applicability of VOG systems. Finally, VOG systems are prone to glasses slippage [92], possibly leading to a degradation of gaze estimation accuracy [93, 57].

To overcome these limitations of VOG systems, we introduce a static laser feedback interferometry (LFI) sensor approach. The LFI sensor itself consists of a tiny (160 μm × 180 μ m) vertical cavity surface emitting laser (VCSEL) with infrared (IR) light at 850 nm (invisible to the user) and a photo detector integrated into the laser cavity [2]. By applying a frequency modulated continuous wave (FMCW) modulation scheme [125], the sensor is capable to measure the distance d towards the eye as well as the eye's rotational speed in beam axis v_T simultaneously with an outstanding update rate of 1 kHz, while consuming only a fraction of the power of VOG systems. Due to the coherent sensing scheme, the sensor in addition is robust against ambient light, as shown in [128]. Our contribution in this work is two-fold:

(i) We introduce the **static LFI sensor** modality for gaze estimation, characterize a static LFI sensor with respect to distance and velocity resolution in a near-eye setting, propose an geometric eye model required for sensor fusion, and build up a simulation tool to generate measurements of multiple static LFI sensors.

(ii) We propose a slippage robust, calibration-free **gaze estimation algorithm**, fusing multiple static LFI sensors in order to reconstruct the gaze vector based on a geometric eye model.

The remainder of this work is organized as follows: Appendix C.1.3 gives an overview over model-based eye tracking approaches of VOG systems as well as IR laser-based eye tracking approaches with focus on LFI sensors. Afterwards in Appendix C.1.6, the LFI sensing principle is introduced together with a geometric eye model in order to link LFI sensor measurements to an eye pose. In addition, a LFI simulation tool is introduced to generate LFI sensor measurements in a multi LFI sensor setting. In Appendix C.1.7, the gaze reconstruction algorithm is introduced to fuse multiple LFI sensors for gaze estimation. In Appendix C.1.8, we characterize distance and velocity noise of the LFI sensor and evaluate the gaze estimation accuracy of the proposed system in presence of glasses slippage. In the final sections, system limitations are discussed, and a final conclusion is drawn.

C.1.3 Related Work

Related work is divided into two parts. During the first part, geometric model-based gaze estimation algorithms used in VOG systems are introduced to highlight the differences between existing geometric model-based gaze estimation algorithms and our novel approach. In the second part, related work using the LFI sensor in a near-eye setting is presented.
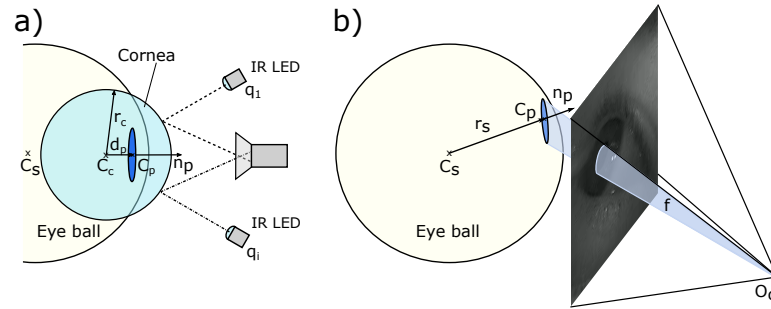
Figure C.1: a) Corneal reflection eye model based on [8]. b) Glint-free eye model based on [9].

C.1.4   Geometric model-based Eye Tracking

Geometric model based eye tracking approaches used in VOG systems can be divided into corneal reflection eye models and glint-free eye models [64]. Corneal reflection models, as shown in Figure C.1a), estimate the gaze vector, described as normal vector of the pupil n_p, by extracting the center of the pupil P_c from a camera image and estimate the center of the cornea C_c. To estimate the cornea center, IR LEDs (q_1 - q_n) are added to the system. The diverging IR light of the IR LEDs leads to specular reflections on the surface of the cornea, so-called glints. The positions of the glints on the cornea are extracted from the camera images in order to estimate the cornea center C_c [63]. According to Hennessey et al. [8] at least two glints are required to estimate the cornea center under the assumption that the system geometry is fixed and known in advance. For reconstruction of the cornea center C_c furthermore the radius of the cornea r_c, the distance between pupil and corneal center d_p and the refraction index of the aqueous humor need to be known or derived from average population [8].

In contrast, glint-free models rely solely on a camera sensor as shown in Figure C.1b) [9]. To estimate the gaze vector n_p, the center of the sclera S_c and the center of the pupil disk P_c in space need to be estimated from camera images. Assuming a pinhole camera model and a projected ellipse representing the pupil in the image plane, the pupil disk with its center P_c is derived from a cone projection f in space. The cone f is constructed from the pupil ellipse representation in the image plane and the camera focal point o_c. As the pupil diameter varies with the portion of ambient light, and the diameter of the pupil ellipse representation depends on the distance between camera sensor and the eye, several singularities exist, which prevent the estimation of the pupil disk and thus the gaze vector from a single image. By introducing geometric constraints like a fixed pupil diameter, possible pupil disk candidates fitting to the projected cone can be derived. By adding further constraints, the correct pupil disk can be estimated from a sequence of frames [9].

Under the assumption that the eye ball diameter r_s is fixed and known in advance, the center of the eye ball S_c is derived from a sequence of pupil disk observations, assuming a stationary relation between camera position and eye ball center S_c. This assumption holds for a short time period, while for longer time periods, glasses tend to slip on the nose and thus the model

parameter S_c needs to be updated continuously [66, 93].

In summary, VOG eye tracking approaches rely on a reduced anatomical model of the human eye with minimum physical dimensions and optical quantities in order to link the sensor features to an eye pose. In addition, fixed model parameters derived from the average population are used as additional system constraints. Hennessey et. al. [8] assume a known system geometry and a rigid glasses frame with mounted IR LEDs and the camera sensor. Swirski et. al. [9] further assume a stationary fixed relation between eye coordinate space and glasses coordinate space. In addition, both model-based approaches estimate the optical axis of the eye as normal vector n_p of the pupil.

C.1.5 Static LFI eye tracking

Static LFI sensors are well known from literature and are used in a wide range of applications, like vibrometry or velocimetry [121, 125]. The first work which proposes a static LFI sensor to measure eye movements was released by Capelli et al. [245]. The authors used a 1310 nm IR laser with an optical power of 4 mW to measure velocity of a rotating disk to simulate eye rotational velocities. Measurements on a real human eye were not performed, as the laser source did not fulfill eye safety requirements. The authors apply a triangular current modulation scheme to the drive current of the laser to be able to measure rotational velocities from 0 °/s to 500 °/s at a sample rate of 60 Hz. They reported a velocity measurement error of 10 %.

Meyer et al. [128] used an unmodulated LFI sensor together with a 2D micro scanner to capture reflectivity images from the eye and to measure the bright pupil response. They reported an outstanding ambient light robustness of LFI sensors due to the coherent sensing scheme of the sensor.

Meyer et al. [4, 5, 133] were also the first to create an eye safety compliant static LFI sensor system and reported measurements on a real human eye. They used two LFI sensors at a wavelength of 850 nm with an optical power below 500 μW to measure distance towards the eye as well as relative eye velocities with an update rate of 1 kHz. The authors restricted their velocity measures to relative eye velocities, because the static LFI sensor was only capable to measure speed in direction of the laser beam axis v_T and thus rotational movements with perpendicular components could not be measured with their setup.

This work builds up on previous works by Meyer et. al. [128, 4, 5, 133] to apply static LFI eye sensors in near-eye setting to develop a high-speed gaze estimation method based on LFI sensor technology.
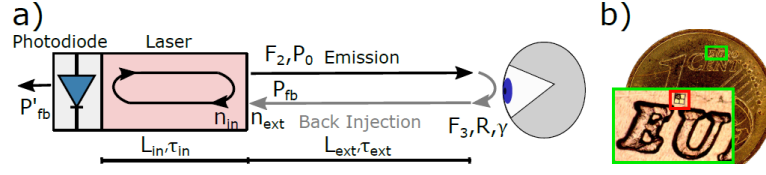
Figure C.2: Left: Sensing principle of an LFI sensor described by the coupled cavity model [6]. Right: LFI sensing element on a coin [4, 5].

C.1.6 Laser Feedback Interferometry Eye Tracking

The basic sensing principle of an LFI sensor can be described by the coupled cavity model as shown in Figure C.2. A single mode VCSEL emits coherent light with optical power P_0 through the laser facet F_2. The light propagates along the optical axis of the laser until it hits a target (the eye) at distance L_{ext}. Dependent on the target's reflectivity R, a portion of the emitted light is back-scattered and injected back into the laser cavity after the round trip time τ_{ext}. The round trip time is dependent on the external refractive index n_{ext} and the speed of light c_0.

The back injected light interferes with the local oscillating field of the laser, leading to a perturbation of the lasers optical power through constructive or destructive interference resulting in a modulation of the optical power P_{fb}

$$P_{fb} = P_0 \left(1 - m \cos\left(\phi_{ext}\right)\right). \tag{C.1}$$

The modulation of optical power depends on the modulation index m, which attributes an amplitude modulation, and the phase modulation ϕ_{ext} through the cosine term. A small fraction of the modulated power P'_{fb} is measured by a photo detector integrated into the laser.

According to Taimre et al. [6], the external phase ϕ_{ext} is linked via the signal phase ϕ_0 of the unperturbed laser with the desired observations (distance and velocity) by

$$\phi_{ext} - \phi_0 + C \sin\left(\phi_{ext} + \arctan\left(\alpha\right)\right) = 0 \tag{C.2}$$

where C describes *Acket's feedback parameter* and α *Henry's linewidth enhancement factor* [120]. Considering operation of the LFI sensor in the weak feedback regime ($C < 1$) and α to be constant Equation (C.2) has only a single solution and ϕ_{ext} is directly linked to

$$\phi_0 = \frac{4\pi n_{ext} L_{ext}}{\lambda}, \tag{C.3}$$

with laser wavelength λ [6]. Assuming operation in free space ($n_{ext} \approx 1$), frequency modulation of optical power through a modulation of phase ϕ_0 requires either a change in distance L_{ext} or a change of the laser's wavelength λ w.r.t. time. To separate both effects, the partial derivatives
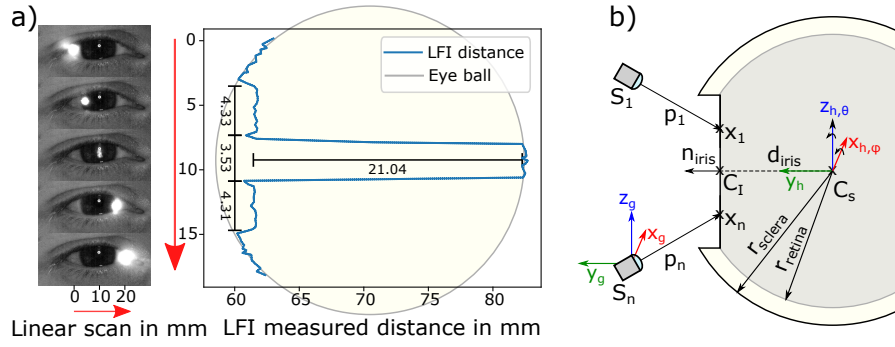
Figure C.3: Left: Images from reference camera. Center: Distance measurements of the LFI sensor for different positions on the eye's surface. Right: Derived geometrical system model.

$\frac{\partial \lambda}{\partial t}$ and $\frac{\partial L_{ext}}{\partial t}$ of Equation (C.3) are calculated leading to

$$f_0 = \frac{2 L_{ext}}{\lambda} \frac{d\lambda}{dt} \tag{C.4}$$

and

$$f_d = \frac{2 v_{ext} \cos(\gamma)}{\lambda}, \tag{C.5}$$

with f_0 as distance corresponding beat frequency, f_d as velocity corresponding Doppler frequency, v_{ext} as eye velocity and γ as incident angle between eye and laser beam [125].

A modulation of the laser's wavelength is achieved through modulation of the laser's drive current while a modulation of the L_{ext} is achieved through a moving target. To measure target distance as well as the target velocity simultaneously, a triangular modulation pattern of the laser drive current is applied similar to FMCW LIDAR systems [126]. Therefore the update rate of the LFI sensor directly corresponds to the triangle modulation rate, which can reach up to 100 kHz for spatial confined semiconductor VCSELs [125].

Geometric Eye Model

Similar to the discussed model based eye tracking approaches in Appendix C.1.3, we derive a geometric model to link the LFI sensor measurements, in particular the distance measures, to the pose of the eye and thus the gaze vector. To derive our eye model, a linear scan across the surface of a healthy human eye is made while the eye is fixating on an object. Figure C.3 a) shows images taken by an IR reference camera while the IR laser beam[1] (white spot) was scanned over the participants eye. The scan direction is indicated by the red arrow. For each scan position, the measured distance between laser and eye surface is shown. The distance measurement follows the curvature of the sclera until it reaches the limbus. At this point, the beam penetrates the cornea and hits the iris. If the beam hits the pupil, light is scattered back from the retina and thus the distance towards the retina is measured. The measured

---

[1] Class 1 laser system according to IEC 60825-1

total diameter of the iris is 12.17 mm, which fits well to average population [246]. From this observation we derived a geometric model as shown in Figure C.3 b). The iris is modeled as plane inside the eyeball with a distance d_{iris} between the center of rotation C_S and the iris center C_I. The eye ball is modeled as two spheres, one sphere with radius r_{sclera} to link measured distances from the outer eye to the model, and a second sphere with radius r_{retina} to link measured distances from the inner eye to the model. In addition, the LFI sensors are modeled as point sources with a pose S_n and a linear laser beam, hitting the eye at an intersection point x_n. The gaze vector is described as normal vector n_{iris} perpendicular to the iris plane. Similar to the geometric eye model proposed by Hennessy et al. [8], d_{iris}, r_{sclera} and r_{retina} are considered as fixed parameters of the model, which are derived from average population.

Multi LFI Sensor Simulation tool
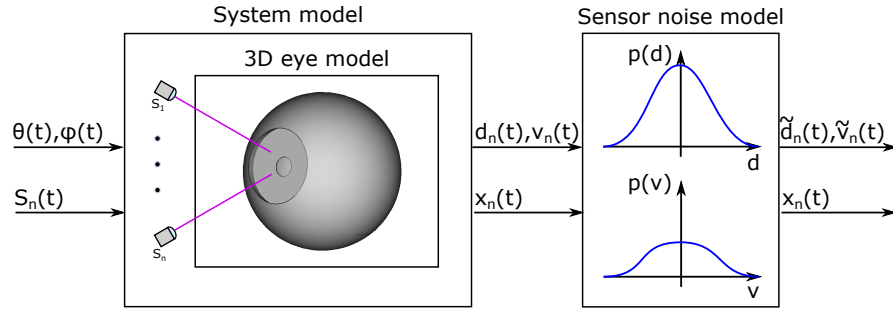
Figure C.4: Block diagram of the LFI simulation tool used to generate LFI measurement data without human error.

In order to evaluate the proposed LFI gaze estimation system, we set up a simulation tool to generate sensor data from arbitrary sensor poses S_n and gaze trajectories denoted by the tuple (θ, ϕ). Compared to a laboratory setup, the simulation allows to evaluate the gaze estimation accuracy while excluding human error e.g. through noise. Thus a simulated environment allows for systematic investigation of the different algorithm steps on gaze accuracy.

Input of the simulation is a trajectory of eye positions described by a tuple of gaze angles $(\theta(t), \phi(t))$ and the sensor pose $S_n(t)$ of n sensors. The sensor pose S_n of the n-th sensor consists of a sensor position s_n and a direction p_n describing the propagation direction of the laser beam in the glasses' coordinate space. The sensor pose can vary over time to investigate the effect of glasses slippage on gaze estimation accuracy of the proposed system.

The input parameters are used to setup our system simulation inside a CAD program. The sensors are modeled as point sources in space with a laser beam propagating towards the 3D eye model. The parameters of the eye model e.g. r_{sclera}, r_{retina} or d_{iris} are derived from a 3D scan of an human eye [247]. To fit the 3D eye model to the proposed geometric eye model derived in Appendix C.1.6, the cornea and the lens are removed.

During the simulation, the eye is rotated by $\theta(t)$ and $\phi(t)$ while the sensor pose is updated for each time step t sequentially according to $S_n(t)$. At each time step, the beam of each LFI sensor is propagated from the laser position s_n along the laser beam direction p_n to calculate the intersection point $x_n(t)$ with the 3D eye model. The corresponding distance measure $d_n(t)$ is calculated as the Euclidean distance between sensor position and intersection with the eye. In addition, based on the intersection point $x_n(t)$, the region of intersection (none, sclera, iris, retina), which is hit by the laser beam, is extracted. The barely occurring region *none* summarizes distance measurements not stemming from the eyeball e.g. distance measurements from the lashes or the eye lid during a blink. To calculate the velocity $v_n(t)$ measured by the LFI sensor S_n, the rotation axis $\omega(t)$ is calculated from the difference between two gaze trajectory points. Afterwards, the velocity is calculated, e.g., for the sclera by

$$v_n(t) = S_n(t) \cdot (\omega(t) \times x_n(t)). \tag{C.6}$$

However, the system model does not take into account sensor noise e.g. due to speckling [248] leading to ideal LFI measurements, leading to ideal senor measures. Therefore we add a sensor noise model to the simulation tool to incorporate sensor noise described by $\mathcal{N}(0, \sigma_d^2)$ and $\mathcal{N}(0, \sigma_v^2)$. The parameters σ_d^2 and σ_v^2 of the sensor noise model are derived from sensor characterization measurements. The final output of the simulation is a set of LFI sensor measurement for each sensor and each time stamp, which is used as input for the gaze reconstruction algorithm.

C.1.7   Gaze Reconstruction Algorithm

The general structure of the gaze estimator consists of three individual stages:

A *Region of Intersection Classification*: For each laser, determine the region of intersection with the eye (none, sclera, iris, retina).

B *LFI-Sensor Position and Pose Estimation*: From measurements of all lasers with known class-labels, estimate the sensor position and orientation of the lasers.

C *Gaze Angle Estimation*: For known sensor positions and with all labeled measurements, estimate the gaze.

In the following sections, each of the three stages is detailed.

Region of Intersection Classification

For the classification, one crucial question is which signal(s) may be used for *robust* classification. From the measurements shown in Figure C.3 a) it is readily seen, that by means of the distance signal, one could easily differentiate between the intersection point being on
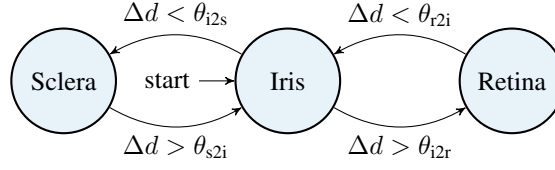
Figure C.5: Deterministic automaton on difference of distance signal for naive classification.

the sclera or on the retina. On the other hand, classification solely based on distance would require some threshold value, which might have to be adapted in case of slippage.

Taking the time difference Δd_n of the distance signal could reveal useful information about the laser position. Because the outer shape of the eye contains discontinuities, these will affect Δd_n significantly. Therefore, a possible naive classifier could be the deterministic automaton shown in Figure C.5, operating solely on Δd_n.

The drawbacks of such automaton are that absolute distance information is neglected completely, and its susceptibility to noise in the distance signal. Also, the transition from sclera to iris will cause a much weaker signal in Δd_n compared to a transition from iris to retina.
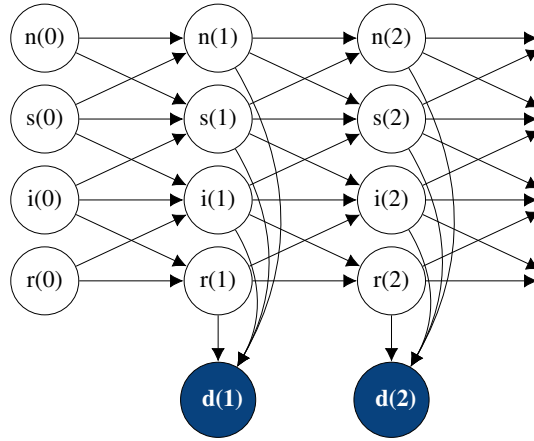
Figure C.6: Hidden Markov Model for classification, depicted as lattice diagram.

We therefore propose to use both signals d_n and Δd_n for classification in a probabilistic framework. To do so, we set up a Hidden Markov Model (HMM) (see, e.g. [249]), which is depicted in Figure C.6. Therein, the hidden (latent) states $z(k) = \begin{bmatrix} n(k) & s(k) & i(k) & r(k) \end{bmatrix}^T$ represent the intersection region (none, sclera, iris, retina) at time k with a 1-of-K representation, i.e., $z_i(k) \in \{0, 1\}$ and $\sum_i z_i(k) = 1$. We assume a multinomial distribution for z with $\mathbb{E}[z_i(k)] = \pi_i$ and $p(z) = \Pi_i \pi_i^{z_i}$.

For the *emission probabilities*, we assume normal conditional distributions

$$p(d \mid z) = \prod_{i=1}^{4} \mathcal{N}\left(d \mid \mu_i, \sigma_i\right)^{z_i} \tag{C.7}$$

211

of the measured distance d given a latent state z. Therein, μ_i, σ_i are treated as unknown parameters, which are to be inferred online. This resembles the fact that the absolute position of the laser sensors are unknown a-priori, and might even change over time due to slippage. With Equation (C.7), the joint distribution of latent and observed state at time instant k follows as Gaussian mixture distribution

$$p(d) = \sum_z p(z)p(d\,|\,z) = \sum_{i=1}^{4} \pi_i \mathcal{N}\left(d\,\middle|\,\mu_i, \sigma_i\right). \tag{C.8}$$

In contrast to the emission probabilities with unknown parameters, the *transition probabilities* $p(z_i(k)|z_i(k-1), \Delta d(k-1))$ can be specified well before-hand, if based on the measured difference of distance. This is mainly due to the fact that eye-size varies only weakly among healthy adults [250]. Hence, distance signal jumps during traversing of the laser signal from one eye part to another may be treated as quasi-constant quantity, akin to the automaton in Figure C.5. The strongest assumption in this regard is the spot size of the laser, which is not infinitesimally small. Hence, for slow eye movements or short sampling times, the laser might hit two neighboring eye-parts simultaneously, which might hinder transition detection.

The conditional probability of a latent variable given its previous value

$$p(z(k)\,|\,z(k-1)) = \prod_{i=1}^{4} \left(\sum_{j=1}^{4} a_{ij} z_j(k-1) \right)^{z_i(k)} \tag{C.9}$$

are assumed to be multinomially distributed with transition probabilities $A = \left[a_{ij}(\Delta d)\right]$ shown in Figure C.7. As can be seen, starting from the laser hitting the *none* region, zero or negative distance change favors the classifier to remain in the *none* region, whereas a significant positive distance change is likely due to a transition to sclera, iris or retina region dependent on the magnitude of the distance change. A direct jump from sclera to retina is given zero probability due to high sampling rate and limited eye movement. Same reasoning holds for the other transition probabilities.
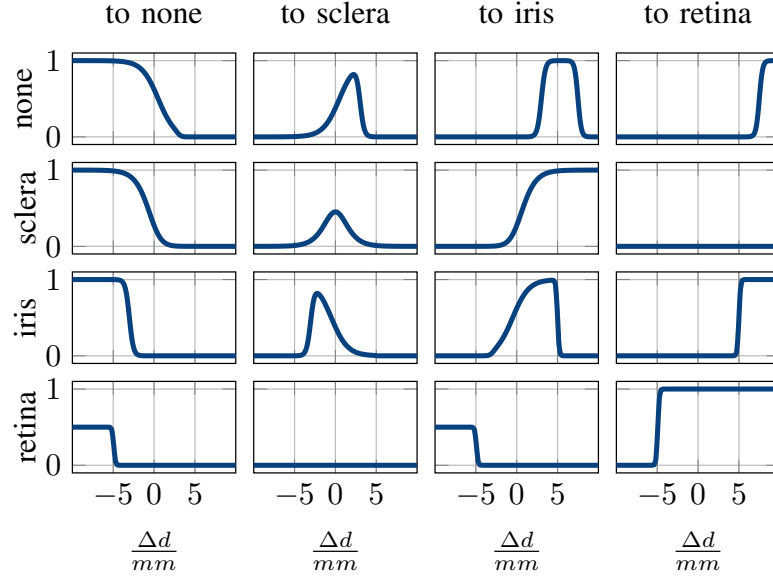
Figure C.7: Modeled transition probabilities.

For the HMM, (approximate) inference has to be carried out online for sequential data, i.e. measured distances $d(k)$ and difference of distances $\Delta d(k)$. We make use of a rather standard maximum likelihood-approach, namely Expectation Maximization (EM). A number of EM-like algorithms have been developed for the case of HMMs, e.g., the *forward-backward algorithm*, and the *Viterbi algorithm* (see [249] and the references therein). However, since the transition probabilities $A(k)$ are assumed to be known, and multinomial and normal distributions are used, the implementation is significantly simplified, which is derived next.

If the complete-data $\{D, Z\} = \{d(k), z(k)\}, k = 1, \ldots, T$, was known, inference of the unknown parameters $\theta = \{\pi_i, \mu_i, \sigma_i\}$ could simply be carried out by maximizing the complete-data log-likelihood function

$$\ln p(D, Z \mid \theta) = \ln \prod_k p(z(k) \mid z(k-1), A(k)) \cdot p(d(k) \mid z(k), \theta) =$$

$$\sum_k \ln p(z(k) \mid z(k-1), A(k)) + \ln p(d(k) \mid z(k), \theta) \quad \text{(C.10)}$$

However, since the latent variables Z are not known a-priori, an iterative EM scheme is used. During the *expectation-step*, the posterior $p(Z \mid D, \theta')$ over latent variables are computed for previous parameter set θ' and measurements D. From Equation (C.10), one may readily derive

the relation

$$p\big(Z \,\big|\, D,\theta'\big) \propto \prod_{k=1}^{T} p(z(k) \,|\, z(k-1)) \, p\big(d(k) \,\big|\, z(k),\theta'\big)$$

$$= \prod_{k=1}^{T} \prod_{i=1}^{4} \left(\pi_i \mathcal{N}\big(d(k) \,\big|\, \mu_i', \sigma_i'\big) \sum_j a_{ij}(k) z_j(k-1) \right)^{z_i(k-1)}. \quad \text{(C.11)}$$

Furthermore, by taking the expected value, the *responsibility*

$$\gamma(z_i(k)) \qquad := \qquad \mathbb{E}\left[z_i(k)\right] \qquad = \qquad \frac{\left(\sum_j a_{ij} z_j(k-1)\right) \pi_i \mathcal{N}\big(d(k) \,\big|\, \mu_i', \sigma_i'\big)}{\sum_h \left(\sum_j a_{ij} z_j(k-1)\right) \pi_h \mathcal{N}\big(d(k) \,\big|\, \mu_h', \sigma_h'\big)} \quad \text{(C.12)}$$

is obtained. Note that for each time instance k Equation (C.12) depends on the current data $d(k)$ and due to the transitions on the latent variable of the previous time step $z(k-1)$.

During the *maximization-step*, the latent variables are fixed in order to recompute the parameters θ. This is achieved by maximizing the log-likelihood function $\ln p(D \,|\, \theta)$, which is obtained by marginalizing over Equation (C.10) for Z, yielding the expectation w.r.t Z:

$$\mathbb{E}_Z\left[\ln p(D, Z \,|\, \theta)\right] = \sum_Z p(Z \,|\, D, \theta) \cdot \ln p(D, Z \,|\, \theta). \quad \text{(C.13)}$$

In order to make the computation tractable, θ' is used within posterior $p\big(Z \,\big|\, X, \theta'\big)$, yielding

$$\mathcal{Q}\big(\theta, \theta'\big) := \sum_Z p\big(Z \,\big|\, D, \theta'\big) \cdot \ln p(D, Z \,|\, \theta). \quad \text{(C.14)}$$

Thus, during iteration of the EM-scheme,

$$\lim_{\theta' \to \theta} \mathcal{Q}\big(\theta, \theta'\big) = \mathbb{E}_Z\left[\ln p(D, Z \,|\, \theta)\right]. \quad \text{(C.15)}$$

Substituting the complete-data log-likelihood Equation (C.10) into Equation (C.14) and marginalizing over Z,

$$\mathcal{Q}\big(\theta, \theta'\big) = \sum_{k=1}^{T} \sum_Z p\big(z(k) \,\big|\, d(k), \theta'\big) \ln p(d(k) \,|\, z(k), \theta)$$

$$+ \sum_{k=1}^{T} \sum_Z p(z(k), z(k-1) \,|\, D) \ln p(z(k) \,|\, z(k-1), D) \quad \text{(C.16)}$$

is obtained, which is to be maximized w.r.t θ. Apparently, the second summand does not depend on θ, hence only the first term has to be considered. Taking the derivative of Equation (C.16) w.r.t θ and solving for optimal parameter values yields known results for Gaussian

mixture distributions [249]:

$$\pi_i = \frac{N_i}{T} = \frac{1}{T} \sum_{k=1}^{T} \gamma(z_i(k)), \tag{C.17a}$$

$$\mu_i = \frac{1}{N_i} \sum_{k=1}^{T} \gamma(z_i(k)) \, d(k), \tag{C.17b}$$

$$\sigma_i = \frac{1}{N_i} \sum_{k=1}^{T} \gamma(z_i(k)) \left(d(k) - \mu_i \right)^2. \tag{C.17c}$$

These update equations for the M-step assume availability of the full data set $d(k), k = 1, \ldots, T$. They can however easily be rewritten as sequential update-equations for the online case:

$$N_i = \underbrace{\sum_{k=1}^{T-1} \gamma(z_i(k))}_{N_i'} + \gamma(z_i(T)) \tag{C.18a}$$

$$\pi_i = \frac{N_i}{T}, \tag{C.18b}$$

$$N_i \mu_i = \underbrace{\sum_{k=1}^{T-1} \gamma(z_i(k)) \, d(k)}_{N_i' \mu_i'} + \gamma(z_i(T)) \, d(T), \tag{C.18c}$$

$$N_i \sigma_i = \underbrace{\sum_{k=1}^{T-1} \gamma(z_i(k)) \left(d(k) - \mu_i \right)^2}_{N_i' \sigma_i'} + \gamma(z_i(T)) \left(d(T) - \mu_i \right)^2. \tag{C.18d}$$

Hence, only the statistics $N_i', N_i' \mu_i, N_i' \sigma_i'$ have to be stored, rather than the entire set of measurements growing over time.

The EM-procedure for an offline-available batch of data usually requires iteration between the E- and M-steps until a certain convergence criterion is fulfilled, e.g., the log-likelihood function or the parameters themselves.

Since we carry out the inference in an online fashion, it is justified to perform only one iteration with each new datum. If θ is quasi-constant over time and sample rate is high, convergence of the parameter estimation may be achieved over time.

Furthermore, the statistics Equation (C.18) are suitable in case of the parameters π_i, μ_i, σ_i being constant over time, since inherently, mean value computations are carried out in Equation (C.18). For the present application, however, parameters might vary over time in case of slippage of the wearable device. In this case, a certain *forgetting property* (see, e.g., [251], [252]), such that the most recent measurements have a more dominant impact on the estimated parameters compared to outdated measurements.

A constant exponential forgetting factor however would require a persistence of excitation.

For the considered application, this requires that each laser to collect measurements from sclera, iris and retina region sufficiently often in order not to completely forget any learned parameters.

We therefore modify the update equations in such a way that the old statistics are weighted by a forgetting factor λ_i depending on the current responsibility $\gamma(z_i(k))$ i.e., number of supporting samples for cluster i and a desired number N^* of past supporting samples, yielding:

$$\lambda_i = \min\left\{\frac{N^* - \gamma(z_i(k))}{N'_i}, 1\right\} \tag{C.19a}$$

$$N_i = \lambda_i N'_i + \gamma(z_i(T)) \tag{C.19b}$$

$$\pi_i = \frac{\lambda_i(T-1)\pi'_i + \gamma(z_i(T))}{\lambda_i(T-1) + 1}, \tag{C.19c}$$

$$\mu_i = \frac{\lambda_i N'_i \mu'_i + \gamma(z_i(T))\, d(T)}{\lambda_i N'_i + \gamma(z_i(T))} \tag{C.19d}$$

$$\sigma_i = \frac{\lambda_i N'_i \sigma'_i + \gamma(z_i(T))\left(d(T) - \mu_i\right)^2}{\lambda_i N'_i + \gamma(z_i(T))} \tag{C.19e}$$

Loosely speaking, these modified update equations collect N^* data points for each cluster and forget about previous ones. Hence, forgetting is explicitly designed for each cluster depending on the actual information, rather than having a constant forgetting independent of the cluster support.

Updating Equation (C.19) will infer the model parameters over time. In contrast to the emission probabilities, the transition probabilities a_{ij} are assumed to be given. This reasonable assumption leads to a low number of parameters to be inferred. One shortcoming of this approach however is the different treatment of sorting of clusters: The transition probabilities assume a sorting $\mu_1 < \mu_2 < \mu_3 < \mu_4$, such that the labeling $l_i\ z \rightarrow \left[\text{none, sclera, iris, retina}\right]$ becomes possible. This sorting however does not hold a priori for the inferred parameters of Equation (C.19). Depending on the initialization of $z(0)$, the means of emission distributions may not be sorted. A solution to this problem is to rearrange the clusters during each time step, such that their means have an ascending order.
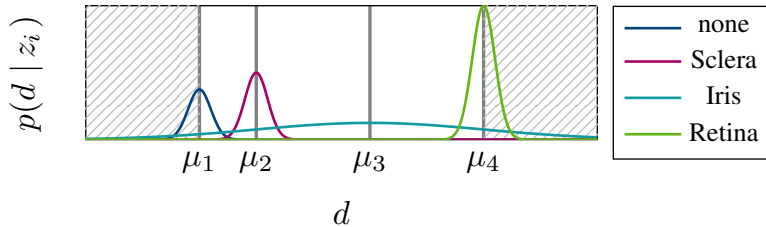
Figure C.8: Example normal distributions for conditional emission probabilities.

Another shortcoming of this classifier is its limitation to the interpolation region. Figure C.8 depicts an exemplary mixture of normal distributions defining the emission probabilities. For this known distribution, one can easily predict the most probable class for a measurement $d(k)$. However in the extrapolation region, shown as shaded area in Figure C.8, this might lead to miss-classification. Consider in this example $d \ll \mu_1$ or $d \gg \mu_4$. Then, one would have highest probability for d belonging to the iris, but it is a *none* or retina point. As an easy solution, clamping the measured distances to $d \in [\mu_1, \mu_4]$ when evaluating the conditional emission probability will lead to the desired behavior.

The classification is summarized as follows:

Algorithm 1 Classification Algorithm.

Init: N^*, z_0

  for k=1, ..., T **do**

      •   get measurements $d(k), \Delta d(k)$

      •   update transition probability matrix $A(\Delta d(k))$

      •   update statistics Equation (C.19)

      •   sort clusters depending on μ_i in ascending order

      •   classify measurement $d(k)$ by means of Equation (C.12) with d limited to the interpolation region.

  end for

As a result, one obtains labels for each measurement, and in addition a vague model of the eye.

This information is exploited in following steps, namely estimation of the sensor positions and actual gaze angle estimation.

LFI-Sensor Position and Pose Estimation

In order to estimate the gaze angle, it is first necessary to determine absolute position and orientation of the laser sensors mounted to the glasses. In the sequel, we denote the absolute sensor position in glasses-fixed coordinates as $s_g \in \mathbb{R}^3$, a unitary orientation vector as p, and denote x_g as intersection point of laser beam with the eye surface in glasses-fixed coordinates. These intersection points

$$x_{g,n} = s_{g,n} + d_n p_n \tag{C.20}$$

can be easily computed from the known sensor position in the glasses frame and the (scalar) distance measurement d_i.

The task is now to find the absolute position of the sensors $s_{g,n}, n = 1, \ldots, N$ in head-fixed

coordinates, with $N \geq 3$, and the origin of head-fixed coordinates being the eye ball center point C_S. The transformation between both coordinate systems may be denoted as

$$x_h = R(\vartheta, \varphi, \varrho) x_g + t, \tag{C.21}$$

with rotation matrix $R \in \mathbb{R}^{3 \times 3}$ and translation $t \in \mathbb{R}^3$. For the rotation matrix, we use the common decomposition in roll, pitch and yaw

$$R(\vartheta, \varphi, \varrho) = R_z(\varrho) R_y(\varphi) R_x(\vartheta). \tag{C.22}$$

The idea is to solve the task by means of trilateration given those measurements were classified to be either a sclera or retina point. For these measurements, we have

$$\|x_{h,n}\| = r = \begin{cases} r_{\text{sclera}}, & l_i = s, \\ r_{\text{retina}}, & l_i = r. \end{cases} \tag{C.23}$$

due to the ball shape of sclera and retina. Substituting Equation (C.21) into Equation (C.23) yields

$$x_{g,n}^T R^T R x_{g,n} + 2 x_{g,n}^T R^T t + t^T t - r^2 = 0, \tag{C.24}$$

which can be further simplied to

$$x_{g,n}^T x_{g,n} + 2 x_{g,n}^T R^T t + t^T t - r^2 = 0, \tag{C.25}$$

since R is unitary (i.e., $R^T R = I$), with the unity matrix I.

Collecting a batch of measurements of all lasers over time, in which the glasses position is constant, Equation (C.25) constitutes a non-linear (possibly over-determined) system of equations in the six unknowns $\vartheta, \varphi, \varrho, t_x, t_y, t_z$.

Although non-linear in nature, one has good chances to solve this root-finding problem, e.g., by means of a Levenberg-Marquardt algorithm [253].

Gaze Angle Estimation

Having carried out classification of measurements and estimation of absolute sensor position, the final step is now the actual gaze estimation. The main idea is to estimate and integrate the angular velocity of the eye. Due to the unknown starting value of the integration, the initial estimation θ', φ' will be corrupted by an offset. In order to compensate for this systematic error, a surrogate model of the eye in eye fixed coordinates is learned over time, by which the quasi-constant offset can be determined. This principle is depicted in Figure C.9.
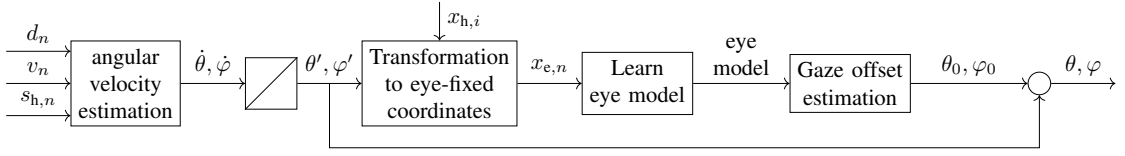
Figure C.9: Gaze estimation principle.

In order to estimate the angular velocity of the eye, we first denote the intersection points of the laser sensors with the eye surface in spherical coordinates

$$x_n(r,\theta,\varphi) = r \begin{bmatrix} \cos\theta\cos\varphi \\ \sin\theta\cos\varphi \\ \sin\varphi \end{bmatrix}. \tag{C.26}$$

In addition to a distance proportional signal, each laser returns a velocity signal, which is the projection of the surface velocity at intersection point x_i in direction p_i of the laser beam. Thus,

$$v_n \quad = \quad -p_n^T\left(\frac{\partial x_n}{\partial r}\dot r + \frac{\partial x_n}{\partial\theta}\dot\theta + \frac{\partial x_n}{\partial\varphi}\dot\varphi\right) \quad = \quad -p_n^T\left(\frac{x_n}{\|x_n\|}\dot r + \frac{\partial x_n}{\partial\theta} + \frac{\partial x_n}{\partial\varphi}\right) \tag{C.27}$$

holds. Obviously, for a laser beam hitting the sphere-like sclera or retina, $\dot r = 0$ follows in case of no slippage, i.e., the measured surface velocity is only due to eye rotation. For a laser hitting the iris however, the change of $r_i = \|x_i\|$ has to be taken into consideration for the velocity signal. In the following, we will regard the change of radius by means of a first order approximation

$$\dot r \approx \|x_n(k) - x_n(k-1)\|/T_s. \tag{C.28}$$

Defining θ and φ as rotations about the x- and z-axes, i.e. $e_\theta = e_x = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T$ $e_\varphi = e_z = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T$ Equation (C.27) constitutes for all lasers $n = 1,\dots,N$ a linear system of equations

$$\begin{bmatrix} -p_n^T(e_\theta \times x_n) & -p_n^T(e_\varphi \times x_n) \end{bmatrix}\begin{bmatrix} \dot\theta \\ \dot\varphi \end{bmatrix} = v_n - \frac{x_n}{\|x_n\|}\dot r_n, \tag{C.29}$$

which can be solved during each time step by means of a (pseudo-) inverse. At this point, we omit a lengthy discussion on existence and uniqueness of the solution of Equation (C.29).

Integrating the estimations of angular velocities yields $\theta' = \int_0^T \dot\theta \, dt$, $\varphi' = \int_0^T \dot\varphi \, dt$, with yet unknown starting values θ_0, φ_0. In order to compensate for these offsets, all eye-surface points $x_{h,i}$ measured in head-fixed coordinates are transformed into eye-fixed coordinates by means of

$$x_e = R_z(\varphi')R_x(\theta')x_h. \tag{C.30}$$

Note that we have neglected rotation around the y_h-axis (rolling) of the eye.

The x_e are then used to parameterize a surrogate model for the eye shape in order to determine the quasi constant offset θ_0, φ_0 with regard to a neutral viewing direction ($\theta = \varphi = 0$). In principle, several surrogate models are applicable to achieve this task. 2D-tensor splines for example could be used to represent the entire outer shape of the eye. We are, however, interested in a minimal representation of the eye, to reconstruct the initial gaze direction. Since information about the viewing direction is mainly given by means of iris-intersection points, it is sufficient to consider only these points. Then, a simple plane representation

$$n_{\text{iris}}^T x = d_{\text{iris}} \tag{C.31}$$

of the iris as depicted in Figure C.10 is a suitable choice. There are four parameters in total, which might be reduced to three if d_{iris} is preset to a reasonable value. In any case, since the problem is parametric linear, it can be solved by means of a recursive least squares (RLS) with exponential forgetting [251].
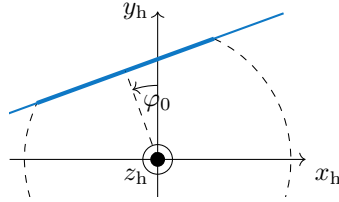
Figure C.10: Simple iris model.

From the estimated parameters of Equation (C.31), the desired offset parameters for the gaze are then given by

$$\theta_0 = \arcsin \frac{n_{\text{iris},z}}{\|n\|}, \tag{C.32a}$$

$$\varphi_0 = \arctan \frac{-n_{\text{iris},x}}{n_{\text{iris},y}}, \tag{C.32b}$$

which in a last step gives the final gaze estimates

$$\theta = \theta' + \theta_0, \tag{C.33a}$$

$$\varphi = \varphi' + \varphi_0. \tag{C.33b}$$

C.1.8 Evaluation

In this section, the proposed static LFI gaze estimation approach is evaluated including the LFI sensor measurement noise characterization, the region of intersection classification accuracy as well as the gaze estimation accuracy in presence of glasses slippage.

LFI Sensor Characterization

For accurate simulation of multi LFI sensor measurements generated by the simulation tool introduced in Appendix C.1.6, the sensor distance and velocity measurement noise need to be characterized. For this purpose, a similar lab setup as proposed by Capelli et al. [245] is used. A disk with a diameter of 24 mm is rotated with a precision rotation stage[2], while the LFI sensor is mounted on a precision linear stage[3] so that the laser beam hits the disk with an incident angle γ of 45 °. The linear stage simulates distance variations of 20 - 30 mm, which is reasonable for a near-eye setting, while the rotating disk simulates rotational eye movements with velocities up to 500 °/s.
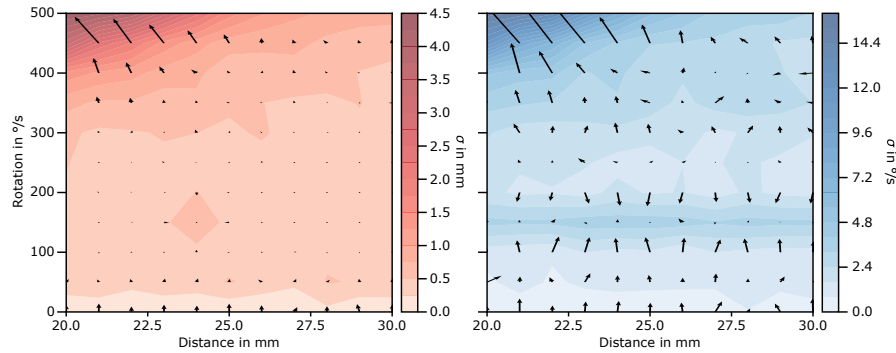
Figure C.11: Sensor noise characterization results over the whole range of eye rotational velocities and distances. Left: Variance of distance measurements, Right: Variance of velocity measurements.

Figure C.11 shows the resulting sensor noise for distance (red) as well as velocity (blue) measurements over the whole parameter space. The LFI sensor is capable of resolving the whole range of eye velocities up to rare occurring saccades with amplitudes up to 500 °/s. Especially for small distances and high velocities above 300 °/s the sensor noise increases as indicated by the gradient vectors in Figure C.11 due to limitations of the current electronics to resolve the signal. Compared to the work of Capelli et al. [245], who reported a maximum velocity error of 10 %, we achieved one of 2.8 % while measuring distance and velocity simultaneously. To exclude errors stemming from electronic limitations, we limit the velocity range to 300 °/s from which σ_d^2 and σ_v^2 are derived. The resulting sensor noise characteristics used for further simulation are 66.85 μm and 2.95°/s for σ_d^2 and σ_v^2, respectively.

Region of Intersection Classification

The region of intersection classification is the first stage of the proposed gaze estimation algorithm, introduced in Appendix C.1.7. To evaluate the performance of the classifier, we simultaneously sample a Pupil Labs V2 Core eye camera at 200 Hz and an LFI sensor at 1 kHz.

---

[2]Jenny Science ROTAX Rxhq 50-12
[3]Jenny Science ELAX Ex 50F20

Both sensors are pointing towards the eye of an participant[4]. From the camera images, ground truth labels denoting the region of intersection hit by the IR laser were derived.
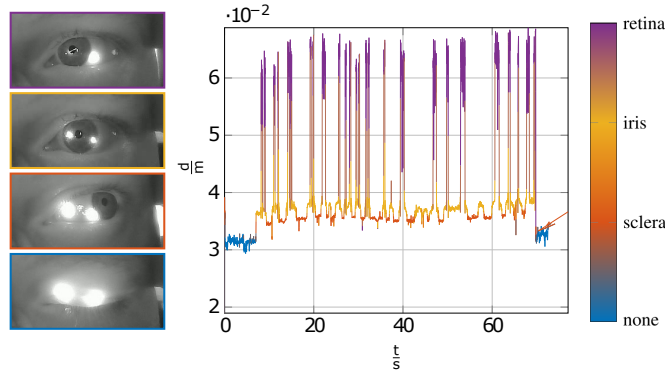
Figure C.12: Left: Reference camera images taken to derive ground truth labels of the left LFI sensor. Right: Measured distance of the left LFI sensor with corresponding label derived by the proposed classifier highlighted in color.

The measured distances d of the LFI sensor are fed into the proposed HMM classifier to derive classification labels. As the sample rates of LFI sensor and camera sensor differ, we up-sample the labels derived from the camera sensor from 200 Hz to 1 kHz by using nearest neighbor (NN) up-sampling. Figure C.12 shows the distances measured by the LFI sensor and the corresponding labels estimated by the classifier highlighted in color. Additionally, images of the VOG eye camera are shown, where the intersection of the LFI sensor with the eye is marked by a bright spot in the IR camera image. Classification was made for the left IR spot in the camera frames. The proposed HMM classifier reaches an macro F1-score of 94.42 %. Classification mismatches occurs mainly between sclera and iris due to the overlapping probability distributions of the two classes. In addition, the ground truth labels are derived from the 200 Hz camera frames and up-sampled using NN up-sampling, leading to an inaccuracy of ground truth labels and thus affect the classification accuracy.

LFI-Sensor Position and Gaze Estimation

To evaluate the sensor position and gaze estimation accuracy of the proposed system, we used the multi LFI sensor simulation introduced in Appendix C.1.6 to generate distance as well as velocity measurement trajectories for a setup of 6 LFI sensors. In order to take into account sensor noise, we initialized the sensor noise model with the parameters obtained by the sensor characterization as described in Appendix C.1.8. To evaluate the robustness of the proposed gaze estimation algorithm, we used a human eye movement trajectory captured by a Pupil Labs V2 Core eye tracker sampled at 200 Hz. We then up-sampled it similar to the ground truth labels in the previous section to 1 kHz by NN up-sampling before feeding the trajectory into the system simulation.

---

[4]Class 1 laser system according to IEC 60825-1
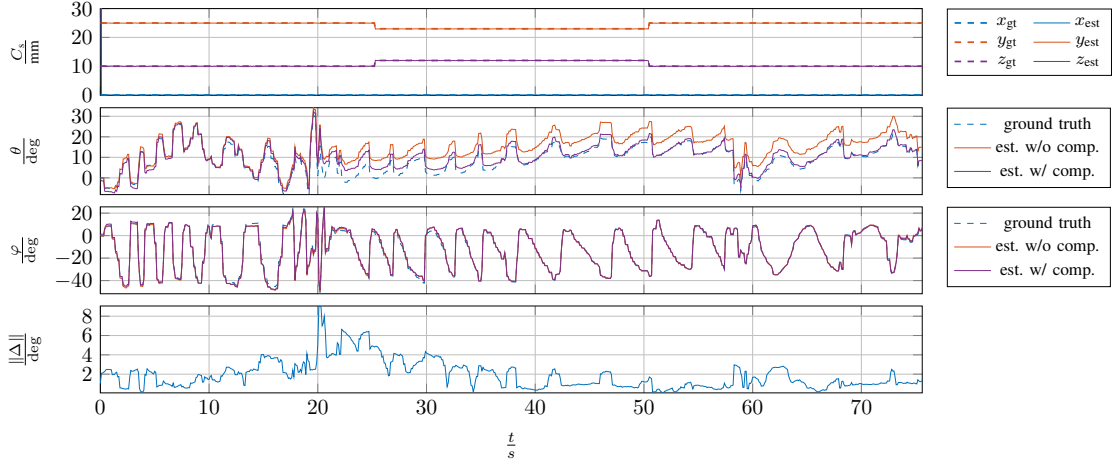
Figure C.13: Top: Original and estimated eye ball center C_S. At $t = 25\,s$ the glasses slip 2 mm towards the eye and 2 mm up the nose. Center: Original gaze angle θ and φ (dashed blue), estimated gaze angle θ' and φ' (orange) and estimated gaze angle θ and φ with offset compensation (purple) Bottom: Absolute gaze angle estimation error in $°$.

In addition, we let the sensors slip at $t = 25\,s\,\Delta C_{Sy} = -2\,mm$ towards the eye and $\Delta C_{Sz} = 2\,mm$ along the nose to investigate robustness against slippage. At t=52 s the glasses slip back to their initial position on the nose. We choose a step like slippage trajectory, as worst case assumption for glasses slippage. Figure C.13 a) shows the original sclera center C_S (dashed) and the corresponding estimation from our algorithm. As the estimation relies on trilateration, our approach is capable of immediately re-estimating the eyeball center in presence of slippage. The root mean squared error (RMSE) of the eye ball center estimation is 76.39 μm and the mean absolute error (MAE) is 123.11 μm, which both are close to the distance noise limit of an individual sensor. Therefore, the eye model can be reconstructed precisely with a minimum error, which is mandatory as inaccuracies in the eye model reconstruction directly impact the gaze estimation accuracy.

Figure C.13 b) and c) show the gaze estimation for both angles θ and φ. The original human eye movement trajectory (dashed blue) starts with random eye movements with different velocities during the first 20 seconds, followed by a reading part between $t = 30\,s$ and $t = 60\,s$ and finally slow eye movements and two fixations at the end of the trajectory. With this eye movement trajectory profile, a variety of eye movements from fixations to saccades are covered.

The reconstruction of the gaze angles by integration of the measured velocity (orange) introduces especially for θ a large offset error. The RMSE and MAE error of the uncompensated gaze estimation is 5.28 $°$ and 6.14$°$ respectively. Therefore, the compensation of the offset is mandatory. For this purpose, we introduced the surrogate model as given by Equation (C.31) to continuously compensate gaze estimation offsets arising from integration of measured rota-

tional velocity. With this additional step, the gaze estimation error declines and the estimated trajectory follows the original trajectory (cyan dashed) closely. The RMSE and MAE error of the compensated gaze estimation is 1.79 ° and 2.32 °, respectively.

The absolute error shown in Figure C.13 d) rises during fast saccadic eye movements and starts declining over time if slower and controlled eye movements dominate the trajectory and the offset compensation starts to correct the velocity integration error.

C.1.9   Discussion and Limitations

The proposed LFI eye tracking approach is capable of tracking the eye robust in presence of glasses slippage. Furthermore, the sensor fusion and gaze estimation approach is able to estimate the optical axis of the eye without additional calibration, which fosters immersion of AR glasses and thus improves user experience.

Compared to a slippage robust VOG eye tracking system like the Tobii Pro Glasses 2 with stereo camera sensors and additional IR LEDs, our LFI eye tracking approach achieves comparable gaze accuracy [93] with a lower power consumption. Furthermore, our system is not limited in the FOV, as we do not solely rely on tracking the pupil, but on the measurement of rotational speed on the entire eyeball.

The main limitation of the proposed LFI eye tracking system is that the surface of the eye must be hit by the laser beams to measure the rotational velocity of the eye. Therefore, it is possible that incorrect velocity measurements stemming from the *none* region (lashes or lids) could lead to a gaze estimation error if the laser sensor e.g., during blinking or squinting the eye. Therefore, measurements classified as *none* need to be handled separately during gaze estimation.

Finally a calibration routine is required to calibrate once the individual sensor poses S_n within the glasses coordinate system, as the proposed algorithm assumes a fixed known sensor pose within the glasses coordinate system.

C.1.10   Conclusion

In this work, we present a novel eye tracking sensor system based on static LFI sensors. In addition to the novel sensing modality, an eye tracking algorithm is proposed, which is capable to fuse distance and velocity measurements of multiple static LFI sensors by means of an optimized geometric eye model. We further focus on a lightweight implementation of the different stages of the algorithm to be able to integrate the whole system into lightweight AR glasses. In addition, we derived a geometric eye model and characterized the sensor noise to build up a multi LFI sensor simulation tool to generate measurement data for a multi LFI sensor setting while excluding human error. With the generated data we evaluated the proposed LFI system and achieved a gaze accuracy of 1.79 ° at an outstanding sampling rate of

1 kHz.

Based on the promising results of this work, we will investigate the achievable gaze estimation accuracy including human error by building up a head worn demonstrator in future work. In addition, we will investigate the combination of the proposed multi LFI system with a low frame rate camera sensor as reference to combine the high accuracy of VOG systems with the outstanding update rate of the proposed LFI sensor system.

Bibliography

[1] L. A. Coldren, S. W. Corzine, and M. L. Mashanovitch, *Diode Lasers and Photonic Integrated Circuits -*.   New York: John Wiley & Sons, 2012.

[2] M. Grabherr, P. Gerlach, R. King, and R. Jäger, "Integrated photodiodes complement the vcsel platform," in *Vertical-Cavity Surface-Emitting Lasers XIII*, vol. 7229.   International Society for Optics and Photonics, 2009, p. 72290E.

[3] W. Fuhl, D. Geisler, T. Santini, T. Appel, W. Rosenstiel, and E. Kasneci, "Cbf: Circular binary features for robust and real-time pupil center detection," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ser. ETRA '18.   New York, NY, USA: ACM, 2018, pp. 8:1–8:6. [Online]. Available: http://doi.acm.org/10.1145/3204493.3204559

[4] J. Meyer, T. Schlebusch, H. Spruit, J. Hellmig, and E. Kasneci, "A compact low-power gaze gesture sensor based on laser feedback interferometry for smart glasses," in *Proc. of SPIE Vol*, vol. 11788, 2021, pp. 117 880D–1.

[5] J. Meyer, T. Schlebusch, H. Spruit, J. Hellmig, and E. Kasneci, "A novel gaze gesture sensor for smart glasses based on laser self-mixing," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts)*, ACM.   New York, NY, USA: ACM, 2021.

[6] T. Taimre, M. Nikolić, K. Bertling, Y. L. Lim, T. Bosch, and A. D. Rakić, "Laser feedback interferometry: a tutorial on the self-mixing effect for coherent sensing," *Adv. Opt. Photon.*, vol. 7, no. 3, pp. 570–631, Sep 2015. [Online]. Available: http://aop.osa.org/abstract.cfm?URI=aop-7-3-570

[7] M. Perslev, M. Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-time: A fully convolutional network for time series segmentation applied to sleep staging," *Advances in Neural Information Processing Systems*, vol. 32, pp. 4415–4426, 2019.

[8] C. Hennessey, B. Noureddin, and P. Lawrence, "A single camera eye-gaze tracking system with free head motion," in *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications*, ser. ETRA '06.   New York, NY, USA: Association for Computing Machinery, 2006, p. 87–94. [Online]. Available: https://doi.org/10.1145/1117309.1117349

[9] L. Świrski and N. A. Dodgson, "A fully-automatic, temporal approach to single camera, glint-free 3d eye model fitting [abstract]," in *Proceedings of ECEM 2013*, Aug. 2013. [Online]. Available: http://www.cl.cam.ac.uk/research/rainbow/projects/eyemodelfit/

[10] A. Petersen, T. Schlebusch, J. Meyer, H. Spruit, and J. Hellmig, "Method for detecting a gaze direction of an eye," US Patent US11 435 578B2, 2022.

[11] A. Petersen, T. Schlebusch, J. Meyer, H. Spruit, and J. Hellmig, "Method for operating a pair of smart glasses," US Patent US20 210 373 659A1, 2022.

[12] E. Viirre, H. Pryor, S. Nagata, and T. Furness 3rd, "The virtual retinal display: a new technology for virtual reality and augmented vision in medicine," *Studies in health technology and informatics*, vol. 50, pp. 252–257, 1998.

[13] M. Sugawara, M. Suzuki, and N. Miyauchi, "Retinal imaging laser eyewear with focus-free and augmented reality," in *2017 24th International Workshop on Active-Matrix Flatpanel Displays and Devices (AM-FPD)*, July 2017, pp. 5–8.

[14] M. Sugawara, M. Suzuki, M. Ishimoto, K. Hasegawa, N. Teshima, K. Yasui, and N. Miyauchi, "Every aspect of advanced retinal imaging laser eyewear: principle, free focus, resolution, laser safety, and medical welfare applications," in *MOEMS and Miniaturized Systems XVII*, W. Piyawattanametha, Y.-H. Park, and H. Zappe, Eds., vol. 10545, International Society for Optics and Photonics. SPIE, 2018, pp. 151 – 168. [Online]. Available: https://doi.org/10.1117/12.2295751

[15] P. Majaranta and A. Bulling, "Eye tracking and eye-based human–computer interaction," in *Advances in physiological computing*. Springer, 2014, pp. 39–65.

[16] A. S. Heiko Drewes, "Interacting with the computer using gaze gestures," in *International Conference on Human Computer Interaction*, 2007.

[17] L. Lee and P. Hui, "Interaction methods for smart glasses: A survey," *IEEE Access*, vol. 6, pp. 28 712–28 732, 2018.

[18] M. Bace, V. Becker, C. Wang, and A. Bulling, "Combining gaze estimation and optical flow for pursuits interaction," in *ACM Symposium on Eye Tracking Research and Applications*, ser. ETRA '20 Full Papers. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3379155.3391315

[19] A. Bulling, J. A. Ward, H. Gellersen, and G. Troster, "Eye movement analysis for activity recognition using electrooculography," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 741–753, 2011.

[20] A. Bulling, D. Roggen, and G. Tröster, "Wearable eog goggles: Seamless sensing and context-awareness in everyday environments," *J. Ambient Intell. Smart Environ.*, vol. 1, no. 2, p. 157–171, Apr. 2009.

[21] C. Braunagel, D. Geisler, W. Rosenstiel, and E. Kasneci, "Online recognition of driver-activity based on visual scanpath classification," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 4, pp. 23–36, 2017.

[22] N. Castner, E. Kasneci, T. Kübler, K. Scheiter, J. Richter, T. Eder, F. Hüttig, and C. Keutel, "Scanpath comparison in medical image reading skills of dental students: Distinguishing stages of expertise development," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ser. ETRA '18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3204493.3204550

[23] N. Castner, T. C. Kuebler, K. Scheiter, J. Richter, T. Eder, F. Huettig, C. Keutel, and E. Kasneci, "Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing," in *ACM Symposium on Eye Tracking Research and Applications*, ser. ETRA '20 Full Papers. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3379155.3391320

[24] S. Kübler, Eivazi and E. Kasneci, "Automated visual scanpath analysis reveals the expertise level of micro-neurosurgeons," in *In MICCAI workshop on interventional microscopy*, 2015.

[25] E. Kasneci, A. A. Black, and J. M. Wood, "Eye-tracking as a tool to evaluate functional ability in everyday tasks in glaucoma," *Journal of ophthalmology*, vol. 2017, 2017.

[26] B. W. Hosp, F. Schultz, O. Höner, and E. Kasneci, "Soccer goalkeeper expertise identification based on eye movements," *PLOS ONE*, vol. 16, no. 5, pp. 1–22, 05 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0251070

[27] B. Hosp, F. Schultz, E. Kasneci, and O. Höner, "Eye movement feature classification for soccer expertise identification in virtual reality," *CoRR*, vol. abs/2009.11676, 2020. [Online]. Available: https://arxiv.org/abs/2009.11676

[28] D. R. Reich, P. Prasse, C. Tschirner, P. Haller, F. Goldhammer, and L. A. Jäger, "Inferring native and non-native human reading comprehension and subjective text difficulty from scanpaths in reading," in *Proceedings of the 2022 ACM Symposium on Eye-Tracking Research and Applications*, ser. ETRA '22. Seattle, WA: Association for Computing Machinery (ACM), 2022.

[29] C. Katsini, Y. Abdrabou, G. E. Raptis, M. Khamis, and F. Alt, *The Role of Eye Gaze in Security and Privacy Applications: Survey and Future HCI Research Directions*. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–21. [Online]. Available: https://doi.org/10.1145/3313831.3376840

[30] L. A. Jäger, S. Makowski, P. Prasse, L. Sascha, M. Seidler, and T. Scheffer, "Deep Eyedentification: Biometric identification using micro-movements of the eye," in *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019*, ser.

Lecture Notes in Computer Science, Brefeld, Fromont, Knobbe, Hotho, Maathuis, and Robardet, Eds., vol. 11907. Cham, Switzerland: Springer International Publishing, 2020, pp. 299–314. [Online]. Available: https://www.springerprofessional.de/en/deep-eyedentification-biometric-identification-using-micro-movem/17945546

[31] D. Lohr and O. V. Komogortsev, "Eye know you too: A densenet architecture for end-to-end biometric authentication via eye movements," 2022.

[32] K. Krejtz, A. T. Duchowski, A. Niedzielska, C. Biele, and I. Krejtz, "Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze," *PloS one*, vol. 13, no. 9, p. e0203629, 2018.

[33] T. Appel, N. Sevcenko, F. Wortha, K. Tsarava, K. Moeller, M. Ninaus, E. Kasneci, and P. Gerjets, "Predicting cognitive load in an emergency simulation based on behavioral and physiological measures," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 154–163.

[34] E. Bozkir, D. Geisler, and E. Kasneci, "Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup," in *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019, pp. 1834–1837.

[35] T. Appel, C. Scharinger, P. Gerjets, and E. Kasneci, "Cross-subject workload classification using pupil-related measures," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 2018, pp. 1–8.

[36] T. Appel, P. Gerjets, S. Hoffman, K. Moeller, M. Ninaus, C. Scharinger, N. Sevcenko, F. Wortha, and E. Kasneci, "Cross-task and cross-participant classification of cognitive load in an emergency simulation game," *IEEE Transactions on Affective Computing*, 2021.

[37] R. J. Leigh and D. S. Zee, *The neurology of eye movements*. Contemporary Neurology, 2015.

[38] A. S. Kaplanyan, A. Sochenov, T. Leimkühler, M. Okunev, T. Goodall, and G. Rufo, "Deepfovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 212:1–212:13, Nov. 2019. [Online]. Available: http://doi.acm.org/10.1145/3355089.3356557

[39] J. Kim, Y. Jeong, M. Stengel, K. Akşit, R. Albert, B. Boudaoud, T. Greer, J. Kim, W. Lopes, Z. Majercik, P. Shirley, J. Spjut, M. McGuire, and D. Luebke, "Foveated ar: Dynamically-foveated augmented reality display," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 99:1–99:15, Jul. 2019. [Online]. Available: http://doi.acm.org/10.1145/3306346.3322987

[40] X. Meng, R. Du, and A. Varshney, "Eye-dominance-guided foveated rendering," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 5, pp. 1972–1980, 2020.

[41] T. Lin, T. Zhan, J. Zou, F. Fan, and S.-T. Wu, "Maxwellian near-eye display with an expanded eyebox," *Opt. Express*, vol. 28, no. 26, pp. 38 616–38 625, Dec 2020. [Online]. Available: http://www.osapublishing.org/oe/abstract.cfm?URI=oe-28-26-38616

[42] K. Ratnam, R. Konrad, D. Lanman, and M. Zannoli, "Retinal image quality in near-eye pupil-steered systems," *Opt. Express*, vol. 27, no. 26, pp. 38 289–38 311, Dec 2019. [Online]. Available: http://www.osapublishing.org/oe/abstract.cfm?URI=oe-27-26-38289

[43] J. Xiong, Y. Li, K. Li, and S.-T. Wu, "Aberration-free pupil steerable maxwellian display for augmented reality with cholesteric liquid crystal holographic lenses," *Opt. Lett.*, vol. 46, no. 7, pp. 1760–1763, Apr 2021. [Online]. Available: http://www.osapublishing.org/ol/abstract.cfm?URI=ol-46-7-1760

[44] J. Xiong, E.-L. Hsiang, Z. He, T. Zhan, and S.-T. Wu, "Augmented reality and virtual reality displays: emerging technologies and future perspectives," *Light: Science & Applications*, vol. 10, no. 1, Oct. 2021. [Online]. Available: https://doi.org/10.1038/s41377-021-00658-8

[45] B. Krajancich, P. Kellnhofer, and G. Wetzstein, "Optimizing depth perception in virtual and augmented reality through gaze-contingent stereo rendering," *ACM Trans. Graph.*, vol. 39, no. 6, nov 2020. [Online]. Available: https://doi.org/10.1145/3414685.3417820

[46] E. G. Mlot, H. Bahmani, S. Wahl, and E. Kasneci, "3d gaze estimation using eye vergence," in *International Conference on Health Informatics*, vol. 6. Scitepress, 2016, pp. 125–131.

[47] T. Zhan, J. Xiong, J. Zou, and S.-T. Wu, "Multifocal displays: review and prospect," *PhotoniX*, vol. 1, no. 1, Mar. 2020. [Online]. Available: https://doi.org/10.1186/s43074-020-00010-0

[48] P. Chakravarthula, D. Dunn, K. Akşit, and H. Fuchs, "Focusar: Auto-focus augmented reality eyeglasses for both real world and virtual imagery," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 11, pp. 2906–2916, 2018.

[49] M. Meißner, J. Pfeiffer, T. Pfeiffer, and H. Oppewal, "Combining virtual reality and mobile eye tracking to provide a naturalistic experimental environment for shopper research," *Journal of Business Research*, vol. 100, pp. 445–458, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148296317303478

[50] R. Piening, K. Pfeuffer, A. Esteves, T. Mittermeier, S. Prange, P. Schröder, and F. Alt, "Looking for info: Evaluation of gaze based information retrieval in augmented reality," in *Human-Computer Interaction – INTERACT 2021*, C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, and K. Inkpen, Eds. Cham: Springer International Publishing, 2021, pp. 544–565.

[51] B. Kress, *Optical Architectures for Augmented-, Virtual-, and Mixed-Reality Headsets*. SPIE, 2020. [Online]. Available: https://books.google.de/books?id=iqcszQEACAAJ

[52] O. Komogortsev, "Virtual and augmented reality, sensors, eye movements, and their impact on healthcare and security," online, 2021, cHI'21 workshop on Eye Movements as an Interface to Cognitive State (EMICS'21). [Online]. Available: https://emics-2021.github.io/EMICS/

[53] M. Kassner, W. Patera, and A. Bulling, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, ser. UbiComp '14 Adjunct.   New York, NY, USA: ACM, 2014, pp. 1151–1160. [Online]. Available: http://doi.acm.org/10.1145/2638728.2641695

[54] I. Hong, K. Bong, and H.-J. Yoo, "Challenges of eye tracking systems for mobile XR glasses," in *Applications of Digital Image Processing XLI*, A. G. Tescher, Ed., vol. 10752, International Society for Optics and Photonics.   SPIE, 2018, pp. 391 – 397. [Online]. Available: https://doi.org/10.1117/12.2322657

[55] M. Juhola, V. Jäntti, and I. Pyykkö, "Effect of sampling frequencies on computation of the maximum velocity of saccadic eye movements," *Biological Cybernetics*, vol. 53, no. 2, pp. 67–72, Dec. 1985. [Online]. Available: https://doi.org/10.1007/bf00337023

[56] W. Fuhl, M. Tonsen, A. Bulling, and E. Kasneci, "Pupil detection for head-mounted eye tracking in the wild: an evaluation of the state of the art," *Machine Vision and Applications*, vol. 27, no. 8, pp. 1275–1288, Nov 2016. [Online]. Available: https://doi.org/10.1007/s00138-016-0776-4

[57] K. Holmqvist, S. L. Örbom, I. T. Hooge, D. C. Niehorster, R. G. Alexander, R. Andersson, J. S. Benjamins, P. Blignaut, A.-M. Brouwer, L. L. Chuang *et al.*, "Eye tracking: empirical foundations for a minimal reporting guideline," *Behavior research methods*, pp. 1–53, 2022.

[58] R. S. Hessels, R. Andersson, I. T. Hooge, M. Nyström, and C. Kemner, "Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research," *Infancy*, vol. 20, no. 6, pp. 601–633, 2015.

[59] S. K. Schnipke and M. W. Todd, "Trials and tribulations of using an eye-tracking system," in *CHI'00 extended abstracts on Human factors in computing systems*, 2000, pp. 273–274.

[60] D. A. Atchison and L. N. Thibos, "Optical models of the human eye," *Clinical and Experimental Optometry*, vol. 99, no. 2, pp. 99–106, 2016. [Online]. Available: https://doi.org/10.1111/cxo.12352

[61] H. Gross, F. Blechinger, and B. Achtner, "Human eye," *Handbook of Optical Systems: Volume 4: Survey of Optical Instruments*, vol. 4, pp. 1–87, 2008.

[62] A. T. T. Duchowski, *Eye Tracking Methodology - Theory and Practice*, 3rd ed.   Cham, Heidelberg, New York, Dordrecht, London: Springer International Publishing, 2017.

[63] E. Guestrin and M. Eizenman, "General theory of remote gaze estimation using the pupil center and corneal reflections," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 6, pp. 1124–1133, 2006.

[64] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16 495–16 519, 2017.

[65] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, 2010.

[66] K. Dierkes, M. Kassner, and A. Bulling, "A novel approach to single camera, glint-free 3d eye model fitting including corneal refraction," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ser. ETRA '18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3204493.3204525

[67] Y.-H. Yiu, M. Aboulatta, T. Raiser, L. Ophey, V. L. Flanagin, P. zu Eulenburg, and S.-A. Ahmadi, "Deepvog: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning," *Journal of Neuroscience Methods*, vol. 324, p. 108307, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027019301578

[68] L. Swirski, A. Bulling, and N. Dodgson, "Robust real-time pupil tracking in highly off-axis images," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ser. ETRA '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 173–176. [Online]. Available: https://doi.org/10.1145/2168556.2168585

[69] W. Fuhl, T. C. Santini, T. Kübler, and E. Kasneci, "Else: Ellipse selection for robust pupil detection in real-world environments," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, 2016, pp. 123–130.

[70] T. Santini, W. Fuhl, and E. Kasneci, "Pure: Robust pupil detection for real-time pervasive eye tracking," *Computer Vision and Image Understanding*, vol. 170, pp. 40–50, 2018.

[71] T. Santini, W. Fuhl, and E. Kasneci, "Purest: Robust pupil tracking for real-time pervasive eye tracking," in *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, 2018, pp. 1–5.

[72] W. Fuhl, S. Eivazi, B. Hosp, A. Eivazi, W. Rosenstiel, and E. Kasneci, "Bore: Boosted-oriented edge optimization for robust, real time remote pupil center detection," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 2018, pp. 1–5.

[73] W. Fuhl, T. Santini, G. Kasneci, and E. Kasneci, "Pupilnet: Convolutional neural networks for robust pupil detection," *arXiv preprint arXiv:1601.04902*, 2016.

Bibliography

[74] W. Fuhl, T. Santini, G. Kasneci, W. Rosenstiel, and E. Kasneci, "Pupilnet v2. 0: Convolutional neural networks for cpu based real time robust pupil detection," *arXiv preprint arXiv:1711.00112*, 2017.

[75] W. Fuhl, D. Geisler, W. Rosenstiel, and E. Kasneci, "The applicability of cycle gans for pupil and eyelid segmentation, data generation and image refinement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[76] A. K. Chaudhary, R. Kothari, M. Acharya, S. Dangi, N. Nair, R. Bailey, C. Kanan, G. Diaz, and J. B. Pelz, "Ritnet: Real-time semantic segmentation of the eye for gaze tracking," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 3698–3702.

[77] R. S. Kothari, A. K. Chaudhary, R. J. Bailey, J. B. Pelz, and G. J. Diaz, "Ellseg: An ellipse segmentation framework for robust gaze tracking," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 5, pp. 2757–2767, 2021.

[78] J. Kim, M. Stengel, A. Majercik, S. De Mello, D. Dunn, S. Laine, M. McGuire, and D. Luebke, "Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–12. [Online]. Available: https://doi.org/10.1145/3290605.3300780

[79] H. Kaur, S. Jindal, and R. Manduchi, "Rethinking model-based gaze estimation," *Proc. ACM Comput. Graph. Interact. Tech.*, vol. 5, no. 2, may 2022. [Online]. Available: https://doi.org/10.1145/3530797

[80] O. Semi, *PYTHON 1.3/0.5/0.3 MegaPixelsGlobal Shutter CMOS Image Sensors*, 2021.

[81] A. Mayberry, P. Hu, B. Marlin, C. Salthouse, and D. Ganesan, "Ishadow: Design of a wearable, real-time mobile gaze tracker," in *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '14. New York, NY, USA: Association for Computing Machinery, 2014, p. 82–94. [Online]. Available: https://doi.org/10.1145/2594368.2594388

[82] K. Lübeck and O. Bringmann, "A heterogeneous and reconfigurable embedded architecture for energy-efficient execution of convolutional neural networks," in *Architecture of Computing Systems – ARCS 2019*, M. Schoeberl, C. Hochberger, S. Uhrig, J. Brehm, and T. Pionteck, Eds. Cham: Springer International Publishing, 2019, pp. 267–280.

[83] A. Garofalo, M. Rusci, F. Conti, D. Rossi, and L. Benini, "Pulp-nn: accelerating quantized neural networks on parallel ultra-low-power risc-v processors," *Philosophical Transactions of the Royal Society A*, vol. 378, no. 2164, p. 20190155, 2020.

[84] W. Fuhl, H. Gao, and E. Kasneci, "Tiny convolution, decision tree, and binary neuronal networks for robust and real time pupil outline estimation," in *ACM Symposium on Eye Tracking Research and Applications*, 2020, pp. 1–5.

[85] E. Kasneci, K. Sippel, K. Aehling, M. Heister, W. Rosenstiel, U. Schiefer, and E. Papageorgiou, "Driving with binocular visual field loss? a study on a supervised on-road parcours with simultaneous eye and head tracking," *PLOS ONE*, vol. 9, no. 2, pp. 1–13, 02 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0087470

[86] W. Fuhl, D. Geisler, T. Santini, W. Rosenstiel, and E. Kasneci, "Evaluation of state-of-the-art pupil detection algorithms on remote eye images," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, 2016, pp. 1716–1725.

[87] T. Santini, H. Brinkmann, L. Reitstätter, H. Leder, R. Rosenberg, W. Rosenstiel, and E. Kasneci, "The art of pervasive eye tracking: Unconstrained eye tracking in the austrian gallery belvedere," in *Proceedings of the 7th workshop on pervasive eye tracking and mobile eye-based interaction*, 2018, pp. 1–8.

[88] D. Geisler, D. Fox, and E. Kasneci, "Real-time 3d glint detection in remote eye tracking based on bayesian inference," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 7119–7126.

[89] F. B. Narcizo, F. E. D. dos Santos, and D. W. Hansen, "High-accuracy gaze estimation for interpolation-based eye-tracking methods," *Vision*, vol. 5, no. 3, 2021. [Online]. Available: https://www.mdpi.com/2411-5150/5/3/41

[90] M. Tonsen, J. Steil, Y. Sugano, and A. Bulling, "Invisibleeye: Mobile eye tracking using multiple low-resolution cameras and learning-based gaze estimation," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, Sep. 2017. [Online]. Available: https://doi.org/10.1145/3130971

[91] G. A. Koulieris, K. Akşit, M. Stengel, R. K. Mantiuk, K. Mania, and C. Richardt, "Near-eye display and tracking technologies for virtual and augmented reality," in *Computer Graphics Forum*, vol. 38, no. 2. Wiley Online Library, 2019, pp. 493–519.

[92] T. Santini, D. C. Niehorster, and E. Kasneci, "Get a grip: Slippage-robust and glint-free gaze estimation for real-time pervasive head-mounted eye tracking," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ser. ETRA '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3314111.3319835

[93] D. C. Niehorster, T. Santini, R. S. Hessels, I. T. Hooge, E. Kasneci, and M. Nyström, "The impact of slippage on the data quality of head-worn eye trackers," *Behavior Research Methods*, vol. 52, no. 3, pp. 1140–1160, 2020.

[94] F. B. Narcizo, "Using priors to improve head-mounted eye trackers in sports," Ph.D. dissertation, IT University of Copenhagen, 2017.

[95] B. R. Pires, M. Hwangbo, M. Devyver, and T. Kanade, "Visible-spectrum gaze tracking for sports," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 1005–1010.

[96] S. Kohlbecher, S. Bardinst, K. Bartl, E. Schneider, T. Poitschke, and M. Ablassmeier, "Calibration-free eye tracking by reconstruction of the pupil ellipse in 3d space," in *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, ser. ETRA '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 135–138. [Online]. Available: https://doi.org/10.1145/1344471.1344506

[97] Tobii, *Pro Glasses 3 Product Description*, 2021. [Online]. Available: https://www.tobiipro.com/siteassets/tobii-pro/product-descriptions/product-description-tobii-pro-glasses-3.pdf/?v=1.4

[98] C. H. Morimoto and M. R. Mimica, "Eye gaze tracking techniques for interactive applications," *Computer Vision and Image Understanding*, vol. 98, no. 1, pp. 4–24, 2005, special Issue on Eye Detection and Tracking. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314204001109

[99] T. Santini, W. Fuhl, and E. Kasneci, "Calibme: Fast and unsupervised eye tracker calibration for gaze-based pervasive human-computer interaction," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 2594–2605. [Online]. Available: https://doi.org/10.1145/3025453.3025950

[100] M. Tonsen, C. K. Baumann, and K. Dierkes, "A high-level description and performance evaluation of pupil invisible," *arXiv preprint arXiv:2009.00508*, 2020.

[101] N. Sarkar, D. Strathearn, G. Lee, M. Olfat, A. Rohani, and R. R. Mansour, "A large angle, low voltage, small footprint micromirror for eye tracking and near-eye display applications," in *2015 Transducers - 2015 18th International Conference on Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS)*, June 2015, pp. 855–858.

[102] N. Sarkar, B. O'Hanlon, A. Rohani, D. Strathearn, G. Lee, M. Olfat, and R. R. Mansour, "A resonant eye-tracking microsystem for velocity estimation of saccades and foveated rendering," in *2017 IEEE 30th International Conference on Micro Electro Mechanical Systems (MEMS)*, Jan 2017, pp. 304–307.

[103] A. microsystems, *Ad-Hawk MindLink Specifications Preliminary Data Sheet*, 05 2021.

[104] K. Holmqvist, S. L. Orbom, M. Miller, A. Kashchenevsky, M. M. Shovman, and M. W. Greenlee, "Validation of a prototype hybrid eye-tracker against the dpi and the tobii spectrum," in *ACM Symposium on Eye Tracking Research and Applications*, ser. ETRA '20 Full Papers. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3379155.3391330

[105] B. Greenberg, "EyeWay Vision: Highly Efficient Immersive AR Glasses Using Gaze-locked Exit Pupil Steering," in *SPIE AVR21 Industry Talks II*, C. Chair, Ed., vol. 11764, International Society for Optics and Photonics. SPIE, 2021. [Online]. Available: https://doi.org/10.1117/12.2598227

[106] B. Greenberg, A. Rotnemer, S. Shwartz, F. Marjieh, Y. Zimmerman, I. Tikhonenkov, and H. Friedman, "Eye tracking systems and methods," WO patent WO21 117 031, 2021.

[107] D. A. Robinson, "A method of measuring eye movement using a scleral search coil in a magnetic field," *IEEE Transactions on Bio-medical Electronics*, vol. 10, no. 4, pp. 137–145, 1963.

[108] A. Bulling, D. Roggen, and G. Tröster, "It's in your eyes: Towards context-awareness and mobile hci using wearable eog goggles," in *Proceedings of the 10th International Conference on Ubiquitous Computing*, ser. UbiComp '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 84–93. [Online]. Available: https://doi.org/10.1145/1409635.1409647

[109] N. Torok, J. Victor Guillemin, and J. M. Barnothy, "Lxxx photoelectric nystagmography," *Annals of Otology, Rhinology & Laryngology*, vol. 60, no. 4, pp. 917–926, 1951, pMID: 14903813. [Online]. Available: https://doi.org/10.1177/000348945106000402

[110] I. Rigas, H. Raffle, and O. V. Komogortsev, "Photosensor oculography: Survey and parametric analysis of designs using model-based simulation," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 6, pp. 670–681, Dec 2018.

[111] I. Rigas, H. Raffle, and O. V. Komogortsev, "Hybrid ps-v technique: A novel sensor fusion approach for fast mobile eye-tracking with sensor-shift aware correction," *IEEE Sensors Journal*, vol. 17, no. 24, pp. 8356–8366, 2017.

[112] D. Katrychuk, H. K. Griffith, and O. V. Komogortsev, "Power-efficient and shift-robust eye-tracking sensor for portable vr headsets," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ser. ETRA '19. New York, NY, USA: ACM, 2019, pp. 19:1–19:8. [Online]. Available: http://doi.acm.org/10.1145/3314111.3319821

[113] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 154–180, 2022.

[114] A. N. Angelopoulos, J. N. Martel, A. P. Kohli, J. Conradt, and G. Wetzstein, "Event-based near-eye gaze tracking beyond 10,000 hz," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 5, pp. 2577–2586, 2021.

[115] T. Stoffregen, H. Daraei, C. Robinson, and A. Fix, "Event-based kilohertz eye tracking using coded differential lighting," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022, pp. 2515–2523.

Bibliography

[116] *THS4567 220 MHz, High Input Impedance, Fully Differential Amplifier with Independent Input and Output Common-Mode Control*, 2020.

[117] R. Michalzik, *VCSELs - Fundamentals, Technology and Applications of Vertical-Cavity Surface-Emitting Lasers*. Springer Berlin Heidelberg, 2013. [Online]. Available: https://doi.org/10.1007/978-3-642-24986-0

[118] G. Acket, D. Lenstra, A. Den Boef, and B. Verbeek, "The influence of feedback intensity on longitudinal mode properties and optical noise in index-guided semiconductor lasers," *IEEE Journal of Quantum Electronics*, vol. 20, no. 10, pp. 1163–1169, 1984.

[119] R. Lang and K. Kobayashi, "External optical feedback effects on semiconductor injection laser properties," *IEEE Journal of Quantum Electronics*, vol. 16, no. 3, pp. 347–355, 1980.

[120] C. Henry, "Theory of the linewidth of semiconductor lasers," *IEEE Journal of Quantum Electronics*, vol. 18, no. 2, pp. 259–264, 1982.

[121] S. D. Guido Giuliani, Michele Norgia and T. Bosch, "Laser diode self-mixing technique for sensing applications," *Journal of Optics A: Pure and Applied Optics*, 2002.

[122] A. Pruijmboom, M. Schemmann, J. Hellmig, J. Schutte, H. Moench, and J. Pankert, "VCSEL-based miniature laser-Doppler interferometer," in *Vertical-Cavity Surface-Emitting Lasers XII*, C. Lei and J. K. Guenter, Eds., vol. 6908, International Society for Optics and Photonics. SPIE, 2008, pp. 157 – 163. [Online]. Available: https://doi.org/10.1117/12.775131

[123] F. Delorme, P. Gambini, M. Puleo, and S. Slempkes, "Fast tunable 1.5 mu m distributed bragg reflector laser for optical switching applications," *Electronics Letters*, vol. 29, no. 1, pp. 41–, 1993.

[124] D. Nordin, "Optical frequency modulated continous wave (fmcw) range and velocity measurement," Ph.D. dissertation, Lulea University of Technology, 2004.

[125] M. Norgia, D. Melchionni, and A. Pesatori, "Self-mixing instrument for simultaneous distance and speed measurement," *Optics and Lasers in Engineering*, vol. 99, pp. 31 – 38, 2017, laser Doppler vibrometry. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0143816616302913

[126] B. Behroozpour, P. A. M. Sandborn, M. C. Wu, and B. E. Boser, "Lidar system architectures and circuits," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 135–142, 2017.

[127] J. Meyer, T. Schlebusch, T. Kuebler, and E. Kasneci, "Low power scanned laser eye tracking for retinal projection ar glasses," in *ACM Symposium on Eye Tracking Research and Applications*, ser. ETRA '20 Adjunct. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3379157.3391995

[128] J. Meyer, T. Schlebusch, H. Spruit, J. Hellmig, and E. Kasneci, "A novel -eye-tracking sensor for ar glasses based on laser self-mixing showing exceptional robustness against illumination," in *ACM Symposium on Eye Tracking Research and Applications*, ser. ETRA '20 Short Papers.   New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3379156.3391352

[129] J. Meyer, T. Schlebusch, W. Fuhl, and E. Kasneci, "A novel camera-free eye tracking sensor for augmented reality based on laser scanning," *IEEE Sensors Journal*, pp. 1–9, 2020.

[130] J. Meyer, T. Schlebusch, and E. Kasneci, "A highly integrated ambient light robust eye-tracking sensor for retinal projection ar glasses based on laser feedback interferometry," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. ETRA, may 2022. [Online]. Available: https://doi.org/10.1145/3530881

[131] J. Meyer, T. Wilm, R. Fiess, T. Schlebusch, W. Stork, and E. Kasneci, "A holographic single-pixel stereo camera sensor for calibration-free eye-tracking in retinal projection augmented reality glasses," in *2022 Symposium on Eye Tracking Research and Applications*, ser. ETRA '22.   New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3517031.3529616

[132] Amazon, "Echo frames - eyeglasses with alexa," online, 2019. [Online]. Available: https://www.amazon.com/Staging-Product-Not-Retail-Sale/dp/B07W72XKPJ

[133] J. Meyer, A. Frank, T. Schlebusch, and E. Kasneci, "A cnn-based human activity recognition system combining a laser feedback interferometry eye movement sensor and an imu for context-aware smart glasses," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 4, dec 2022. [Online]. Available: https://doi.org/10.1145/3494998

[134] J. Meyer, A. Frank, T. Schlebusch, and E. Kasneci, "U-har: A convolutional approach to human activity recognition combining head and eye movements for context-aware smart glasses," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. ETRA, May 2022. [Online]. Available: https://doi.org/10.1145/3530884

[135] F. Mulvey, A. Villanueva, D. Sliney, R. Lange, S. Cotmore, and M. Donegan, *Exploration of safety issues in Eyetracking*.   COGAIN EU Network of Excellence, 2008, iST-2003-511598.

[136] J. Vos and D. van Norren, "Retinal damage by optical radiation. an alternative to current, acgih-inspired guidelines," *Clinical and Experimental Optometry*, vol. 88, no. 4, pp. 200–211, 2005. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1444-0938.2005.tb06698.x

[137] O. Tsutom, "Thermal effect of visible light and infra-red radiation (i.r.-A, i.r.-B and i.r.-C) on the eye: a study of infra-red cataract based on a model," *The Annals of Occupational Hygiene*, vol. 38, no. 4, pp. 351–359, 08 1994. [Online]. Available: https://doi.org/10.1093/annhyg/38.4.351

[138] D. Sliney, F. Mulvey, J. Charlier, D. Cleveland, G. Daunys, M. Donegan, D. Droege, M. Joos, K. Schulmeister, P. Söderberg, O. Stephankova, and A. Villanueva, "CIE 245:2021 optical safety of infrared eye trackers applied for extended durations," Various institutions, Tech. Rep., 2021. [Online]. Available: https://doi.org/10.25039/tr.245.2021

[139] *IEC 60825-1:2014 Safety of laser products - Part 1: Equipment classification and requirements*, International Electrotechnical Commission Std., 2014.

[140] M. S. Obaidat, I. Traore, and I. Woungang, Eds., *Biometric-Based Physical and Cybersecurity Systems.* Springer International Publishing, 2019. [Online]. Available: https://doi.org/10.1007/978-3-319-98734-7

[141] *ISO/IEC 19794 Information technology — Biometric data interchange formats*, ISO Std., 2014.

[142] D. Todorov, *Mechanics of user identification and authentication: Fundamentals of identity management.* Auerbach Publications, 2007.

[143] S. Eberz, K. B. Rasmussen, V. Lenders, and I. Martinovic, "Looks like eve: Exposing insider threats using eye movement biometrics," *ACM Trans. Priv. Secur.*, vol. 19, no. 1, jun 2016. [Online]. Available: https://doi.org/10.1145/2904018

[144] P. Prasse, L. A. Jäger, S. Makowski, M. Feuerpfeil, and T. Scheffer, "On the relationship between eye tracking resolution and performance of oculomotoric biometric identification," *Procedia Computer Science*, vol. 176, pp. 2088–2097, 2020, knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920321487

[145] C. J. HUN, L. H. KYUNG, and L. I. JAE, "Method for providing advertising using eye-gaze," May 2018.

[146] D. J. Liebling and S. Preibusch, "Privacy considerations for a pervasive eye tracking world," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, ser. UbiComp '14 Adjunct. New York, NY, USA: Association for Computing Machinery, 2014, p. 1169–1177. [Online]. Available: https://doi.org/10.1145/2638728.2641688

[147] J. L. Kröger, O. H.-M. Lutz, and F. Müller, *What Does Your Gaze Reveal About You? On the Privacy Implications of Eye Tracking.* Cham: Springer International Publishing, 2020, pp. 226–241. [Online]. Available: https://doi.org/10.1007/978-3-030-42504-3_15

[148] J. Steil, M. Koelle, W. Heuten, S. Boll, and A. Bulling, "Privaceye: Privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ser. ETRA '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3314111.3319913

[149] J. Z. Lim, J. Mountstephens, and J. Teo, "Emotion recognition using eye-tracking: taxonomy, review and current challenges," *Sensors*, vol. 20, no. 8, p. 2384, 2020.

[150] A. T. Duchowski, K. Krejtz, N. A. Gehrer, T. Bafna, and P. Bækgaard, *The Low/High Index of Pupillary Activity*. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–12. [Online]. Available: https://doi.org/10.1145/3313831.3376394

[151] C. Braunagel, W. Rosenstiel, and E. Kasneci, "Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 4, pp. 10–22, winter 2017.

[152] E. Bozkir, O. Günlü, W. Fuhl, R. F. Schaefer, and E. Kasneci, "Differential privacy for eye tracking with temporal correlations," *PLoS ONE*, aug 2021.

[153] S. Bazrafkan, A. Kar, and C. Costache, "Eye gaze for consumer electronics: Controlling and commanding intelligent systems." *IEEE Consumer Electronics Magazine*, vol. 4, no. 4, pp. 65–71, Oct 2015.

[154] P. M. Corcoran, F. Nanu, S. Petrescu, and P. Bigioi, "Real-time eye gaze tracking for gaming design and consumer electronics systems," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 2, pp. 347–355, May 2012.

[155] A. Bissoli, D. Lavino-Junior, M. Sime, L. Encarnacao, and T. Bastos-Filho, "A human machine interface based on eye tracking for controlling and monitoring a smart home using the internet of things," *Sensors*, vol. 19, no. 4, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/4/859

[156] H. Dalal, *Implementing Technology as Mother Nature intended to be - its impact on a wide range of experiences and market opportunities*, Jul. 2019. [Online]. Available: https://blog.tobii.com/introducing-tobii-spotlight-technology-767b79959ee3

[157] J. Lin, D. Cheng, C. Yao, and Y. Wang, "Retinal projection head-mounted display," *Frontiers of Optoelectronics*, vol. 10, no. 1, pp. 1–8, Mar 2017. [Online]. Available: https://doi.org/10.1007/s12200-016-0662-8

[158] C. Jang, K. Bang, S. Moon, J. Kim, S. Lee, and B. Lee, "Retinal 3d: Augmented reality near-eye display via pupil-tracked light field projection on retina," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 190:1–190:13, Nov. 2017. [Online]. Available: http://doi.acm.org/10.1145/3130800.3130889

[159] Y. Shoji, "Device for tracking pupil of eye ball using intensity changes of reflected light from eyeball and image display using the same," U.S. Patent Patent, US20 070 705 471, 2009, uS20070705471.

[160] C. Ahlstrom, K. Kircher, and A. Kircher, "A gaze-based driver distraction warning system and its effect on visual behavior," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 965–973, June 2013.

[161] H. Cecotti, "A multimodal gaze-controlled virtual keyboard," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 4, pp. 601–606, Aug 2016.

[162] D. Bonino, E. Castellina, F. Corno, and L. D. Russis, "Dogeye: Controlling your home with eye interaction," *Interacting with Computers*, vol. 23, no. 5, pp. 484 – 498, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0953543811000701

[163] R. H. Webb, G. W. Hughes, and O. Pomerantzeff, "Flying spot tv ophthalmoscope," *Appl. Opt.*, vol. 19, no. 17, pp. 2991–2997, Sep 1980. [Online]. Available: http://ao.osa.org/abstract.cfm?URI=ao-19-17-2991

[164] C. K. Sheehy, Q. Yang, D. W. Arathorn, P. Teeruveedhula, J. F. de Boer, and A. J. Roorda, "High-speed, image-based eye tracking with a scanning laser ophthalmoscope," *Biomedical Optics Express*, vol. 3, no. 10, pp. 2611–2622, 2012.

[165] S. Stevenson, C. Sheehy, and A. Roorda, "Binocular eye tracking with the tracking scanning laser ophthalmoscope," *Vision Research*, vol. 118, pp. 98 – 104, 2016, fixational eye movements and perception. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S004269891500036X

[166] D. Carpentras and C. Moser, "See-through ophthalmoscope for retinal imaging," *Journal of Biomedical Optics*, vol. 22, no. 5, pp. 1 – 5, 2017. [Online]. Available: https://doi.org/10.1117/1.JBO.22.5.056006

[167] T. DuBose, D. Nankivil, F. LaRocca, G. Waterman, K. Hagan, J. Polans, B. Keller, D. Tran-Viet, L. Vajzovic, A. N. Kuo, C. A. Toth, J. A. Izatt, and S. Farsiu, "Handheld adaptive optics scanning laser ophthalmoscope," *Optica*, vol. 5, no. 9, pp. 1027–1036, Sep 2018. [Online]. Available: http://www.osapublishing.org/optica/abstract.cfm?URI=optica-5-9-1027

[168] O. Sakhno, Y. Gritsai, H. Sahm, J. Stumpe, and M. Wegener, "Bragg polarization gratings used as switchable elements in AR/VR holographic displays," in *Digital Optics for Immersive Displays*, B. C. Kress, W. Osten, and H. Stolle, Eds., vol. 10676, International Society for Optics and Photonics. SPIE, 2018, pp. 91 – 107. [Online]. Available: https://doi.org/10.1117/12.2309788

[169] W. R. Folks, J. M. López-Alonso, B. Monacelli, A. R. Weeks, G. Zummo, D. Mullaly, and G. D. Boreman, "Characterization of digital-micromirror device-based infrared scene projector," *Optical Engineering*, vol. 44, no. 8, pp. 1 – 7, 2005.

[170] L. Y. Lin and E. G. Keeler, "Progress of mems scanning micromirrors for optical bio-imaging," *Micromachines*, vol. 6, no. 11, pp. 1675–1689, 2015. [Online]. Available: https://www.mdpi.com/2072-666X/6/11/1450

[171] B. O'Bard and K. George, "Classification of eye gestures using machine learning for use in embedded switch controller," in *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, May 2018, pp. 1–6.

[172] H. Salehifar, P. Bayat, and M. A. Majd, "Eye gesture blink password: a new authentication system with high memorable and maximum password length," *Multimedia Tools and Applications*, vol. 78, no. 12, pp. 16 861–16 885, Jun 2019. [Online]. Available: https://doi.org/10.1007/s11042-018-7043-9

[173] G. A. Koulieris, K. Akşit, M. Stengel, R. K. Mantiuk, K. Mania, and C. Richardt, "Near-eye display and tracking technologies for virtual and augmented reality," *Computer Graphics Forum*, vol. 38, no. 2, pp. 493–519, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13654

[174] A. Maimone, A. Georgiou, and J. S. Kollin, "Holographic near-eye displays for virtual and augmented reality," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 85:1–85:16, Jul. 2017. [Online]. Available: http://doi.acm.org/10.1145/3072959.3073624

[175] J. Kim, M. Stengel, A. Majercik, S. De Mello, D. Dunn, S. Laine, M. McGuire, and D. Luebke, "Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: ACM, 2019, pp. 550:1–550:12. [Online]. Available: http://doi.acm.org/10.1145/3290605.3300780

[176] A. Russell, "Osram os infrared in ar/vr," online, Sep. 2019. [Online]. Available: http://microvision.blogspot.com/2019/10/st-micro-blog-and-osram-powerpoint.html

[177] M. von Waldkirch, P. Lukowicz, and G. Tröster, "Effect of light coherence on depth of focus in head-mounted retinal projection displays," *Optical Engineering*, vol. 43, no. 7, pp. 1552 – 1560, 2004. [Online]. Available: https://doi.org/10.1117/1.1748209

[178] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[179] A. T. Duchowski, "Gaze-based interaction: A 30 year retrospective," *Computers & Graphics*, vol. 73, pp. 59 – 69, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0097849318300487

[180] W. Fuhl, D. Hospach, T. C. Kübler, W. Rosenstiel, O. Bringmann, and E. Kasneci, "Ways of improving the precision of eye tracking data: Controlling the influence of dirt and dust on pupil detection," *Journal of Eye Movement Research*, vol. 10, no. 3, May 2017. [Online]. Available: https://bop.unibe.ch/JEMR/article/view/3657

[181] W. Fuhl, "1000 pupil segmentations in a second using haar like features and statistical learning," *arXiv preprint arXiv:2102.01921*, 2021.

[182] T. Wilm, S. Höckh, R. Fiess, and W. Stork, "Holographic combiners for augmented reality applications fabricated by wave front recording," in *Novel Optical Systems, Methods, and Applications XXIV*, C. F. Hahlweg and J. R. Mulley, Eds., vol. 11815,

International Society for Optics and Photonics. SPIE, 2021, pp. 13 – 22. [Online]. Available: https://doi.org/10.1117/12.2596838

[183] J. Zhao, B. D. Chrysler, and R. K. Kostuk, "Design of a waveguide eye-tracking system operating in near-infrared with holographic optical elements," *Optical Engineering*, vol. 60, no. 8, 2021. [Online]. Available: https://doi.org/10.1117/1.OE.60.8.085101

[184] W. A. Donnelly, "The advanced human eye model (ahem): a personal binocular eye modeling system inclusive of refraction, diffraction, and scatter." *Journal of refractive surgery*, vol. 24 9, pp. 976–83, 2008. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.518.4567&rep=rep1&type=pdf

[185] K. Nguyen, C. Wagner, D. Koons, and M. Flickner, "Differences in the infrared bright pupil response of human eyes," in *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, ser. ETRA '02. New York, NY, USA: Association for Computing Machinery, 2002, p. 133–138. [Online]. Available: https://doi.org/10.1145/507072.507099

[186] J. W. Goodman, *Speckle Phenomena in Optics - Theory and Applications*. Bellingham, Washington: SPIE Press, 2020.

[187] J. M. Schmitt, S. H. Xiang, and K. M. Yung, "Speckle in optical coherence tomography," *Journal of Biomedical Optics*, vol. 4, no. 1, pp. 95 – 105, 1999. [Online]. Available: https://doi.org/10.1117/1.429925

[188] B. Sensortec, *Smartglasses Light Drive*, 2020.

[189] S. De Ma, "Conics-based stereo, motion estimation, and pose determination," *Int. J. Comput. Vision*, 1993. [Online]. Available: https://doi.org/10.1007/BF01440844

[190] M. Lidegaard, D. W. Hansen, and N. Krüger, "Head mounted device for point-of-gaze estimation in three dimensions," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ser. ETRA 14, New York, NY, USA, 2014, p. 83–86. [Online]. Available: https://doi.org/10.1145/2578153.2578163

[191] C. Lander, M. Löchtefeld, and A. Krüger, "Heyebrid: A hybrid approach for mobile calibration-free gaze estimation," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 4, 2018. [Online]. Available: https://doi.org/10.1145/3161166

[192] S. T. S. Holmström, U. Baran, and H. Urey, "Mems laser scanners: A review," *Journal of Microelectromechanical Systems*, vol. 23, no. 2, pp. 259–275, 2014.

[193] F. Bruder, T. Fäcke, and T. Rölle, "The Chemistry and Physics of Bayfol® HX Film Holographic Photopolymer," *Polymers*, vol. 9, 2017.

[194] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 1, 1999.

[195] B. Sensortec, "Bml050 microscanner," online, 2017. [Online]. Available: https://www.datasheetarchive.com/whats_new/2c340e9c53ac33ba46736a3016892aec.html

[196] D. K. Maj Isabelle Olsson, Mitchell Joseph Heinrich and J. Lapetina, "Wearable device with input and output structures," US Patent US2 013 044 042A, 2013.

[197] S. Natsume, "Virtual reality glasses," US Patent USD795 952S, 2017.

[198] R. Bednarik, T. Gowases, and M. Tukiainen, "Gaze interaction enhances problem solving: Effects of dwell-time based, gaze-augmented, and mouse interaction on problem-solving strategies and user experience," *Journal of Eye Movement Research*, vol. 3, no. 1, Jun. 2009. [Online]. Available: https://bop.unibe.ch/JEMR/article/view/2287

[199] R. D. Findling, L. N. Nguyen, and S. Sigg, "Closed-eye gaze gestures: Detection and recognition of closed-eye movements with cameras in smart glasses," in *Advances in Computational Intelligence*, I. Rojas, G. Joya, and A. Catala, Eds. Cham: Springer International Publishing, 2019, pp. 322–334.

[200] T. Santini, W. Fuhl, T. Kübler, and E. Kasneci, "Bayesian identification of fixations, saccades, and smooth pursuits," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, ser. ETRA '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 163–170. [Online]. Available: https://doi.org/10.1145/2857491.2857512

[201] M. Startsev, I. Agtzidis, and M. Dorr, "1d cnn with blstm for automated classification of fixations, saccades, and smooth pursuits," *Behavior Research Methods*, vol. 51, no. 2, pp. 556–572, 2019.

[202] S. Hoppe and A. Bulling, "End-to-end eye movement detection using convolutional neural networks," 2016.

[203] Google LLC, "Glass - Glass," https://www.google.com/glass/start/, 2021, Online; accessed 15 April 2021.

[204] Epson, "Moverio bt-300 smart glasses," 2020. [Online]. Available: https://www.epson.com.hk/For-Home/Wearables/Smart-Glasses/Moverio-BT-300-Smart-Glasses/p/V11H756060

[205] M. Islam, S. Sakamoto, Y. Yamada, A. Vargo, M. Iwata, M. Iwamura, K. Kise *et al.*, "Self-supervised deep learning for reading activity classification," *arXiv preprint arXiv:2012.03598*, 2020.

[206] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Comput. Surv.*, vol. 46, no. 3, Jan. 2014. [Online]. Available: https://doi.org/10.1145/2499621

[207] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep learning models for real-time human activity recognition with smartphones," *Mobile Networks and Applications*, vol. 25, 04 2020.

[208] Microsoft Corporation, "Kinect - Windows app development," https://developer.microsoft.com/en-gb/windows/kinect/, 2021, Online; accessed 14 April 2021.

[209] Microsoft Corporation, "Azure Kinect DK - Develop AI models | Microsoft Azure," https://azure.microsoft.com/en-gb/services/kinect-dk/, 2021, Online; accessed 14 April 2021.

[210] Nintendo Co., Ltd., "Nintendo Switch | Nintendo Switch Family | Nintendo," https://www.nintendo.co.uk/Nintendo-Switch/Nintendo-Switch-1148779.html, 2021, Online; accessed 14 April 2021.

[211] Runtopia Technology Ltd, "RUNTOPIA|REACH. Faster, Longer, Safer." https://landing.runtopia.net/, 2021, Online; accessed 14 April 2021.

[212] Sensoria Inc., "Sensoria Home Page," https://www.sensoriafitness.com/, 2021, Online; accessed 14 April 2021.

[213] SALTED Ltd., "SALTED," https://www.salted.ltd/en/insole, 2021, Online; accessed 14 April 2021.

[214] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges and opportunities," 2020.

[215] S. M. Azar, M. G. Atigh, A. Nickabadi, and A. Alahi, "Convolutional relational machine for group activity recognition," 2019.

[216] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *Pervasive Computing*, A. Ferscha and F. Mattern, Eds.   Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 1–17.

[217] X. Li, Y. Zhang, I. Marsic, A. Sarcevic, and R. S. Burd, "Deep learning for rfid-based activity recognition," in *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*, ser. SenSys '16.   New York, NY, USA: Association for Computing Machinery, 2016, p. 164–175. [Online]. Available: https://doi.org/10.1145/2994551.2994569

[218] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *SIGKDD Explor. Newsl.*, vol. 12, no. 2, p. 74–82, Mar. 2011. [Online]. Available: https://doi.org/10.1145/1964897.1964918

[219] F. Wahl, M. Freund, and O. Amft, "Wiseglass: Smart eyeglasses recognising context," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 2, no. 5, 12 2015.

[220] T. Hayashi, M. Nishida, N. Kitaoka, T. TODA, and K. Takeda, "Daily activity recognition with large-scaled real-life recording datasets based on deep neural network using multi-modal signals," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E101.A, pp. 199–210, 01 2018.

[221] A. Bulling, J. A. Ward, H. Gellersen, and G. Tröster, "Robust recognition of reading activity in transit using wearable electrooculography," in *Pervasive Computing*, J. Indulska, D. J. Patterson, T. Rodden, and M. Ott, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 19–37.

[222] A. Bulling, C. Weichel, and H. Gellersen, "Eyecontext: Recognition of high-level contextual cues from human visual behaviour," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 305–308. [Online]. Available: https://doi.org/10.1145/2470654.2470697

[223] J. Steil and A. Bulling, "Discovery of everyday human activities from long-term visual behaviour using topic models," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 75–85. [Online]. Available: https://doi.org/10.1145/2750858.2807520

[224] G. Lan, B. Heit, T. Scargill, and M. Gorlatova, *GazeGraph: Graph-Based Few-Shot Cognitive Context Sensing from Human Visual Behavior*. New York, NY, USA: Association for Computing Machinery, 2020, p. 422–435. [Online]. Available: https://doi.org/10.1145/3384419.3430774

[225] A. Bulling, J. A. Ward, and H. Gellersen, "Multimodal recognition of reading activity in transit using body-worn sensors," *ACM Trans. Appl. Percept.*, vol. 9, no. 1, Mar. 2012. [Online]. Available: https://doi.org/10.1145/2134203.2134205

[226] S. Ishimaru, K. Kunze, K. Kise, J. Weppner, A. Dengel, P. Lukowicz, and A. Bulling, "In the blink of an eye: Combining head motion and eye blink frequency for activity recognition with google glass," in *Proceedings of the 5th Augmented Human International Conference*, ser. AH '14. New York, NY, USA: Association for Computing Machinery, 2014. [Online]. Available: https://doi.org/10.1145/2582051.2582066

[227] S. Ishimaru, K. Kunze, Y. Uema, K. Kise, M. Inami, and K. Tanaka, "Smarter eyewear: Using commercial eog glasses for activity recognition," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, ser. UbiComp '14 Adjunct. New York, NY, USA: Association for Computing Machinery, 2014, p. 239–242. [Online]. Available: https://doi.org/10.1145/2638728.2638795

Bibliography

[228] C. Braunagel, E. Kasneci, W. Stolzmann, and W. Rosenstiel, "Driver-activity recognition in the context of conditionally autonomous driving," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, 2015, pp. 1652–1657.

[229] J. Steil, P. Müller, Y. Sugano, and A. Bulling, "Forecasting user attention during everyday mobile interactions using device-integrated and wearable sensors," in *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, ser. MobileHCI '18.   New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3229434.3229439

[230] J.-B. Yang, N. Nhut, P. San, X. li, and P. Shonali, "Deep convolutional neural networks on multichannel time series for human activity recognition," *IJCAI*, 07 2015.

[231] A. Ignatov, "Real-time human activity recognition from accelerometer data using convolutional neural networks," *Applied Soft Computing*, vol. 62, pp. 915–922, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494617305665

[232] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Systems with Applications*, vol. 59, pp. 235–244, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417416302056

[233] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, 2016. [Online]. Available: https://www.mdpi.com/1424-8220/16/1/115

[234] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[235] B. Chikhaoui, F. Gouineau, and M. Sotir, "A cnn based transfer learning model for automatic activity recognition from accelerometer sensors," in *Machine Learning and Data Mining in Pattern Recognition*, P. Perner, Ed.   Cham: Springer International Publishing, 2018, pp. 302–315.

[236] B. Sensortec, "Bmi270 6-axis, smart, low-power inertial measurement unit for high-performance applications," 2020. [Online]. Available: https://www.bosch-sensortec.com/products/motion-sensors/imus/bmi270

[237] A. Kumar, S. Goyal, and M. Varma, "Resource-efficient machine learning in 2 KB RAM for the internet of things," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70.   PMLR, 06–11 Aug 2017, pp. 1935–1944. [Online]. Available: http://proceedings.mlr.press/v70/kumar17a.html

[238] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.

[239] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum, and C. Igel, "U-sleep: resilient high-frequency sleep staging," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–12, 2021.

[240] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, 2020.

[241] J. Riemensberger, A. Lukashchuk, M. Karpov, W. Weng, E. Lucas, J. Liu, and T. J. Kippenberg, "Massively parallel coherent laser ranging using a soliton microcomb," *Nature*, vol. 581, no. 7807, pp. 164–170, 2020.

[242] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention.* Springer, 2015, pp. 234–241.

[243] P. Elspas, Y. Klose, S. T. Isele, J. Bach, and E. Sax, "Time series segmentation for driving scenario detection with fully convolutional networks." in *VEHITS*, 2021, pp. 56–64.

[244] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[245] G. Capelli and G. Giuliani, "Laser velocimeter for the measurement of eye movements," in *2011 International Workshop on Biophotonics*, 2011, pp. 1–3.

[246] S. Thainimit, L. A. Alexandre, and V. M. N. de Almeida, "Iris surface deformation and normalization," in *2013 13th International Symposium on Communications and Information Technologies (ISCIT)*, 2013, pp. 501–506.

[247] N. Mitsuhashi, K. Fujieda, T. Tamura, S. Kawamoto, T. Takagi, and K. Okubo, "Body-Parts3D 3D structure database for anatomical concepts," *Nucleic Acids Research*, 10 2008.

[248] R. Kliese and A. Rakić, "Spectral broadening caused by dynamic speckle in self-mixing velocimetry sensors," *Optics express*, vol. 20, no. 17, pp. 18 757–18 771, 2012.

[249] C. M. Bishop, "Pattern recognition," *Machine learning*, vol. 128, no. 9, 2006.

[250] I. Bekerman, P. Gottlieb, and M. Vaiman, "Variations in eyeball diameters of the healthy adults," *Journal of ophthalmology*, vol. 2014, 2014.

[251] R. M. Johnstone, C. R. Johnson, R. R. Bitmead, and B. D. O. Anderson, "Exponential convergence of recursive least squares with exponential forgetting factor," in *1982 21st IEEE Conference on Decision and Control*, 1982, pp. 994–997.

[252] V. Shaferman, M. Schwegel, T. Glück, and A. Kugi, "Continuous-time least-squares forgetting algorithms for indirect adaptive control," *European Journal of Control*, vol. 62, pp. 105–112, 2021.

[253] J. J. Moré, "The levenberg-marquardt algorithm: implementation and theory," in *Numerical analysis.* Springer, 1978, pp. 105–116.