Towards Everyday Virtual Reality through Eye Tracking

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M. Sc. Efe BOZKIR

aus Karşıyaka, Türkei

Tübingen

2021

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

Tag der mündlichen Qualifikation:	25.02.2022
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatterin:	Prof. Dr. Enkelejda Kasneci
2. Berichterstatter:	Jun.-Prof. Dr. Michael Krone

"Our true mentor in life is science."
– Mustafa Kemal Atatürk

To my family

Acknowledgments

Firstly, I would like to thank my supervisor, Prof. Dr. Enkelejda Kasneci for accepting me to the Chair for Human-Computer Interaction (former Perception Engineering), guiding me throughout my doctoral studies, and for being one of the most supportive people that I have ever worked with through my entire academic and professional career. Special and sincere thanks to Jun.-Prof. Dr. Michael Krone, Prof. Dr. Oliver Bringmann, and Prof. Dr. Richard Göllner for evaluating my work. Furthermore, I thank Margot Reimold for helping me with all the bureaucratic burdens and making my life easier.

During this work, I was quite lucky to have plenty of pleasant and fruitful collaborations, including interdisciplinary ones. I particularly thank Prof. Dr. Richard Göllner, Dr. Lisa Hasenbein, Philipp Stark, and Joseph Ferdinand for the engaging collaborations on the classroom studies. Furthermore, I thank Prof. Dr. Rafael F. Schaefer, Prof. Dr. Nico Pfeifer, Dr. Onur Günlü, Dr. Mete Akgün, and Ali Burak Ünal for the delightful discussions and work on the privacy-related topics. In addition, I thank our collaborators in Greece in the DAAD project, especially Prof. Dr. Athanassios Skodras.

Without an enjoyable work environment, I would not be successful. Therefore, I thank the past and present members of Chair for Human-Computer Interaction, many of them becoming my collaborators in the meantime, including Dr. Wolfgang Fuhl, Dr. Shahram Eivazi, Dr. Thomas Kübler, Dr. Nora Castner, Dr. David Geisler, Dr. Thiago Santini, Yao Rong, Benedikt Hosp, Daniel Weber, Björn Severitt, and others. Special thanks go to my co-workers in the office, Dr. Ömer Sümer, Hong Gao, and Babette Bühler, for the privilege of beneficial discussions and for sharing nice moments. Furthermore, I thank the members of the Data Science & Analytics Research Group for the nice ambiance and especially Martin Pawelczyk for fruitful brainstormings on different ideas.

A great appreciation goes to my friends in Tübingen outside the lab for all the fun, silly jokes, and unforgettable memories together. I particularly thank Caner Bağcı, Dr. Murat Seçkin Ayhan, Dr. Mete Akgün, Ali Burak Ünal, Mehmet Direnç Mungan, Emre Barış Karaaslan, Dr. Ezgi Süheyla Karaaslan, and Bilge Sürün. Moreover, I thank my friends in Munich, especially Selin Kenet, Umut Kaya, and Ümit Suat Mayadalı. Even though we were physically away from each other during all that time, I very much enjoyed our online games and chats, especially during the pandemic. You made my life more liveable during weird times.

Acknowledgments

I have been living away from my family for a long time but thanks to their warm welcome in Tübingen from the very first day, I have again remembered the family feeling with them. My thanks go to my cousin Dr. Acun Papakçı, Dr. Elif Kösedağı, Arinna, and Taru.

Being the only child of my parents led me to have the most incredible friends ever. Even though I live far away from all of you, I feel your sister- and brotherhood all the time. My thanks go to Esra İçer, Dr. Arda Erdut, Utku Barış Özsoy, Kutay Yücetürk, Yuşa Öz, Ömer Kaya, Emin Melih Özsoy, Levent Budakoğlu, Altuğ Bayram, and Mertcan Yığın. Moreover, I sincerely thank my cousin Batuhan Sarıoğlu for being there all the years, even though our paths fell physically apart after studying together in Istanbul. Additionally, I thank uncle Ercüment and Ayhan, for their encouragement and trust even starting from my high school years. I really appreciate it.

Lastly and most importantly, I thank my mother Neşe and my father Fevzi for their unconditional love, support, encouragement, and confidence. These literally mean the world to me.

Efe BOZKIR

Abstract

With developments in computer graphics, hardware technology, perception engineering, and human-computer interaction, virtual reality and virtual environments are becoming more integrated into our daily lives. Head-mounted displays, however, are still not used as frequently as other mobile devices such as smart phones and watches. With increased usage of this technology and the acclimation of humans to virtual application scenarios, it is possible that in the near future an everyday virtual reality paradigm will be realized.

When considering the marriage of everyday virtual reality and head-mounted displays, eye tracking is an emerging technology that helps to assess human behaviors in a real time and non-intrusive way. Still, multiple aspects need to be researched before these technologies become widely available in daily life. Firstly, attention and cognition models in everyday scenarios should be thoroughly understood. Secondly, as eyes are related to visual biometrics, privacy preserving methodologies are necessary. Lastly, instead of studies or applications utilizing limited human participants with relatively homogeneous characteristics, protocols and use-cases for making such technology more accessible should be essential.

In this work, taking the aforementioned points into account, a significant scientific push towards everyday virtual reality has been completed with three main research contributions. Human visual attention and cognition have been researched in virtual reality in two different domains, including education and driving. Research in the education domain has focused on the effects of different classroom manipulations on human visual behaviors, whereas research in the driving domain has targeted safety related issues and gaze-guidance. The user studies in both domains show that eye movements offer significant implications for these everyday setups. The second substantial contribution focuses on privacy preserving eye tracking for the eye movement data that is gathered from head-mounted displays. This includes differential privacy, taking temporal correlations of eye movement signals into account, and privacy preserving gaze estimation task by utilizing a randomized encoding-based framework that uses eye landmarks. The results of both works have indicated that privacy considerations are possible by keeping utility in a reasonable range. Even though few works have focused on this aspect of eye tracking until now, more research is necessary to support everyday virtual reality. As a final significant contribution, a blockchain- and smart contract-based eye tracking data collection protocol for virtual reality is proposed to make virtual reality more accessible. The findings present valuable insights for everyday virtual reality and advance the state-of-the-art in several directions.

Zusammenfassung

Durch Entwicklungen in den Bereichen Computergrafik, Hardwaretechnologie, Perception Engineering und Mensch-Computer Interaktion, werden Virtual Reality und virtuelle Umgebungen immer mehr in unser tägliches Leben integriert. Head-Mounted Displays werden jedoch im Vergleich zu anderen mobilen Geräten, wie Smartphones und Smartwatches, noch nicht so häufig genutzt. Mit zunehmender Nutzung dieser Technologie und der Gewöhnung von Menschen an virtuelle Anwendungsszenarien ist es wahrscheinlich, dass in naher Zukunft ein alltägliches Virtual-Reality-Paradigma realisiert wird.

Im Hinblick auf die Kombination von alltäglicher Virtual Reality und Head-Mounted-Displays, ist Eye Tracking eine neue Technologie, die es ermöglicht, menschliches Verhalten in Echtzeit und nicht-intrusiv zu messen. Bevor diese Technologien in großem Umfang im Alltag eingesetzt werden können, müssen jedoch noch zahlreiche Aspekte genauer erforscht werden. Zunächst sollten Aufmerksamkeits- und Kognitionsmodelle in Alltagsszenarien genau verstanden werden. Des Weiteren sind Maßnahmen zur Wahrung der Privatsphäre notwendig, da die Augen mit visuellen biometrischen Indikatoren assoziiert sind. Zuletzt sollten anstelle von Studien oder Anwendungen, die sich auf eine begrenzte Anzahl menschlicher Teilnehmer mit relativ homogenen Merkmalen stützen, Protokolle und Anwendungsfälle für eine bessere Zugänglichkeit dieser Technologie von wesentlicher Bedeutung sein.

In dieser Arbeit wurde unter Berücksichtigung der oben genannten Punkte ein bedeutender wissenschaftlicher Vorstoß mit drei zentralen Forschungsbeiträgen in Richtung alltäglicher Virtual Reality unternommen. Menschliche visuelle Aufmerksamkeit und Kognition innerhalb von Virtual Reality wurden in zwei unterschiedlichen Bereichen, Bildung und Autofahren, erforscht. Die Forschung im Bildungsbereich konzentrierte sich auf die Auswirkungen verschiedener Manipulationen im Klassenraum auf das menschliche Sehverhalten, während die Forschung im Bereich des Autofahrens auf sicherheitsrelevante Fragen und Blickführung abzielte. Die Nutzerstudien in beiden Bereichen zeigen, dass Blickbewegungen signifikante Implikationen für diese alltäglichen Situationen haben. Der zweite wesentliche Beitrag fokussiert sich auf Privatsphäre bewahrendes Eye Tracking für Blickbewegungsdaten von Head-Mounted Displays. Dies beinhaltet Differential Privacy, welche zeitliche Korrelationen von Blickbewegungssignalen berücksichtigt und Privatsphäre wahrende Blickschätzung durch Verwendung eines auf randomisiertem Encoding basierenden Frameworks, welches Augenreferenzunkte verwendet. Die Ergebnisse beider Arbeiten zeigen, dass die Wahrung der Privatsphäre möglich ist und gleichzeitig der Nutzen in einem akzeptablen Bereich bleibt. Wenngleich es bisher nur

wenig Forschung zu diesem Aspekt von Eye Tracking gibt, ist weitere Forschung notwendig, um den alltäglichen Gebrauch von Virtual Reality zu ermöglichen. Als letzter signifikanter Beitrag, wurde ein Blockchain- und Smart Contract-basiertes Protokoll zur Eye Tracking Datenerhebung für Virtual Reality vorgeschlagen, um Virtual Reality besser zugänglich zu machen. Die Ergebnisse liefern wertvolle Erkenntnisse für alltägliche Nutzung von Virtual Reality und treiben den aktuellen Stand der Forschung in mehrere Richtungen voran.

Contents

List of Figures

List of Tables

List of Abbreviations

ADHD	Attention-Deficit/Hyperactivity Disorder
AMT	Amazon Mechanical Turk
ANOVA	Analysis of Variance
AR	Augmented Reality
ART	Aligned Rank Transform
CFPA	Chunk-based Fourier Perturbation Algorithm
CPT	Continuous Performance Task
DARE	Decomposable and Affine Randomized Encoding
DCFPA	Difference- and Chunk-based Fourier Perturbation Algorithm
DFT	Discrete Fourier Transform
DP	Differential Privacy
DT	Decision Tree
EEG	Electroencephalography
ETH	Ether
FOV	Field of View
FPA	Fourier Perturbation Algorithm
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
HCI	Human-Computer Interaction
HDD	Head-down Display
HMAC	Keyed-Hashing for Message Authentication
HMD	Head-mounted Display
HUD	Head-up Display
I-VT	Velocity-Threshold Identification
IDFT	Inverse Discrete Fourier Transform
IVR	Immersive Virtual Reality
k-NN	k-Nearest Neighbor
LPA	Laplace Perturbation Algorithm
MR	Mixed Reality
NMSE	Normalized Mean Square Error
PDF	Probability Density Function
OOI	Object-of-Interest
RBF	Radial Basis Function

RE	Randomized Encoding
RF	Random Forest
SHA	Secure Hash Algorithm
SMC	Secure Multi-party Computation
SVM	Support Vector Machine
SVR	Support Vector Regression
TTC	Time-to-Collision
VR	Virtual Reality
XR	Extended Reality

1 List of Publications

Accepted Publications Relevant to This Thesis

1. **Efe Bozkir**, Onur Günlü, Wolfgang Fuhl, Rafael F. Schaefer, and Enkelejda Kasneci. Differential privacy for eye tracking with temporal correlations. *PLoS ONE*, 16(8):e0255979, 2021. doi: 10.1371/journal.pone.0255979.
2. Hong Gao, **Efe Bozkir**, Lisa Hasenbein, Jens-Uwe Hahn, Richard Göllner, and Enkelejda Kasneci. Digital transformations of classrooms in virtual reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2021. ACM. doi: 10.1145/3411764.3445596.
3. **Efe Bozkir**, Philipp Stark, Hong Gao, Lisa Hasenbein, Jens-Uwe Hahn, Enkelejda Kasneci, and Richard Göllner. Exploiting object-of-interest information to understand attention in VR classrooms. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, New York, NY, USA, 2021. IEEE. doi: 10.1109/VR50410.2021.00085.
4. **Efe Bozkir**, Shahram Eivazi, Mete Akgün, and Enkelejda Kasneci. Eye tracking data collection protocol for VR for remotely located subjects using blockchain and smart contracts. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR) Work-in-progress papers*, New York, NY, USA, 2020. IEEE. doi: 10.1109/AIVR50618.2020.00083.
5. **Efe Bozkir**, Ali Burak Ünal, Mete Akgün, Enkelejda Kasneci, and Nico Pfeifer. Privacy preserving gaze estimation using synthetic images via a randomized encoding based framework. In *ACM Symposium on Eye Tracking Research and Applications*, New York, NY, USA, 2020. ACM. doi: 10.1145/3379156.3391364.
6. **Efe Bozkir**, David Geisler, and Enkelejda Kasneci. Assessment of driver attention during a safety critical situation in VR to generate VR-based training. In *ACM Symposium on Applied Perception 2019*, New York, NY, USA, 2019. ACM. doi: 10.1145/3343036.3343138.
7. **Efe Bozkir**, David Geisler, and Enkelejda 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) Workshops*, New York, NY, USA, 2019. IEEE. doi: 10.1109/VR.2019.8797758.

Other Publications not Relevant to This Thesis

8. Wolfgang Fuhl, **Efe Bozkir**, and Enkelejda Kasneci. Reinforcement learning for the privacy preservation and manipulation of eye tracking data. In *Artificial Neural Networks and Machine Learning – ICANN 2021*, Cham, Switzerland, 2021. Springer. doi: 10.1007/978-3-030-86380-7_48
9. Ömer Sümer, **Efe Bozkir**, Thomas Kübler, Sven Grüner, Sonja Utz, and Enkelejda Kasneci. FakeNewsPerception: An eye movement dataset on the perceived believability of news stories. *Data in Brief*, 35, 2021. doi: 10.1016/j.dib.2021.106909.
10. Wolfgang Fuhl, **Efe Bozkir**, Benedikt Hosp, Nora Castner, David Geisler, Thiago C. Santini, and Enkelejda Kasneci. "Encodji: Encoding gaze data into emoji space for an amusing scanpath classification approach". In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, New York, NY, USA, 2019. ACM. doi: 10.1145/3314111.3323074.

1.1 Scientific Contribution

This work advances the state-of-the-art in multiple directions and is considered a significant research step towards the achievement of everyday virtual reality through eye tracking. Contributions include valuable insights on human visual attention and cognition studied in multiple domains, namely education and driving, privacy preserving manipulation of eye movements with differential privacy and a randomized encoding-based framework, and a versatile protocol employing blockchain and smart contracts for eye tracking data collection suitable for subjects that are located remotely in virtual reality.

Chapter 2 introduces each topic under the umbrella subject of everyday virtual reality. Seven scientific publications at renowned venues from 2019 to 2021 served as motivation for this work and are introduced in Chapter 3 along with their fundamental methodologies and findings. Chapter 4 discusses the outcomes and provides an outlook on how to make virtual reality collectively more available in human everyday life.

2 Introduction

With recent developments in different fields of computer science, such as computer graphics, sensing technology, and artificial intelligence, and with the decreased cost of smart glasses and head-mounted displays (HMDs) [1], virtual and augmented reality (VR/AR) are fast becoming integrated into our daily lives. It is estimated that the VR headset market size will grow at an annual rate of ≈ 28% from 2021 to 2028 [2]. This indicates that humans will use such devices more regularly in variety of daily routines, including entertainment, education, and training. Currently, different devices with wide range of technical capabilities are available, from cheap, low-end options like Google Cardboard [3] to high-end devices like HTC Vive Pro Eye [4].

Virtual reality is defined by LaValle [5, p. 1] as "inducing a targeted behavior in an organism by using sensory stimulation, while the organism has little or no awareness of the interference.". In his definition, there are four main components including targeted behavior, organism, sensory stimulation, and awareness. Targeted behavior is essentially defined as the experience that the living organism is having. For instance, it could be the experience of a student attending a lecture in an immersive VR classroom or a novice driver training in VR for safety critical situations that could occur in the real world. According to LaValle [5], the organism could be any life form; however, in the context of this work, the humans are the focus. In addition, while sensory stimulation is carried out by integrating regular or alternative senses for humans, awareness of the experience is related to sense of presence.

Sense of presence is an important issue for virtual reality. It can be discussed through a philosophic perspective by taking Matrix[1]-like utopic scenarios and environments into account. Alternative and virtual realities have been traced back to the 18^{th} century and are thought to originate with the writings of Immanuel Kant [7] who discussed the realities that occur in one's mind, but differ from the real world [5]. Later in 20^{th} century, Antonin Artaud used the term "la réalité virtuelle" to describe theatre as being similar to a second reality in his work [8]. The terms "artificial reality" and later "virtual reality" in the technical domain were used by Myron Krueger and Jaron Lanier in 1960s and 1980s, respectively [9, 10, 11].

[1] *The Matrix Trilogy* is a science fiction trilogy that was written and directed by the Wachowskis [6].

Philosophical discussions of terminology and the utmost possibilities of virtual realities and their effects on humans aside, VR researchers and practitioners have been working on different aspects of this technology, such as hardware, software, and human perception, for the last several decades. While we are still far from an extreme sense of presence in VR with the current technology due to reasons such as low resolutions in VR HMDs, limited field-of-view (FOV), or possible cybersickness when HMDs are being used, Jason Paul of NVIDIA has estimated that in 2017 we may be only two decades away from generating resolutions that match human eyes [12]. Working towards such a goal, research and development of VR hardware and software go hand in hand with research of human physiology and perception, thus connecting the perception engineering discipline as a whole [5].

With continuous efforts in perception engineering, it is likely that VR technology will be used more frequently in human daily life. Garner et al. [13] define everyday VR as any kind of activity that people are linked to once a day. The authors provide real world examples as well as use-cases for classrooms, skills training, and workplaces. Apart from the applications, there is also a push in the VR research community in this direction. In 2021, the Workshop on Everyday Virtual Reality was organized at the IEEE VR for the 7^{th} year in a row [14].

A number of applications and research questions in the VR domain are evaluated by using pre- or post-tests and questionnaires such as presence and realism [15, 16], simulator sickness [17], and mental workload [18]. While these evaluation methods are well established and psychologically relevant, they do not offer insights on the temporal dynamics of visual attention and cognition during the use of VR systems, a useful approach when everyday use-cases are considered. In particular, eye, head, and hand tracking sensors could be used for immersive VR environments that are realized with HMDs. Furthermore, the data acquired from these sensing modalities could be combined with self-reported measurements to gain more insights from users.

Eyes and their movements are of a particular interest since it is possible to analyze where people look at specific points in time along with the presented stimulus as long as an accurate estimation of eye regions and gaze is carried out. While in-the-wild scenarios for such tasks are more challenging due to illuminations or occlusions [19], HMD-based VR setups offer more controlled conditions as the eye tracking sensors are located within the HMDs. This enables not only eye images that can be recorded closer to eye regions, but also presents the possibility of configurations with controlled illumination that could be convenient for evaluating pupil related measurements for mental workload [20]. With recent developments in sensing technologies, it is possible to have integrated eye trackers in modern high-end HMDs (e.g., HTC Vive Pro Eye) or plug eye tracker sensors to such HMDs [21]. These have further eased the building of data collection pipelines for eye movements and the understanding of human visual attention using data from immersive virtual environments.

While it is possible to infer valuable information using eye movements in VR environments, such as salient regions of the scene [22], human intentions [23], or forecasting eye fixations [24]

which can be used for user assistive tasks, eyes include biometric information. For instance, personal authentication via iris textures is a well known approach [25, 26]. Additionally, eye movements [27] or the combination of aggregated eye movement features [28] assist with biometric authentication. Zhang et al. [27] report that eye movement-based authentication schemes could be used with VR devices for applications such as in-app purchases [29]. In addition, Steil et al. [28] found that people agree on sharing eye tracking data if a governmental health agency is available within the process or the data is used for research purposes. Taking authentication and personal identification possibilities into account, eye tracking data should be maintained in such a way that if authentication is not required for a task performed in the virtual environment, personal identities are still protected from adversaries. Still, other tasks which require eye movement data such as gaze guidance or foveated rendering should not be significantly affected due to privacy protection. Recently, in both Europe and in the US, data protection regulations have been legally introduced [30, 31]. With the increased amount of daily VR applications and their usage in the commercial domain, it is foreseeable that the applications and scenarios that enable daily usage of VR HMDs along with biological signals leading to authentication such as eye movements will require dedicated privacy protection mechanisms.

Apart from the potential of eye movements in VR and accompanying privacy concerns, another direction that would collectively increase the accessibility of VR environments is the enabling of remote interaction including multiple people as opposed to single player-like application scenarios. This is still a challenge as VR HMDs are not used as frequently as other personal devices like mobile phones or smart watches. One reason may be that such devices are perceived as being solely for entertainment purposes and, even with their decreased costs, it is not straightforward to use them in the daily life. However, especially with occurrences such as the COVID-19 pandemic, the value of remote work and collaboration has become more appreciated. With the immersion that VR setups provide, many domains could be digitally transformed. High-end HMDs along with native VR applications provide high quality resolutions and immersive and realistic environments. When multi-person setups and corresponding communication between people are considered, however, it is necessary to have communication via web-services and often by introducing third-party entities in order to enable such conditions. This creates opportunity for additional parties to manipulate data if the collected data is shared across parties. In addition, implementing and adapting such applications and systems requires a lot of effort. In terms of development efforts, accessibility, and cost, web-based VR is an emerging area that enables people to use VR environments at a lower cost with solutions such as Google Cardboard. While easier implementation and integration is inherent in such applications, it is not very straightforward to obtain complex 3D scenes and high quality sensor readings such as eye tracking (i.e., due to unavailability of standardized sensors and low sampling frequencies of such setups.). Overall, in the near or distant future both paradigms will be developed and utilized while VR environments and HMDs become mainstream in our daily lives. The issues with both approaches will be solved or mitigated depending on the prerequisites of the application domains and trade-offs will be

decided.

Taking all into consideration, the rest of this chapter is organized as follows. As the content of this thesis is research on everyday VR through the leveraging of eye movements in VR, privacy preserving eye tracking, and the accessibility of VR using biological signals such as eye movements, relevant introductions to each topic have been covered. The possibilities of eye tracking signals in VR are discussed in Section 2.1. Then, privacy considerations with an emphasis on authentication are explored in Section 2.2. Later, the accessibility of VR for everyday setups with a focus on eye movements is examined in Section 2.3. The final section, Section 2.4, introduces how these topics could be combined within the framework of everyday virtual reality.

2.1 Eye Tracking in Virtual Reality

Assessing eye movements could yield a plethora of possibilities for human-computer interaction. Eye movements represent the information gathered when humans look at specific areas of the presented stimulus. Such movements do not happen fully consciously. In the last several decades, researchers have used eye movements to evaluate human behaviors in different tasks or domains such as during reading [32], multimedia learning [33], web search [34], driving [35], in medicine [36], linguistics [37], user experience design [38], psychology [39], education [40], programming [41], or marketing [42]. This use in such a variety of applications and domains in fact shows the potential of eye movements for future directions. In the context of VR, eye movements are also used to assess human attention in an offline way or to actively provide gaze related interaction during virtual experiences. In this section, a proportion of works related to eye tracking and virtual reality are analyzed and reviewed.

To apply fine grained eye movement analyses in the context of immersive virtual environments with HMDs, first eye regions and gaze vectors are estimated using eye tracking sensors that are located within HMDs. Some HMDs provide the opportunity to use the integrated eye trackers (e.g., HTC Vive Pro Eye) or there is also the potential of integrating eye tracking plug-ins (e.g., Pupil-Labs Eye Trackers [21]) without significant effort. One could deploy custom and low-cost sensors [43] along with gaze estimation models and track the eyes as well. These approaches include the steps of estimating eye regions such as pupil and iris and detecting gaze directions using computer vision and machine learning techniques. While gaze estimation is especially important for detecting eye movements such as fixations or saccades, semantic segmentation of eye regions is also convenient for privacy reasons. One might, for example, want to obfuscate iris texture [44] if eye videos are saved during the experience. Kansal and Nathan [45] proposed a convolutional encoder-decoder network for eye region segmentation, whereas Chaudhary et al. [46] used combination of U-Net [47] and DenseNet [48] to carry out real-time semantic segmentation for eyes. Boutros et al. [49] have proposed a baseline multi-scale segmentation network in which the number of parameters are significantly diminished while reducing the overall accuracy marginally. Cycle GANs [50]

for eye region segmentation [51], ellipse segmentation framework for robust gaze tracking [52], fast and efficient few-shot segmentation for eyes [53], domain adaptation for eye segmentation [54], and identity-preserving eye image generation from semantic segmentation [55] have also been proposed in the literature. These works focused on practical usability, especially in terms of real-time working functionality or privacy-preservation which are optimal for VR/AR use-cases.

In addition to the segmentation tasks, different approaches have also been introduced for gaze estimation by using pupil detection [56, 57, 21, 58, 59, 60, 61, 62, 63, 64, 65] and recently by mainly machine learning [66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76]. While some of the aforementioned works are not directly designed for VR and HMDs, they provide implications for gaze estimation. Affordable and low-cost solutions are also introduced in this context [77, 78]. Estimating the pupil and eye gaze opens up the possibility of detecting eye movement events such as fixations and saccades which can be linked to visual attention and other user states. Fixations are periods during which users fix their gaze on a certain area in the stimulus for a significant amount of time. On the other hand, saccades represent a high-speed shift of eye gaze from one fixation to another. Together, they generate the visual scanpath [79, 80]. Previous literature has found that long fixations are related to individuals engaging more with the content in the stimulus or that they represent an increased amount of cognitive processes [81]. Additionally, longer saccade durations indicate inefficient scanning or searching [82], whereas large saccade amplitudes are related to distant attention shift [83]. Detection of such events could be done by applying threshold-based algorithms [84, 85] or by using probabilistic approaches [86, 87]. These measures might represent different concepts with presented stimulus; however, they are a valuable source of information for VR setups and for related design decisions. In addition to fixation and saccade related features, with the detection of eye regions such as pupil area, one could also use pupil diameter for assessing cognitive load [20, 88].

In the VR context, using such generated features or raw signals provides a variety of online and offline opportunities. While not being limited to the following, a few of these opportunities include foveated/gaze-contingent rendering, saliency prediction, eye-based interaction, user intent analysis and prediction, assessment of cognitive load, and visual attention analysis. Foveated rendering concentrates on rendering the content that users visually focus in high quality while presenting the remaining stimulus in relatively lower quality. From a computational perspective, this might contribute to making VR HMDs more accessible in daily scenarios. To this end, a lot of effort has been concentrated on researching different aspects [89, 90, 91, 92, 93, 94, 95]. Arabadzhiyska et al. [91] studied prediction of saccade ending points in the context of foveated rendering by considering quality mismatch is not noticeable during saccadic supression. Griffith and Komogortsev [93], on the other hand, proposed a data augmentation strategy to improve saccade landing point prediction. Hsu et al. [92] analyzed the perceivability of foveated rendering in different settings in which the technical assessments could prove beneficial for the VR community. Meng et al. [89] proposed a foveated rendering approach based on eye-dominance and indicated that their approach

offers a better rendering performance than conventional approaches. The task of predicting future gaze locations is not only related to foveated rendering, but also to saliency prediction, and has recently been studied for virtual environments as well [96, 97, 24]. While saliency maps and fixation predictions on the images are studied extensively in the literature [98, 99, 100], humans explore immersive virtual environments with HMDs differently. Thus, attention models also differ compared to conventional setups [22]. Several works have focused on saliency related tasks in 3D virtual environments for head movement prediction [101], salient object detection [102], or visualization of 3D heatmaps [103]. In most of the works, eye tracking data provides a variety of benefits both from the research and practical perspective.

While it is likely that estimating saliency for VR or improving foveated rendering configurations will lead people to use this technology more comfortably in daily life, understanding human behavior and interaction during the use of such technology is another key factor and more related to the focus of this work. To this end, Hirzle et al. [104] discuss the foundations of gaze interaction in HMDs. In terms of eye-based interaction, many different aspects including gaze- and blink-based inputs [105, 106, 107], object interactions via gaze [108, 109], and navigation in VR [110, 111] have been studied. In these works, there are several findings that concern the use of eye features. For instance, Lu et al. [107] have reported that according to users' subjective feedback, for hands-free interaction in VR blinking is preferred as opposed to dwelling over content for a specific amount of time, which is considered more common. Nguyen and Kunz [110] have found that users do not detect the scene rotations up to a certain degree during blinks due to visual supression, which is helpful for redirected walking algorithms in VR. Furthermore, Sidenmark and Lundström [108] have indicated that interactions with stationary objects during hand interactions in VR might be favorable in terms of attaining fixations.

User intention analysis, efforts to understand human visual attention, and cognitive load assessment via eyes during immersive VR experiences are of a particular interest for many researchers. This is due to the possibility of supporting and assisting users during the VR experience when such information is known. Additionally, while it is not possible, at present, to create identical configurations with the real world due to technological limitations, one can create hypothetical or utopic scenarios by using human behaviors obtained from different situations. Taking these into account, this direction differs from others such as foveated rendering, saliency estimation, or interactions within virtual spaces. Additionally, understanding human visual attention and perception can also improve the interaction experience during the virtual experience. In the human-aware interaction direction, prediction of touch intentions using gaze [112], prediction of interaction intentions using gaze information [23], prediction of tasks using eye movements [113] are possible. From a visual attention perspective, a data-driven and eye tracking based approach for locating elements in 3D virtual spaces [114], immersion preserving attention guidance [115], and improvement of driving habits with the help of visual attention analyses [116] have been studied. As mental and cognitive load can be also assessed using eyes, particularly via pupil sizes [20, 117], such information may be helpful for context sensitive assistance and support for users. Recently, Luong et al. [118] have studied the real

time recognition of mental workload using physiological features, including the features related to pupillary activities in a VR flight simulator, and obtained up to 65% accuracy. Another aviation use-case was demonstrated by Wilson et al. [119] with deep learning including eye features such as pupillary responses and blinks. The authors have shown that it is possible to classify two level cognitive load over 81% accuracy. Similarly, Kübler et al. [120] have studied the pupillary responses to hazard perception in a 360-degree virtual reality setup and found that pupil dilations are helpful for perception.

Apart from the aforementioned directions and use-cases, eye tracking and gaze measurements have been used in VR research in different settings such as in medicine [121], expert-novice analysis based training [122], or education [123]. The majority of the studies provide implications through eye movements or pupil related activities, which shows the overall usefulness of eye tracking for VR and its future potential.

2.2 Privacy and its Considerations for Eye Tracking

Despite the advantages of using sensor data and eye tracking, privacy risks exist. Some of the privacy risks of extended reality (XR) that also mention eye tracking data were discussed by Mhaidli and Schaub [124]. Silva et al. focused on eye tracking support for visual analytics and identified that ensuring privacy [125, p. 8] is a major theme in terms of oppotunities and challenges for eye movements. Katsini et al. [126] discussed the aspects of eye gaze in terms of security and privacy by focusing more on authentication schemes. Similarly, Liebling and Preibusch [127] have argued the need for privacy mechanisms in pervasive eye tracking. Recently, there has already been a push for privacy in both the eye tracking and VR communities. With legal regulations like General Data Protection Regulation (GDPR) [30] or similar, it is possible that there will be more emphasis in this direction especially with VR devices being used more frequently in everyday life.

One of the most straightforward use-cases of eye data in the biometrics domain is iris authentication. Iris textures can be used reliably for biometric authentication [25, 26] and iris recognition systems are already used widely, for instance at airports (e.g., in the UK, The Netherlands, and in Canada) [128, pp. 2-3]. Security protocols such as multi-party computations or homomorphic encryption schemes have recently been applied for such purposes as well [129, 130]. While not providing the same level of authentication accuracy as iris recognition schemes, multiple studies have shown that eye movements can also be used for biometric authentication [131, 132, 133]. Due to the lower success rates, Komogortsev et al. [133, p. 4] have argued that such methods could be integrated with biometric authentication systems that use face or iris recognition as an additional security layer. Eberz et al. [134] have suggested that eye movement-based authentication could be applicable with settings available in consumer level equipment. Similarly, Zhang et al. [27] have argued for possible use-cases for VR setups with eye movement-based authentication. Furthermore, Zhu et al. [135] proposed a two-factor user authentication method for VR HMDs based on blinking behaviors and pupil

sizes. Such schemes are considerably helpful for authentication-requiring use-cases such as in-app purchases or login scenarios in VR and are resilient to shoulder surfing attacks especially in HMD consisting situations. However, apart from the authentication-requiring scenarios, collecting and aggregating eye movement data without privacy protection could introduce a privacy breach given the wider use of VR devices and the works that map eye movements to user authentication and identification.

Computational privacy and related algorithms could be developed for and applied to any kind of time-series data which is similar to eye movement observations obtained from VR displays. However, if the intent is to make privacy preserving mechanisms work practically, one should consider the real time working capability of such mechanisms. From the privacy perspective, multiple aspects should be covered. These include differential privacy (DP) [136, 137, 138], secure multi-party computation [139, 140] (SMC)-based solutions, and more practical use-cases such as data masking. Differential privacy is an overall scheme for sharing data without compromising the information by which individuals participate in a corresponding dataset by introducing noise on the data [136]. The main issue is to find a proper utility-privacy trade-off. In the SMC-based works, the main idea is to compute an output without compromising the raw data of the input parties, usually by sharing secrets. Yao [139, p. 1] gives the example of two millionaires who want to know who is richer without providing information about their own wealth. Other solutions such as randomized encoding [141, 142] could be also applied; however, computation complexity or communication costs play an important role in terms of the practical usability of the solutions. From an eye tracking in VR perspective, multiple input parties could calculate intentions, activities, or estimate gaze without compromising the raw eye movement information that is obtained from their eye trackers.

In the literature, privacy aspects of eye tracking data have not been researched in-depth yet. Recently, Steil et al. [143] and Liu et al. [144] applied standard differential privacy mechanisms on eye movement features (e.g., rate of fixations, mean saccade amplitudes) and on heatmaps, respectively. While these works are the first in the literature to introduce differential privacy mechanisms on the eye tracking data, they do not address the correlations and privacy loss due to them [145] in the differential privacy context. Li et al. [146] proposed a formal approach for area-of-interests that works in real time. There are several approaches that focus on iris obfuscation [44] by degrading iris authentication [147, 148] or removing iris biometric on the eye images [149, 150]. Adversarial attacks have also been performed on the classifiers based on eye tracking [151]. Furthermore, while not being directly related to VR, Steil et al. [28] have studied an automated shutter mechanism on the field camera of an eye tracker based on the scene privacy which could be applied to AR setups. In this work, the authors found that scene privacy, namely privacy of the stimulus, could be detected to some extent solely by using eye movement features.

2.3 Accessibility of Virtual Reality for Everyday Scenarios

It is likely that VR applications are going to be used by wider communities going forward, taking the market size estimations [2] and current research into account. Even today, users can access VR applications through application stores (e.g., Steam [152], Oculus Go Store [153]), VR supported platforms (e.g., Mozilla Hubs [154]), or even through YouTube [155] with 360-degree videos. The access to VR content through web browsers and services is indeed a valuable contribution since more people can potentially access VR. However, native applications designed for a specific system allow for higher quality graphical stimulus and usually a better human-computer interaction experience due to already available physics engines and potential easier use of sophisticated sensors.

Considering either native or web-based VR applications, a lot of research and development in computer hardware, software, and perception engineering has been carried out to make different kinds of VR systems and applications more available in daily life. On the research side, particularly in perception engineering, human attention and behaviors are important. Developed systems are usually evaluated with human data for many reasons. For instance, if the purpose of a system is entertainment, practitioners may focus on eliminating cybersickness during the virtual experience. In another example, for a system that is designed for training, practitioners may target for scenarios that users might encounter in their real lives. Overall, these approaches require collection of human data during confrontation with developed simulations. The knowledge that is obtained through research on human data is used to provide more advanced virtual experiences. With the many iterations of this loop, everyday usage of VR and HMDs will be more feasible in the future.

On the research side, collection of human data in VR experiments is mostly done in laboratory settings, limiting number of participants in the experiments. This can also lead to homogeneous characteristics of the participants. When the variety of users that can actually own HMDs and experience VR applications is considered, there is a vicious cycle that can be broken through the development of remote data collection routines and protocols. Such data covers both questionnaires, which is usually easier to collect by using web services, and other sensor data such as eye tracking. To tackle this issue, Ma et al. [156] have proposed a solution using Amazon Mechanical Turk (AMT) [157] and a web-based VR application. In their solution, users have to validate the existence of their VR devices by taking photos including their worker IDs. While they did not collect any eye movements, the effort is a valid solution for the problem of enabling the crowd for VR experiments. Very recently, Rivu et al. [158] addressed a very similar issue and reported four approaches to conducting remote VR studies including providing a standalone application through direct download or an application store (e.g., Steam), uploading a standalone application to a VR platform (e.g., Mozilla Hubs), or directly setting up the VR application on a VR platform. The authors have indicated that providing the standalone VR application through a direct download option is preferable for the most advanced functionalities and data collection options; however, recruitment would need to be done through social media or forums. In addition, they have also emphasized the

importance of clear instructions for the experiment and possible ways for the user to connect with the experimenter if needed. In general, the options Rivu et al. [158] have proposed should go hand-in-hand and used by practitioners depending on the requirements. For instance, if VR platforms do not have an option to collect a specific type of data (e.g., eye movements, hand tracking), then it is more reasonable to provide the applications via a direct link. On the other hand, if it is too much effort to advertise the application with the use of direct download, one could pursue recruitment through application stores by choosing the second option proposed by Rivu et al. [158].

In summary, researchers should strive not only to propose novel solutions in perception engineering and VR, but to make these solutions accessible for wider communities during research cycles and to evaluate the solutions with a greater number of participants. Until now, this aspect, in particular, is researched by few and remains an open research direction.

2.4 Towards Everyday Virtual Reality

Developments in eye tracking methodology, usage of eye tracking data in VR studies and applications, privacy preserving considerations for such data, and efforts to make VR more accessible will help integrate virtual experiences in everyday life. To this end, in line with the everyday VR definition provided by Garner et al. [13], this work proposes novel solutions to a variety of problems facing everyday VR and presents an overall framework.

Multi-modal gaze, i.e., head pose and eye gaze, in VR and pupillary measurements could help both for design considerations of virtual environments and for online user assistive tasks during virtual experiences. Design considerations for some of the everyday tasks and setups are more important than the others. For instance, one could design direct replicas of real world configurations based on scientific findings in real world settings for some of the everyday setups in VR. However, as attention models could differ in real and virtual worlds, a dedicated attention assessment and related cognitive processes should be carried out. A classroom environment for learning falls directly in the category of everyday VR and relevant knowledge transfer can be done from real world studies as comparison. With the increasing popularity of e-learning platforms and, very recently, the necessity for online learning due to the COVID-19 pandemic, learning in remote setups has emerged as standard. However, most of the setups lack immersion and provide limited interaction. Learning in VR might solve such disadvantages, but, at the same time, human behaviors should be analyzed before offering VR-based learning to make it highly accessible.

Another potential, but unconventional direction for VR environments that take place more frequently in everyday life is training. This may be considered unconventional in the umbrella term of everyday VR since once required expertise is gained after regular use of a dedicated virtual environment one may not necessarily need to use it moving forward. In terms of training, there are multiple ways to proceed. A straightforward application is providing novices in some specific domain with an immersive environment in which to gain new information.

An alternative is replicating unexpected and unusual occasions that can happen in everyday life. Training in the real world for such occasions might be dangerous or even impossible depending on the domain. For instance, safety related situations that happen in maritime applications [159] could benefit from virtual training. More related to everyday VR is the example of driver training. Skill training apart, according to Bialkova and Ettema [160, p. 2], driving and transportation scenarios are considered within the everyday VR context. While actual driving instruction is performed with real vehicles, there are some safety critical situations during the training of novice drivers that are ethically impossible to create in the real world. For example, the scenario of a pedestrian crossing the street unexpectedly. Additionally, with the growing number of semi-autonomous vehicles and the assumption that fully autonomous vehicles will be available in the future, human-machine interaction is likely to be crucial in traffic scenarios. While not as unique as the maritime example, the number of different interaction scenarios with semi-autonomous vehicles that one may encounter is limited in real world driving learning. With VR and pre-programmed routines, novices can train for such situations in VR in large part thanks to gaze-based assistance and behavior analysis. However, before all of these, it should be researched whether such gaze-based assistance for safety critical situations has a positive effect on drivers.

The growing importance of privacy-related topics, legally and technically, will require eye tracking-enabled VR experiences and configurations to take privacy issues into account. Data conventionally collected before and after the experiments with questionnaires can be anonymized without too much effort. Users at least are usually more accustomed to such data collection protocols. Furthermore, the usage of questionnaires or similar methods does not make much sense when users' everyday experiences with virtual environments (e.g., at their homes or personal spaces) are taken into account. As suggested by Steil et al. [143, p. 8], users might not be extensively aware of the kind of inference generated using eye tracking in the VR and AR context. Differential privacy [136, 137] is especially common in the database applications area, and could be used effectively in eye tracking signals by calibrating the required amount of noise to add to signals while providing data for further inference. Well-established formal methods in the differentially private eye tracking field will further help usage of individual protected eye tracking data collected in everyday scenarios and setups. Additionally, VR HMDs are on the way to becoming personal devices like mobile phones and smart watches. For providing assistance during experience, machine learning models are trained with huge amounts of data, collected from various people. Such models can also be trained in cloud for scalability and better processing power. In a very primitive and naive setup, users can upload their collected eye tracking data directly to the cloud. However, one may not share the raw data due to aforementioned inference possibilities especially by the commercial applications. In such cases, use of secure multi-party computation [139, 140], randomized encoding [141, 161], or similar procedures are reasonable as long as real time capability of such solutions in the test time is empirically evidenced. The solution can be anything from privacy preserving gaze estimation to gaze prediction in VR; however, such setups depend on the classifiers that are used and real time working capability can be affected by the effectiveness

of the communication with the cloud instances. Overall, the use of cloud-based predictive models by preserving the privacy computationally is actually very relevant for everyday VR setups when the contemporary cloud infrastructure is utilized [162].

To enable everyday usage, the inclusion of more people in VR systems, from a data perspective, is necessary. While the decreased cost of HMDs, marketing, and other relevant attempts could affect this, one important factor is the availability of large scale data collection possibilities and related protocols and applications. For behavioral data other than sensors such as eye tracking, AMT has been used widely for crowdsourcing purposes. However, there are several issues with such paradigms if they are applied to VR/AR sensor data collection, particularly eye tracking. Either forwarding from AMT (or similar) to web-based custom VR services [156] or the supplying of stand-alone applications [158] should be done. As web-based VR currently faces major challenges such as lower resolutions or lack of integrated high quality eye tracker devices, providing standalone applications via different channels such as forums or social media is more reasonable. However, in this approach, one should solve the issue of participant compensation and data sharing optimally without any centralized third-parties due to the behavioral biometrics that are inherent in eye movements. If third-parties are involved in the process, an extra layer of protection e.g., by using cryptography, can be applied, but this would increase the overall complexity. Furthermore, validation of data authenticity should be performed to check for adversarial participant behavior.

3 Motivation and Main Findings

In this chapter, the papers I have authored in the direction of everyday VR during my doctoral studies have been summarized in terms of motivation, methodology, and findings according to the open directions that are laid out in Chapter 2. The published papers are available in Chapters A, B, and C.

Eye tracking in VR for visual attention and cognitive processes has been researched and studied in two different domains, namely education and driving. In the education-related studies, the research focus is relating conventional eye movements and pupillometry and virtual objects of interests with cognitive processes and visual attention, respectively. In the driving studies, the effect of gaze-aware assistance on human behavior and the feasibility of cognitive load estimation during a safety critical pedestrian crossing have been researched. These are introduced in Section 3.1.

Privacy preserving eye tracking has been researched in two directions as well. First, differential privacy mechanisms for eye movement signals that have been collected from VR and AR-related setups taking temporal correlations into account are discussed. Next, the privacy preserving gaze estimation task for VR using two input and one function parties is introduced in Section 3.2.

Lastly, for the accessibility of VR along with eye tracking, a blockchain- and smart contract-based eye tracking data collection protocol for remote participants is introduced in Section 3.3.

Overall, three sub-directions that are researched in this thesis introduce novel use-cases while advancing the state-of-the-art in multiple directions, and propose together a framework that is depicted in Figure 3.1 which further pushes for everyday VR.

Figure 3.1: Overall framework of contributions in the thesis.

3.1 Visual Attention and Cognitive Processes in Virtual Reality

The first two subsections (Sections 3.1.1, 3.1.2) focus on the education content in VR for the design considerations of virtual spaces in everyday educational VR, whereas the latter two (Sections 3.1.3, 3.1.4) focus on the driving domain with the purpose of assessing the feasibility of VR for safety critical scenario training.

3.1.1 Eye Movements in Virtual Classrooms

This subsection is based on the paper 2 in Chapter 1, *Digital transformations of classrooms in virtual reality* at *2021 CHI Conference on Human Factors in Computing Systems*.

Motivation and Main Methodology

How to design and visualize interaction related content is crucial when everyday virtual learning environments are taken into consideration, as these might affect motivation, engagement, performance, and eventually learning outcomes of students in the long term. While there are many studies which concentrate on a variety of issues for the real world classroom [163, 164], few conclusions are drawn for VR environments. In the real world studies, human behaviors are usually extracted and analysed based on head and body movements. In VR, it is possible to extract eye movements and pupillary information, which may be linked to cognitive processes, given the fact that modern high-end HMDs integrate eye trackers. In fact, virtual

environments, especially immersive ones, could have a large range of benefits for humans when environmental and physical inaccessibilities are considered. For instance, during the COVID-19 pandemic, many institutions temporarily switched to remote teaching and learning. While this switch was particularly motivated by the pandemic, remote teaching and learning can also be useful for handicapped people or when in-person meetings are not feasible due to the restrictions of physical distance. For example, while during the pandemic, many of the scientific conferences and meetings were organized remotely or as hybrid events. After the pandemic, such a hybrid model could remain to accommodate difficulties due to remoteness. Virtual environments have already been realized using relatively trivial tools such as Mozilla Hubs [154]. Before more sophisticated tools are employed for such teaching, learning, and interaction purposes, the effects of different configurations on humans should be researched in several virtual environment types, including halls, auditoriums, and classrooms. The major motivation of this work is to understand the effects of different virtual classroom configurations on students including different avatar representation styles of virtual characters, namely realistic and cartoon, different participant locations in the classroom, including back and front, and different hand-raising behaviors of virtual peer-learners, particularly 20%, 35%, 65%, and 80%. An example view from the used VR classroom is depicted in Figure 3.2.

Figure 3.2: An example view from the used VR classroom.

Using raw eye and head tracking data that was collected from 381 sixth-grade students in a between-subjects design during a computational thinking lecture that took approximately 15 minutes, eye fixation, saccade, and pupil diameter related features were extracted. The raw data only provides angular information that was reported by the eye tracker within the HMD. Therefore, a processing pipeline is needed to extract features such as durations of fixations and saccades, amplitudes of saccades, number of fixations, and pupil diameter values. Apart from the pupil diameter values, fixation and saccade extraction algorithms should be tailored especially for the VR, as in VR one should take head movements into account compared to conventional eye tracking experiments that include chin-rests. For VR setups, Agtzidis et

al. [85] have proposed a Velocity-Threshold Identification (I-VT) similar thresholding approach for 360-degree videos, and, in this work, a similar approach was followed by setting velocity and duration thresholds for fixations and saccades. For the pupil diameters, a standard processing pipeline including smoothing [165] and baseline correction [166] components was employed. After pre-processing and feature extraction phases, full factorial ANOVAs ($\alpha = 0.05$) were applied for different features to find out whether different classroom manipulations have significant effects on human visual behaviors. For multiple comparisons and non-parametric versions of the analyses, Tukey-Kramer and Aligned Rank Transform (ART) [167] were applied, respectively.

Main Findings

Different classroom manipulations have statistically significant effects on human visual behaviors, including durations of the fixations and saccades, number of fixations, saccade amplitudes, and pupil diameters. More specifically, in terms of participant locations in the virtual classroom, participants sitting in the back of the classroom had significantly longer fixations than those sitting in the front. Furthermore, the front sitting participants had longer saccade durations and larger saccade amplitudes than those sitting in the back.

In terms of avatar representations, the analyses have yielded beneficial, but less significant results. The participants that experienced cartoon-styled avatars in the virtual classroom had longer fixation durations compared to the participants who observed realistic-styled avatars. In addition, participants that experienced the VR classroom with realistic-styled avatars had longer saccadic durations and larger pupil diameters during the virtual lecture than the participants that observed cartoon-styled avatars.

Lastly, in terms of attention towards different hand-raising behaviors, mixed results were found. In particular, participants had significantly larger pupil diameters with 80% hand-raising peer-learners compared to 35% of the virtual peers raising their hands. Additionally, in the 65% hand-raising condition, participants had significantly more fixations than in the 80% hand-raising configuration. These results have important implications for the design of everyday virtual and interactive classrooms, especially in terms of cognitive processes.

3.1.2 Visual Attention on Virtual Objects in Virtual Classrooms

This subsection is based on the paper 3 in Chapter 1, *Exploiting object-of-interest information to understand attention in VR classrooms* at *2021 IEEE Virtual Reality and 3D User Interfaces*.

Motivation and Main Methodology

The findings that are introduced in Section 3.1.1 provide implications primarily for cognitive processes during the virtual classroom experience. Another aspect that is important in virtual

environments in general is the assessment of interactions with the virtual objects and the provided 3D stimulus. Virtual objects are essentially gathered together in the virtual environments to create the overall 3D stimulus. Thanks to modern game engines that are mostly used to design such environments, it is also possible to associate physics-related components with 3D virtual objects. For instance, each peer-learner, instructor, and any other static or dynamic content can be represented as objects in the classroom environment and volumetric details can be fetched for further analysis. In terms of human-computer interaction, several interaction models with objects during the experience are possible, such as with hand-held controllers, solely by hand tracking, audio-based input, or gaze-based methods. However, before providing such interaction models, preferably online and real time, one should study visual attention on different objects under different manipulations. Therefore, the main focus of this work is studying the gaze-based attention mainly on the most important objects and object groups in the VR classroom that is studied in Section 3.1.1. The most important objects are considered virtual instructor, virtual peer-learners by aggregating attention from each peer-learner, and the lecture screen where the lecture content is visualized. Such analysis and features differ from conventional eye tracking features such as fixations or saccades because they are more related to cognitive processing. Studying the focus on 3D objects in the environment is more related to visual attention and interactions in particular parts of the VR classroom.

To do such analysis, the head and eye gaze information should be mapped to the 3D environment. To this end, using the head pose reported by the HMD and the gaze vector provided by the eye tracker, an invisible ray is traced to the classroom environment. As virtual objects have invisible geometric colliders around them, the 3D hit point of the traced ray [168] is calculated. Participants can attend some objects very shortly in an insignificant amount of time and this does not indicate much about the visual attention. To overcome this issue, a duration threshold of 200 milliseconds was used for minimum attention duration. This scheme was applied for each frame and attention times were obtained for each object. The set threshold is greater than fixation detection thresholds in the previous literature. This was done intentionally since participants could have any kind of eye movements on the attended virtual objects. Afterwards, to analyze the implications of each classroom manipulation, full factorial ANOVAs ($\alpha = 0.05$) for attention time on each relevant object, namely virtual peer-learners, instructor, and screen were applied. Similar to the method in Section 3.1.1, Tukey-Kramer post-hoc test and ART [167] were used for multiple comparisons and non-parametric analyses, respectively.

Main Findings

Analyses showed that the participants that were located in the back of the classroom had significantly longer attention time on the virtual peer-learners than the participants who were located in the front. On the contrary, the front sitting participants had significantly longer attention time on the virtual instructor and lecture screen than the participants who sat in the

back.

Cartoon-styled avatars attracted more attention time on the virtual peers than the realistic-styled avatars, while in the realistic-styled avatar configuration, the virtual instructor drew more attention time than in the cartoon-styled avatar configuration. In terms of attention time on the screen, realistic- or cartoon-styled avatar configurations did not differ significantly.

Mixed effects were obtained in the analyses for different hand-raising configurations of the virtual peers as is the case for the conventional eye movement features discussed in Section 3.1.1. The attention time on the peer-learners was the most in the extreme hand-raising configurations, namely 20% and 80%. More particularly, the attention time on the peers was significantly greater in the 80% hand-raising configuration than in the 65%. Furthermore, 20% and 65% configurations differed statistically significantly in terms of focus on the peer-learners, with the 20% configuration being longer. In the analyses for the attention time on the virtual instructor, the only significant difference was found between 65% and 80%, with focus time in the 65% hand-raising configuration being longer. Furthermore, focus on the lecture screen was significantly longer with 65% of the virtual peers raising their hands than with 80% of the peers raising their hands. The focus time on the screen in the 65% hand-raising configuration was also significantly higher than the 35% hand-raising configuration; however, the effect was smaller compared to the effect in the 80% hand-raising configuration.

3.1.3 Driver Attention Analysis during a Safety Critical Situation in Virtual Reality

This subsection is based on the paper 6 in Chapter 1, *Assessment of driver attention during a safety critical situation in VR to generate VR-based training* at *2019 ACM Symposium on Applied Perception*.

Motivation and Main Methodology

Whether VR setups could help training for a variety of scenarios that humans may encounter in their real lives and cannot train for due to safety reasons or a low likelihood of occurrence is an open research direction in the context of everyday VR. In the driving domain, safety critical scenarios are of a particular interest since by using VR novice drivers can train at their homes to get accustomed to these scenarios and improve their skills. However, before generating the training packages for these purposes, it is important to understand whether VR configurations with gaze-based assistance help and draw the attention of drivers in safety critical situations. To this end, an unexpected pedestrian crossing scenario in a VR setup has been designed, with two conditions. The conditions include an experimental condition where participants are informed about the criticality of the pedestrian by gaze-aware red cues around the figure, and a control condition where participants are not informed. A view from the driving vehicle cockpit is depicted in Figure 3.3 [169, p. 2].

The study involved 16 participants with between-subjects design. The critical pedestrian

Figure 3.3: An example view of the road that includes critically crossing pedestrian from the driving vehicle's cockpit. © 2019 IEEE.

started crossing the street when driving vehicle was closer than approximately 45 meters as this distance simulates an expected time-to-collision between approximately 1.8-5 seconds. In such conditions, Rasouli et al. [170] report the high occurrence likelihood of joint attention between pedestrians and drivers. However, in the opposite case, the outcome might be fatal. Since such a scenario cannot be generated and validated in the real world, VR setups stand alone for applying such scenarios with low costs. The gaze-aware pedestrian warning was activated when the distance to the crossing pedestrian was approximately 77 meters and deactivated if the participant's gaze was within 5 meters of the crossing pedestrians for at least 0.85 seconds. Intersection between participant gaze and the pedestrian 3D object was carried out with the help of ray-casting [168], similar to the study in Section 3.1.2. To assess the usefulness of the setup along with gaze-aware cues, the closest distance between vehicles and crossing pedestrians, driver inputs on accelerator and brake, and participant pupil diameters were evaluated. Pupil diameters were smoothed [165] and baseline corrected [166] in a similar manner to the study in Section 3.1.1. The experimental and control conditions were compared with two sample T-test in terms of closest distances to crossing pedestrians. Furthermore, within each condition, paired T-tests were applied to driver accelerator inputs and pupil diameters for baseline driving and risky driving timeframes. Baseline driving corresponds to driving without any intervention such as gaze-aware warnings or pedestrian crossing, and is calculated using the observations just before the gaze-aware warnings or start of the pedestrian crossing for experimental and control conditions, respectively. Risky driving timeframe corresponds to the time after the start of the critical pedestrian crossing.

Main Findings

Analysis on the closest distances to the crossing pedestrian showed that the participants who received the gaze-aware critical pedestrian warnings before the crossing passed statistically significantly distant to the pedestrian than the participants that did not receive the warnings. This indicates that the warnings helped the participants drive safer.

According to the analysis of inputs on accelerator and brake pedals, significant differences on accelerator inputs between baseline and risky driving timeframes started earlier in the participants who received gaze-aware warning cues, indicating that they realized the criticality earlier. In addition, five of the participants who did not receive pedestrian cues performed full braking whereas no participant receiving the cues performed a full brake. In terms of pupil diameters, a trend similar like in the accelerator inputs was observed. Overall results indicated safe and smooth driving experiences for the participants who were provided with gaze-aware risky pedestrian warnings.

3.1.4 Cognitive Load Estimation during Virtual Driving

This subsection is based on the paper 7 in Chapter 1, *Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup* at *2019 IEEE Conference on Virtual Reality and 3D User Interfaces Workshops*.

Motivation and Main Methodology

VR setups have potential to provide timely benefits to users if their states can be estimated in real time during virtual experiences, which could be helpful in the context of everyday VR. As aforementioned, one of the non-intrusive ways to do this is through the eyes of the users. Considering the controlled illumination VR provides unlike real world configurations, user cognitive load estimation based on pupillometry is one direction that could be investigated.

Based on the VR driving experiment discussed in Section 3.1.3, data annotations were carried out using pupil diameter measurements for low and high cognitive load, based on the time points that pedestrian warnings and pedestrian crossing started for experimental and control groups, respectively. An empirical timeshift of 0.8 seconds was introduced as well, considering that pupils did not dilate directly on time as the manipulations were introduced due to biological factors. Once the data of each participant was annotated for binary classification purposes, classifiers including Support Vector Machine (SVM), Decision Tree (DT), k-Nearest Neighbors (k-NN), and Random Forest (RF) were trained in a leave-one-participant-out cross-validation configuration and validated accordingly. The feature vectors included driver performance measurements incorporating accelerator and brake inputs as well as pupil diameter values. In the validation phase, accuracy, precision, recall, and F1-scores were calculated to assess performance of the classification tasks. In addition, to evaluate the real time working capability of each classifier, the cognitive load prediction time span of each test sample was calculated

and averaged using all samples.

Main Findings

The assessment of the accuracy, precision, recall, and F1-scores revealed that cognitive load estimation for VR setups has potential and could be applied in a person-independent way successfully. Particularly, SVM performed the best with over 80% accuracy, whereas DT and RF worked comparably with accuracies between 70-80%. 1-NN, 5-NN, and 10-NN were evaluated for the k-NN with 10-NN working the best with approximately 79% accuracy.

Since these estimations should be applied in real time for user assistive tasks, real time working capabilities were assessed with a VR-capable computer. On average, SVMs and DTs worked the fastest with approximately 0.3 milliseconds for estimating cognitive load per sample. Estimation times for k-NNs with different k values took relatively longer with approximately 0.74 and 0.76 milliseconds, whereas RF-based estimation took the most time per sample with approximately 5.4 milliseconds.

3.2 Privacy Preserving Eye Tracking for Virtual Reality

In this section, differential privacy for eye tracking and privacy preserving gaze estimation based on eye landmark data are introduced in Section 3.2.1 and Section 3.2.2, respectively.

3.2.1 Differential Privacy for Eye Tracking

This subsection is based on the paper 1 in Chapter 1, *Differential privacy for eye tracking with temporal correlations* at *PLoS ONE* in 2021.

Motivation and Main Methodology

Differential privacy is a rigorous framework for protecting the information about whether an individual participated in a dataset or not [136, 137]. Considering a database that includes incomes of individuals, when data is queried for the mean value of the total number of individuals, the mean income for N individuals is obtained. If the same database is queried for N-1 individuals, using two mean values, an adversary could automatically infer the remaining individual's income. In differential privacy, privacy protection is achieved by adding randomly generated noise to the query outcomes based on a privacy parameter ϵ using, for instance, Laplace or Gaussian distributions, so that the query answers do not significantly change based on the participation of individuals. Provided privacy is increased when the ϵ value is decreased. On the one hand, the privacy of individuals is preserved thanks to differentially private mechanisms. One the other hand, due to the added noise data quality and utility are decreased. Therefore, it is crucial to find reasonable trade-offs. Formally, ϵ-Differential Privacy

(ϵ-DP) is defined as follows.

Definition 1. ϵ-Differential Privacy (ϵ-*DP [136, 137]*). A randomized mechanism M satisfies ϵ differential privacy for all databases D_1 and D_2 which differ at most in one element for every $S \subseteq Range(M)$ with the following.

$$\Pr[M(D_1) \in S] \leq e^{\epsilon} \Pr[M(D_2) \in S]. \tag{3.1}$$

The amount of noise added depends on query sensitivities which are defined in the following.

Definition 2. *Query sensitivity [136].* For any random query of X^n and $w \in \{1,2\}$, the query sensitivity (Δ_w) of X^n is defined as the smallest value for all databases D_1 and D_2 that differ maximum in one element with

$$||X^n(D_1) - X^n(D_2)||_w \leq \Delta_w(X^n) \tag{3.2}$$

The standard Laplace mechanism of differential privacy (i.e., Laplace perturbation algorithm (LPA)) achieves differential privacy for $\lambda = \Delta_1(X^n)/\epsilon$ [136] with noisy observations generated according to $\widetilde{X}^n = X^n(D) + Lap^n(\lambda)$, where $Lap^n(\lambda)$ consists of n independent random variables from a Laplace distribution with a zero mean and variance $2\lambda^2$.

While this framework is widely used in database applications, it is not very straightforward to apply it to time-series data. Firstly, in time-series data, observations from different time points are temporally correlated. When standard differential privacy mechanisms are applied, an adversary can use this background information and make inferences on the data with the assumption of independent noise realizations on each time point. According to Zhao et al. [171] and Cao et al. [172], there exist privacy leaks with the correlations and actual ϵ value for a pre-defined ϵ increases with such effect. Secondly, the required amount of noise to make time-series data differentially private significantly increases with long signals. As eye movements and the features that are extracted from raw eye movements are temporally correlated and can have long durations depending on the stimuli or individuals, standard differential mechanisms might perform poorly. In the previous work, there are privacy frameworks for correlated or sensor data such as Pufferfish [173] or Olympus [174]; however, these require different assumptions such as the necessity of a domain expert or modeling the privacy as adversarial networks. The Fourier perturbation algorithm (FPA) proposed by Rastogi and Nath [175] deals with temporally correlated time-series data, and noisy observations are generated according to Algorithm 1 [175, 176], where DFT, IDFT, and PAD correspond to discrete Fourier transform, inverse discrete Fourier transform, and zero padding, respectively. Unlike the value claimed by Rastogi and Nath [175], the FPA achieves ϵ-differential privacy for $\lambda = \frac{\sqrt{n}\sqrt{k}\Delta_2(X^n)}{\epsilon}$, with its proof in the Section B.1.

To solve these issues and make the eye movement features obtained from VR/AR setups differentially private, LPA [136] and corrected FPA [175] are evaluated. Furthermore, two

Algorithm 1: Fourier Perturbation Algorithm (FPA).

Inputs: X^n, λ, k
Output: $\widetilde{X}^n$
$F^k = DFT^k(X^n)$.
$\widetilde{F}^k = LPA(F^k, \lambda)$.
$\widetilde{X}^n = IDFT(PAD^n(\widetilde{F}^k))$.

additional mechanisms, particularly chunk-based FPA and difference- and chunk-based FPA have been proposed and named as CFPA and DCFPA, respectively. The major purposes of these extensions are decreasing temporal correlations, the sensitivities required to achieve differential privacy, and ideally computational complexities as the chunk sizes are selected power of 2 [177], namely 32, 64, and 128. The CFPA applies the FPA mechanism to each chunk, whereas the DCFPA applies the FPA to the consecutive difference signals within the chunks. Difference signals are observed to be significantly decorrelated, which implies that in this privacy mechanism, the privacy reduction due to the temporal correlations is less than the others. The CFPA and DCFPA processes are summarized in Figure 3.4 [176, p. 9].

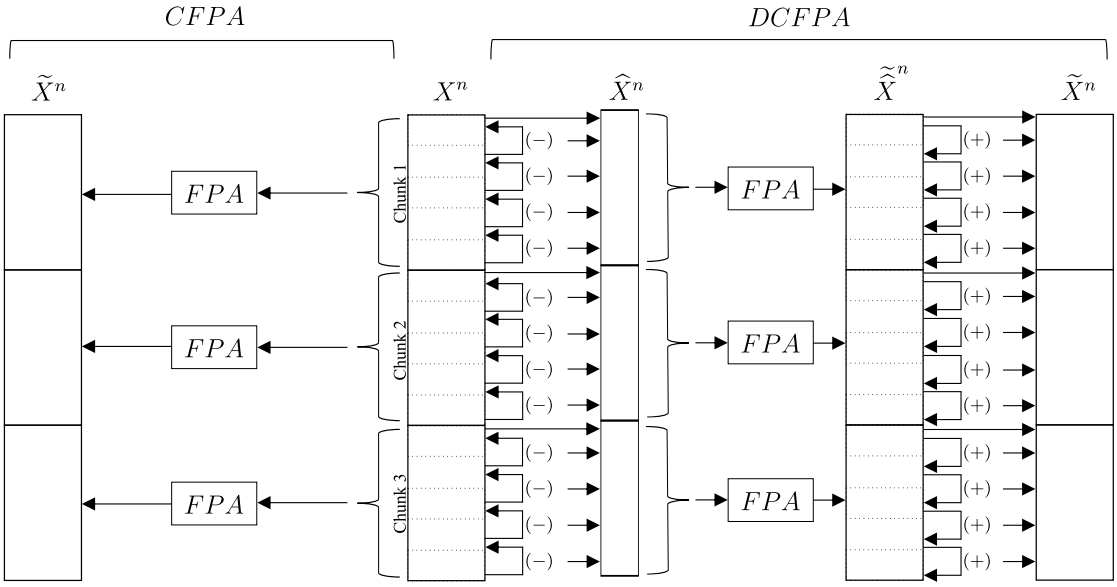


Figure 3.4: Workflow of the CFPA and DCFPA.

To evaluate proposed mechanisms, MPIIDPEye [143] and MPIIPrivacEye [28] datasets have been used with five different privacy levels, namely $\epsilon = [0.48, 2.4, 4.8, 24, 48]$. The former dataset includes 52 eye movement features extracted from a reading task of three different document types (i.e., comics, news, and textbook) in VR, whereas the latter dataset includes the same eye movement features from an AR similar setup. As a significant amount of noise is introduced to the extracted features to achieve differential privacy, it is important to validate the usefulness of the private data. To do this, absolute normalized mean square error (NMSE) was used especially for comparison of different privacy mechanisms namely, LPA, FPA, CFPA, and

DCFPA. However, while this metric shows the divergence trend of noisy data from the original data, it does not directly show how usable the private data is for different tasks. To find this out, for the MPIIDPEye, gender and document type classification tasks were employed. For the MPIIPrivacEye, privacy sensitivity detection of the viewed scene was carried out. In addition, person identification tasks were applied to evaluate whether it is possible to recognize the individuals using machine learning classifiers for both MPIIDPEye and MPIIPrivacEye. For these purposes, Support Vector Machines (SVMs), k-Nearest Neighbors (k-NNs), Decision Trees (DTs), and Random Forests (RFs) were used. Except for person identification tasks, all classifiers were trained and evaluated in a person-independent manner to ensure generic outcomes.

Main Findings

As the analyses are split into two groups, including utility based on the NMSE and classification accuracies, the findings are reported separately as well. However, before the usability of the data, data correlations are also analyzed. The difference signals used by the DCFPA are observed to be less correlated than the original observations for both datasets. This means that the DCFPA method is less vulnerable to temporal correlations in terms of privacy.

Utility evaluations based on the absolute NMSE showed that the CFPA and DCFPA outperform the standard Laplace mechanism of differential privacy. All CFPA evaluations outperform the FPA as theoretically assumed due to reduced sensitivities. Different chunk sizes with CFPA perform very similarly, therefore it is reasonable to use larger chunks as they better reduce correlations. The DCFPA, especially with smaller chunks, surpasses other methods in the most private settings (i.e., $\epsilon = 0.48$). In the DCFPA, smaller chunks (e.g., 32) perform significantly better than others in terms of absolute NMSE.

For the MPIIDPEye dataset, three classification tasks, namely document type, gender classifications, and person identification, were applied. Ideally, individuals and their genders should not be recognized through private data. At the same time, since document type classification is treated as a utility task, the accuracy of this task should be as high as possible. Analyses have shown that when the FPA is applied along with majority voting for person identification, very high accuracies (i.e., 100%) are obtained. This is an indication that even with the addition of noise, it is still possible to recognize individuals. With the DCFPA, it is not possible to identify individuals accurately either in high (e.g., $\epsilon = 0.48$) or low privacy ($\epsilon = 48$) regions. When the CFPA is applied, it is possible to identify individuals accurately starting from $\epsilon = 24$. However, when high privacy regions are considered, classifiers fail to identify individuals as is the case for the DCFPA. In terms of gender classification, it is only possible to detect gender to some extent (e.g., up to accuracy of 0.68) with the CFPA in the lowest privacy regions. Hence, almost all the methods are able to hide gender information unlike person identification. Document types are identified accurately with the FPA in all privacy regions with accuracy over 0.85 with random forests. The CFPA works with an accuracy over 0.7 in the low privacy regions, whereas the DCFPA performs better with accuracies of 0.64 and 0.69 for high ($\epsilon = 0.48$) and middle

($\epsilon = 4.8$) privacy regions, respectively. Even though the FPA works better than other methods for document type classification since it is possible to identify individuals very accurately when this method is used, one should consider either the CFPA or DCFPA when differential privacy mechanisms for eye movements are considered.

For the MPIIPrivacEye, privacy sensitivity classification works very similarly for the CFPA and DCFPA with accuracies in the vicinity of 0.6 in all privacy regions. Similar to the MPIIDPEye dataset, when the FPA is applied, it is possible to identify individuals very accurately with majority voting. The CFPA and DCFPA are able to hide personal identifiers successfully as in all privacy regions the person identification accuracies are close to random guessing probability. The accuracies without majority voting follow a similar trend with either higher or lower accuracies depending on the actual values not only for the MPIIPrivacEye dataset, but also for the MPIIDPEye dataset.

3.2.2 Privacy Preserving Gaze Estimation based on Eye Landmarks

This subsection is based on the paper 5 in Chapter 1, *Privacy preserving gaze estimation using synthetic images via a randomized encoding based framework* at *2020 ACM Symposium on Eye Tracking Research and Applications*.

Motivation and Main Methodology

Eye movements and the features extracted from raw data provide a lot of insight into individuals as discussed in the previous section. However, the focus of privacy mechanisms might not be hiding individual participation in a dataset in all setups like differential privacy. Instead, the focus may be protecting the data as whole, especially if the data is related to health status, medicine and etc. As eye movements in VR can even be related to diseases [121], one should consider such methods for privacy as well. Still, as it is beneficial to use eye movements for user assistive systems applied in VR environments, for such cases machine learning models should be trained and tested without providing raw data with setups that include multiple parties. This may be due to a lack of processing power for each individual leading to classifier training in the cloud or a lack of data for specific tasks.

The main purpose of this work is to show the applicability of eye tracking data analytics collected from VR setups with several parties. While the main focus is VR HMDs, data collected from other equipment such as smart glasses or optical see through displays can be used as well. One of the main prerequisites is the real time working capability of such frameworks. For validation, a gaze estimation task using a baseline model based on Support Vector Regression (SVR), with three parties including two input and one function party, is employed. Input parties are considered data providers for the function party to create and train gaze estimation models. The function party could be thought of as a cloud instance that processes data. For input data, UnityEyes [69] has been used to generate 20k synthetic eye images in total, similar

to those obtained via eye tracker sensors inside HMDs and two samples are visualized in Figure 3.5.

(a) A synthetic eye gazing up. (b) A synthetic eye gazing down.

Figure 3.5: Generated sample eye images with UnityEyes.

From the generated images, 36 eye landmark-based features [74] were used. After each input party extracts eye landmarks locally, communication between each party and the cloud instance starts. First input party, named "Alice", generates two random vectors and a value and sends these to the second input party, named "Bob". Both parties mask their extracted raw eye landmark features with these randomly generated values, and send them to the function party with the gram matrix of their samples. After obtaining the inputs from Alice and Bob, the function party computes the dot product of Alice's and Bob's samples and completes the remaining part of the gram matrix. Later, using the gram matrix, the function party trains the SVR to estimate the gaze. The security framework employed in this work is inspired by Ünal et al. [178], and its security analysis is available in Section B.2. As input parties primarily send the masked data and gram matrix of their samples, and as the function party does not know the generated values from Alice for masking, apart from training the SVR, it cannot infer the raw data of Alice or Bob. Similarly, as input parties do not collude, it is not possible for them to make inferences about each other's raw data.

Main Findings

Evaluations on different amounts of data, namely 5k, 10k, and 20k samples, show that when the number of samples are increased, mean angular error slightly decreases even with synthetic data. In particular, mean angular error decreased from 0.21 to 0.18 in the test time when number of samples increased from 5k to 20k.

As this work is considered as a proof-of-concept for the applicability of a randomized encoding [141, 161]-based framework for the VR and eye tracking domains, one of the most important issues is evaluating the execution time during testing. In the experiment with 20k samples, it took approximately 4.5 seconds to predict gaze direction of 4k test samples with a standard computer, which corresponds to approximately 1.1 milliseconds per sample.

3.3 Accessibility of Virtual Reality

This section introduces the eye tracking data collection protocol suitable for remotely located participants in the context of accessible VR.

3.3.1 Remote Eye Tracking Data Collection Protocol for Virtual Reality

This subsection is based on the paper 4 in Chapter 1, *Eye tracking data collection protocol for VR for remotely located subjects using blockchain and smart contracts* at *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality Work-in-progress papers*.

Motivation and Main Methodology

Analyses on visual attention, cognitive processes using eye movements, and privacy preserving manipulation of eye movement signals have great potential to make VR more accessible and available in everyday life. However, to make VR collectively available, more people need access to VR HMDs and there should be potential to access more people's data, particularly eye tracking data as in the context of this work. With the recent COVID-19 pandemic, this issue has become more prominent than ever before. To do this, remote data collection protocols are needed for VR setups.

Collecting such data remotely is not trivial. Firstly, it is important to keep the data quality high and comparable with laboratory studies. Secondly, data collectors should utilize mechanisms which guarantee that collected data is not altered by malicious users. In the laboratory studies, this is straightforward and trivial since subjects that participate in the experiments do not have access to applications and data collection pipelines. However, when remote collection is considered, subjects carry out tasks on their own computers and thus have opportunity to analyze any kind of application that maybe of interest. Thirdly, in such user studies, subjects are provided with small amount of money or gifts as compensation. In studies of human behavior that are carried out online such as crowdsourcing, services like AMT [157] are used for subject compensation. However, AMT or similar services do not have direct support for VR experiments and eye tracking data collection. Furthermore, such services introduce an extra layer between the experimenter and subjects, which is undesirable from a privacy perspective.

To enable remote eye tracking data collection, a protocol using white-box cryptography [179], blockchain [180], and smart contracts [181] has been designed. The overall workflow is as follows. Firstly, subjects obtain the VR application from the data collector. After gaining access to the application, the experiment is carried through following instructions. The application informs subjects about the data quality and whether it is good enough for reporting it to the data collector. After this confirmation, subjects initiate data collection smart contract staking double the amount of compensation that they will acquire in the end. Then, the data collector

approves the data collection process, staking double the amount of compensation that will be given. At this point, four units of compensation are locked in the blockchain and, as long as both parties do not fulfill their obligations, the staked amounts get stuck in the blockchain. Afterwards, the subject stores the hash value reported by the VR application in the blockchain then sends the collected data to the data collector along with the transaction ID of storing the hash value in the blockchain. The hash value reported to the subject is calculated using the data collector's secret key and white-box cryptography [182, 179] so that even though subjects infer the hash function, they do not have the ability to construct a fake hash value for altered data. Therefore, they must behave honestly which means they are not expected to alter the collected data. After obtaining the data and transaction ID of the hash storage on the blockchain, the data collector checks whether the hash value stored by the subject in the blockchain and the local hash value calculated by using the selected secret key overlap. When they match it means that the subject is honest, hence the data collector can confirm the data collection in the smart contract. At this point, the smart contract unlocks the compensation and distributes three units of compensation to the subject and one unit to the data collector. At the end of the process, collected data is transferred to the data collector and subjects receive their compensation without any centralized third party service processing their data or intervening in the compensation management. The aforementioned protocol was realized on the Ropsten Testnet of Ethereum platform [181] with dummy synthetic data, and is available as follows: https://ropsten.etherscan.io/address/0x0e937a4a4618dd8d5a12ec4a9f8fd61d6bfd13e4.

Main Findings

Unlike the other studies, this study has shown that remote VR data collection for eye tracking is possible with just a little more effort than laboratory studies. For this purpose, VR applications should calculate the quality of eye tracking data (e.g., tracking ratios) at the end of each experimental session and report it for further processes. In addition, cryptographic implementations are needed to guarantee that subjects do not alter the data. Lastly, use of blockchain and smart contracts shows that, with a little more effort, compensation distribution along with data transfers can be realized without a centralized service and are a credit to the proposed protocol.

4 Discussion

In this chapter, the papers that are summarized in Chapter 3 and presented in Chapters A, B, and C are discussed within the umbrella term of "Everyday Virtual Reality". Findings on visual attention and cognitive processes based on eye movements in education and driving domains are discussed in Section 4.1, based on the papers 2, 3 and 6, 7 in Chapter 1, respectively. Implications of privacy preserving eye tracking for VR focusing on differential privacy and a randomized encoding-based framework are explained in Section 4.2, based on the papers 1 and 5, respectively. Then, how these mechanisms could be combined with the everyday VR setups in terms of remote data collection is discussed in Section 4.3, standing on the paper 4 in Chapter 1. Finally, the outlook in the realm of all covered topics is drawn in Section 4.4.

4.1 Visual Attention and Cognitive Processes

Visual attention and cognitive process related research questions were investigated through two experiments in the education and driving domains. Therefore, the discussion is split into two subsections.

4.1.1 Virtual Reality in the Classroom Context

VR-based classrooms not only offer the possibility to make online and remote learning more immersive and interactive, they also support the studying different classroom manipulations that are difficult to generate in the real world. While some manipulations are directly related to VR environments rather than real world scenarios, such as the visualization styles of avatars (e.g., cartoon or realistic), it is also possible to control, for instance, the number of hand-raising peer-learners during the experience, which could affect student self-concept [183] in the long run. It is also relatively straightforward to create conditions which typically happen in the real world such as locating students attending to a virtual lecture in the back or front of the virtual classroom.

In the aforementioned studies, these three classroom manipulations have mainly been

studied by taking human head and eye movements into account. Locating students in the back or front of the virtual classroom has yielded different implications. The participants sitting in the back had longer fixation durations during the lecture, which implies that they spent more time processing information. This may be related to the relationship between mean fixation durations and task difficulty [184]. While the task is essentially the same for all participants, participants sitting in the back view the lecture content through a smaller field of view and may have difficulty extracting information. On the contrary, the back sitting participants had shorter saccades and smaller saccade amplitudes, meaning that they shifted their attention less than those sitting in the front which could also be related to content size in their field of view. Furthermore, the back sitting participants engaged significantly more with the virtual peer-learners whereas the front sitting participants engaged more with the virtual instructor and lecture screen. Considering that the virtual instructor and screen are more related to learning lecture content and considering the increased fixation time in the back sitting condition, while designing virtual learning spaces, one might locate students in the frontal regions of the classrooms if visually attending the lecture content is important. If engagement with virtual peers is more important, one might favor either locating the students in the back of the classroom or even organizing the desks in a U-, V-, or O-shape so students can have a better view of their virtual peers. In summary, according to visual attention and cognition findings, the location of the students in the classroom should be determined based on the goal of the virtual lecture.

The avatar representation styles are directly related to VR environments because it is not possible to have such configurations in real classrooms. According to the findings of this manipulation, students engaged more with the environment in general with longer fixation and shorter saccade durations when cartoon-styled avatars were presented. While the opposite trend between fixation and saccade durations fits with theoretical expectations, the students that encountered cartoon-styled avatars also engaged more with the peer-learners than the students encountered with realistic-styled avatars. Considering that the students who attended the virtual lecture were small children, engaging more with the cartoon-styled peer-learners is a reasonable and explainable outcome. On the contrary, the students that encountered realistic avatars had larger pupil diameters indicating a higher cognitive load in general. Taking the reasonably controlled illumination of the students' sitting positions into account, the findings indicate that the lecture with realistic characters increased focus and concentration in the learning space. Similar to sitting positions, when such environments are designed, together with the target groups' demographic information, visualization styles should be tailored to the dedicated lecture type, such as interacting with either engaging peers or realistic instructors.

The hand-raising behaviors of the peer-learners can theoretically be manipulated in virtual and real classrooms. However, it is very challenging to have a controlled experiment for such a condition in the real world. Secondly, if these behaviors are manipulated in the real classroom students are already biased by knowing the typical performance of their classmates when reacting to different topics. The behaviors related to such "artificial" manipulation might not

provide naturalistic conclusions. Therefore, it is more feasible to study these manipulations virtually. However, according to the findings on visual behavior, much research is needed. The results indicate that the extreme hand-raising behaviors of peer-learners yielded higher cognitive load in students. It is likely that when a moderate number of peers raise their hands for questions during the lecture, students find these actions to be straightforward and their level of focus is less than in the extreme levels of hand-raising behaviors. In terms of visual distraction that could occur with many peer-learners raising their hands, the 80% hand-raising condition gathered the most attention on the peer-learners. This indicates that the efforts of virtual peer-learners to participate in the lecture were noticed by the students. As theoretically expected, this yielded the least attention on the virtual instructor and on the virtual lecture screen. However, the effects of this manipulation are relatively mixed and should be further investigated. In the case of continuous VR class attendance, hand-raising should be calibrated carefully due to the possible impact on student self-concepts. Furthermore, one might consider an adaptive manipulation depending on the individual, subject, or topic in the context of everyday VR.

In summary, the findings indicate that human visual attention and cognition differ significantly when such educational manipulations are introduced, contributing to the state-of-the-art. While the optimal configurations may depend on multiple factors like the target group characteristics (e.g., ages, VR experience, and etc.), lecture content, and the mainstream drawbacks of virtual environments with HMDs, such as limited field of view and possible cybersickness when confronted with long durations, such environments have great potential in the digital era. Considering the switch to digital teaching during the recent COVID-19 pandemic, the increasing number of online classes, and even the efforts to generate classroom twins for VR [123], it seems that new teaching and learning paradigms are on the way in our daily life.

4.1.2 Virtual Reality in Driving

The driving studies conducted have shown that multiple implications can be drawn using eye movements not only for driving configurations, but also for time dependent and critical tasks that can be carried out in the everyday VR context.

While driving in the real world, many modern vehicles provide driver assistance features such as collision warning and lane keeping. However, in real life, due to human safety, it is not possible to train people for safety critical situations. Humans can train for interactions with pedestrians, driving with automated vehicles, or overtaking scenarios with low-cost VR HMDs easily at their homes without any significant consequences. To this end, a first step towards interactions with critically crossing pedestrians has shown that even the gaze-aware and minimalistic warning cues for critical pedestrians, which are located around the periphery, help drivers to drive safer and smoother according to assessment of pupil diameters, driver pedal inputs, and distances to the critically crossing pedestrians. Furthermore, during these

scenarios cognitive load estimation can be carried out accurately and in real time. While cognitive load assessment for real world situations may be viewed skeptically due to effect of illumination on human pupil sizes, thanks to the relatively controlled illumination that can be provided in VR, it is possible to make use of pupil dilations in such setups. Considering that even the Formula 1 drivers practice with simulators [185] (e.g., Lando Norris practicing with a simulator [186]), and their visual attention through eye movements is considered almost superhuman [187], and as such similar setups may be used in daily lives for different purposes such as entertainment, it can be argued that ordinary people and particularly novice drivers can train for many different traffic scenarios with the help of relatively low-cost VR simulators and eye movements.

As the shift of visual attention in a time dependent context is not necessarily related to driving, there are implications for other everyday VR scenarios. Currently, even though VR and HMDs have several disadvantages such as low resolutions, vergence-accomodation conflict [188], and significant weight of the HMDs, with small cues it is possible to shift attention quickly towards important regions of the presented 3D stimulus. In the driving context, this is validated with pedestrians. Other scenarios could include an attention shift for a student attending a class in VR to support the overall learning process, for a video gamer to notify important milestones during the game, or for novices in more or less any domain.

4.2 Privacy Preserving Eye Tracking

Differential privacy mechanisms applied to eye movement features and privacy preserving gaze estimation based on a randomized encoding-based framework are discussed in the following subsections.

4.2.1 Differential Privacy

Findings on the application of differential privacy mechanisms such as FPA, CFPA, or DCFPA indicate that it is not very trivial to have high utility while preserving privacy due to high amount of noise required by the differential privacy mechanisms applied to correlated time-series data. These effects are discussed in the following based on two evaluation metrics, namely the utility metric based on the NMSE and classification accuracies of different tasks.

Firstly, from a higher utility aiming perspective application of chunking in the CFPA and DCFPA significantly decreased the amount of noise needed to make the eye movement query outcomes differentially private. This can been seen when the CFPA is compared with FPA and the different chunk sizes (i.e., 32, 64, and 128) within the DCFPA are evaluated using the NMSE-based utility metric. While chunking could be considered a method from the signal processing domain, according to the Parallel Composition Theorem [189], since the chunks are not overlapping, the differential privacy is preserved. Using larger chunks decorrelates the data more effectively therefore, it should be preferable when the utilities are similar between various

chunk sizes. However, when the DCFPA is considered, apart from the chunking mechanism since differences between consecutive observations are used, and when the noisy values are propagated within each chunk to obtain final noisy observations, the Sequential Composition Theorem [189] is applied for overall privacy and ϵ calculations. Therefore, when the chunk sizes are larger more noise is needed to achieve differential privacy in this method. While this is a disadvantage due to the divergence of overall eye movement signals from the original signals, since it has been empirically determined that the eye movement difference signals are less correlated compared to the original signals, the privacy leak for this method due to data correlations is less than in others methods. Overall, based on the NMSE-based utility metric, since the divergence of the differentially private eye movements is less for the CFPA and DCFPA compared to the FPA and the standard Laplace mechanism of the differential privacy, it is better to use these methods for eye movements. However, there are many trade-offs such as different chunk sizes and ϵ values, namely different privacy regions. The methods should be tailored according to the eye movement feature generation pipelines, and possibly further tasks that are applied in the everyday VR context.

According to the discussed utility perspective, it is optimal that differentially private eye movement signals are less diverged from original eye movement signals. When it comes to classification tasks, however, the overall goal is more complex than the NMSE-based utility metric because there are different tasks such as the classification of document types or person identification. For instance, while it is desirable to have high accuracies in document type or privacy sensitivity classification (e.g., as applied for MPIIDPEye [143] and MPIIPrivacEye [28] datasets), low accuracies in gender prediction and person identification are preferred when the privacy of individuals is considered. Comparing differential privacy mechanisms that are appropriate for time-series data, namely FPA, CFPA, and DCFPA, it is possible to have almost perfect accuracies for person identification tasks with the FPA in both datasets. Even if other classification accuracies are obtained in a preferable success, it is likely that the FPA is vulnerable to person identification attacks. On the contrary, both the CFPA and DCFPA significantly decrease person identification accuracies towards guessing probabilities. All the mechanisms successfully hide gender information in a person-independent cross-validation setup in the high privacy regions, which is expected since, even with the clean data, the accuracies are in the vicinity of 70% also according to the previous work [143]. The FPA works over 85% accuracy in the document type classification task which is comparably higher than the DCFPA with 64% accuracy in the most private regions (i.e., $\epsilon = 0.48$); however, due to its lack of resistance to attacks on person identification, one should determine on a trade-off when using the FPA. As the human reading behaviors consist of "Z"-type (or similar) patterns and the CFPA and the DCFPA perturbs the eye movement data with chunks, it is suspected that such patterns are removed easily with these mechanisms unlike the FPA. Therefore, this is a task-specific outcome of the evaluated methods. In the privacy sensitivity detection solely based on the differentially private eye movements, both the CFPA and DCFPA outperform the FPA, which performs state-of-the-art in terms of differential privacy perspective in the eye tracking domain. Overall, there are multiple trade-offs to consider before applying differential privacy

mechanisms on the eye movements. These include further tasks, the stimulus information from the original eye movements that were collected, data correlations, and the amount of background information that an adversary may have on the data. Especially when more practical use-cases in everyday life are considered, practitioners that design privacy mechanisms should take the latter issue into account and propose privacy solutions accordingly.

4.2.2 Randomized Encoding

Unlike differential privacy, if the complete raw data needs to be private cryptographic approaches should be employed. The randomized encoding (RE)-based framework that is utilized falls into this category because the raw data should not be available, for example, to function parties (e.g., third party cloud or server instances). In principle, since eye movements and eye tracking data obtained from VR HMDs represents visual biometrics, any type of additional task working with encrypted raw eye tracking data could be employed. As aforementioned, when such a system is built for training and testing machine learning models, and a real time interaction mechanism is needed, test times should fit this expectation. For this reason, instead of evaluating more sophisticated tasks such as cognitive load detection, gaze prediction, foveated rendering or similar tasks for everyday VR setups, a fundamental gaze estimation task was chosen. This is because gaze estimation is the starting point for all other eye movement related tasks in the VR setups.

The use-case of privacy preserving gaze estimation via the randomized encoding based framework has shown that it is possible to estimate gaze using the baseline SVR model identically to the non-private version (See the proof in Section B.2). While the decreasing trend of mean angular error is anticipated with a higher amount of data, the most important discussion point is its real time working capability. A prediction time of approximately 1.1 milliseconds falls in the range of a real time capable system in the VR domain. However, since this is only the prediction time on the function party instance, a possible communication latency is introduced during an everyday application scenario. It is assumed that as long as efficient communication between input and function parties is available, the proposed work will function in real time. In addition, from a machine learning perspective the input feature size is 36 for this work and is dedicated to the baseline gaze estimation task. Feature set size might also be important depending on the task and configuration. In the eye tracking literature, there are generic features based on fixations, saccades, pupil diameters, and blinks. For instance, in the works of Steil et al. [143, 28], 52 eye movement features are used for estimating stimulus type, gender, or scene privacy accurately based on SVMs. As the used feature vector sizes and machine learning models overlap for different tasks, our results imply that all these tasks could be done with multiple input parties as well while preserving the privacy. Such approaches not only protect data privacy, but also help to increase the training data as the total training data consists of data from multiple parties. Therefore, in the case of a lack of data for training models, these frameworks can be utilized as well. The proposed scenario and use-case stands as the only work in the current literature on the intersection of cryptography, randomized

encoding, and eye tracking in everyday VR.

4.3 Remote Eye Tracking Data Collection Possibilities for Virtual Reality

Obtaining eye movement behaviors of human subjects with heterogeneous backgrounds is indeed an important challenge for data driven VR systems such as user-assistive or gaze-guidance based on machine learning. Usually these types of studies are conducted within a small group of human subjects of similar ages and backgrounds who might not be representative of different populations. The protocol proposed in Section 3.3 offers a first step to solve this issue and make crowdsourcing possible with high-end VR HMDs.

As the proposed protocol assumes that the data collectors provide subjects with the VR application via a secure and direct method, the advertisement of the experiment should be completed externally via mailing groups, online forums, and etc. While these actions require extra effort to attract HMD users for the experiments, Rivu et al. [158, pp. 20-21] have reported that this method has the potential for independent studies that require different technical functionalities. However, one should be aware that attracting HMD users in such a way will likely create a sample set of experienced VR and HMD users which may introduce a confounder on user studies that evaluate eye movements.

Another important discussion point is the increased effort required by researchers and developers. For instance, the white-box cryptography [179] paradigm should be implemented within the VR application to prevent data altering attacks. This requires the selection of secret keys, a hashing algorithm, and the actual implementation of a cryptographic approach. However, in the future it is likely that software packages will be available for such purposes that every VR application can benefit from. In addition, in order to increase the data quality coming from remote participants, practitioners should implement processing pipelines for quality checking and filtering of eye movement data. For instance, one might filter the data by setting thresholds using tracking ratios or eye openness as reported by eye tracking sensors. Although doing these actions within the VR application might be seen as overhead by some in general, they are necessary to guarantee the validity of remote data collection.

Lastly, our protocol makes use of blockchain and smart contracts. In particular, we have used Ethereum platform [181] and its smart contract functionality for our use-case. Blockchain is needed for its immutability for recording the hash value of the collected eye movement data. Smart contracts are utilized for compensation in experiments without a centralized authority between data collector and subjects. While any blockchain-based platform that supports smart contract functionality could be used for this protocol, compensations are distributed in cryptocurrencies in any case. Due to the fact that even well-established cryptocurrencies such as Bitcoin [180] and Ether [181] are highly volatile in value, one might be skeptical about using them for such purposes. In addition, they have not yet been embraced for daily

usage by large communities. However, going forward it is possible that such currencies will be used in daily life more frequently (e.g., recently El Salvador have accepted Bitcoin as legal tender along with American dollar [190]), that communities will soon be more familiar with them. Additionally, such technologies disable the extra layer of centralized institutions between parties, particularly for financial transactions, and decrease the cuts applied by these centralized entities. Another advantage is that, in the case of public blockchain usage, the compensation distribution process is directly transparent via the web. Therefore, while some may consider the use of cryptocurrencies in everyday VR experiments utopic or futuristic, there is great potential.

4.4 Outlook

Virtual reality and eye tracking research requires interdisciplinary work. Both research areas consist of components related to computer hardware, computer graphics, human-computer interaction, cognitive science, artificial intelligence, and psychology. On top of these, fields including cryptography and security are involved in the privacy preservation of eye movement data collected from HMDs. This work is one of the first that combines these multiple aspects to create an overall framework towards everyday virtual reality. In terms of visual attention and cognition, conventional measures related to fixations, saccades, pupil diameters, object-of-interests similar to area of interests, etc., have been customized and used in the evaluation of VR scenarios in education and driving domains along with machine learning and other sensing modalities depending on the context. Privacy preserving paradigms cover two areas: differential privacy and the utilization of a randomized encoding-based framework. In differential privacy, the privatized features are aggregated statistics of fixations, saccades, pupil diameters, or blinks similar to those used in education and driving studies, apart from the feature extraction timespans. The main reason for this is that in the attention and cognition studies, the focus was human visual behaviors throughout the complete experiments, whereas in the differential privacy context, the main goal was time-series representations of such statistics. Contrary to the features used in the differential privacy work, gaze estimation utilizing an RE-based framework focused on eye landmarks for a straightforward reason. Estimation of gaze is the initial point of extracting features related to fixations and saccades. At the same time, if the eye movement data is completely encrypted, it is reasonable to encrypt the initial step rather than moderate or final steps of the data processing pipeline. After privacy preserving gaze estimation, one can extract features locally and use them for further analyses with privacy guarantees. Such works on preserving the privacy of individuals that intend to use VR related technologies, along with eye tracking, will likely enable usage from wider communities and help VR become more accessible in everyday life. Lastly, the remote data collection protocol proposed for eye tracking and VR spans all the works as it is possible to collect either raw eye tracking or feature-related data with such protocols. It is even possible to collect data outside of eye tracking, as long as the data quality can be controlled within VR applications. As the presented work has multiple aspects, the outlook for each component is

discussed separately. Each of these aspects also contribute technically to recent discussions on Metaverse [191, 192, 193] and they should be taken into consideration.

In the education works, the observed virtual classroom was pre-scripted and did not include a real interaction experiencing students' perspectives. In addition, the peer-learners in the classroom were simple pre-scripted bots. While the results are very important and the first for such setups from a visual attention and cognition perspective, more sophisticated setups could be employed. For instance, pre-scripted peer-learners could be replaced with smart agents, namely computer games like intelligent bots to study attention towards such agents. Furthermore, while synchronization may not be trivial using these custom environments, to generate a more realistic setup each peer-learner and the instructor could be connected to a real person using web-based technologies to replicate a complete virtual and remote classroom environment with actual classmates and teacher. In this case, one should employ other sensing modalities such as hand tracking and audio to ensure interactive virtual environments. Such configurations will also help make VR viable in everyday life considering the immersion and real person interactions it provides similar to conventional classrooms in the remote teaching and learning scenario.

The driving research that was conducted is related to setups that one might encounter in real daily driving. The end goal is to enable humans to train for similar time critical situations by making use of eye movements not only in driving, but also in other domains. In terms of driving, future works could focus on interactions within other critical situations such as take-over scenarios, complex traffic situations that include multiple critical pedestrians or vehicles, and even interactions with semi- and fully-automated vehicles. There are works in the direction of automated vehicles, in particular, with more sophisticated and expensive driving simulators (e.g., [194, 195]); however, it is an open question as to how visual behaviors and cognition would appear in the case of everyday VR with HMD scenarios. At the same time, with the growing market size for autonomous vehicles [196], interactions between the human driver and semi-autonomous vehicles will be key, and there is a research gap in this direction. With what VR provides, it is likely that such interactions could be studied to ensure a deeper understanding of these relationships in daily life. Additionally, the effects of gaze-aware and minimalistic human visual support, as featured in the aforementioned driving work, could be studied for domains other than driving. Lastly, VR-based training packages for critical scenarios could be utilized and evaluated over time.

Differential privacy has been studied in many domains in addition to human-computer interaction. What is proposed in this work is mainly for temporally correlated eye movement feature signals and for reducing the effects of temporal correlations on individual privacy. However, considering substantial improvements on various tasks with deep neural networks, one could employ differential privacy along with deep learning [197], similar to Abadi et al.'s work [198] while using the eye movements and tasks for VR setups. Taking into account the recent open source libraries for privacy such as Opacus [199], which enables model training with differential privacy, this direction may be more relevant and focused in the near future. In

addition, local differential privacy [200] scenarios could be employed for not only eye tracking, but also for any type of biometric data collected from the VR HMDs. As it is required to noise the data locally in local differential privacy, utility and noise dynamics are different and should be studied explicitly.

Privacy preserving gaze estimation work has focused on two input parties and a function party which could be thought of as a cloud instance. However, in real world scenarios the number of input parties may be more than two. Considering many HMD users wish to train models while preserving their privacy, this scenario has similarities to N number of hospitals that share their data privately for further processes as proposed by Chen et al. [201]. Such directions are possible for eye tracking data collected from VR displays as well. Furthermore, instead of focusing on SVMs, different classifiers such as decision trees, random forests, and artificial neural networks could be considered. While the proposed gaze estimation method based on a randomized encoding framework is the first for the eye tracking domain, privacy preserving eye tracking and human-computer interaction is indeed a green field in terms of these research directions in the context of everyday VR.

The blockchain-based work for collecting eye movement data from remote participants is a proof-of-concept protocol. Therefore, there are multiple ways to extend this work. The first step is to actually implement the complete workflow within a VR application and test its usability by employing self-reported measures provided by participants. While interactions with blockchains and smart contracts are trivial in principle, they are considerably new technologies. The blockchain concept was introduced with Bitcoin [180] and smart contracts for blockchains with Ethereum [181] in the last two decades. Therefore, the usability of such technologies in VR applications is a relevant direction to explore. Even though there are some works that explore the user experience of cryptocurrency wallets with augmented reality [202] and general use-cases for interaction design [203], not much research has been conducted for the aforementioned. GazeCoin [204], a cryptocurrency for payments based on the eye gaze for VR/AR was recently introduced as a token. This also shows the potential of such technologies for VR and everyday use-cases. For the proposed protocol, while the Ethereum platform and its blockchain were used, newer platforms such as Avalanche [205] and Polkadot [206] could be employed along with their extended functionalities. Apart from these, one might strive for more intuitive ways to guarantee data integrity and a decentralized method of experiment compensation in future works.

4.5 Conclusion

A significant contribution to the scientific research of everyday VR using eye tracking was carried out in the context of this thesis. This encompasses research on human attention and cognition understanding based on eye movements and features in multiple domains, including education and driving, privacy preserving manipulations of eye movement features and signals with their algorithmic foundations, and a versatile protocol that may help make

VR more accessible to a wide range of human participants with different socio-demographic backgrounds.

A Visual Attention and Cognition in VR through Eye Tracking

This chapter includes the following publications:

1. Hong Gao*, **Efe Bozkir***, Lisa Hasenbein, Jens-Uwe Hahn, Richard Göllner, and Enkelejda Kasneci. Digital transformations of classrooms in virtual reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI)*, New York, NY, USA, 2021. ACM. doi: 10.1145/3411764.3445596.
2. **Efe Bozkir***, Philipp Stark*, Hong Gao, Lisa Hasenbein, Jens-Uwe Hahn, Enkelejda Kasneci, and Richard Göllner. Exploiting object-of-interest information to understand attention in VR classrooms. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, New York, NY, USA, 2021. IEEE. doi: 10.1109/VR50410.2021.00085.
3. **Efe Bozkir**, David Geisler, and Enkelejda Kasneci. Assessment of driver attention during a safety critical situation in VR to generate VR-based training. In *ACM Symposium on Applied Perception (SAP)*, New York, NY, USA, 2019. ACM. doi: 10.1145/3343036.3343138.
4. **Efe Bozkir**, David Geisler, and Enkelejda 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) Workshops*, New York, NY, USA, 2019. IEEE. doi: 10.1109/VR.2019.8797758.

* indicates equal contribution.

Figure A.1: Immersive virtual reality classroom.

A.1 Digital Transformations of Classrooms in Virtual Reality

A.1.1 Abstract

With rapid developments in consumer-level head-mounted displays and computer graphics, immersive VR has the potential to take online and remote learning closer to real-world settings. However, the effects of such digital transformations on learners, particularly for VR, have not been evaluated in depth. This work investigates the interaction-related effects of sitting positions of learners, visualization styles of peer-learners and teachers, and hand-raising behaviors of virtual peer-learners on learners in an immersive VR classroom, using eye tracking data. Our results indicate that learners sitting in the back of the virtual classroom may have difficulties extracting information. Additionally, we find indications that learners engage with lectures more efficiently if virtual avatars are visualized with realistic styles. Lastly, we find different eye movement behaviors towards different performance levels of virtual peer-learners, which should be investigated further. Our findings present an important baseline for design decisions for VR classrooms.

A.1.2 Introduction

Recently, many universities and schools have switched to online teaching due to the COVID-19 pandemic. Online and remote learning may become more prevalent in the near future. However, one of the disadvantages of teaching and learning in such ways compared to conventional classroom-based settings is the limited social interaction with teachers and peer-learners. As this may demotivate learners in the long term, better social engagement providing solutions such as immersive virtual reality (IVR) can be used for teaching and learning. Next-generation VR platforms such as Engage[1] or Mozilla Hubs[2] may offer better social engagement for learners in the virtual environments; however, the effects of such environments on learners have to be better investigated. In addition to the opportunity to provide more efficient social engagement configurations, VR also enables building and evaluating situations that are difficult to set up

[1] https://engagevr.io/
[2] https://hubs.mozilla.com/

in real life (e.g., due to the privacy-related concerns or current availability).

While VR technology has a long history in the education domain [207, 208], the current availability of consumer-grade head-mounted displays (HMDs) allows for the creation of immersive experiences at a reasonable cost, making it possible to employ immersive personalized VR experiences in classrooms in the near future [209]. However, the digital transformations of classrooms reflect an important and critical step when developing VR environments for learning purposes and require further research. A unique opportunity to understand the gaze-based behavior, and consequently, attention distribution of learners in such VR settings is provided through the analysis of the eye movement of learners [210]. Since some of the high-end HMDs already consist of integrated eye trackers, it does not require extensive effort to extract eye movement patterns during simulations in VR. A thorough analysis of the eye movements allows to infer information on the users going beyond the gaze position, for example stress [211], cognitive load [169], visual attention [212], evaluation and diagnosis of diseases [121], future gaze locations [96], or training evaluation [116]. In the virtual classroom, this rich source of information could even be combined with the virtual teachers' attention, similar to real-world classrooms [213, 163], to design more responsive and engaging learning environments.

In this study, we design an immersive VR classroom that is similar to a real classroom, enabling students to perceive an immersive virtual classroom experience. We focus on exploring the impact of the digital transformation from the classroom to immersive VR on learners by analyzing their eye movements. For this purpose, three design factors are studied, including sitting positions of the participating students, different visualization styles of the virtual peer-learners and teachers, and different performance levels of virtual peer-learners with different hand-raising behaviors. Figure A.1 shows the overall design of the virtual classroom. Consequently, our main contributions are as follows.

- We design an immersive VR classroom and conduct a user study to enable students to virtually perceive "interactive" learning.
- We analyze the effect of different sitting positions on learners, including sitting in the front and back. We find significantly different effects in fixation and saccade durations, and saccade amplitudes in relation to the sitting position.
- We evaluate the effect of different visualization styles of virtual avatars on learners including cartoon and realistic styles and find significantly different effects in fixation and saccade durations, and pupil diameters.
- We assess the effect of different performance levels of virtual peer-learners on learners by evaluating various hand-raising percentages, and find significant effects particularly in pupil diameters and number of eye fixations.

A.1.3 Related Work

As head-mounted displays (HMDs) and related hardware become more accessible and affordable, VR technology may become an important factor in the educational domain, particularly

given its provided immersion and potential for teaching [214, 215]. Various recent works on VR and education indicate that VR may offer significant advantages for learning and teaching. For instance, based on the post-session knowledge tests, both augmented and virtual reality (AR/VR) are found to promote intrinsic benefits such as increasing learners' immersion and engagement when used for learning structural anatomy [216]. In [217], the impact of VR systems on student achievements in engineering colleges was investigated by evaluating the results of post-quizzes and the results show that VR conditions present significant advantages when compared to no-VR conditions since students improve their performance, which indicates that VR can successfully support teaching engineering classes. Additionally, VR was also evaluated to help teachers develop specific skills that can be helpful in their teaching processes [218]. In addition to teaching and learning processes, another aspect under evaluation concerns the types of virtual environment configurations that are used not only for learning, but also for exploring immersion, motivation, and interaction. To this end, different types of VR setups have been studied. [209] introduced an immersive VR tool to support teaching and studying art history, which indicates, when used for high-school students, an increased motivation towards art history. [219] explored the possibility of using low-cost VR setups to improve daily classroom teaching by using a smartphone-based VR system. According to the evaluations using pre- and post tests, the proposed VR setup helps students perform better compared to traditional teaching using whiteboard and slides. Furthermore, HMD-based VR environment was studied in an elementary classroom for teachers to guide their students in exploring learning elements in immersive virtual field trips [220]. It has been concluded that students' motivation was enhanced after the virtual field trips. Overall, such works imply that while increasing motivation and engagement, different types of VR environments provide plenty of benefits and can be used to assist learning and teaching processes by providing users with immersive experiences.

One disadvantage of such VR and online learning tools is that learners' motivation and performance may be affected by lack of social interaction [221], peer accompaniment [222], or immersion [223]. Furthermore, realism in immersive environments can have various implications [224], related to both learning and interaction. To address these issues, several works have focused on how to provide more realistic and immersive environments. For example, [225] discusses the design of the VR environments for classrooms by replicating real learning conditions and enhancing learning through real-time interaction between learners and instructors. Furthermore, [226] constructed virtual classmates by synthesizing previous learners' time-anchored comments and indicates that when students are accompanied by a small number of virtual peer-learners built with prior learners' comments, their learning outcomes are improved. In addition to virtual peer-learners, the presence of virtual instructors may also have an impact on learning in VR. [227] investigated this and reports that learners engaged more with the environment and progressed further with the interaction prompts when a virtual instructor was provided. These works and findings indicate that the styles and types of virtual agents in the virtual environments may have several effects on students' attention and perception during immersion and should be taken into account. The evaluation

of real-time visual attention towards similar configurations, which could be carried out using sensors such as eye trackers, may not only help to understand learning processes but also provide empirical insights about interactions during virtual classes for digital transformations of classrooms in VR.

From immersion and interaction point of view, video teleconferencing systems share similar goals with the VR classrooms as such systems enable people to experience highly immersive and interactive environments [228] and have been studied in the VR context as well. For example, [229] proposed a video teleconference experience using a VR headset and found that the sense of immersion and feeling of presence of a remote person increases with VR. Furthermore, different mixed reality (MR)-based 3D collaborative mediums were studied in terms of teleconference backgrounds and user visualization styles [230]. The real background scene and realistically constructed avatars promote a higher sense of co-presence. Low-cost setups were investigated also for real-time VR teleconferencing [231], as it was done for VR learning environments and it is found that it is possible to improve image quality using headsets in these setups. The possibility of having low-cost setups may become an important factor in the future when accessibility and extensive usage of everyday VR environments for learning [217] and interaction [225] are considered.

In general, while the visualization styles and rendering are considered to affect learners' perception and attention, in virtual learning environments particularly in IVR classrooms, other design factors are also important for attention-related tasks. For instance, [232] has studied the effect of being closer to the teacher, being in the teacher's field of view (FOV), and the availability of virtual co-learners in virtual classrooms. In particular, the authors found that students learn more if they are closer to the teacher and by being in the center of the teacher's FOV. In addition, when no co-learners or co-learners who have positive attitudes towards the lecture (e.g., looking at the teacher or taking notes) are available, students learn more information about the lecture instead of the virtual room. Gazing time was approximated according to the time students kept the virtual teacher in their FOVs; however, real-time gaze information was missing during the experiments. Exact gazing patterns and different eye movement events during learning are particularly needed for understanding moment-to-moment visual behaviors of students. In another work, [233] studied the effect of the sitting position on attention-deficit/hyperactivity disorder (ADHD) experiencing students in such classrooms and found indications that front-seated students are affected positively by this configuration in terms of learning. However, similar to [232], the authors did not have gaze information available but identified that the evaluation of eye movements may provide additional insights during learning, particularly in terms of real-time visual interaction, when learning and cognitive processes are taken into consideration. In addition, eye movements are also considered as choice of measurements to study visual perception during learning [234, 40]. [235] and [236] have studied attention measures and social interaction in similar setups using continuous performance tests and head movements, respectively. The latter work has used head movements as a proxy for visual attention and found that head movements shift between target and interaction partner. This finding partly supports the finding of [227] that the learners' engagement

increases when a virtual instructor is presented. However, both works lack eye movement measurements. As also reported by [236], eye movements should be examined along with head movements to understand attention and interaction more in-depth, since eyes can move differently. In addition, [237] studied the relationship between performance, sense of presence, and cybersickness, whereas [238] examined attention, more particularly ADHD with continuous performance task in a virtual classroom. However, both works are more in the clinical domain, which are relatively different from an everyday classroom setup. [239] provides a general overview more from clinical perspective. Lastly, although has not been studied extensively in VR yet, peer-learners' engagement expressed by hand-raising behavior [240] may also affect the attention and visual behaviors of learners in the VR classrooms, which could be further studied.

In summary, while showing that VR could be a useful technology to support education, the aforementioned works primarily focused on the importance of used mediums and configurations, visualization styles, participant locations for visual attention, engagement, motivation, and learning of participants in VR classrooms. Yet, real-time and moment-to-moment interactions with the environment and visual behaviors of students in an everyday VR classroom setup were not studied in depth. Although obtaining such information in real-time is challenging, analyzing eye-gaze and eye movement features can provide valuable understanding into visual attention and interaction in a non-intrusive way, especially for designing such classroom configurations. For instance, long fixations can be related to the increased amount of cognitive process [81], whereas long saccadic behaviors are related to inefficient search behavior [82]. Furthermore, pupillometry is highly related to cognitive workload [241, 242]. Such information is also argued for consideration in IVR environments [243, 244]. In fact, when designing immersive VR environments for digital transformations of classrooms in virtual worlds, such features can be key to understand visual attention, cognitive processes, and visual interactions towards different classroom manipulations, which may also affect learning and teaching processes. To address this research gap, we study three configurations in an everyday VR classroom setup including different visualization styles of virtual avatars, sitting positions of participants, and hand-raising based performance levels of peer-learners by using eye movement features.

A.1.4 Methodology

The main purpose of our study is to investigate the effects of digital transformations of the classrooms to VR settings on learners. Therefore, we designed a user-study to study these effects. In this section, we discuss the participant information, apparatus, experimental design, experiment procedure, measurements, data pre-processing steps, and our hypotheses. Our study and data collection were approved by the institutional ethics committee at the University of Tübingen (date of approval: 25/11/2019, file number: A2.5.4-106_aa) as well as the regional council responsible for educational affairs at the district of Tübingen.

Participants

Participants were recruited from local academic track schools via e-mails and invitation letters. After obtaining written informed consent from both students and their parents or legal guardians, all students who indicated interest were admitted to the study. 381 volunteer sixth-grade students (179 female, 202 male), whose ages range from 10 to 13 (M = 11.51, SD = 0.56), were recruited to participate in the experiment. Due to hardware problems or incorrect calibration, data from 32 participants were removed. In addition, data from 61 participants were also removed due to eye tracker related issues including low eye tracking ratio (lower than 90%). Therefore, data from 288 participants (137 female, 151 male), whose ages range from 10 to 13 (M = 11.47, SD = 0.51), were used for evaluations. We had 16 different conditions in the experiment and the average number of participants for each condition was 18 (SD = 5.3). In addition to the actual study and data collection, we successfully piloted both our technical setup and the experimental workflow with 55 similar aged (M = 11.35, SD = 0.52) sixth-grade students (20 female, 35 male).

Apparatus

In our experiments we employed HTC Vive Pro Eye devices with a refresh rate of 90 Hz and a field of view of 110°. The VR environment was designed and rendered using the Unreal Game Engine[3] v4.23.1. The screen resolution for each eye was set to 1440 × 1600. To collect eye movement data, we used the integrated Tobii eye tracker with a 120 Hz sampling rate and a default calibration with $0.5° - 1.1°$ accuracy.

Experimental Design

The virtual classroom designed in our study has 4 rows and 2 columns of desks along with chairs, as well as other objects which typically exist in the conventional classrooms such as a board and display. In total, there are 24 virtual peer-learners sitting on the chairs. A virtual teacher standing in front of the classroom teaches a ≈ 15-minute virtual lecture to the students about computational thinking [245]. During the lecture, the virtual teacher walks around the podium. The virtual peer-learners and participants sit on the chairs throughout the lecture. The lecture has four phases including **(a) topic introduction** (≈ 3 minutes), **(b) knowledge input** (≈ 4.5 minutes), **(c) exercises** (≈ 5.5 minutes), and **(d) summary** (≈ 1.5 minutes). There are distracting behaviors from virtual peer-learners (e.g., raising hands, turning around) in the first, second, and third phases of the lecture.

In the beginning of the first phase, the teacher enters the classroom, stays in the classroom for a while, and then leaves for ≈ 20 seconds, giving participants the opportunity to look around and adjust to the virtual environment. The topic of the lecture is displayed on the board as *"Understanding how computers think"*. During the first phase, the teacher asks five

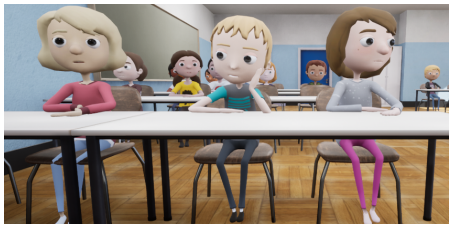
[3] https://www.unrealengine.com/

simple questions to interact with the students. Some of the peer-learners raise their hands and answer the questions. In the second phase, the teacher explains two terms to the students, namely, the terms *"loop"* and *"sequence"*. These terms are also shown on the display. Then, the teacher asks four questions about each term and the peer-learners raise their hands to answer the questions. In the third phase, the teacher gives the students two exercises to evaluate whether or not they understand the terms correctly. For each exercise, the students have some time to think. Then, the teacher provides the answers for each exercise, and the peer-learners vote for the correct answer by raising their hands. In the last phase, the teacher stands in the middle of the classroom to summarize the lecture. No questions are asked in this phase; therefore, none of the peer-learners raise their hands.

Our study is in between-subjects design. The participants are located either in the front or back region of the virtual classroom. The participants that sit in the front of the virtual classroom have one row in front of them, whereas the participants that sit in the back have three rows in front of them. The visualization styles of the avatars have two levels as well, in particular cartoon and realistic. Lastly, the hand-raising percentages, which are intended to show the performance levels of the virtual peer-learners, have four different levels, including 20%, 35%, 65%, and 80%. Combining all, we have a $2 \times 2 \times 4$ factorial design that forms 16 different conditions in total. Participants' views from back and front sitting positions, cartoon- and realistic-styled avatars are depicted in Figures A.2 (a), (b), (c), and (d), respectively.

(a) Back sitting participant experiencing the VR classroom.

(b) Front sitting participant experiencing the VR classroom.



(c) Cartoon-styled avatars.

(d) Realistic-styled avatars.

Figure A.2: Views from the immersive virtual reality classroom.

Procedure

Each experimental session took ≈ 45 minutes including preparation time. We conducted the experiments in groups of ten participants by assigning each participant randomly to one of the sixteen conditions. Before the data assessment took place at the participating schools, students were informed that they could drop out of the study at any time without consequences. After a brief introduction to the experiment and the data collection process, participants had the opportunity to acclimate with the hardware and the VR environment.

The experiment started with the eye tracker calibration. After calibration success, the experimenters pressed the "Enter" button to start the actual experiment and data collection process, wherein participants experienced the immersive virtual environment and the lecture. The experiments were supposed to be carried out in one session without breaks, mimicking thus a real classroom teaching session, lasting about 15 minutes. At the end of the experiment, the VR application displayed a message telling the participants to take off their HMDs. Lastly, participants filled out questionnaires about their experienced presence and perceived realism.

Measurements

For this work, our main focus was eye-gaze, head-pose, and pupil related activities of the participants as these are considered to be rich information sources, especially in VR. Fixations are the periods during which eyes are stationary within the head while fixated on an area of interest. Saccades, on the other hand, are the high-speed ballistic eye movements that shift eye-gaze from one fixation to another.

Using fixations, saccades, and pupil diameters, plenty of eye movement features are extracted. In this study, we extracted the number of fixations, fixation durations, saccade durations, saccade amplitudes, and normalized pupil diameters to analyze different conditions of the experiment. In the eye tracking literature, longer fixation durations correspond to engaging more with the object or increased cognitive process [81]. Fixation durations are mainly related to cognition and attention; however, it is argued that they are affected by the procedures that lead to learning and it is reported that fixation durations can be used to understand learning processes as well [246]. For instance, [247] has studied fixation patterns during learning in simulation- and microcomputer-based laboratory and found that simulation group had longer fixation duration, which means more attention and deeper cognitive processing. In addition to the fixations, longer saccade durations correspond to less efficient scanning or searching [82], whereas longer saccade amplitudes mean that attention is drawn from a distance [83]. Furthermore, a larger pupil diameter is related to higher cognitive load [20]. In addition, while being task dependent, [248] has indicated that pupil diameter measurements in high task load correlate with individual's performance. However, as pupil diameter values are also affected by the illumination, a controlled environment is needed to assess it. In our VR setup, the illumination is controlled across different conditions. Besides, a general overview of considering eye tracking as a tool to enhance learning with graphics is provided in [249].

Additionally, the self-reported presence and realism were assessed by questionnaires. The items in the questionnaires were based on the conceptualizations of [15] and [16] which were developed particularly to assess students' perception of the VR classroom situation. The experienced presence and perceived realism were assessed via using a 4-point Likert scales ranging from 1 ("do not agree at all") to 4 ("completely agree") with nine (e.g., "I felt like I was sitting in the virtual classroom." or "I felt like the teacher in the virtual classroom really addressed me.") and six items (e.g., "What I experienced in the virtual classroom, could also happen in a real classroom." or "The students in the virtual classroom behaved similarly to real classmates."), respectively.

Data Pre-processing

As the raw eye tracking data collected from the VR device does not include fixations, saccades or similar eye movements, we first pre-processed the data to identify these events. Detecting different eye movements in the VR setup is a challenging task and different from the traditional eye tracking experiments that include equipment such as chin-rests, as participants have opportunity to move their heads freely in VR. In the eye tracking literature, Velocity-Threshold Identification (I-VT) method is used to classify fixations based on velocities [84]. In the VR context, [85] applied a similar method to detect eye movement events. We opted for a similar approach.

Before applying the I-VT, we first applied linear interpolation for the missing gaze vectors. After the interpolation, we identified the fixations when the HMD was stationary. However, the identification of saccades was not restricted by the HMD movement. The used velocity and duration thresholds for the HMD movement states, fixations, and saccades are depicted in Table A.1, where the velocities and durations are given as v and Δ, respectively. Unlike the fixations and saccades, the pupil diameter values are reported by the eye tracker. As raw pupil diameter values are affected by blinks and noisy sensor readings, we smoothed and normalized the pupil diameter readings using Savitzky-Golay filter [165] and divisive baseline correction using a baseline duration of ≈ 1 seconds [166], respectively.

Table A.1: Head and eye movement event identification thresholds.

Event	Conditions for velocity (v)	Conditions for duration (Δ)
Stationary HMD	$v_{head} < 7°/s$	-
Fixation	$v_{head} < 7°/s$ and $v_{gaze} < 30°/s$	$100ms < \Delta_{fixation} < 500ms$
Saccade	$v_{gaze} > 60°/s$	$30ms < \Delta_{saccade} < 80ms$

Hypotheses

We developed three hypotheses, each corresponds to one design factor.

- **Hypothesis-1 (H1)**: We hypothesize that the different sitting positions of the participants yield different effects on the eye movements. As the participants that sit in the front are closer to the board, displays, and the teacher, we assume that they can attend the virtual lecture more efficiently than participants in the back and have less difficulty extracting information about the lecture. However, as they have a narrower field of view, particularly towards the frontal part of the classroom, they need to shift their attention more than the participants sitting in the back.
- **Hypothesis-2 (H2)**: We hypothesize that different visualization styles of virtual avatars affect student visual behaviors differently. More particularly, as students are familiar with realistic styles in the conventional classrooms, we claim that compared to cartoon-styled visualization condition, they attend the scene shorter during fixations in the realistic-styled visualization setting as cartoon-styled avatars are more attractive to the students. Therefore, students engage with the environment more in the cartoon-styled visualization condition than in the realistic-styled condition.
- **Hypothesis-3 (H3)**: We hypothesize that different hand-raising percentages of virtual peer-learners can distinctively affect the behaviors of participants. Specifically, we anticipate that when relatively higher percentages of hand-raising levels are provided, such as 65% or 80%, the participant's cognitive load will be higher due to the fact that many of the peer-learners attend the lecture with a high focus. Similarly, participants have more fixations in the classroom in the higher hand-raising percentage conditions as a higher number of hand-raising percentage creates an opportunity for various attention and distraction points.

A.1.5 Results

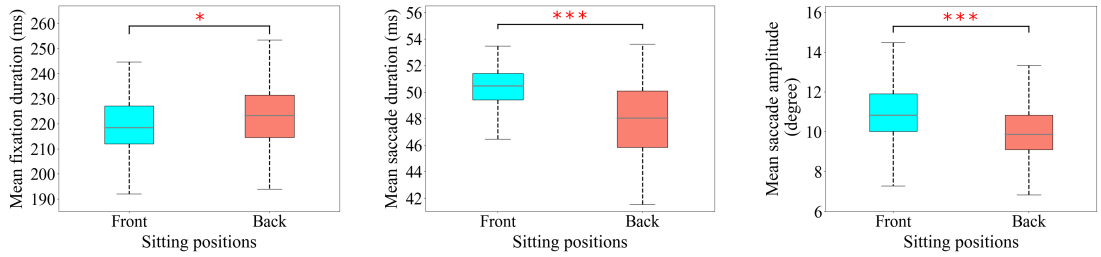
As we have three factors that form 16 different conditions, we applied 3-way full-factorial analysis of variance (ANOVA) by setting the level of significance to $\alpha = 0.05$ with Tukey-Kramer post-hoc test. For the non-parametric factorial analysis, we used the Aligned Rank Transform (ART) [167] before applying ANOVA procedures.

Analysis on Different Sitting Positions

Different sitting positions have an impact on the mean fixation and saccade durations, and mean saccade amplitudes. The mean fixation durations of the front and back sitting participants are illustrated in Figure A.3 (a). The participants that sit in the back have significantly longer mean fixation durations ($M = 222.6ms, SD = 14.57ms$) than the participants that sit in the front ($M = 218.75ms, SD = 13.11ms$), with $F(1,272) = 6.7$, $p = .01$.

Both saccade durations and amplitudes are influenced by the sitting positions and are depicted in Figures A.3 (b) and (c), respectively. The results reveal significantly longer saccade durations in the front condition ($M = 50.23ms, SD = 1.7ms$) than in the back condition ($M = 47.9ms, SD = 2.62ms$), with $F(1,272) = 73.76$, $p < .001$. Similarly, the mean saccade amplitude

is significantly larger in the front condition ($M = 10.93°, SD = 1.54°$) than in the back condition ($M = 10.05°, SD = 1.38°$), with $F(1,272) = 22.6$, $p < .001$.

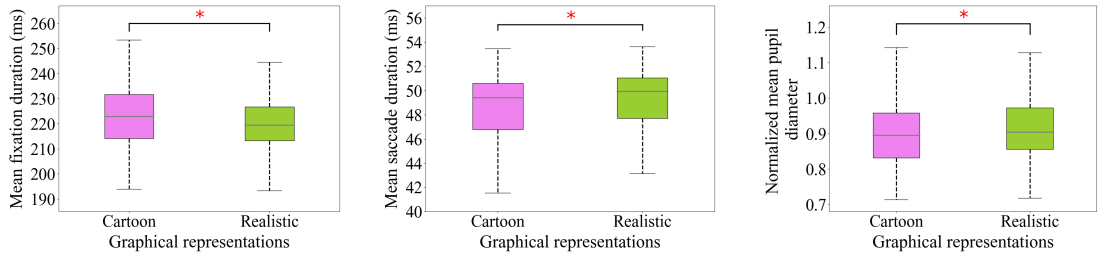


(a) Mean fixation durations. (b) Mean saccade durations. (c) Mean saccade amplitudes.

Figure A.3: Results for different sitting positions. Significant differences are highlighted with * and *** for $p < .05$ and $p < .001$, respectively.

Analysis on Different Avatar Styles

Different avatar visualization styles affect the mean fixation and saccade durations, and pupil diameters. The results are depicted in Figures A.4 (a), (b), and (c), respectively. The mean fixation durations are significantly longer in the cartoon-styled avatar condition ($M = 222.88ms, SD = 14.06ms$) than in the realistic-styled avatar condition ($M = 218.6ms, SD = 13.76ms$), with $F(1,272) = 5.27$, $p = .022$. By contrast, the mean saccade durations are significantly shorter in the cartoon-styled avatar condition ($M = 48.58ms, SD = 2.66ms$) than in the realistic-styled condition ($M = 49.3ms, SD = 2.35ms$), with $F(1,272) = 6.22$, $p = .013$.



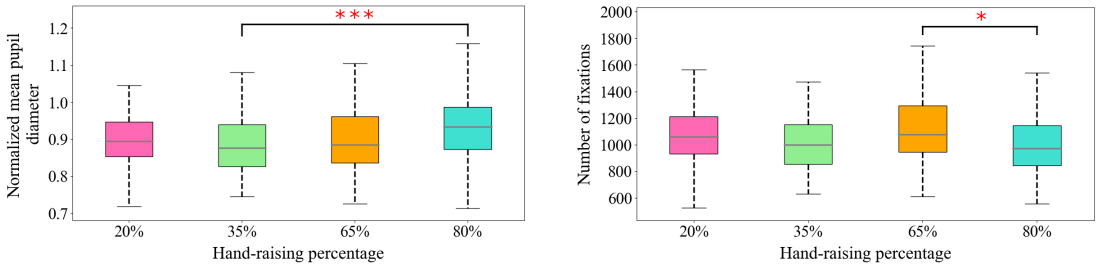
(a) Mean fixation durations. (b) Mean saccade durations. (c) Pupil diameters.

Figure A.4: Results for different avatar visualization styles. Significant differences are highlighted with * for $p < .05$.

The normalized mean pupil diameter, which reflects the cognitive load, is significantly larger in the realistic-styled avatar condition ($M = 0.94, SD = 0.16$) than in the cartoon-styled avatar condition ($M = 0.91, SD = 0.13$), with $F(1,272) = 3.94$, $p = .048$.

Analysis on Different Hand-raising Behaviors

The hand-raising behaviors of virtual peer-learners have significant impacts on the pupil diameters and number of fixations as depicted in Figures A.5 (a) and (b), respectively. We found significant effects on normalized mean pupil diameter values with $F(3,272) = 4.78$, $p = .003$. Particularly, mean pupil diameter in the 80% hand-raising condition ($M = 0.96, SD = 0.16$) is significantly larger than in the 35% hand-raising condition ($M = 0.9, SD = 0.12$), with $F(3,272) = 4.78$, $p < .001$. In addition, we found significant effects on number of fixations with $F(3,272) = 3.01$, $p = .03$. More specifically, there are notably more fixations in the 65% hand-raising condition ($M = 1112.92, SD = 245.07$) than in the 80% hand-raising condition ($M = 995.49, SD = 211.98$), with $F(3,272) = 3.01$, $p = .028$.



(a) Pupil diameters.

(b) Number of fixations.

Figure A.5: Results for different hand-raising percentages. Significant differences are highlighted with * and *** for $p < .05$ and $p < .001$, respectively.

Analysis on Experienced Presence and Perceived Realism

We did not find significant effects of different experimental conditions on the self-reported experienced presence and perceived realism. Overall, the self-reported experienced presence and perceived realism values are in the vicinity of highest values with ($M = 2.91$, $SD = 0.55$) and ($M = 2.91$, $SD = 0.57$), respectively. These mean that even though we did not obtain statistically significant differences between conditions, the participants experienced high levels of presence and realism in the IVR classroom environment.

A.1.6 Discussion

The results show that there are significant differences in the eye movement features between front and back sitting position conditions. Firstly, participants had longer fixations in the back sitting condition. This indicates that they had more processing time than the participants sitting in the front, which can be related to difficulty extracting information, similar to the relationship between task difficulty and mean fixation duration [184]. Secondly, the participants that sit in the front had longer saccade durations and amplitudes, which suggests that they needed to shift their attention more during the virtual lecture. While being located closer to

the lecture content, longer saccade durations indicate that the participants sitting in the front had less efficient scanning behavior [82] during the lecture. We assume that this was due to the narrower field of view. These results support our **H1**. When designing virtual classes, these results should be taken into account, particularly when determining where students should be located in the classroom, depending on the context.

Our results show consequential effects in the eye movement features in different avatar style conditions. As mean fixation durations are longer in the cartoon-styled visualization condition, we assume participants found the cartoon-styled avatars more attractive and attention-grabbing. Therefore, their fixation behaviors were longer during the virtual lecture. On the contrary, the mean saccade durations are longer in realistic-styled conditions as the fixation durations are shorter, which is theoretically expected. Furthermore, the pupil diameters of the participants in the realistic-styled condition are larger, indicating that the cognitive load of these participants was significantly higher during the lecture, which is suggested by the previous work [20]. This is an indication that participants may have taken the lecture more seriously and in a more focused manner when the visualization was realistic. These findings support our **H2**. Rendering realistic-styled avatars may be computationally expensive depending on the configuration. Therefore, an optimal trade-off should be decided, taking the behavioral results into account while designing the virtual classrooms.

Furthermore, we observe significant effects in attention towards different hand-raising based performance levels of the peer-learners. Particularly, the pupil diameters of the participants in the 80% condition are significantly larger than the pupil diameters of the participants in the 35% condition. We interpret this to mean that when the performance and attendance level of peer-learners was relatively higher, the participants' cognitive load became higher, indicating that they might pay more attention to the lecture content. This partially supports our **H3**. In addition, a greater number of fixations are observed in the 65% condition than in the 80% condition. We claim that when almost all of the peer-learners participated in hand-raising behaviors during the lecture, participants acknowledged this information without significantly shifting their gaze. However, this claim requires further investigation. Manipulation of different hand-raising conditions may affect student self-concept [183], which should be further studied as well.

In our study, the interaction and perception in the immersive VR classroom were assessed mainly by using eye-gaze and head-pose information. However, while the virtual teacher and peer-learners talk in the simulations, no response or interaction by means of audio or gestures was expected from the participants. Combining visual perceptions and interactions with such data may provide additional insights particularly for better interaction design in VR classrooms. A future iteration can also evolve into an everyday virtual classroom platform where each virtual agent is actually connected to a real person, similar to in platforms such as Mozilla Hubs. To this end, further design settings such as optimal seating arrangement (e.g., U-shape, circle shape) in addition to the sitting positions should be investigated. Evaluation

of similar configurations in online learning platforms such as Coursera[4], Udemy[5], or MOOCs[6] could provide additional implications for interaction modeling. Furthermore, gaze-based attention guidance can be considered for more interactive VR classroom experience and it can be achieved by fine-grained eye movement analysis focusing on short time windows instead of complete experiments. While being out of the scope of this paper, assessing learning outcomes and combining them with visual interaction and scanpath behaviors from immersive VR classroom could also offer insights for optimal VR classroom design.

A.1.7 Conclusion

In this work, we evaluated three major design factors of immersive VR classrooms, namely different participant locations in the virtual classroom, different visualization styles of virtual peer-learners and teachers, including cartoon and realistic, and different hand-raising behaviors of peer-learners, particularly through the analysis of eye tracking data. Our results indicate that participants located in the back of the virtual classroom may have difficulty extracting information during the lecture. In addition, if the avatars in the classroom are visualized in realistic styles, participants may attend the lecture in a more focused manner instead of being distracted by the visualization styles of the avatars. These findings offer valuable insights about design decisions in the VR classroom environment. Few indicators were obtained from the evaluation of the different hand-raising behaviors of peer-learners, providing a general understanding of attention towards peer-learner performance. However, these indicators should be further investigated and remain a focus of future work.

Acknowledgments

This research was partly supported by a grant to Richard Göllner funded by the Ministry of Science, Research and the Arts of the state of Baden-Württemberg and the University of Tübingen as part of the Promotion Program of Junior Researchers. Lisa Hasenbein is a doctoral candidate and supported by the LEAD Graduate School & Research Network, which is funded by the Ministry of Science, Research and the Arts of the state of Baden-Württemberg within the framework of the sustainability funding for the projects of the Excellence Initiative II. Authors thank Stephan Soller, Sandra Hahn, and Sophie Fink from the Hochschule der Medien Stuttgart for their work and support related to the immersive virtual reality classroom used in this study.

[4] https://www.coursera.org/
[5] https://www.udemy.com/
[6] https://www.mooc.org/

A.2 Exploiting Object-of-Interest Information to Understand Attention in VR Classrooms

A.2.1 Abstract

Recent developments in computer graphics and hardware technology enable easy access to virtual reality headsets along with integrated eye trackers, leading to mass usage of such devices. The immersive experience provided by virtual reality and the possibility to control environmental factors in virtual setups may soon help to create realistic digital alternatives to conventional classrooms. The importance of such settings has become especially evident during the COVID-19 pandemic, forcing many schools and universities to provide the digital teaching. Researchers foresee that such transformations will continue in the future with virtual worlds becoming an integral part of education. Until now, however, students' behaviors in immersive virtual environments have not been investigated in depth. In this work, we study students' attention by exploiting object-of-interests using eye tracking in different classroom manipulations. More specifically, we varied sitting positions of students, visualization styles of virtual avatars, and hand-raising percentages of peer-learners. Our empirical evidence shows that such manipulations play an important role in students' attention towards virtual peer-learners, instructors, and lecture material. This research may contribute to understanding of how visual attention relates to social dynamics in the virtual classroom, including significant considerations for the design of virtual learning spaces.

A.2.2 Introduction

Everyday use of head-mounted displays (HMDs) is increasing as virtual reality (VR) technology and virtual environments are already being used in various domains such as gaming and entertainment. In addition, some of the consumer-grade HMDs are coming to market with integrated eye trackers that may help to assess human attention during immersion and allow for more interactive virtual environments. It is likely that, in the near future, such tools will become widely used mobile devices similar to today's mobile phones or smart watches. To this end, not only should researchers strive to improve the capabilities of these devices, but scrutiny should also be given to understanding human behavior and attention while using such technology.

Measures of eye movements obtained through eye-tracking are effective indicators of human states and visual behavior to some extent; however, they are dependent on application or task [250]. Analyzing and modeling human attention using this data in a specific domain may not be transferable to other domains. Thus, when assessing human attention in digital environments, or more particularly in VR for the application in educational technology, specific domain knowledge and configurations should be considered. There is already some history of training and teaching in digital or virtual setups [251, 214]. Today, due to the COVID-19 pandemic, virtual or digital education has become more popular and even a necessity in many

cases. Currently, many schools and universities are carrying out their teaching responsibilities remotely via platforms such as Zoom[7] or Webex[8]. Such platforms lack the possibility of instructor-student interaction beyond audio and video features and encounter privacy concerns if videos are recorded and stored during classes. VR setups offer the immersion, interaction, and privacy preservation that current remote learning platforms lack. In addition, as VR allows users to easily control the environmental settings, it is possible to evaluate different classroom manipulations and subsequent effects on human behavior, a step that is exponentially more difficult in real world classrooms.

In this work, we exploit object-of-interest information by using eye-gaze and three main sets of objects in immersive VR. We focus on virtual peer-learners, virtual instructor, and screen to understand visual attention through the design of a virtual classroom and a lecture about computational thinking. We choose these objects-of-interests since they are of particular interest with regard to attention towards social dynamics and learning. Our study has three different design factors: Different sitting positions of participating students, different visualization styles of virtual avatars including an instructor and peer-learners, and different hand-raising behaviors of virtual peer-learners. Different sitting positions include seating participating students in the front or back of the virtual classroom. In addition, different visualization styles of avatars consists of two conditions that are cartoon- and realistic-styled avatars. Lastly, different hand-raising behaviors include 20%, 35%, 65%, and 80% of the peer-learners raising their hands to answer questions during the lecture. To the best of our knowledge, this is the first work that assesses students' attention by using object-of-interest information in an immersive VR classroom through the manipulation of sitting positions of students, visualization styles of peer-learners and instructor, and hand-raising behaviors of peer-learners collectively. Such manipulations may be important indicators of students' visual attention towards lecture contents and social dynamics in the classroom and should be taken into consideration when designing VR classrooms.

A.2.3 Related Work

Since our work benefits from VR in education and in eye tracking research, we discuss the state-of-the-art along these two lines. Various studies using VR in education settings assess the mechanisms of attention or social dynamics by using pre- or post-tests or by relying on head movement behavior as a proxy for gaze. Using eye tracking in addition to such information presents the possibility of a deeper understanding of visual and situational attention during immersive experiences.

[7] https://www.zoom.us/
[8] https://www.webex.com/

Virtual Reality in Education and Classrooms

VR offers great promise for supporting teaching and learning procedures, especially when digital learning, physical inabilities, ethical concerns, and situational limitations are considered. An extensive review of immersive VR in education and its pedagogical foundations are discussed in [214] and [252], respectively. We focus on research on VR in education and immersive VR classrooms in this section.

The effectiveness of learning in virtual and augmented reality (VR/AR) compared to tablet-based applications and the impact of VR-based systems on students' achievements are studied in [216] and [217], respectively, and these works indicate several advantages of VR-based conditions. In addition, it has been found that students' motivation increases when VR is used as a teaching tool in art history [209] and social studies [220]. VR not only supports the effectiveness of learning, but also can improve instructor teaching skills [218].

Apart from VR applications in teaching and learning, the design and degree of realism in VR classrooms have also been studied. Presence of a virtual instructor was found to increase the engagement and progress of users [227]. Furthermore, the processes of synthesizing virtual peer-learners by using previous learner comments [226] and designing VR classrooms by replicating real conditions [225] which may affect learning are considered.

Several works focused on understanding visual attention and behavior in immersive VR classrooms. Bailenson et al. [232] and Blume et al. [233] studied learning outcomes according to sitting positions and offer compelling evidence that students seated in the front have better learning outcomes. Few studies, however, took head movements into consideration [253, 235, 237, 236] in such setups. In [235], the immersive VR classroom was used as a tool to study attention measures for attention deficit/hyperactivity disorder (ADHD), whereas in [237] reliability of virtual reality and attention was studied with continuous performance task (CPT) for clinical research. Social interaction using head movements was studied in [236] with users' head movements found to shift between the interaction partner and target. Some studies argued for eye tracking measurements, especially in clinical research for diagnosis or attention related tasks [239, 238]. However, none of the previous works have focused on social interactions and dynamics in the immersive VR classroom in an everyday setting by using object-of-interest information and eye movements.

Eye Tracking in Virtual Reality

Eye tracking and gaze estimation are considered challenging tasks in a real world setting because it is difficult to control factors such as occlusions or illumination changes [19, 254]. However, in most of the VR setups, eye trackers are located inside of HMDs. This creates not only a more controlled and reliable environment for eye tracking, but also provides a unique opportunity to analyze and process human visual behavior during the VR experience.

Eye tracking has been used in many applications and shown to be helpful for various tasks

in VR such as guiding attention in panoramic videos using central and peripheral cues [255], predicting motion sickness by using 3D Convolutional Neural Networks [256], synthesizing personalized training programs to improve skills [116], foveated rendering using saccadic eye movements and eye-dominance [91, 89], evaluation and diagnoses of diseases such as Parkinson's disease [121], re-directed walking using blinking behavior [257], or continuous authentication using eye movements [27]. While these works have used either the eye tracking or gaze data to derive more meaningful information for related tasks, assessing visual attention via eyes and gaze-based interaction is more relevant for classroom setups in particular. Bozkir et al. [212] assessed visual attention using gaze guidance and pupil dilations in a time-critical situation, whereas Khamis et al. [258] discussed gaze-based interaction using smooth pursuit eye movements in VR. In addition, Sidenmark and Lundström [108] analyzed eye fixations on interacted objects during hand interaction in VR and found that interaction with stationary objects may be favorable. Aforementioned works indicate that eye movements can be used reliably in VR setups. Moreover, considering that the majority of objects in a classroom are stationary or have limited spatial movement, visual attention extracted from such data may provide valuable insight into human behavior. While exploiting objects-of-interests could be considered as a primitive task, it forms the foundation of more complex tasks necessary to understand visual attention.

A.2.4 Methodology

The main focus of this work is to investigate object-of-interest information in different manipulations of an immersive VR classroom. We focus on three objects that may be considered as the most important objects in the current setup, namely peer-learners, instructor, and screen.

Participants

381 volunteer sixth-grade students (179 female and 202 male) between 10 to 13 years old (M = 11.5, SD = 0.6) were recruited for the experiment. In this age group, students are able to use an HMD, but do not have much experience with VR. They also had no background knowledge about the lecture content. Data from 101 participants were removed due to hardware related problems, incorrect calibration, low eye tracking ratio (lower than 90%), and synchronization issues. The average number of participants per condition was 17.5 (SD = 5.2). Finally, we used the data of 280 participants (140 female and 140 male) with the aforementioned average age and standard deviation. For each condition group separately, participants' gender was also equally distributed (M = 0.58, SD = 0.08). The study was approved by the ethics committee of the University of Tübingen prior to the experiments. Participants and their parents or legal guardians provided written informed consent in advance.

Apparatus

For the experiments, HTC Vive Pro Eye devices with integrated Tobii eye trackers were used. The HTC Vive Pro Eye has a refresh rate of 90 Hz and field of view of 110°. The integrated eye tracker has 120 Hz sampling rate. The screen resolution per eye was set to 1440 × 1600. Unreal Game Engine v4.23.1[9] was used to render the virtual classroom.

Experimental Design

The virtual classroom consists of 4 rows of desks organized in 2 columns. Next to each desk, chairs are located to let virtual peer-learners sit. There are 24 virtual peer-learners in the environment and all of them sit on chairs during the entirety of the lecture. Some of the chairs are kept empty so as not to overcrowd the virtual classroom. In addition, the virtual classroom includes other objects, which exist in real classrooms such as board, screen, cupboard, clock, and windows. The lecture content is visualized on the white screen. Additionally, the virtual instructor walks around the podium, replicating behavior similar to that of a real instructor. Figures A.6 (a), (b), (c), and (d) show the overall design, hand-raising peer-learners, realistic-styled peer-learners, and cartoon-styled peer-learners, respectively.

(a) Overall virtual classroom design.

(b) Hand-raising cartoon-styled peer-learners from back.

(c) Realistic-styled peer-learners.

(d) Hand-raising cartoon-styled peer-learners.

Figure A.6: Views from the virtual classroom.

The content of the virtual lecture is about computational thinking [245] and the lecture takes ≈ 15 minutes in total, including 4 phases. These four phases are grouped as “Introduction to

[9] https://www.unrealengine.com/

the topic", "Knowledge input", "Exercises", and "Summary" and take ≈ 3, ≈ 4.5, ≈ 5.5, and ≈ 1.5 minutes, respectively. The topic of the virtual lecture is visible on the board as "Understanding how computers think". The first phase starts with the virtual instructor entering the classroom. After staying for a while, the instructor leaves the classroom for about 20 seconds. During this time, participants have the opportunity to explore the classroom, look around, and acclimate themselves with the virtual environment. During the initial phase of the lecture, the instructor asks five questions, and some of the virtual peer-learners raise their hands to interact. In the second phase, the instructor describes two terms, "sequence" and "loop", and shows these terms on the white screen. After the descriptions, the instructor asks four questions about each term and some of the peer-learners raise their hands to answer them. In the third phase, the instructor assigns two exercises and allows students some time to think about them. Later, choices for each exercise are provided by the instructor and, this time, peer-learners raise their hands to vote on the correct answer out of the presented options. In the fourth phase, the instructor summarizes the lecture without asking any questions, which means that peer-learners do not raise their hands. In addition, no hand-raise is expected from the participants as hand poses are not measured during the experiments.

Our study is conceptualized in a between-subjects design. We evaluated three design factors, namely sitting positions of the participants, visualization styles of virtual avatars, and hand-raising percentages of virtual peer-learners. Participants were seated either in the front or back rows, which means that the participants seated in the front had one row in front of them, whereas participants seated in the back had three rows between them and the screen. Both conditions were aligned in the aisle side of the desks that were on the right side of the classroom. This manipulation can give insights about students' attention during a lecture, when they have either the overview over whole class and see most of their virtual peer-learners or when they are positioned closer to instructor and screen the lecture is presented on. Participants encountered either cartoon- or realistic-styled virtual avatars in the environment, including the virtual instructor and peer-learners. The cartoon-styled avatars have larger heads and tinier arms and legs as compared to the realistic-styled avatars. Since the animation and design of more realistic looking avatars is time and cost expensive, it should be interesting to investigate the impact of such manipulation. In addition, various hand-raising percentages of virtual peer-learners consist of four levels, namely 20%, 35%, 65%, and 80%. This means that when a question is asked during the lecture by the virtual instructor, a corresponding percentage of virtual peer-learners raise their hands to answer the question. The last two manipulations are of particular interest, regarding the question how social avatars should be designed in a virtual classroom and how they are perceived by students. Under which condition do students use social information and how does visualization and certain behaviour influence students attention. This helps to simulate and evaluate social dynamics and engagement during the virtual lecture using visual attention. In total, our 2 (factor 1) ×2 (factor 2) ×4 (factor 3) between-subjects design leads to 16 treatment groups.

Procedure

In the beginning of the experiment, the assistants introduced the experiment and its process to the participants. Participants had the opportunity to familiarize themselves with the hardware and the VR environment. Afterwards, the actual experiment and data collection began. Firstly, the eye tracker was calibrated. Then, the experiment was started with assistants pressing a start button. At the end of the virtual lecture, the participants were told to take the HMD off by a message which was displayed in the virtual environment. Virtual lectures were carried out without any breaks. After watching the virtual lecture, participants filled out questionnaires about their perceived realism and experienced presence which were conceptualized for the VR classroom according to [16, 15].

Each session took ≈ 45 minutes in total. The experiments were carried out in groups of ten participants who were randomly allocated to one of the 16 treatment groups by using a random number generator to ensure the random distribution of conditions within groups. To maintain natural behavior, participants selected the physical seat in the experiment room freely without being informed about experimental conditions. Although research assistants helped with technical issues regarding the use of the HMD, participants were blinded to the true purpose and design of the study, as it was solely introduced as a learning experience.

Data Processing and Measurements

During the experiments, head location and pose, gaze, and eye related data along with experimental condition were collected. Head movements are particularly helpful for mapping eye-gaze in the virtual environment. These were saved in data sheets for each participant using anonymous identifiers which ensured the privacy of the participants.

As gaze data reported by the eye tracker can be affected negatively by blinks or noisy sensor measurements, we applied a linear interpolation on the gaze vectors to clean the data. Afterwards, using head pose and interpolated gaze data, we applied ray-casting [168] to map the gaze into the 3D virtual environment. The objects in the 3D environment are surrounded by dedicated colliders; therefore, we were able to calculate 3D gaze points and gazed objects using the procedure visualized in Figure A.7.

However, gazed objects may not directly represent visual attention as participants can gaze on some objects unconsciously for a very short time. To overcome this issue, we set an attention threshold of 200 ms, meaning that we count the objects as object-of-interest if participants stay with their gaze on the objects for at least the amount of the attention threshold. As we assume that both fixations and saccades can occur during attending one object, the selected threshold is larger than classical fixation thresholds applied in eye tracking literature for both conventional [84] or VR eye tracking [85] setups. While we also experimented with various threshold values, our results show similar trends across different thresholds.

In addition to the data related to visual attention, self-reported perceived realism and

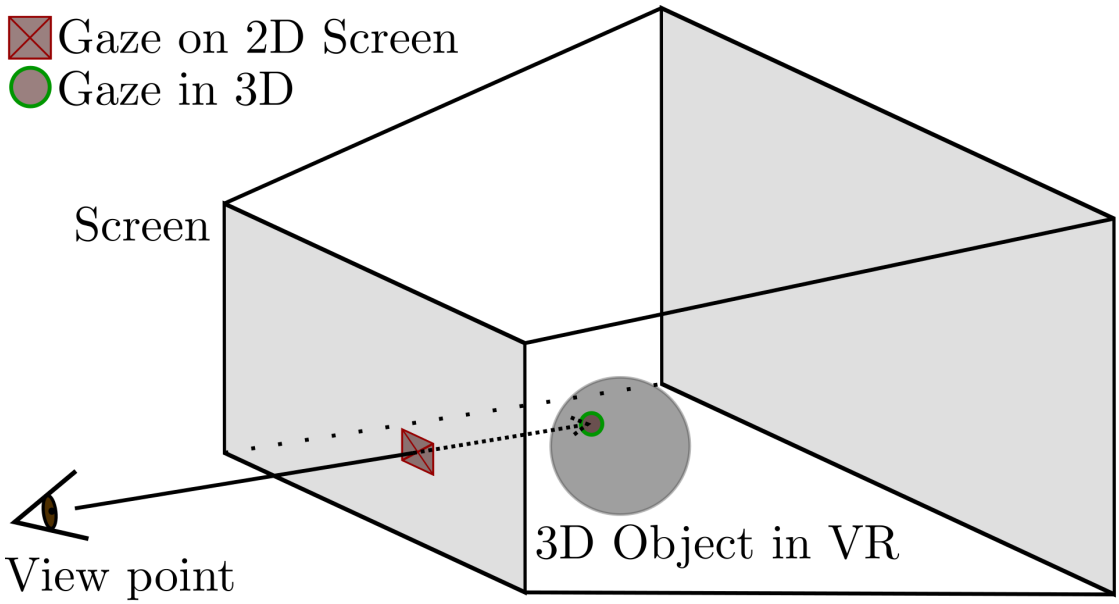


Figure A.7: Ray-casting procedure to obtain 3D gazed object.

experienced presence were obtained at the end of the experiments with 4-point Likert scales ranging from 1 ("completely disagree") to 4 ("completely agree") with 6 (e.g., "I felt like the teacher and the classmates could be real people") and 9 (e.g., "During the virtual lecture, I almost forgot that I was wearing the VR glasses") items, respectively.

In this study, we focused on three main objects in the virtual classroom, namely peer-learners, virtual instructor, and screen, when we extracted object-of-interest information. We decided that these objects may have a significant impact on social dynamics in the classrooms and for overall course of lecture. In our analyses, the attention time on each peer-learner is aggregated and the object of "peer-learners" represents the aggregated object and related attention. In addition, in our classroom setup there is one board and one white screen behind the instructor as depicted in Figure A.6 (a). The lecture content is provided on the white screen only; therefore, in our analysis we refer to the white screen when mentioning screen object.

Research Hypotheses

Our hypotheses correspond to the experimental factors of sitting positions, avatar visualization styles, and various hand-raise percentages of virtual peer-learners, respectively. Furthermore, since we analyze behaviors towards three different objects in the virtual classroom, namely peer-learners, instructor, and screen, for simplicity we call attention to attending these objects-of-interests for the rest of the paper.

Visual Attention in Different Sitting Positions (H1)
We expect that participants seated in the front condition have less attention on peer-learners,

naturally because they do not have as many peer-learners sitting in front of them as opposed to the participants sitting in the back. In addition, the participants that are located in the front are closer to the virtual instructor and the screen that visualizes lecture content. Due to the proximity and having fewer moving and occluding objects in their field of view (FOV), we hypothesize that these participants have more attention time on both virtual instructor and screen than the participants sit in the back.

Visual Attention in Different Visualization Styles of Virtual Avatars (H2)
We hypothesize that attention time on peer-learners in the cartoon-styled visualization is longer than in the realistic-styled visualization as cartoon-styled peer-learners are more exciting for participants when ages of our interest group are taken into consideration. In addition, we assume that participants look at the realistic-styled instructor for longer than at cartoon-styled instructor as participants may consider the realistically rendered instructor more credible in a learning environment. Lastly, we do not expect any differences in terms of attention towards virtual screen that lecture content is visualized, as the visualization style of the screen does not change.

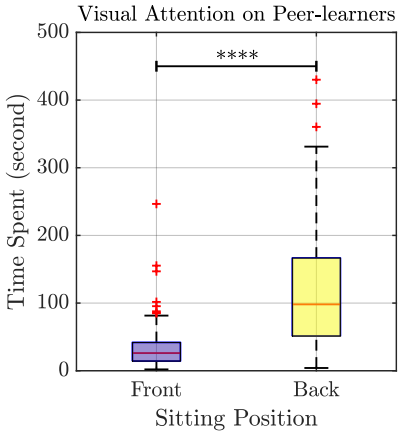
Visual Attention in Different Hand-raising Behaviors of Peer-learners (H3)
We hypothesize that attention time on peer-learners increases with a higher number of virtual peer-learners raising their hands when questions are asked, as this would create a visually more dynamic classroom. Additionally, we expect that if fewer virtual peer-learners raise their hands, this will lead participants to keep their attention either on the instructor or the lecture screen due to having less amount of visual distractors when questions are provided by the virtual instructor.

A.2.5 Results

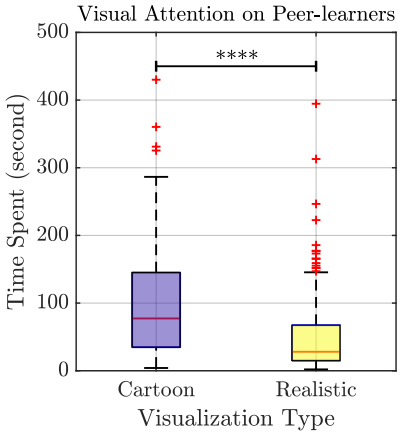
In this section, we analyze the total amount of time spent on each object-of-interest (OOI), which we call visual attention, between different conditions. For each OOI, we applied a 3-way full factorial ANOVA for statistical comparison using alpha level of 0.05. For non-parametric analysis, we transformed the data using the aligned rank transform (ART) [167] before applying ANOVAs. For the pairwise comparisons, we used Tukey-Kramer post-hoc test as the sample sizes were not equal. While the main focus of this work is to assess visual attention using OOI information, here we report experienced presence and perceived realism questionnaires to support our main results. We obtained mean values of 2.91 for experienced presence and perceived realism with $SD = 0.55$ and $SD = 0.57$, respectively, without any significant differences between conditions.

Visual Attention on Peer-learners

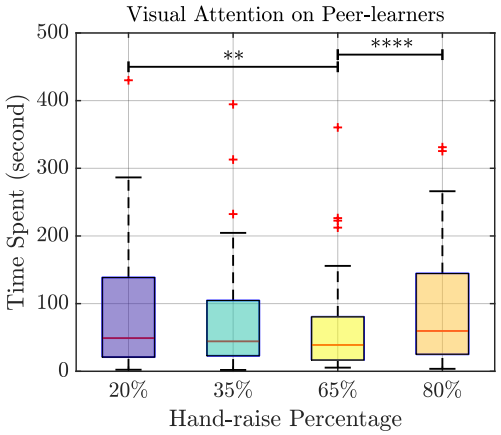
Total time spent on peer-learners for different sitting positions, avatar visualization styles, and various hand-raising behaviors are depicted in Figures A.8 (a), (b), and (c), respectively. Total time spent on peer-learners is significantly longer in the back seated condition ($M = 115.07$ sec, $SD = 85.28$ sec) than it is in the front seated condition ($M = 33.59$ sec, $SD = 32.45$ sec) with ($F(1,264) = 156.23$, $p < .0001$, $\eta^2 = .36$).



(a) Comparison between sitting positions.



(b) Comparison between visualization types.



(c) Comparison between hand-raising behaviors.

Figure A.8: Attention towards virtual peer-learners for different classroom manipulation configurations. $**$ and $****$ correspond to the significance levels of $p < .01$ and $p < .0001$, respectively.

Attention towards peer-learners as different visualization styled avatars differs significantly. Cartoon-styled peer-learners ($M = 98.67$ sec, $SD = 82.79$ sec) drew significantly more attention than the realistic-styled peer-learners ($M = 55.28$ sec, $SD = 65.65$ sec) with ($F(1,264) = 54.13$, $p < .0001$, $\eta^2 = .17$).

Furthermore, for different hand-raising manipulations, attention time on the peer-learners differs significantly with ($F(3,264) = 6.93$, $p < .001$, $\eta^2 = .07$). Particularly, the total time spent on peer-learners in the 80% condition ($M = 88.95$ sec, $SD = 78.15$ sec) is significantly longer than in the 65% condition ($M = 59.23$ sec, $SD = 65.19$ sec) with ($F(3,264) = 6.93$, $p < .0001$, $\eta^2 = .07$). In addition, the total time spent in the 20% condition ($M = 88.62$ sec, $SD = 87.53$ sec) is significantly longer than in the 65% condition ($M = 59.23$ sec, $SD = 65.19$ sec) with ($F(3,264) = 6.93$, $p = .005$). In summary, attention time towards extreme levels of hand-raising percentages are longer than for intermediate levels.

Additionally, we found some significant interaction effects regarding the attention time on the peer-learners. The time on peer-learners in the hand-raising condition depends on the sitting position of the students with ($F(3,264) = 3.88$, $p = .0097$, $\eta^2 = .041$), as well as the attention time on peer-learners in the avatar visualization styles condition depends on the sitting position with ($F(1,264) = 11.37$, $p < .001$, $\eta^2 = .039$) and vice versa. A small interaction effect was found between the hand-raising condition and the avatar visualization styles with ($F(3,264) = 3.36$, $p = .02$, $\eta^2 = .036$).

Visual Attention on Instructor

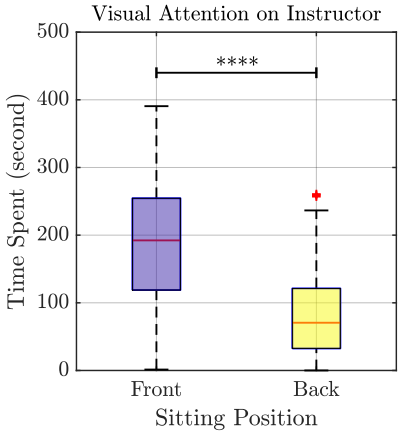
Total time spent on instructor for different sitting positions, avatar visualization styles, and various hand-raising behaviors are depicted in Figures A.9 (a), (b), and (c), respectively. The participants that are seated in the front ($M = 190.07$ sec, $SD = 93.13$ sec) attended to the virtual instructor significantly more than the participants seated in the back ($M = 80.37$ sec, $SD = 60.78$ sec) with ($F(1,264) = 144.16$ $p < .0001$, $\eta^2 = .34$).

The virtual instructor drew significantly more attention in the realistic-styled avatar condition ($M = 145.98$ sec, $SD = 96.63$ sec) than in the cartoon-styled avatar condition ($M = 114.82$ sec, $SD = 89.83$ sec) with ($F(1,264) = 11.81$, $p < .001$, $\eta^2 = .04$).

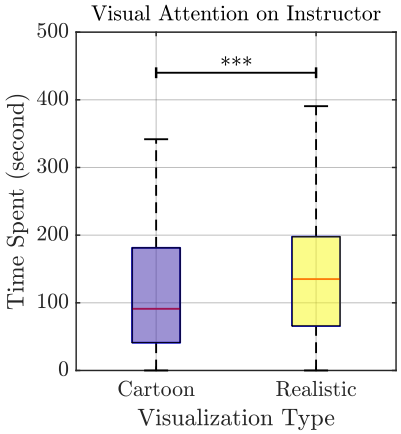
Furthermore, attention time on the instructor is found to differ significantly between different hand-raising behaviors of the peer-learners with ($F(3,264) = 3.54$, $p = .015$, $\eta^2 = .04$). In particular, the total time spent on virtual instructor in the 65% condition ($M = 152.46$ sec, $SD = 91.48$ sec) is significantly longer than the 80% condition ($M = 117.39$ sec, $SD = 91.12$ sec) with ($F(3,264) = 3.54$, $p = .009$, $\eta^2 = .04$). Overall, more attention is drawn by the virtual instructor in the intermediate levels of hand-raising than the extreme levels. There were no interaction effects found for attention time on instructor.

Visual Attention on Screen

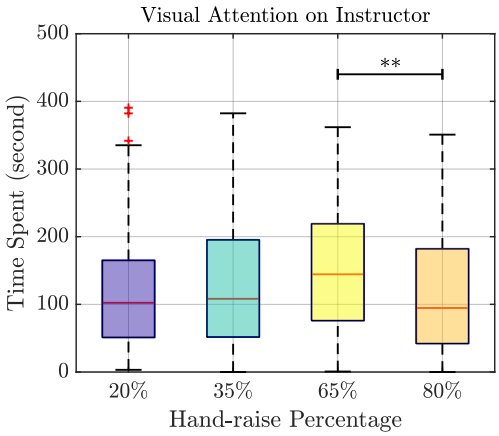
Total time spent on the screen, where the lecture content visualized for different sitting positions, avatar visualization styles, and various hand-raising behaviors are depicted in Figures A.10 (a), (b), and (c), respectively. The participants that are seated in the front ($M = 218.65$ sec, $SD = 78.70$ sec) attended to the lecture screen for a significantly longer period of



(a) Comparison between sitting positions.



(b) Comparison between visualization types.



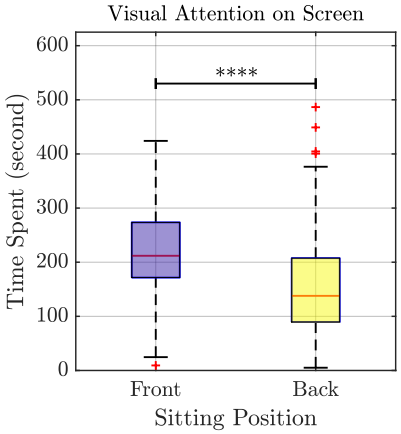
(c) Comparison between hand-raising behaviors.

Figure A.9: Attention towards virtual instructor for different classroom manipulation configurations. $**$, $***$, and $****$ correspond to the significance levels of $p < .01$, $p < .001$, and $p < .0001$, respectively.

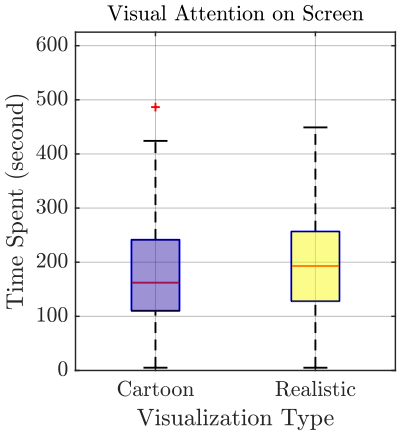
time than the back seated participants ($M = 154.21$ sec, $SD = 96.88$ sec) with ($F(1,264) = 42.5$, $p < .0001$, $\eta^2 = .14$).

We did not find significant effects on screen attention between cartoon- and realistic-styled avatar conditions ($F(1,264) = 1.9$, $p = .17$, $\eta^2 < .01$); however, attention time in realistic style ($M = 193.35$ sec, $SD = 92.30$ sec) was slightly longer than cartoon style ($M = 173.95$ sec, $SD = 96.11$ sec).

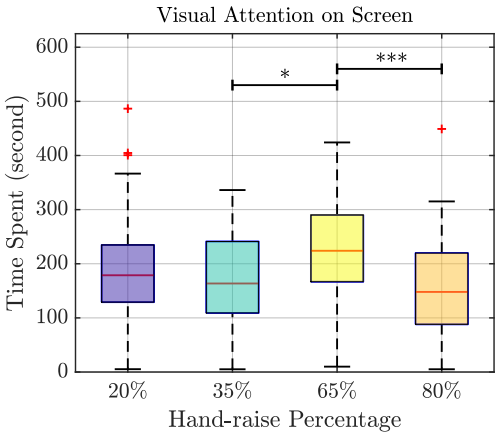
In addition, the total attention time on the screen is found to differ significantly between different hand-raising conditions with ($F(3,264) = 5.74$, $p < .001$, $\eta^2 = .06$). In particular, attention time on screen is longer in the 65% hand-raising condition ($M = 222.03$ sec, $SD = 94.90$ sec) than in the 80% condition ($M = 156.06$ sec, $SD = 88.25$ sec) with ($F(3,264) = 5.74$,



(a) Comparison between sitting positions.



(b) Comparison between visualization types.



(c) Comparison between hand-raising behaviors.

Figure A.10: Attention towards screen for different classroom manipulation configurations. $*$, $***$, and $****$ correspond to the significance levels of $p < .05$, $p < .001$, and $p < .0001$, respectively.

$p < .001$, $\eta^2 = .06$). In addition, attention time in the 65% condition is also significantly longer than in the 35% hand-raising condition ($M = 174.87$ sec, $SD = 81.28$ sec) with ($F(3,264) = 5.74$, $p = .025$). The overall trend of attention on the lecture screen is similar to virtual instructor with the intermediate conditions being higher than the extreme conditions. There were no interaction effects found for attention time on screen.

A.2.6 Discussion

We discuss experimental results particularly for social interaction and dynamics in VR classrooms, usability of eye tracking data, and the advantages of such classrooms along with their limitations.

Social Dynamics in VR Classroom

We discuss our findings about social dynamics in the VR classroom in three parts, particularly based on **H1**, **H2**, and **H3** which are related to different sitting positions, different avatar visualization styles, and different hand-raise behaviors of peer-learners, respectively.

In our analyses, we found that the participants seated in the front of the classroom attended less on the peer-learners than the participants in the back, which was expected because they had fewer peers in their FOV, unless they turn back of the classroom. Assuming that during the course of the lecture, participants are supposed to listen and pay attention to the topics told by the instructor, the visual attention we observed is normal. Briefly, this is an indication that participants focus on the lecture content or instructor instead of visually interacting with their peers when seated in the front. Further, as a supporting evidence to aforementioned result, front seated participants had spent significantly more time visually attending the instructor and the screen than the participants seated in the back. We assume that these results are due to being closer to them and having fewer occluding objects in the frontal participants' FOV. These findings confirm our **H1**. Additionally, the results from the interaction effects support this hypothesis. The differences in visual attention on their virtual peer-learners for the avatar visualization style and hand-raising depend on the sitting position. Participants located in the back of the classroom have more peer-learners in their line of sight and therefore recognize the behaviour of the virtual peer-learners more, than participants seated in the front.

Our results indicate that students visually attended for longer on the peer-learners when avatars in the classroom were presented in cartoon styles. Considering the number of peer-learners in the environment and the ages of our participants being between 10-13, we argue that participants may have felt like engaging more with their peer-learners due to the emotional reasons as cartoon-styled peers are more appropriate to their ages. Realistic-styled peer-learners may be too ordinary for student engagement with peers in our setup, which led to less amount of attention. On the contrary, participants visually spent more time on the instructor when realistic-styled avatars were used. We conceive that if the avatar styles are ordinary, then the visual attention shifts to the instructor instead of interacting with the peer-learners. Lastly, as we did not find any statistical difference in attention time on the screen between different avatar visualization styles, we conclude that visual attention on the screen is not affected by such avatar visualization styles. Realism that is provided by the avatar styles may introduce additional computational complexity as such visualizations can be computationally expensive or can require additional effort to implement in advance. If the interaction with peer-learners is the main focus of the lecture, then practitioners can opt for cartoon-styled avatars. This also decreases the effort of generating the avatars. Overall, these findings confirm our **H2**.

In the analysis on different hand-raising behaviors of the peer-learners, we found mixed effects. In the attention time towards peer-learners, we found a clear evidence that attention time in the extreme hand-raising conditions, namely when 80% or 20% of the virtual peer-

learners raise their hands after the questions were asked by the virtual instructor is longer than in the intermediate conditions (35% and 65%). The extreme conditions may represent either more or less capable groups of peer-learners in the learning environment and participants may have a higher self-concept when surrounded by a less capable group and the other way around, which is related to the Big-fish-little-pond effect [259]. Having reasonably higher attention on peer-learners on these conditions also indicates that VR can present an opportunity to create digital environments to further study students' self-concept. On the other hand, intermediate hand-raising conditions may help students to focus more on learning related objects in the classroom instead of peer-learners such as lecture content or instructor as experimentally indicated. However, we expected an approximately linear increase in terms of attention time towards higher hand-raising conditions in the attention time on peer-learners. While we obtained an expected result between the 65% and 80% hand-raising conditions, the results regarding the 20% hand-raising condition do not support our hypothesis **H3**. This might be due to a moment of surprise when only a handful of peer-learners raises their hands indicating that few number of peer-learners know the answers of the questions. Furthermore, we found that attention time on the instructor tended to be longer in the intermediate levels of hand-raising than in the extreme conditions. Statistically significant results are only found for the difference between the 65% and 80% condition. While a decreasing linear trend towards the higher hand-raising percentages exists between the 65% and 80% for attention on the instructor, the overall trend is against our hypothesis, even though they are aligned with the attention time on peer-learners. Lastly, the experimental results on attention time on the screen is similar as compared to the attention time on the instructor. However, the 35% hand-raising condition drew significantly less attention than the 65% condition, which does not support our hypothesis. Overall, while some of our expectations are verified, **H3** is not confirmed. Still, the resulting behaviors should be further investigated with regard to effects on students' self-concepts during VR learning and considered when creating a classroom students are habituated to.

In summary, the three different manipulations that we studied have important effects on students' visual behavior in immersive VR classrooms in terms of social dynamics. For instance, in practice, students' self-concept can be affected by consistent hand-raising behaviors of virtual avatars over the time. While this may be less problematic in real classrooms as peer students may have different capabilities in different themes, it should carefully considered in the virtual setting, because we could present always the same behavior of the peer-learners. An adaptive strategy for hand-raising behaviors of the virtual peer-learners may be considered in practice. In addition, seating the students in the front along with realistic-styled avatars may help to increase visual attention on the lecture content. However, if a more interactive classroom environment is focused on visual interaction, practitioners can either seat students in locations where they can see their peer-learners clearly or design VR classrooms differently in terms of seating plans.

Usability of Eye Tracking Data

As eye tracking data is considered a noisy data source, we discuss our insights into the usability of this data, for particularly the immersive VR classroom setups. As aforementioned, we defined the visual attention on the different objects by using an attention threshold, which was 200 ms. In the end, in almost all conditions, the total amount of time that was spent on only the three types of objects was in the vicinity of half of the complete experiment duration despite having a relatively higher attention threshold value compared to fixation detection algorithms in the eye tracking literature. Such amount of total attention time on these three objects empirically validates our assumption of independence between them as well. We removed a significant number of samples from eye movement data due to sensory issues (e.g., lower eye tracking ratio) in order to obtain high-quality data and accurate attention mapping on the objects in the virtual classroom. While this may not be necessary for larger objects such as virtual screen in the classroom, it might cause mapping the attention wrongly for the smaller objects such as virtual avatars if the data quality is low. Considering that the participants were children in our experiments and they did not have experience with virtual reality and eye tracking, number of data removals due to such issues would be more than the experiments that are carried out with adults. In addition, unlike pre- or post-tests, eye tracking allows researchers to analyze time-dependent and temporal visual behavior changes, which can help assess students' states during virtual lectures and adapt to the environment accordingly. Therefore, despite the drawbacks, we suggest using eye movement data in such classrooms as long as an accurate calibration is applied in advance. A further iteration could take relationship of eye movement-based visual attention into consideration or analyze perceived relevance of lecture content along with eye-gaze behaviors such as in [260] and [261], respectively.

Advantages and Limitations

One of the advantages of immersive VR classroom setups is the opportunity of simulating different classroom manipulations in remote settings, which are difficult to do in real world, and evaluate students' behaviors and learning under such manipulations. Another advantage of such setups is the possibility of preserving the privacy of students since the videos that include faces are not recorded in such settings. In real world classrooms, it is troublesome to record and store videos of the class while lecturing, even though there are some efforts supporting the automated anonymization [262] of such data. In contrast, data collected from virtual classrooms can be pseudo-anonymized. However, one should be aware of the amount of personal information that can be extracted from eye movement data and how to manipulate it [176, 263, 264]. Furthermore, one should take the relationship between iris texture and biometrics into account and how to preserve privacy in case eye videos are recorded and stored [150]. In addition, we observed during experiments that some of the students intended to raise their hands when seeing the hand-raising behaviors of the virtual peer-learners. While we did not record hand tracking data in our study, it is possible to accurately assess the intentions of students towards questions asked by the virtual instructor by using a hand tracker

device on the HMD, which is another advantage of VR setups compared to real classrooms. Although, hand-raising is a good indicator of children's participation during a lecture, we do not know if students interpret this behaviour of their virtual peers as a sign of competence, engagement, or motivation.

Despite the advantages, there are other technical limitations regarding the use of VR classrooms. Long periods of exposure to VR lectures can lead to immense levels of cybersickness. In addition, a vast amount of HMD movement on the head may cause a drift in eye tracker calibration, leading to incorrect sensor readings. This can affect interaction experience if gaze-aware features are included in virtual environments. These should be taken into consideration when designing a virtual classroom and lecture. Particularly, the duration of the lecture should be chosen carefully to minimize these effects.

A.2.7 Conclusion

To understand the visual attention in VR classrooms in different manipulations, we analyzed object-of-interest information based on eye-gaze. We found that participants seated in the front attended more time to the virtual instructor and the screen displaying lecture content. In addition, participants focused on the cartoon-styled peer-learners more than realistic-styled ones, whereas in the realistic-styled avatar manipulation the virtual instructor drew more visual attention. The extreme conditions of hand-raising behaviors drew more attention towards virtual peer-learners, whereas in the intermediate conditions visual attention was focused more on the instructor and screen. These findings are based on the eye movements of the participants and correspond to the social dynamics of VR classrooms such as students' self-concept or peer-learner interaction; however, such manipulations may also affect learning outcomes. While our results provide primitive but fundamental cues about how to design immersive VR classrooms by taking students' visual behaviors into account for different goals in digital teaching, effects of such manipulations on the learning outcome should be further investigated.

As future work, we plan to specifically investigate the relationship between different manipulations with temporal gaze dynamics as an immediate response to asked questions and related students' performances.

Acknowledgments

This research was partly supported by a grant to Richard Göllner funded by the Ministry of Science, Research and the Arts of the state of Baden-Württemberg and the University of Tübingen as part of the Promotion Program of Junior Researchers. Lisa Hasenbein and Philipp Stark are doctoral candidates and supported by the LEAD Graduate School & Research Network, which is funded by the Ministry of Science, Research and the Arts of the state of Baden-Württemberg within the framework of the sustainability funding for the projects of the

Excellence Initiative II. Authors thank Stephan Soller, Sandra Hahn, and Sophie Fink from the Hochschule der Medien Stuttgart for their work and support related to the immersive virtual reality classroom used in this study.

A.3 Assessment of Driver Attention during a Safety Critical Situation in VR to Generate VR-based Training

A.3.1 Abstract

Crashes involving pedestrians on urban roads can be fatal. In order to prevent such crashes and provide safer driving experience, adaptive pedestrian warning cues can help to detect risky pedestrians. However, it is difficult to test such systems in the wild, and train drivers using these systems in safety critical situations. This work investigates whether low-cost virtual reality (VR) setups, along with gaze-aware warning cues, could be used for driver training by analyzing driver attention during an unexpected pedestrian crossing on an urban road. Our analyses show significant differences in distances to crossing pedestrians, pupil diameters, and driver accelerator inputs when the warning cues were provided. Overall, there is a strong indication that VR and Head-Mounted-Displays (HMDs) could be used for generating attention increasing driver training packages for safety critical situations.

A.3.2 Introduction

Having safe driving experiences and decreasing the number of crashes are two of the most important issues when it comes to driving safety. Every year, many fatal crashes occur on roads all over the world. According to the Road Safety Annual Report in International Transport Forum 2018, most of the fatal crashes occurred on rural roads; however, the number of fatal crashes in urban roads has been increasing in more than half of the countries since 2000 [265].

Apart from road or weather conditions, distracted driving can cause fatal crashes. While a total prevention is almost impossible, many crashes can be prevented by training drivers using driver assistant systems. With recent developments in the field of augmented reality (AR) and head-up display (HUD) technology, new means have become available to overlay different warnings to the driver, such as pedestrian warnings or road signs. In fact, many modern cars already employ this technology to a certain degree. The majority of studies that concentrated on driver training and the interaction between these technologies and drivers in safety critical situations used driving simulators. With the recent developments in VR and HMDs, it is possible to apply these scenarios and trainings in VR with lower cost. However, it is an open question whether VR and HMDs can be used in studying driver training and interaction for safety critical situations.

In order to assess whether VR, HMDs, and gaze-aware cues can be useful and driver attention can be increased properly in this context, we focused on an unexpected pedestrian crossing behavior at non-designated crosswalks on urban roads when the Time-to-Collision (TTC) between the driving vehicle and crossing pedestrian is very short ($\approx$ 1.8-5 seconds). [170] mentioned that in this range of TTC, there is a high likelihood that joint attention between crossing pedestrian and driver happens. However, in case it does not happen, due to

distracted pedestrian or driver, it is more likely that a crash will happen. In our experiments, control group did not receive any critical pedestrian warning cues, whereas the experimental group had the gaze-aware critical pedestrian warning cues. By analyzing closest distances between driving vehicles and crossing pedestrians, pupil diameter changes of drivers between baseline and risky driving timeframes, and driver performance measurements, we found that there is a strong indication that gaze-aware visual warnings for critical pedestrians help increasing the driver attention earlier in VR. Therefore, low-cost VR setups along with realistic and gaze-aware warnings can be introduced to train drivers for safety critical scenarios. Major contributions of our work are as follows: **(a)** Demonstrating a very critical scenario in terms of collision risk between driver and pedestrian with and without risky pedestrian warning cues in VR and **(b)** Evaluation of gaze-aware critical pedestrian warning cues in VR whether they increase driver attention earlier so that attention increasing VR-based training packages can be proposed and further evaluated. Since the dedicated driving scenario is highly dynamic and time-critical, the outcome of the current study can be taken as a basis for any study that includes time-dependent and safety-critical scenarios in VR.

A.3.3 Related Work

Driving simulation studies have been conducted in various domains. Two of the most common issues addressed were safety and driver assistance. [266] introduced a novel interface for HUD over head-down display (HDD). HUDs and AR cues have been used for various purposes. [267] discussed that specificity of visual warnings provided advantages in gaze, brake reaction times, passing speeds, and collision rates. [268] showed the benefits of HUDs while turning left, whereas [269] presented positive effects of AR cues in terms of time-to-contact and gap response variation to assist elderly drivers during left-turns. In addition, [270] showed the navigational AR aid for recognizing turn locations earlier via 3D volumetric HUD. [271] discussed that adaptive support in HUD for lane keeping helped drivers drive more centrally and with less lateral variation. The effect of in-car AR system for reducing collisions caused by other vehicles' movements was presented by [272]. Additionally, increase in situational awareness using AR in automated driving for take-over scenarios was studied by [273] and [274], whereas classification of drivers' take-over readiness was studied by [194].

While numerous studies can be counted in the context of driver assistance, the studies include pedestrian safety, hazard anticipation, and driver training are more relevant to our work. [275] showed that AR cueing increased the response rate for pedestrian and warning sign detection in directing driving attention to roadside hazards. The study of [276] in a driving simulator with a maximum speed of about 30km/h showed that gaze guidance reduced number of pedestrian collisions. [277] studied three driver awareness levels of a pedestrian in a driving simulator: Perception, vigilance, and anticipation. They showed that AR cues were capable of enhancing the driver awareness in all levels. The outdoor study conducted by [278] showed that AR pedestrian warnings provided positive results on measures such as braking, distances to pedestrians, and gaze-on pedestrian travel distances. The study of [279]

on eye movements showed that the scanning patterns of novice drivers reflected their failure to recognize potential risks. Driving simulator studies have been used in driver training and VR as well. [280] found out that drivers who were trained in a simulator improved their driving skills in turning into correct lane and proper signal use. Furthermore, [281] evaluated hazard anticipation and found that trained drivers recognized the risks more often. [116] showed the effect of improvement of bad driving habits via synthesizing personalized training programs in VR. [282] assessed drivers' hazard anticipation across VR and driving simulators to evaluate the usage of VR headsets and justified that VR headsets could be used for measuring driving performance. [283] studied personality traits on sacrifice decisions including pedestrians during VR-based driving. While the studies which include driving simulators and hazardous situations showed great potential for driver training, it is an open question whether visual cues for critical situations in VR can increase driver attention properly, so that VR-based training packages can be proposed and synthesized for safety critical situations.

A.3.4 Experiment

We focused on driver behavior in a very critical scenario when pedestrians tried to cross the road with TTC was between ≈ 1.8-5 seconds in VR. In this range of TTC, there is a high likelihood that pedestrian or joint attention occurs [170]. However, if it does not occur, the outcome can be fatal. Our experiment included a control group that did not receive any cues, and an experimental group that received gaze-aware critical pedestrian cues. Our major hypothesis is that if the gaze-aware warning cues can successfully increase the driver attention earlier in the safety critical situation in VR, similar low-cost VR setups along with adaptive warnings could be proposed for driver training for these situations.

Participants

16 volunteer participants (4 female, 12 male) whose ages range from 25 to 50 ($M \approx 31$) and driving experiences range from 5 to 30 years ($M = 12$) participated in the experiment. Participants were separated into two groups. A control group, receiving no critical pedestrian warning cues, and an experimental group, receiving the warning cues.

Apparatus

HTC-Vive along with Pupil-Labs Binocular Add-on [21], which has binocular 120hz eye tracking cameras and clip-on rings, Logitech G27 Steering Wheel and Pedals, and Phillips headphones were used to create driving setup. Eye tracking was measured using the open-source hmd-eyes of Pupil-Labs with Pupil Service version 1.7. Virtual city was created using Unity3D game engine. For the environment, vehicles, and pedestrians, we purchased and used models from Urban City Pack, City Park Exterior Props, Traffic Sign Set, Modern People, Traffic Cars, Realistic Car HD 02, Realistic Car Controller, Simple Waypoint System, and Playmaker

asset packages. We designed the main roads long and straight so that the drivers would have opportunity to speed up as they want and drive naturally. Example scenes from our virtual environment are shown in Figure A.11.

(a) Cockpit of driving car.

(b) Main road.

Figure A.11: Example scenes from VR environment.

The dedicated setups were run on a PC equipped with an NVIDIA Titan X graphics card with 12GB memory, a 3.4GHz Intel i7-6700 processor, and 16GB of RAM.

Since the visual warning cues for experimental group are very important in our setup, Figure A.12 shows a pedestrian model with and without warning cues.

(a) With cue.

(b) Without cue.

Figure A.12: Pedestrian with and without warning cues.

Procedure

In the beginning of the experiment, participants were informed about the purpose and scope of the experiment orally. They had the opportunity to stop and cancel the experiment anytime. At the end of the experiment, participants filled a small questionnaire about demographic and qualitative information. The experiment consisted of two phases. For both phases, participants were given written instructions before starting. In the first phase, participants acclimated the setup. This phase did not include any pedestrians or dynamic objects apart from the driver's car; no data were collected during this phase. Generally, this phase lasted in 5-10 minutes, although if participants had not felt comfortable, they could have continued

driving. Once they felt comfortable with the setup, they continued to the second phase.

In the beginning of the second phase, 2D calibration with 16 points using hmd-eyes of Pupil-Labs was performed. After calibration success, participants started the experiment. The starting location of the driving vehicle was in the beginning of the main road, where a critical pedestrian crossing happened. Since there was no intersection until the end of this road, all of the participants were required to drive until the end. At the end of the road, they could have turned left or right and continued driving, however our data analyses did not concentrate on the data acquired after the turn, since they could have encountered with different scenarios. The speed limit of the driven road was 90km/h, and participants were supposed to realize this by traffic signs. The driving vehicle was also equipped with maximum speed warning.

The critically crossing pedestrian scenario was as follows. At the beginning of each run, two occurrences of a critical pedestrian were generated along with other non-critical pedestrians on the side walks. The critically crossing pedestrian was determined at random, as active and proceeded to dangerously cross the street before the driving vehicle. Both of these occurrences had dedicated gaze-aware warning cues. Pedestrians were not located in the beginning of the road, so that the drivers had the opportunity to speed-up or slow-down until the crossing. Pedestrian warnings were activated for the experimental group when the distance between front of the driving vehicle and critical pedestrians became $\approx 77m$. The crossing pedestrian started crossing the road from the right side when the distance between vehicle and pedestrian was $d_{critical} \approx 45m$. We assumed that drivers would obey the speed limit (90km/h) and also drive faster than 30km/h. This way, parameter of $d_{critical}$ helps to map expected TTC to $\approx 1.8s \leq TTC \leq 5s$ interval. Ray-casting [168] method was used to map gaze signal, which was obtained from Pupil-Labs software, from 2D canvas to 3D environment by the help of Unity3D colliders [284] that were attached to virtual objects. Once the drivers' gaze signal in 3D environment was closer than 5 meters to the pedestrians for ≈ 0.85 seconds, the cues were deactivated. Therefore, the cues became gaze-aware. Since the control group did not receive cues, the timeframes consisted of different milestones for each group. t_w and t_m correspond to start of the critical pedestrian warning and start of the pedestrian movement respectively. For the control group, baseline driving corresponds to $[t_m - \delta t, t_m]$, whereas for the experimental group, it is $[t_w - \delta t, t_w]$. $[t_m, t_m + \delta t]$ is the risky driving timeframe for both groups. Setting different values of δt means changing the durations of the timeframes.

Measurements

The metrics analyzed were the closest distances between the crossing pedestrians and the driving vehicles, driver performance measurements including inputs on accelerator and brake pedals, and pupil diameter changes between baseline and risky driving timeframes. Particularly, since the critically crossing pedestrian is only safety critical for the driving vehicle inside of the driven lane, we took the closest distance in this lane. Driver inputs on pedals are also indicators of perception and reflect the smoothness of the driving experience as well. Lastly, pupil enlargement corresponds to increase in cognitive load [241]. Pupil diameter

values were fetched from Pupil-Labs software in pixel units. For smoothing and normalization, we applied Savitzky-Golay filter [165], and divisive baseline correction using baseline duration of 0.5 seconds and median [166].

Hypotheses

Our hypotheses are based on the driver attention and actions. Since the experimental group was provided with the risky pedestrian cues, we expected that the closest distances between the crossing pedestrians and the driving vehicles for the experimental group would be more than the control group. In addition, when the visual cues were provided to the drivers, we expected that they would understand the criticality earlier, and their cognitive load would increase earlier. Pupil dilation is one of the indicators of the cognitive load increase, therefore we expected that pupil dilation of the experimental group would happen earlier. Furthermore, cues would affect driver inputs on accelerator and brake pedals, hence it was expected that experimental group drivers would take their foot off the accelerator earlier and perform smoother braking behavior than the control group drivers. In all, we expected that the experimental group would perform safer and smoother driving experience than the control group.

A.3.5 Results

Analyses for the distances, driver performance measurements and pupil diameters during baseline driving and risky driving timeframes were calculated using MATLAB and are as follows.

Closest Distance to Crossing Pedestrian

We measured the closest distances between the crossing pedestrians and driving vehicles until pedestrians completed half of their trajectories, since during the second half, the pedestrians were not safety critical to the driving vehicles anymore. Figure A.13 shows the results for this metric. We applied two sample T-test with alpha level of 0.05 and found significant difference between two groups with $p \approx 0.00059$ (Cohen's $d \approx 2.21$). One of the participants in the control group hit the pedestrian, and the experiment was terminated at that moment. In addition, since the velocities of the vehicles were not fixed, the difference in distances could vary. However, the deceleration trend in the experimental group started from t_w, which is a strong indication that they acknowledged the critical situation earlier than the control group and behaved accordingly. Overall, it is clear that the experimental group participants drove safer than the control group participants.

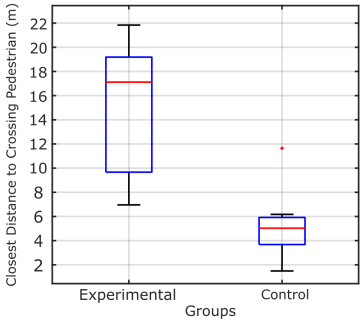
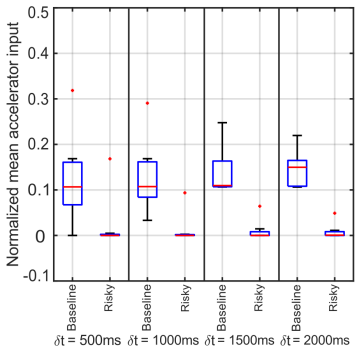


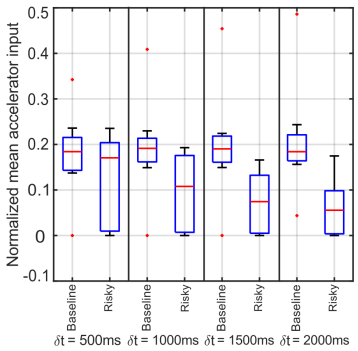
Figure A.13: Closest distance to crossing pedestrian - Experiment group relationship.

Driver Performance Measurements

Driver inputs on accelerator and brake pedals are the two main indicators of safe and smooth driving. Therefore, we analyzed the normalized driver inputs on accelerator and brake during different durations of baseline and risky driving timeframes. First, we applied paired T-test with alpha level of 0.05 between baseline and risky driving timeframes using normalized mean accelerator inputs. As expected, significant differences for experimental group even for very short durations (e.g. $\delta t \approx 50ms$, $p = 0.0158$, Cohen's $d \approx 1.12$) were found. However, significant differences were found for the control group starting from $\delta t \approx 1.4s$ ($p = 0.0495$, Cohen's $d \approx 0.84$). Figure A.14 shows the dedicated analyses. Finding significant differences in shorter δt values means that the drivers acknowledged the critical situation earlier. Therefore, it is a significant indicator that visual pedestrian cues helped drivers drive safely even during a very dangerous situation. Furthermore, we analyzed braking behaviors by analyzing whether participants performed full brake, since the braking happens in very short time. In total, five of the participants in the control group performed full brake, whereas none of the participants in the experimental group did this. This indicates that visual cues also helped to have smoother driving experience.



(a) Accelerator input - Baseline & Risky driving relationship for the experimental group.

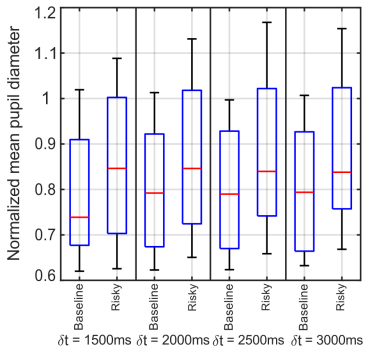


(b) Accelerator input - Baseline & Risky driving relationship for the control group.

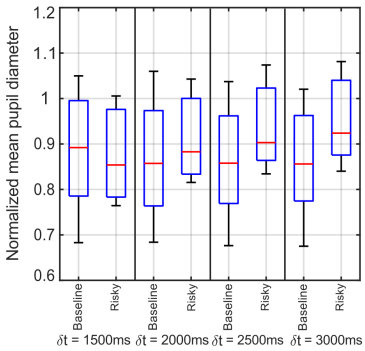
Figure A.14: Accelerator inputs - Driving timeframe relationship.

Pupil Diameter

Since pupil dilation is one of the indicators of cognitive load increase, we analyzed normalized pupil diameters of the drivers in the same way as accelerator inputs between baseline and risky timeframes using paired T-test with alpha level of 0.05. Since HMDs and VR offer controlled illumination, we expected that pupil dilation would happen due to the increase in cognitive load, and pupil diameters of the experimental group would increase earlier than the control group. Analyses showed that significant difference in pupil diameters between baseline and risky timeframes for the experimental group starts from $\delta t \approx 1.4s$ ($p = 0.048$, Cohen's $d \approx 0.85$), whereas it starts from $\delta t \approx 2.4s$ ($p = 0.0489$, Cohen's $d \approx 0.84$) for the control group. Figure A.15 shows the results. Overall, there is a strong indication that cues for the critical pedestrians increased cognitive load of the experimental group earlier so that they behaved accordingly.



(a) Pupil diameter - Baseline & Risky driving relationship for the experimental group.



(b) Pupil diameter - Baseline & Risky driving relationship for the control group.

Figure A.15: Pupil diameter - Driving timeframe relationship.

A.3.6 Conclusion

We introduced a VR driving simulation environment and a safety critical pedestrian crossing to study whether VR setups and gaze-aware cues can increase driver attention in critical situations despite the prevalent disadvantages, such as narrow field-of-view, low resolution or weight of HMDs, so that low-cost VR-based training for safety critical situations can be proposed and further evaluated. To the best of our knowledge, this is the first work that assesses VR setups using gaze-aware cues for safety critical situations in driving by analyzing eye tracking and performance metrics. We found significant differences in the distances to crossing pedestrians, accelerator inputs, and pupil diameters between baseline and risky timeframes. Results indicate that driver attention can be increased earlier with minimalistic gaze-aware cues properly in safety critical situations in VR. Most of the previous work on driving simulation and training were done using physical driving simulators. However, VR setups can decrease cost of implementation and time. Overall, we suggest that driver attention increasing training packages can be introduced in VR. Since many modern cars have different

warnings for safety critical situations, VR could be used to assess these systems and train people to get acclimated with them as well.

As future work, detailed eye-tracking analyses, a study to generate better attention grabbing cues, and a driver training study for critical situations to assess whether drivers improve their bad driving habits by VR-based training can be done.

A.4 Person Independent, Privacy Preserving, and Real Time Assessment of Cognitive Load using Eye Tracking in a Virtual Reality Setup

A.4.1 Abstract

Eye tracking is handled as key enabling technology to VR and AR for multiple reasons, since it not only can help to massively reduce computational costs through gaze-based optimization of graphics and rendering, but also offers a unique opportunity to design gaze-based personalized interfaces and applications. Additionally, the analysis of eye tracking data allows to assess the cognitive load, intentions and actions of the user. In this work, we propose a person-independent, privacy-preserving and gaze-based cognitive load recognition scheme for drivers under critical situations based on previously collected driving data from a driving experiment in VR including a safety critical situation. Based on carefully annotated ground-truth information, we used pupillary information and performance measures (inputs on accelerator, brake, and steering wheel) to train multiple classifiers with the aim of assessing the cognitive load of the driver. Our results show that incorporating eye tracking data into the VR setup allows to predict the cognitive load of the user at a high accuracy above 80%. Beyond the specific setup, the proposed framework can be used in any adaptive and intelligent VR/AR application.

A.4.2 Introduction

Cognitive load is referred to as the amount of information processing activity that is imposed on working memory [285]. Cognitive load recognition is important and beneficial for many applications. It has been studied extensively in various domains, such as in education, psychology, or driving, since information on the cognitive load of an individual can be helpful to design user-adaptive interfaces. Various ways have therefore been proposed to assess the cognitive load of a subject, such as by means of N-back tasks (e.g., Appel et al. [241]), through the analysis of electroencephalography (EEG) signals (e.g., Zarjam et al. [286], Walter et al. [287]), by means of eye movements studies or through assessment of facial expressions (e.g., Hussain et al. [288]). Eye tracking offers a particularly non-invasive way of cognitive load assessment, especially through the measurement and analysis of the pupil diameter.

Meanwhile, eye tracking has also found its way into the driving domain, not only as a means to study driving behavior, but also as a powerful input modality for advanced driver assistance systems (e.g., Kübler et al. [289]) or even as a means of driver observation on context of automated driving (e.g. Braunagel et al. [290, 194]). Modern cars are already capable of tasks such as lane following, traffic sign and light detection, automated parking, and collision warning. However, the full autonomous driving task is still too complex without human input and guidance. For this reason, current cars employ a variety of multi-modal warning systems for many different purposes to ensure driving safety and provide smooth driving experience.

Augmented reality (AR) and head-up-display (HUD) technologies have been used as interfaces to such systems both in practice and driving simulation research. In the following, we will briefly review related work in this context.

Many driving simulation studies have been conducted in driving simulators or virtual reality (VR) environments in order to analyze driving behavior, safety, performance and training using HUDs or virtual warnings. For example, HUDs for blind spot detection and warning were discussed in a related work by Kim et al. [291]. Tran et al. [268] addressed the usage and benefits of HUDs during left turns. Moreover, benefits and improvement of driving behavior for lane keeping using adaptive warnings were discussed by Dijksterhuis et al. [271]. The effect of improving bad driving habits using VR in a user-study was discussed by Lang et al. [116]. In the context of eye tracking and driving, there are several studies with various goals. For example, Konstantopoulos et al. [292] studied eye movements during day, night, and rainy driving in a driving simulator. Braunagel et al. [195] introduced a novel approach for driver activity recognition using head pose and eye tracking data. Furthermore, Braunagel et al. [194] proposed a classification scheme to recognize driver take-over readiness using gaze, traffic, and a secondary task in conditional automated driving. Pomarjanschi et al. [276] showed that gaze guidance reduced the number of pedestrian collisions in a driving simulator.

In the driving context, there are many studies that focus on cognitive load and driving. Engström et al. [293] analyzed the effect of cognitive load on driving performance and found out that the effects of cognitive load on driving are task dependent. Yoshida et al. [294] proposed an approach to classify driver cognitive load to improve in-vehicle information service using real world driving data. Gabaude et al. [295] conducted a study in a driving simulator to understand the relationship between mental effort and driving performance using cardiac activity, driving performance and subjective data measurements. Mizoguchi et al. [296] proposed an approach to identify cognitive load of the driver using inductive logic programming with eye movement and driver input measurements in real driving situations. Fridman et al. [297] proposed a scheme to estimate cognitive load in a 3-class problem in the wild for driving scenarios using convolutional neural networks.

Driving simulation studies for safety critical situations using warnings and cognitive load recognition in driving exist in the literature. However, it is still an open question whether it is possible to recognize cognitive load of the driver in safety-critical situations and especially when the driver is confronted with visual gaze-aware warnings. In order to tackle this issue, we used the data collected using a VR setup from our previous work [212] where drivers encountered a dangerously crossing pedestrian in an urban road. In order to keep the situation safety critical, Time-to-Collision (TTC) between driving vehicle and crossing pedestrian was kept $1.8sec < TTC < 5sec$. Rasouli et al. [170] discussed that in this range of TTC, there is a high likelihood that pedestrian or joint attention between driver and pedestrian happens. However, if it does not happen, the outcome can be fatal. Our study was conducted with 16 participants. Half of them received gaze-aware pedestrian warning cues, whereas the other half did not receive any cue.

In our proposed scheme, the cognitive load of the drivers are recognized using critical and non-critical time frames of driving for each participant. Since critical time frames are very short, we kept non-critical time frames also short in order to have a uniform distribution in the training data. We trained multiple classifiers and evaluated them leave-one-person-out fashion in order to obtain person-independent results. Furthermore, since the time frames that are used in training and testing are very short, they do not reflect the complete intention of driver during driving. Therefore, we obtained a privacy-preserving scheme. In addition to person-independence and privacy-preserving features, our system also works in real time, which brings the opportunity to implement the same system in real life.

In general, when the cognitive load of the driver is recognized in a safety critical situation, visual cues and support can be adapted accordingly in order to provide safer, smoother, and less stressful driving experience even in very risky situations. In this work, we focused on a proof-of-concept in the driving scenario due to its highly dynamic and uncertain nature. However, our results show that the same methodology can be applied to any adaptive and gaze-aware application, especially in VR/AR.

A.4.3 Proposed Approach

Since the proposed system depends on the driving data which were collected using a VR setup, Section A.4.3 describes first the VR setup and the collected data from our previous work [212]. Then in Section A.4.3, data processing, training, and testing procedures are discussed.

VR Setup and Environment

In our previous work [212], we conducted a user-study to evaluate safety during driving in VR.

The hardware setup was created using HTC-Vive, Logitech G27 Steering Wheel and Pedals, Phillips headphones and Pupil-Labs HTC-Vive Binocular Add-on. Figure A.16 shows the dedicated setup.

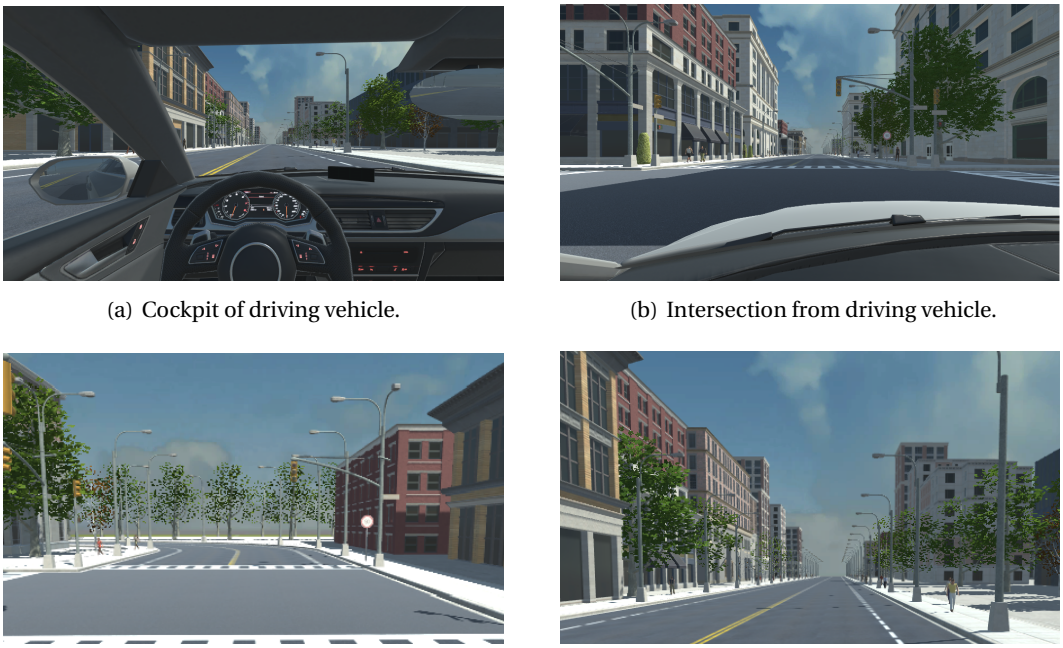
The hardware setup was used in a virtual environment created using Unity3D. We used 3D models from Urban City Pack, City Park Exterior, and Traffic Sign Sets packages to design the virtual city. Since we had not only critically crossing pedestrians, but also other pedestrians, we used Modern People asset packages for pedestrian models. Vehicle models and helper scripts were obtained from Realistic Car HD02, Traffic Cars, and Realistic Car Controller asset packages. Lastly, Playmaker and Simple Waypoint System packages were used to make pedestrian and vehicle movements smoother. For the eye tracking measures, Pupil Service version 1.7 of open source hmd-eyes[10] from Pupil-Labs was used. Examples scenes from VR environment are shown in Figure A.17.

The user-study consisted of acclimation and data collection phases. In the acclimation

[10] https://github.com/pupil-labs/hmd-eyes

Figure A.16: Experimental setup for VR.

phase, no data was collected. In the data collection phase, participants encountered with a dangerously crossing pedestrian. Two critical pedestrians on the side walk of main road were generated. In the beginning of the experiment, one of them was marked as crossing pedestrian. This critical pedestrian started crossing the road when the distance between driving vehicle and pedestrian was ($d_{critical} \approx 45m$). Every participant encountered with critically crossing pedestrian due to the start position of the vehicle in the data collection phase. They had the opportunity to speed up or slow down until the pedestrian crossing. Half of the participants started observing critical pedestrian warning cues around the pedestrian model with red color ≈ 32 meters in advance to pedestrian crossing. These parameters helped to keep TTC as $1.8s < TTC < 5s$, since the speed limit of main road was $90km/h$. Participants were supposed to realize the speed limit via speed signs on the road. Otherwise, the vehicle was equipped with maximum speed warning on a small in-car board. The pedestrian cues were made gaze-aware and were deactivated when gaze signal of the driver was closer than 5 meters to pedestrian for ≈ 0.85 seconds. Gaze signal on 2D canvas was obtained from Pupil Service from Pupil-Labs and then mapped from 2D to 3D with the help of ray-casting and Unity colliders. The hyper-parameters were determined empirically. The measurements, which changed over the time, were recorded in real time and were available for offline analysis. Since the pupil diameter values are very important for recognizing cognitive load, we post-processed pupil diameter



(a) Cockpit of driving vehicle.

(b) Intersection from driving vehicle.

(c) Intersection.

(d) Main road.

Figure A.17: Example scenes from VR environment.

measurements to remove the noise and normalize the data. For smoothing and normalization, we applied Savitzky-Golay filter and divisive baseline correction using a baseline duration of 0.5 seconds respectively.

Corresponding setup and experiments were run on a PC equipped with NVIDIA Titan X graphics card with 12GB memory, a 3.4GHz Intel i7-6700 processor and 16GB of RAM.

Cognitive Load Recognition

The data we obtained from the experiment (mentioned in Section A.4.3) is not annotated with regard to the cognitive load levels. Therefore, we first annotated our data with two levels of cognitive load: Low and high. We set $t_{critical}$ for both with-and without-pedestrian cue scenarios. The purpose of $t_{critical}$ is to separate the time domain into low and high cognitive load levels. It is taken as $t_{warning}$ and $t_{movement}$ for with-and without-warning scenarios respectively. The reason of taking two different $t_{critical}$ values is that cognitive load of the drivers who receive critical pedestrian cues starts increasing from $t_{warning}$, whereas cognitive load of others who do not receive any cue increases after the start of pedestrian movement.

In order to find the time frames to annotate exactly, we applied T-test using the pupil diameter data of each participant between $[t_{critical} - \delta t, t_{critical}]$ and $[t_{critical}, t_{critical} + \delta t]$. We used pupil diameter measurements due to the fact that pupil diameter is one of the main indicators of cognitive load. Once a significant difference in T-test was found with $p < 0.05$,

we assumed that we found a proper δt value. In order to keep the distributions significantly different but rather close to each other, we did not accept distributions where $p < 0.01$. Since cognitive load also depends on biological factors, which do not happen immediately, we shifted $t_{critical}$ by $+\theta t_{shift} = 0.8s$. In the end, we annotated each frame in the dedicated time frames with low or high cognitive load as it is shown in Table A.2:

Table A.2: Cognitive load annotations for time-frames.

Time Frame	Cognitive Load
$[t_{critical} + \theta t_{shift} - \delta t, t_{critical} + \theta t_{shift}]$	Low
$[t_{critical} + \theta t_{shift}, t_{critical} + \theta t_{shift} + \delta t]$	High

In order to recognize cognitive load of the driver, we trained different classifiers including Support Vector Machines (SVM), decision trees, random forests, and k-Nearest Neighbors (k-NN) using each frame. For the feature set, we used pupil diameters, and driver inputs on accelerator and brake pedals and steering wheel. Min-max normalization was applied to input data. In order to make our approach person-independent, we evaluated the data of each driver against the trained model using rest of the drivers. For example, in order to evaluate the first participant, we trained classifiers with other 15 participants and then evaluated the first participant using the data and its labels. This approach assures that we obtain person-independent results in the end.

Offline analyses offer many insights from the collected data. However, real time working capability is as important as the accuracy of the system especially in VR/AR fields. With this motivation, we evaluated whether our proposed scheme is capable of working in real time.

A.4.4 Results

In the following, we report results of our automated cognitive load recognition and its real time working capabilities that was conducted using MATLAB on a PC which is equipped with NVIDIA GeForce GTX 1070 mobile graphics card with 8GB of RAM, a 2.2GHz Intel i7-8750 processor, and 32GB of RAM.

In our dataset, there are 1171 frames in total and from each frame, maximum four features were used in training and testing. In addition, there are ≈ 73 frames (Mean) per participant (SD=12.5). We trained SVM, decision trees, random forests, and k-NN and tested according to the discussed setup in Section A.4.3 and used different combinations of features along with pupil diameter. We observed that using steering wheel input of driver did not lead to more accurate recognition. Since participants did not need to change steering wheel angle too much during the encountered scenarios, it is acceptable. Taking into account that cognitive load does not change very dramatically in short amount of time and each participant was evaluated against the trained models using the rest of the participants, the cognitive load recognition results are reasonable. The highest accuracy of 80.7% was achieved by SVM. Adding more

training data and participants has a great potential to increase the accuracy of predictions. Accuracy, precision, recall, and F1-score results which were obtained using these classifiers and feature set of pupil diameter and driver inputs on accelerator and brake are shown in Table A.3.

Table A.3: Results of cognitive load recognition.

Method	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	0.8070	0.7671	0.8574	0.8098
Decision Tree	0.7344	0.7332	0.7005	0.7165
Random Forest	0.7436	0.7372	0.7230	0.7299
1-Nearest Neighbor	0.6968	0.6846	0.6809	0.6828
5-Nearest Neighbor	0.7566	0.7473	0.7433	0.7453
10-Nearest Neighbor	0.7882	0.7947	0.7522	0.7729

During the training of classifiers which are mentioned in Table A.3, we set some hyper-parameters. For SVM, we used linear kernel function. For k-NN approach, we evaluated 1-NN, 5-NN and 10-NN. The accuracy results increase by increasing the k value. For random forest classifier, we used five trees to train for classification purposes.

Since it is important to apply the proposed approach in real life scenarios, we evaluated whether cognitive load recognition can be done in real time. Table A.4 shows the mean time spent for one prediction in each method.

Table A.4: Evaluation of mean prediction durations.

Method	Mean Prediction Duration (ms)
Support Vector Machine	0.319
Decision Tree	0.305
Random Forest	5.42
1-Nearest Neighbor	0.741
5-Nearest Neighbor	0.742
10-Nearest Neighbor	0.764

It is clear that all methods can be used for real time purposes. However, under this setup, it is reasonable to use SVM due to its higher accuracy and low prediction duration. In addition, if the dataset size increases, the real time working capability of k-NN is affected negatively. The same applies when the number of trees in random forests is increased.

A.4.5 Conclusion and Discussion

We proposed a scheme to recognize cognitive load of the drivers in safety critical situations using data collected during a driving study in VR. The scheme is person-independent because it generalizes well cross-subject. With more training data, there is a high potential for this scheme to work in a generic way. If person-specific setup is requested, the same scheme can be applied by adjusting the training data. In this case, even a more accurate cognitive load recognition can be obtained.

Due to the fact that we concentrated on very short time frames, complete driving data of participants were not exposed and only small amount of frames was used in training and testing. Only, the pupil diameter measurements were baseline-corrected using the first 0.5 seconds of driving. Therefore, it is a privacy-preserving scheme. Lastly, our scheme is capable of working real time. This outcome is very important and means that same scheme can be used in real driving studies and vehicles. It will enable more adaptive and intelligent feedbacks and inputs in driver warning systems; and eventually lead to safer and smoother driving experiences. We strongly suggest that similar schemes should be applied to real vehicles.

While this study is in driving domain, the outcome shows that our approach can be applied in similar adaptive user studies in VR and AR fields. The results indicate that there is a unique opportunity to design eye-tracking enabled interfaces and applications. Since we think that eye tracking has a great potential to transform VR and AR into another level, the outcome is valuable.

Despite the advantages and reasonable outcomes, there are some limitations as well. Firstly, since data were collected under VR setup, there is a likelihood that drivers do not behave naturally in VR. Virtual environment, weight of Head-Mounted-Display (HMD), or different dynamics of pedals or steering wheels can cause different behaviors than the real life. While we assume that participants became familiar with these in the acclimation phase, one should not ignore this possibility. Secondly, since the safety critical situations during driving happen in very short amount of time, it is difficult to collect big data in this context both using simulations or in real world.

As future work, more data and features can be used. There is a high likelihood that the accuracy of cognitive load recognition increases with more data. The same scheme can be applied to real driving simulators along with safety critical scenarios. Therefore, the findings in VR experiment can be compared with the future driving simulator experiments in terms of cognitive load recognition. Secondly, using raw eye videos along with other extracted features can be used to train deep models to estimate cognitive load. Furthermore, markov models or recurrent neural networks can be used to predict the cognitive load since they are suitable for time dependent data.

B Privacy Preserving Eye Tracking

This chapter includes the following publications:

1. **Efe Bozkir***, Onur Günlü*, Wolfgang Fuhl, Rafael F. Schaefer, and Enkelejda Kasneci. Differential privacy for eye tracking with temporal correlations. *PLoS ONE*, 16(8):e0255979, 2021. doi: 10.1371/journal.pone.0255979.
2. **Efe Bozkir***, Ali Burak Ünal*, Mete Akgün, Enkelejda Kasneci, and Nico Pfeifer. Privacy preserving gaze estimation using synthetic images via a randomized encoding based framework. In *ACM Symposium on Eye Tracking Research and Applications (ETRA)*, New York, NY, USA, 2020. ACM. doi: 10.1145/3379156.3391364.

* indicates equal contribution.

B.1 Differential Privacy for Eye Tracking with Temporal Correlations

B.1.1 Abstract

New generation head-mounted displays, such as VR and AR glasses, are coming into the market with already integrated eye tracking and are expected to enable novel ways of human-computer interaction in numerous applications. However, since eye movement properties contain biometric information, privacy concerns have to be handled properly. Privacy-preservation techniques such as differential privacy mechanisms have recently been applied to eye movement data obtained from such displays. Standard differential privacy mechanisms; however, are vulnerable due to temporal correlations between the eye movement observations. In this work, we propose a novel transform-coding based differential privacy mechanism to further adapt it to the statistics of eye movement feature data and compare various low-complexity methods. We extend the Fourier perturbation algorithm, which is a differential privacy mechanism, and correct a scaling mistake in its proof. Furthermore, we illustrate significant reductions in sample correlations in addition to query sensitivities, which provide the best utility-privacy trade-off in the eye tracking literature. Our results provide significantly high privacy without any essential loss in classification accuracies while hiding personal identifiers.

B.1.2 Introduction

Recent advances in the field of head-mounted displays (HMDs), computer graphics, and eye tracking enable easy access to pervasive eye trackers along with modern HMDs. Soon, the usage of such devices might result in a significant increase in the amount of eye movement data collected from users across different application domains such as gaming, entertainment, or education. A large part of this data is indeed useful for personalized experience and user-adaptive interaction. Especially in virtual and augmented reality (VR/AR), it is possible to derive plenty of sensitive information about users from the eye movement data. In general, it has been shown that eye tracking signals can be employed for activity recognition even in challenging everyday tasks [298, 290, 299], to detect cognitive load [241, 300], mental fatigue [301], and many other user states. Similarly, assessment of situational attention [212], expert-novice analysis in areas such as medicine [302] and sports [303], detection of personality traits [304], and prediction of human intent during robotic hand-eye coordination [305] can also be carried out based on eye movement features. Additionally, eye movements are useful for early detection of anomias [306] and diseases [307]. More importantly, eye movement data allow biometric authentication, which is considered to be a highly sensitive task [308]. A task-independent authentication using eye movement features and Gaussian mixtures is, for example, discussed by Kinnunen et al. [131]. Additionally, biometric identification based on an eye movements and oculomotor plant model are introduced by Komogortsev and Holland [132] and by Komogortsev et al. [133]. Eberz et al. [134] discuss that eye movement features can be used reliably also for authentication both in consumer level devices and var-

ious real world tasks, whereas Zhang et al. [27] show that continuous authentication using eye movements is possible in VR headsets. While authentication via eye movements could be useful in biometric applications, the applications that do not require any authentication step possess privacy risks for the individuals if such information is not hidden in the data. In addition, if such information is linked to personal identifiers, the risk might be even higher.

As biometric content can be retrieved from eye movements, it is important to protect them against adversarial behaviors such as membership inference. According to Steil et al. [143, p. 3], people agree to share their eye tracking data if a governmental health agency is involved in owning data or if the purpose is research. Therefore, privacy-preserving techniques are needed especially on the data sharing side of eye tracking considering that the usage of VR/AR devices with integrated eye trackers increases. As removing only the personal identifiers from data is not enough for anonymization due to linkage attacks [309], more sophisticated techniques for achieving user level privacy are necessary. Differential privacy [136, 137] is one effective solution, especially in the area of database applications. It protects user privacy by adding randomly generated noise for a given sensitivity and desired privacy parameter, ϵ. The differentially private mechanisms provide aggregate statistics or query answers while protecting the information of whether an individual's data was contained in a dataset. However, high dimensionality of the data and temporal correlations can reduce utility and privacy, respectively. Since eye movement features are high dimensional, temporally correlated, and usually contain recordings with long durations, it is important to tackle utility and privacy problems simultaneously. For eye movement data collected from HMDs or smart glasses, both local and global differential privacy can be applied. Applying differential privacy mechanisms to eye movement data would optimally anonymize the query outcomes that are carried out on such data while keeping data utility and usability high enough. As opposed to global differential privacy, local differential privacy adds user level noise to the data but assumes that the user sends data to a central data collector after adding local noise [310, 311]. While both could be useful depending on the application use-case, for this work, we focus on global differential privacy, considering that in many VR/AR applications which collect eye movement data, there is a central trusted user-level data collector and publisher.

To apply differential privacy to the eye movement data, we evaluate the standard Laplace Perturbation Algorithm (LPA) [136] of differential privacy and Fourier Perturbation Algorithm (FPA) [175]. The latter is suitable for time series data such as the eye movement feature signals. We propose two different methods that apply the FPA to chunks of data using original eye movement feature signals or consecutive difference signals. While preserving differential privacy using parallel compositions, chunk-based methods decrease query sensitivity and computational complexity. The difference-based method significantly decreases the temporal correlations between the eye movement features in addition to the decorrelation provided by the FPA that uses the discrete Fourier transform (DFT) as, e.g., in the works of Günlü and İşcan [312] and Günlü et al. [313]. The difference-based method provides a higher level of privacy since consecutive sample differences are observed to be less correlated than original consecutive data. Furthermore, we evaluate our methods using differentially private eye

movement features in document type, gender, scene privacy sensitivity classification, and person identification tasks on publicly available eye movement datasets by using similar configurations to previous works by Steil et al. [143, 28]. To generate differentially private eye movement data, we use the complete data instead of applying a subsampling step, used by Steil et al. [143] to reduce the sensitivity and to improve the classification accuracies for document type and privacy sensitivity. In addition, the previous work [143] applies the exponential mechanism for differential privacy on the eye movement feature data. The exponential mechanism is useful for situations where the best enumerated response needs to be chosen [138]. In eye movements, we are not interested in the "best" response but in the feature vector. Therefore, we apply the Laplace mechanism. In summary, we are the first to propose differential privacy solutions for aggregated eye movement feature signals by taking the temporal correlations into account, which can help provide user privacy especially for HMD or smart glass usage in VR/AR setups.

Our main contributions are as follows. (1) We propose chunk-based and difference-based differential privacy methods for eye movement feature signals to reduce query sensitivities, computational complexity, and temporal correlations. Furthermore, (2) we evaluate our methods on two publicly available eye movement datasets, i.e., MPIIDPEye [143] and MPIIPrivacEye [28], by comparing them with standard techniques such as LPA and FPA using the multiplicative inverse of the absolute normalized mean square error (NMSE) as the utility metric. In addition, we evaluate document type and gender classification, and privacy sensitivity classification accuracies as classification metrics using differentially private eye movements in the MPIIDPEye and MPIIPrivacEye datasets, respectively. Classification accuracy is used in the literature as a practical utility metric that shows how useful the data and proposed methods are. Our utility metric also provides insights into the divergence trend of differentially private outcomes and is analytically trackable unlike the classification accuracy. For both datasets, we also evaluate person identification task using differentially private data. Our results show significantly better performance as compared to previous works while handling correlated data and decreasing query sensitivities by dividing the data into smaller chunks. In addition, our methods hide personal identifiers significantly better than existing methods.

Previous Research

There are few works that focus on privacy-preserving eye tracking. Liebling and Preibusch [127] provide motivation as to why privacy considerations are needed for eye tracking data by focusing on gaze and pupillometry. Practical solutions are; therefore, introduced to protect user identity and sensitive stimuli based on a degraded iris authentication through optical defocus [147] and an automated disabling mechanism for the eye tracker's ego perspective camera with the help of a mechanical shutter depending on the detection of privacy sensitive content [28]. Furthermore, a function-specific privacy model for privacy-preserving gaze estimation task and privacy-preserving eye videos by replacing the iris textures are proposed by Bozkir and Ünal et al. [264] and by Chaudhary and Pelz [150], respectively. In addition,

solutions for privacy-preserving eye tracking data streaming [314] and real-time privacy control for eye tracking systems using area-of-interests [146] are also introduced in the literature. These works lack studying effects of temporal correlations.

For the user identity protection on aggregated eye movement features, works that focus on differential privacy are more relevant for us. Recently, standard differential privacy mechanisms are applied to heatmaps [144] and eye movement data that are obtained from a VR setup [143]. These works do not address the effects of temporal correlations in eye movements over time in the privacy context. In the privacy literature, there are privacy frameworks such as the Pufferfish [173] or the Olympus [174] for correlated and sensor data, respectively. These works, however, have different assumptions. For instance, the Pufferfish requires a domain expert to specify potential secrets and discriminative pairs, and Olympus models privacy and utility requirements as adversarial networks. As our focus is to protect user identity in the eye movements, we opt for differential privacy by discussing the effects of temporal correlations in eye movements over time and propose methods to reduce them. It has been shown that standard differential privacy mechanisms are vulnerable to temporal correlations as such mechanisms assume that data at different time points are independent from each other or adversaries lack the information about temporal correlations, leading an increased privacy loss of a differential privacy mechanism over time due to the temporal correlations [315, 172]. The aggregated eye movement features over time might end up in an extreme case of such correlations due to various user behaviors. Therefore, in addition to discussing the effects of such correlations on differential privacy over time, we propose methods to reduce the correlations so that the privacy leakage due to the temporal correlations are minimal.

B.1.3 Materials and Methods

In this section, the theoretical background of differential privacy mechanisms, proposed methods, and evaluated datasets are discussed.

Theoretical Background

Differential privacy uses a metric to measure the privacy risk for an individual participating in a database. Considering a dataset with weights of N people and a mean function, when an adversary queries the mean function for N people, the average weight over N people is obtained. After the first query, an additional query for $N-1$ people automatically leaks the weight of the remaining person. Using differential privacy, noise is added to the outcome of a function so that the outcome does not significantly change based on whether a randomly chosen individual participated in the dataset. The amount of noise added should be calibrated carefully since a high amount of noise might decrease the utility. We next define differential privacy.

Definition 3. *ϵ-Differential Privacy (ϵ-DP) [136, 137].* A randomized mechanism M is ϵ-differentially private if for all databases D and D' that differ at most in one element for all

$S \subseteq Range(M)$ with

$$\Pr[M(D) \in S] \leq e^{\epsilon} \Pr[M(D') \in S]. \tag{B.1}$$

The variance of the added noise depends on the query sensitivity, which is defined as follows.

Definition 4. *Query sensitivity [136].* For a random query X^n and $w \in \{1,2\}$, the query sensitivity Δ_w of X^n is the smallest number for all databases D and D' that differ at most in one element such that

$$||X^n(D) - X^n(D')||_w \leq \Delta_w(X^n) \tag{B.2}$$

where the L_w-distance is defined as

$$||X^n||_w = \sqrt[w]{\sum_{i=1}^{n} \left(|X_i|\right)^w}. \tag{B.3}$$

We list theorems that are used in the proposed methods.

Theorem 1. Sequential composition theorem [189]. Consider n mechanisms M_i that randomization of each query is independent for $i = 1,2,...,n$. If $M_1, M_2, ..., M_n$ are $\epsilon_1, \epsilon_2, ..., \epsilon_n$-differentially private, respectively, then their joint mechanism is $\left(\sum_{i=1}^{n} \epsilon_i\right)$-differentially private.

Theorem 2. Parallel composition theorem [189]. Consider n mechanisms as M_i for $i = 1,2,...,n$ that are applied to disjoint subsets of an input domain. If $M_1, M_2, ..., M_n$ are $\epsilon_1, \epsilon_2, ..., \epsilon_n$-differentially private, respectively, then their joint mechanism is $\left(\max_{i \in [1,n]} \epsilon_i\right)$-differentially private.

We define the Laplace Perturbation Algorithm (LPA) [136]. To guarantee differential privacy, the LPA generates the noise according to a Laplace distribution. $Lap(\lambda)$ denotes a random variable drawn from a Laplace distribution with a probability density function (PDF): $\Pr[Lap(\lambda) = h] = \frac{1}{2\lambda} e^{-|h|/\lambda}$, where $Lap(\lambda)$ has zero mean and variance $2\lambda^2$. We denote the noisy and differentially private values as $\widetilde{X}_i = X_i(D) + Lap(\lambda)$ for $i = 1,2,\ldots,n$. Since we have a series of eye movement observations, the final noisy eye movement observations are generated as $\widetilde{X}^n = X^n(D) + Lap^n(\lambda)$, where $Lap^n(\lambda)$ is a vector of n independent $Lap(\lambda)$ random variables and $X^n(D)$ is the eye movement observations without noise. The LPA is ϵ-differentially private for $\lambda = \Delta_1(X^n)/\epsilon$ [136].

We define the error function that we use to measure the differences between original X^n and noisy $\widetilde{X}^n$ observations. For this purpose, we use the metric normalized mean square error (NMSE) defined as

$$\text{NMSE} = \frac{1}{n} \sum_{i=1}^{n} \frac{(X_i - \widetilde{X}_i)^2}{\overline{X}\,\overline{\widetilde{X}}} \tag{B.4}$$

where

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \qquad \overline{\widetilde{X}} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{X}_i. \tag{B.5}$$

We define the utility metric as

$$\text{Utility} = \frac{1}{|\text{NMSE}|}. \tag{B.6}$$

As differential privacy is achieved by adding random noise to the data, there is a utility-privacy trade-off. Too much noise leads to high privacy; however, it might also result in poor analyses on the further tasks on eye movements. Therefore, it is important to find a good trade-off.

Methods

Standard differential privacy mechanisms are vulnerable to temporal correlations, since the independent noise realizations that are added to temporally correlated data could be useful for adversaries. However, decorrelating the data without the domain knowledge before adding the noise might remove important eye movement patterns and provide poor results in analyses. Many eye movement features are extracted by using time windows, as in previous work [143, 28], which makes the features highly correlated. Another challenge is that the duration of eye tracking recordings could change depending on the personal behaviors, skills, or personalities of the users. The longer duration causes an increased query sensitivity, which means that higher amounts of noise should be added to achieve differential privacy. In addition, when correlations between different data points exist, ϵ' is defined as the actual privacy metric instead of ϵ [171] that is obtained considering the fact that correlations can be used by an attacker to obtain more information about the differentially private data by filtering. In this work, we discuss and propose generic low-complexity methods to keep ϵ' small for eye movement feature signals. To deal with correlated eye movement feature signals, we propose three different methods: FPA, chunk-based FPA (CFPA) for original feature signals, and chunk-based FPA for difference based sequences (DCFPA). The sensitivity of each eye movement feature signal is calculated by using the L_w-distance such that

$$\begin{aligned}\Delta_w^f(X^n) &= \max_{p,q} \left|\left| X^{n,(p,f)} - X^{n,(q,f)} \right|\right|_w \\ &= \max_{p,q} \sqrt[w]{\sum_{t=1}^{n} \left(\left| X_t^{(p,f)} - X_t^{(q,f)} \right|\right)^w}\end{aligned} \tag{B.7}$$

where $X^{n,(p,f)}$ and $X^{n,(q,f)}$ denote observation vectors for a feature f from two participants p and q, n denotes the maximum length of the observation vectors, and $w \in \{1,2\}$.

Fourier Perturbation Algorithm (FPA)

In the FPA [175], the signal is represented with a small number of transform coefficients such

that the query sensitivity of the representative signal decreases. A smaller query sensitivity decreases the noise power required to make the noisy signal differentially private. In the FPA, the signal is transformed into the frequency domain by applying Discrete Fourier Transform (DFT), which is commonly applied as a non-unitary transform. The frequency domain representation of a signal consists of less correlated transform coefficients as compared to the time domain signal due to the high decorrelation efficiency of the DFT. Therefore, the correlation between the eye movement feature signals is reduced by applying the DFT. After the DFT, the noise sampled from the LPA is added to the first k elements of $DFT(X^n)$ that correspond to k lowest frequency components, denoted as $F^k = DFT^k(X^n)$. Once the noise is added, the remaining part (of size $n-k$) of the noisy signal $\widetilde{F}^k$ is zero padded and denoted as $PAD^n(\widetilde{F}^k)$. Lastly, using the Inverse DFT (IDFT), the padded signal is transformed back into the time domain. We can show that ϵ-differential privacy is satisfied by the FPA for $\lambda = \frac{\sqrt{n}\sqrt{k}\Delta_2(X^n)}{\epsilon}$ unlike the value claimed in previous work [175], as observed independently by Kellaris and Papadopoulos [316]. The procedure is summarized in Figure B.1, and the proof is provided below. Since not all coefficients are used, in addition to the perturbation error caused by the added noise, a reconstruction error caused by the lossy compression is introduced. It is important to determine the number of used coefficients k to minimize the total error. We discuss how we choose k values for FPA-based methods below.

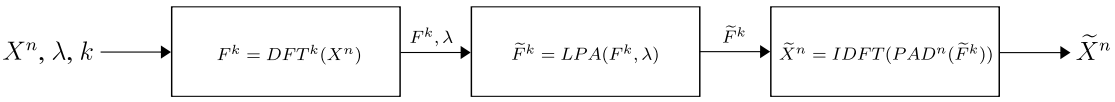


Figure B.1: Flow of the Fourier Perturbation Algorithm (FPA).

Proof of FPA being differentially private. We next prove that the FPA is ϵ differentially private for $\lambda = (\sqrt{n}\sqrt{k}\Delta_2(X^n))/\epsilon$. First, we prove the inequalities (a) and (b) in the following.

$$\Delta_1(\widehat{F}^n) \overset{(a)}{\leq} \sqrt{k}\cdot\Delta_2(\widehat{F}^n) \overset{(b)}{\leq} \sqrt{n}\cdot\sqrt{k}\cdot\Delta_2(X^n) \tag{B.8}$$

where $\widehat{F}^n(I) = [\widehat{F}^k(I), 0, 0, \ldots, 0]$ such that $n-k$ zeros are padded. Consider (B.8)(a), which follows since we have

$$\begin{aligned}\Delta_1(\widehat{F}^n) &= \max_{I,I'} \left|\left| \widehat{F}^n(I) - \widehat{F}^n(I') \right|\right|_1 = \max_{I,I'} \sum_{j=1}^{n} \left| \widehat{F}_j(I) - \widehat{F}_j(I') \right| \\ &= \max_{I,I'} \sum_{j=1}^{k} \left| \widehat{F}_j(I) - \widehat{F}_j(I') \right| \cdot 1 \end{aligned} \tag{B.9}$$

so that by applying Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\max_{I,I'}\sum_{j=1}^{k}\left|\widehat{F}_j(I)-\widehat{F}_j(I')\right|\cdot 1 &\leq \max_{I,I'}\Big(\sum_{j=1}^{k}\left|\widehat{F}_j(I)-\widehat{F}_j(I')\right|^2\Big)^{1/2}\cdot\Big(\sum_{j=1}^{k}1^2\Big)^{1/2}\\ &\leq \max_{I,I'}\left|\left|\widehat{F}^n(I)-\widehat{F}^n(I')\right|\right|_2\cdot\sqrt{k}\\ &\leq \sqrt{k}\cdot\Delta_2(\widehat{F}^n). \end{aligned} \tag{B.10}$$

Consider next (B.8)(b), which follows since we obtain

$$\Delta_2(\widehat{F}^n)=\max_{I,I'}\left|\left|\widehat{F}^n(I)-\widehat{F}^n(I')\right|\right|_2=\max_{I,I'}\Big(\sum_{j=1}^{n}\left|\widehat{F}_j(I)-\widehat{F}_j(I')\right|^2\Big)^{1/2} \tag{B.11}$$

and since F^n has more non-zero elements than $\widehat{F}^n$, we have

$$\Delta_2(\widehat{F}^n)\leq\max_{I,I'}\Big(\sum_{j=1}^{n}\left|F_j(I)-F_j(I')\right|^2\Big)^{1/2}. \tag{B.12}$$

Recall that $F^n(I)=DFT(X^n(I))$, $F^n(I')=DFT(X^n(I'))$, and DFT is linear, so we have

$$DFT(X^n(I)-X^n(I'))=F^n(I)-F^n(I'). \tag{B.13}$$

By applying Parseval's theorem to the DFT, we obtain

$$\Big(\frac{1}{n}\cdot\sum_{j=1}^{n}\left|F_j(I)-F_j(I')\right|^2\Big)^{1/2}=\Big(\sum_{j=1}^{n}\left|X_j(I)-X_j(I')\right|^2\Big)^{1/2}. \tag{B.14}$$

Combining (B.12) and (B.14), we prove (B.8)(b) since we have

$$\begin{aligned}\Delta_2(\widehat{F}^n) &\leq \max_{I,I'}\sqrt{\sum_{j=1}^{n}\left|X_j(I)-X_j(I')\right|^2}\cdot\sqrt{n}\\ &\leq \max_{I,I'}\left|\left|X^n(I)-X^n(I')\right|\right|_2\cdot\sqrt{n}\\ &\leq \Delta_2(X^n)\cdot\sqrt{n}. \end{aligned} \tag{B.15}$$

Finally, since the LPA that is applied to $\widehat{F}^k$ is ϵ-DP for $\lambda=\frac{\Delta_1(\widehat{F}^n)}{\epsilon}$ [136], (B.8) proves that the FPA is ϵ-DP for $\lambda=\dfrac{\sqrt{n}\sqrt{k}\Delta_2(X^n)}{\epsilon}$.

Chunk-based FPA (CFPA)

One drawback of directly applying the FPA to the eye movement feature signals is large query sensitivities for each feature f due to long signal sizes. To solve this, Steil et al. [143] propose to subsample the signal using non-overlapping windows, which means removing many data points. While subsampling decreases the query sensitivities, it also decreases the amount of data. Instead, we propose to split each signal into smaller chunks and apply the FPA to each chunk so that complete data can be used. We choose the chunk sizes of 32, 64, and 128 since there are divide-and-conquer type tree-based implementation algorithms for fast DFT calculations when the transform size is a power of 2 [177]. When the signals are split into chunks, chunk level query sensitivities are calculated and used rather than the sensitivity of the whole sequence. Differential privacy for the complete signal is preserved by Theorem 2 [189] since the used chunks are non-overlapping. As the chunk size decreases, the chunk level sensitivity decreases as well as the computational complexity. However, the parameter ϵ' that accounts for the sample correlations might increase with smaller chunk sizes because temporal correlations between neighboring samples are larger in an eye movement dataset. On the other hand, if the chunk sizes are kept large, then the required amount of noise to achieve differential privacy increases due to the increased query sensitivity. Therefore, a good trade-off between computational complexity, and correlations is needed to determine the optimal chunk size.

Difference- and chunk-based FPA (DCFPA)

To tackle temporal correlations, we convert the eye movement feature signals into difference signals where differences between consecutive eye movement features are calculated as

$$\widehat{X}_t^{(f)} = \left\{ X_t^{(f)} - X_{t-1}^{(f)} \right\} \Big|_{t=2}^{n} \quad , \quad \widehat{X}_1^{(f)} = X_1^{(f)}. \tag{B.16}$$

Using the difference signals denoted by $\widehat{X}^{n,(f)}$, we aim to further decrease the correlations before applying a differential privacy method. We conjecture that the ratio ϵ'/ϵ decreases in the difference-based method as compared to the FPA method. To support this conjecture, we show that the correlations in the difference signals decrease significantly as compared to the original signals. This results in lower ϵ' and better privacy for the same ϵ. The difference-based method is applied together with the CFPA. Therefore, the differences are calculated inside chunks. The first element of each chunk is preserved. Then, the FPA mechanism is applied to the difference signals by using query sensitivities calculated based on differences and chunks. For each chunk, noisy difference observations are aggregated to obtain the final noisy signals. This mechanism is differentially private by Theorem 1 [189], and described in Algorithm 2.

Since Theorem 1 can be applied to the DCFPA when the consecutive differences are assumed to be independent, which is a valid assumption for eye movement feature signals as we illustrate below, there is also a trade-off between the chunk sizes and utility for the DCFPA. If a large chunk size is chosen, then the total ϵ value could be very large, which reduces privacy.

Algorithm 2: DCFPA.

Inputs: X^n, λ, k
Output: $\widetilde{X}^n$
1) $\widehat{X}_t = \left\{ X_t - X_{t-1} \right\}\Big|_{t=2}^{n}$, $\widehat{X}_1 = X_1$.
2) $\widehat{\widetilde{X}}^n = FPA(\widehat{X}^n, \lambda, k)$.
3) $\widetilde{X}_t = \left\{ \widehat{\widetilde{X}}_t + \widehat{\widetilde{X}}_{t-1} \right\}\Big|_{t=2}^{n}$, $\widetilde{X}_1 = \widehat{\widetilde{X}}_1$.

Therefore, we choose chunk sizes of 32, 64, and 128 for the DCFPA as well for evaluation We illustrate the CFPA and DCFPA in Figure B.2, for instance with three chunks.

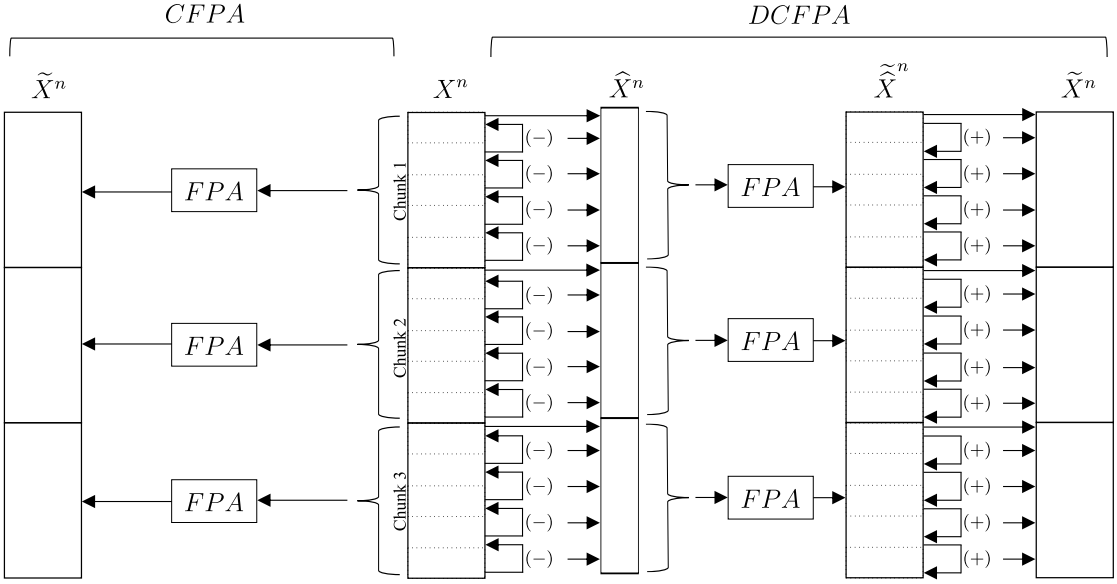


Figure B.2: Flow of the CFPA and DCFPA.

Choice of the Number of Transform Coefficients

The proposed methods require a selection of a value for k. A small k value increases the reconstruction error, while a large k value results in an increase in the perturbation error. Therefore, it is important to find an optimal k value that minimizes the sum of the two errors. In this work, we compare a large set of possible k values to choose the best values.

We apply the aforementioned differential privacy mechanisms by using 100 noisy evaluations to find optimal k values applied to features or chunks. Optimal k values have the minimum absolute NMSE for each chunk, eye movement feature, and document or recording type. In a distributed setting, each party should know the k values in advance. However, in a centralized setting, it is crucial to choose the k values in a differentially private manner. To evaluate the differential privacy in the eye tracking area while taking the temporal correlations into account, we focus on optimal k values for this work. One shortcoming of this approach is that the optimal k value compromises some information about the data, which leaks pri-

vacy [175]. Our observation is that the information leaked by optimizing the parameter k is negligible as compared to the privacy reduction due to temporally correlated data. Thus, we illustrate the results with optimal k values.

Datasets

We evaluate our methods on two different publicly available eye movement datasets namely, MPIIDPEye and MPIIPrivacEye that are dedicated to privacy-preserving eye tracking. Both datasets consist of aggregated and timely eye movement feature signals related to eye fixations, saccades, blinks, and pupil diameters which are commonly used in VR/AR applications as they represent individual user behaviors. As all minimum values of wordbook features ranging from 1 to 4 are zeros in both datasets, we exclude them from the utility and privacy calculations. In addition, we remark that both datasets are available for non-commercial scientific purposes.

MPIIDPEye [143]: A publicly available eye movement dataset consisting of 60 recordings that is collected from VR devices for a reading task of three document types (comics, newspaper, and textbook) from 20 (10 female, 10 male) participants. Each recording consists of 52 eye movement feature sequences computed with a sliding window size of 30 seconds and a step size of 0.5 seconds.

MPIIPrivacEye [28]: A publicly available eye movement dataset consisting of 51 recordings from 17 participants for 3 consecutive sessions with a head-mounted eye tracker and a field camera, which is similar to an AR setup. Each recording consists of 52 eye movement feature sequences computed with a sliding window size of 30 seconds and a step size of 1 second, and each observation is annotated with binary privacy sensitivity levels of the scene that is being viewed. The dataset also consists of scene features extracted with convolutional neural networks. We do not evaluate the last part of the recording 1 of the participant 10, as the eye movement features are not available for this region. To detect the privacy level of the scene that is being viewed, we remark that the scene is very important [317]; however, an individual's eye movements can improve the detection rate when they are fused with the information from the scene.

B.1.4 Results

This section discusses data correlations in addition to evaluations using utility and classification metrics. The utility and classification results are averaged over 100 noisy evaluations with the optimal k values in MATLAB. We evaluate and compare the utility of differentially private eye movement feature signals by using absolute NMSE, as this metric provides analytically trackable results. However, it does not provide implications regarding the practical usability of the private eye movement signals. Therefore, we also report classification accuracies of document type, scene privacy sensitivity, gender prediction, and person identification tasks in order to show the usability of the private data and proposed methods. An optimal trade-off

between utility tasks (e.g., low absolute NMSE, high classification accuracy in document type prediction) and privacy (e.g., low ϵ, low classification accuracy in person identification or gender prediction tasks) is favorable.

Correlation Analysis

Using the correlation coefficient as the metric, we first illustrate high temporal correlation between eye movement feature data. Since there are 52 eye movement features in both datasets, it is not feasible to illustrate all correlation results. Thus, in the following we illustrate the correlations for the features *ratio large saccade* and *blink rate* in the MPIIDPEye and MPIIPrivacEye datasets, respectively. The correlation coefficients of *ratio large saccade* and *blink rate* for three document and recording types over a time difference Δt w.r.t. the signal samples at, e.g., the fifth time instance for original eye movement feature signals and difference signals for all participants for both datasets are depicted in Figures B.3, B.4 and Figures B.5, B.6, respectively. As correlations between the difference signals are significantly smaller than correlations between the original eye movement feature signals, the DCFPA is less vulnerable to privacy reduction due to temporal correlations, thus ensuring that the value of ϵ' is close to the differential privacy design parameter ϵ.

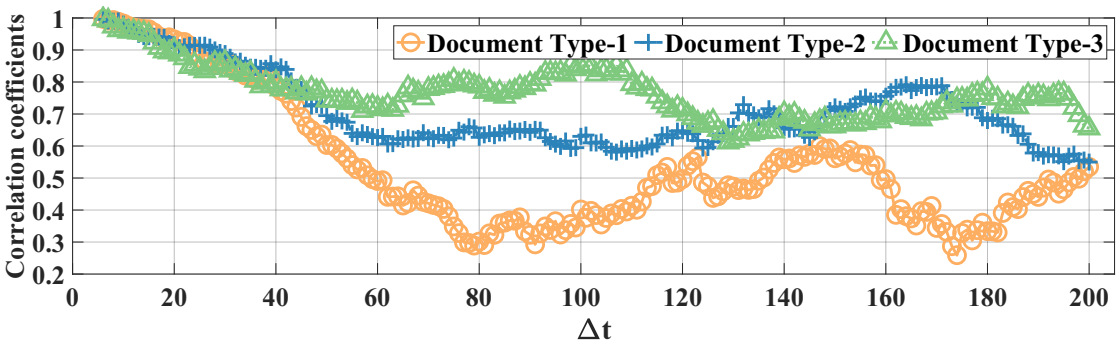


Figure B.3: Correlation coefficients of the raw signals of feature *ratio large saccade* in the MPIIDPEye dataset. The values are calculated over a time difference of Δt (Each time step corresponds to 0.5s) w.r.t. the samples at the fifth time instance.

Utility Results

We evaluate the utility defined in Eq (B.6) by applying our methods separately to different document and recording types; therefore, we report the utility results separately. As we apply the proposed methods separately to each eye movement feature, we first calculate the mean utility of each feature and then calculate the average utility over all features. The utility results for various ϵ values for aforementioned methods on the MPIIDPEye and MPIIPrivacEye datasets are given in Figures B.7, B.8, B.9 and Figures B.10, B.11, B.12, respectively.

While a high absolute NMSE, i.e., low utility, does not necessarily mean that a mechanism is

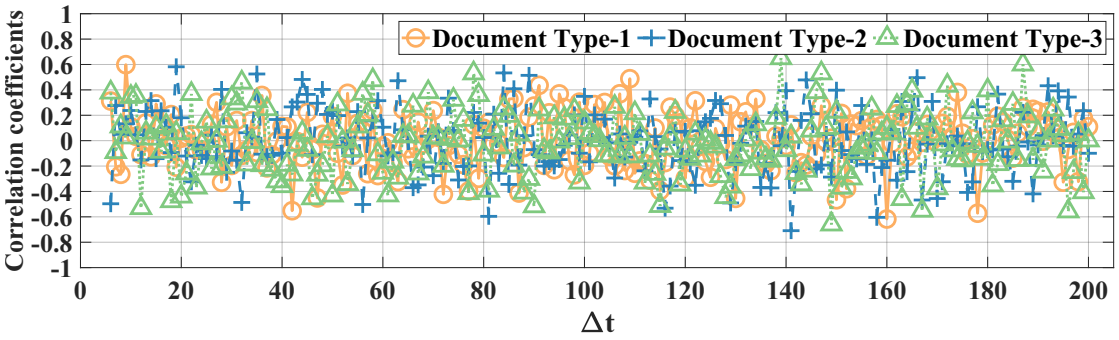


Figure B.4: Correlation coefficients of the difference signals of feature *ratio large saccade* in the MPIIDPEye dataset. The values are calculated over a time difference of Δt (Each time step corresponds to 0.5s) w.r.t. the samples at the fifth time instance.

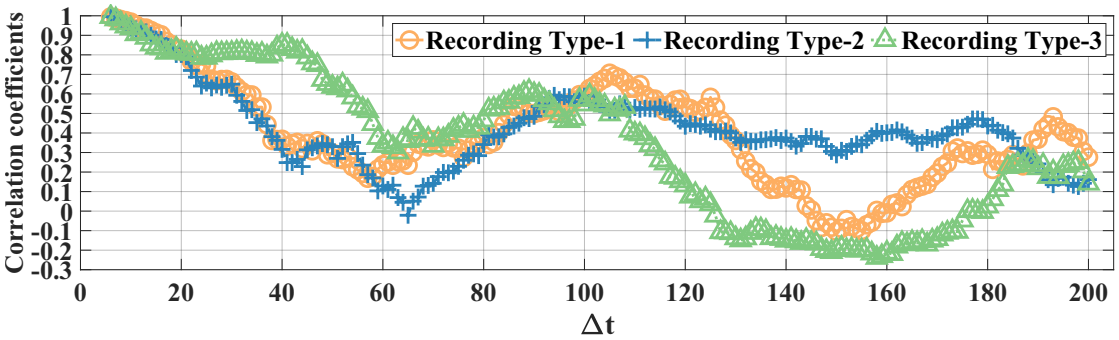


Figure B.5: Correlation coefficients of the raw signals of feature *blink rate* in the MPIIPrivacEye dataset. The values are calculated over a time difference of Δt (Each time step corresponds to 1s) w.r.t. the samples at the fifth time instance.

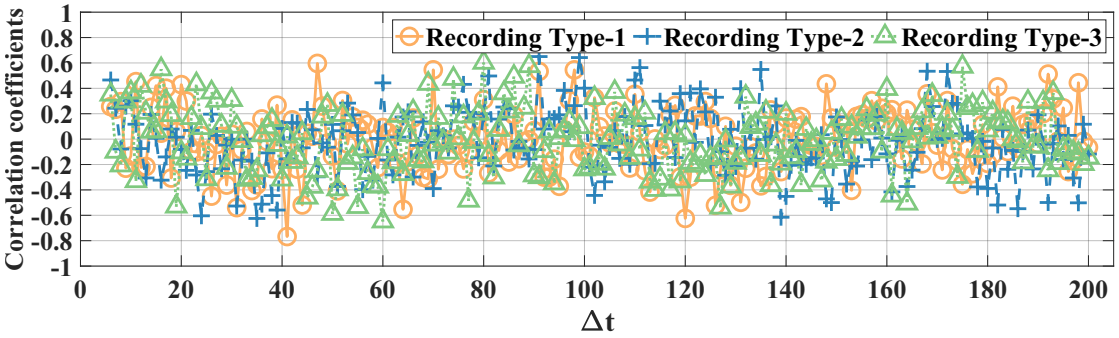


Figure B.6: Correlation coefficients of the difference signals of feature *blink rate* in the MPIIPrivacEye dataset. The values are calculated over a time difference of Δt (Each time step corresponds to 1s) w.r.t. the samples at the fifth time instance.

completely useless, higher utility means that the mechanism would perform more effectively than low utility in various tasks. The trend in the utility results of both evaluated datasets are similar. As the query sensitivities are lower in CFPA, utilities of CFPA are always higher than the utilities of the FPA as theoretically expected. The DCFPA particularly with small

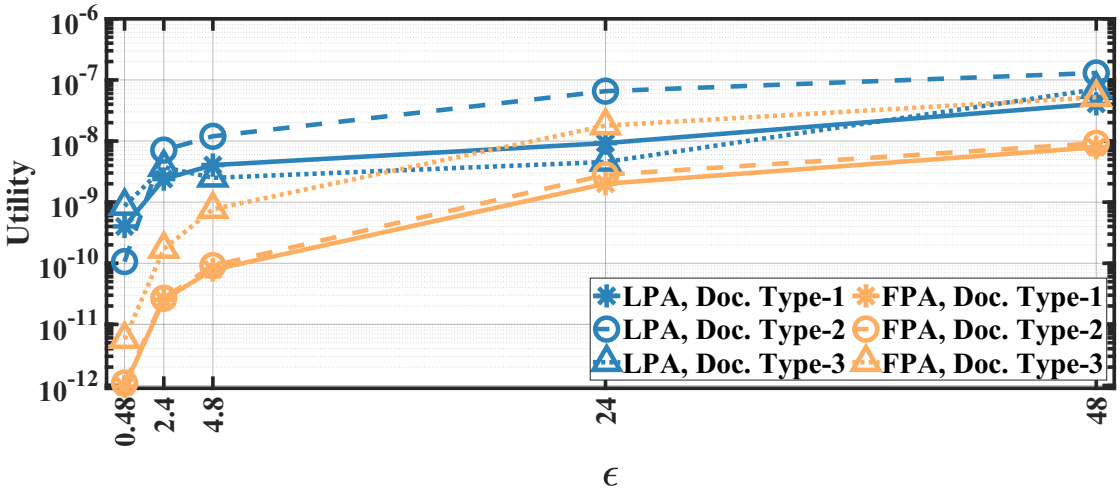


Figure B.7: Utility of the LPA and FPA for MPIIDPEye.

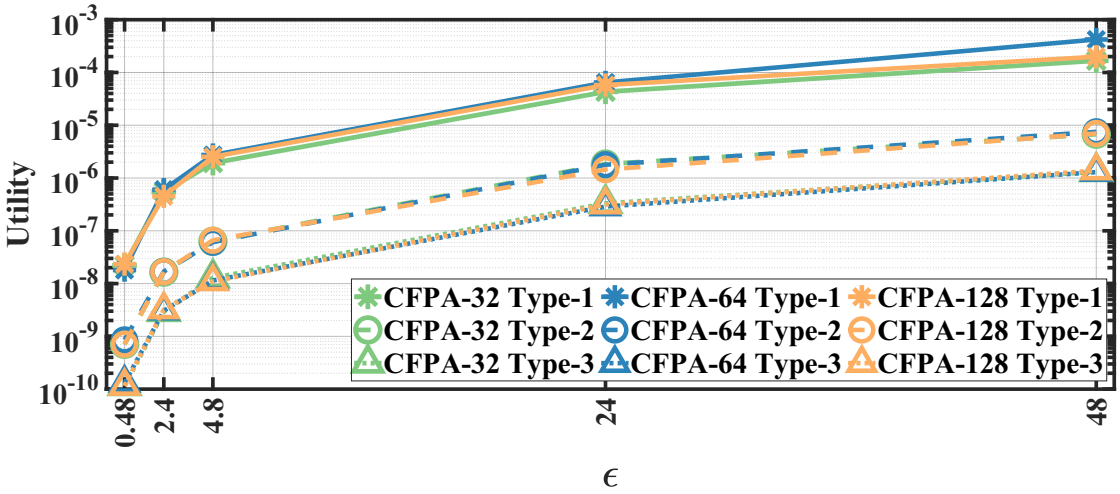


Figure B.8: Utility of the CFPA for MPIIDPEye.

chunks outperforms other methods in the most private settings, namely in the lowest ϵ regions. When different chunk sizes are compared within the CFPA and DCFPA, different chunk sizes perform similarly for the CFPA method. For the DCFPA, there is a significant trend for better utilities when the chunk sizes are decreased. However, as temporal correlations in the smaller chunk sizes higher and since a higher chunk size reduces the temporal correlations better, it is ideal to use a higher chunk size if the utilities are comparable. In general, while the LPA, namely the standard Laplace mechanism used for differential privacy, is vulnerable to temporal correlations [172], our methods also outperform it in terms of utilities. In addition to high utilities, the calculation complexities are decreased with the CFPA and DCFPA which is another advantage of chunk-based methods.

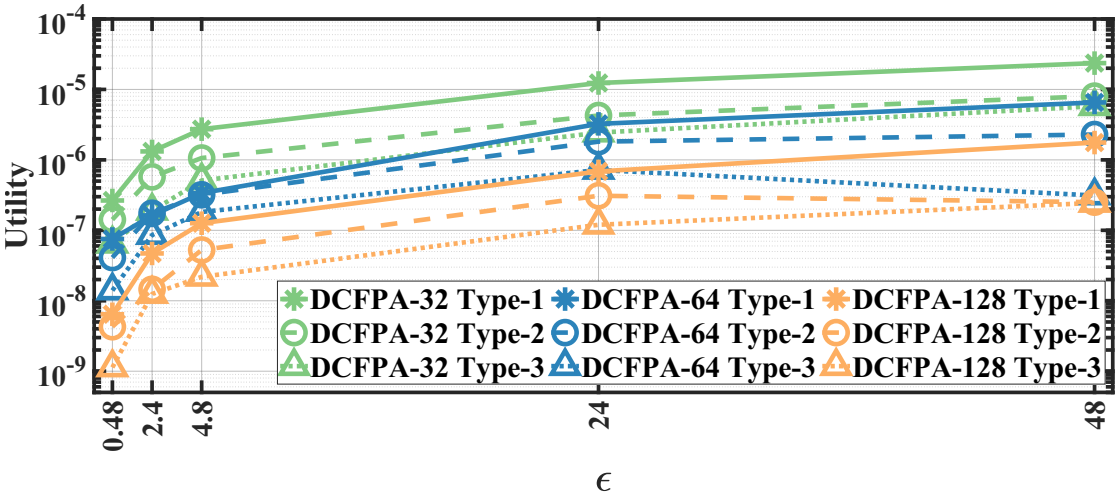


Figure B.9: Utility of the DCFPA for MPIIDPEye.

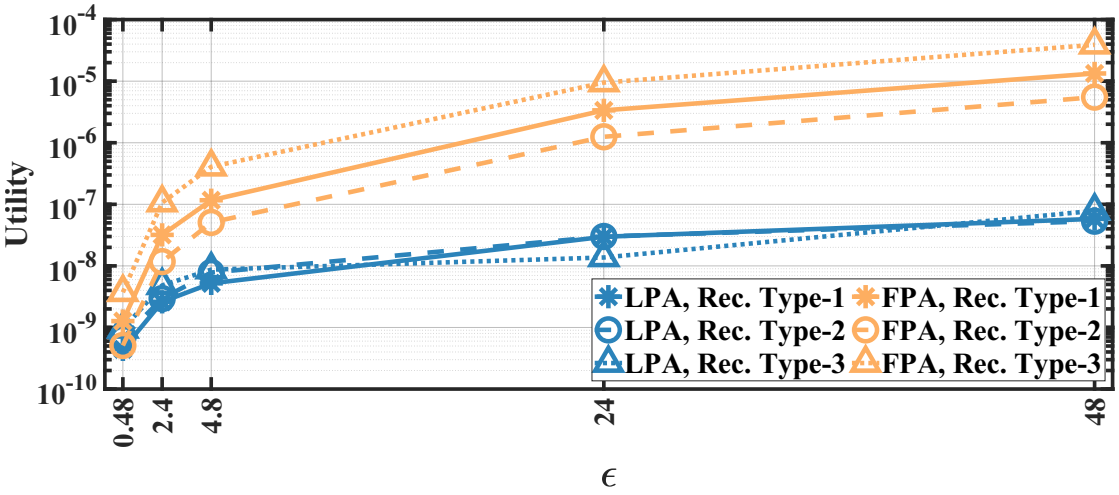


Figure B.10: Utility of the LPA and FPA for MPIIPrivacEye.

Classification Accuracy Results

We evaluate document type and gender classification results for the MPIIDPEye and privacy sensitivity classification results for the MPIIPrivacEye by using differentially private data generated by the methods which handle temporal correlations in the privacy context. In addition, for both datasets, we evaluate person identification tasks. While a NMSE-based utility metric provides analytically trackable way for comparison, evaluating private data using classification accuracies give insights about the usability of the noisy data in practice. Instead of only using Support Vector Machines (SVM) as in previous works [143, 28], we evaluate a set of classifiers including SVMs, decision trees (DTs), random forests (RFs), and k-Nearest Neighbors (k-NNs). We employ a similar setup as in previous work [143] with radial basis function (RBF) kernel, bias parameter of $C = 1$, and automatic kernel scale for the SVMs. For

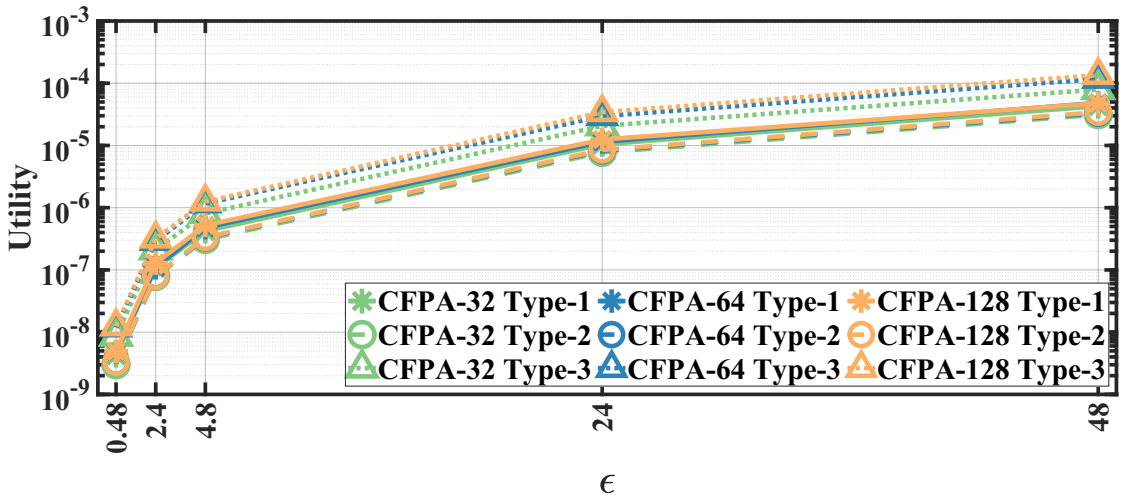


Figure B.11: Utility of the CFPA for MPIIPrivacEye.

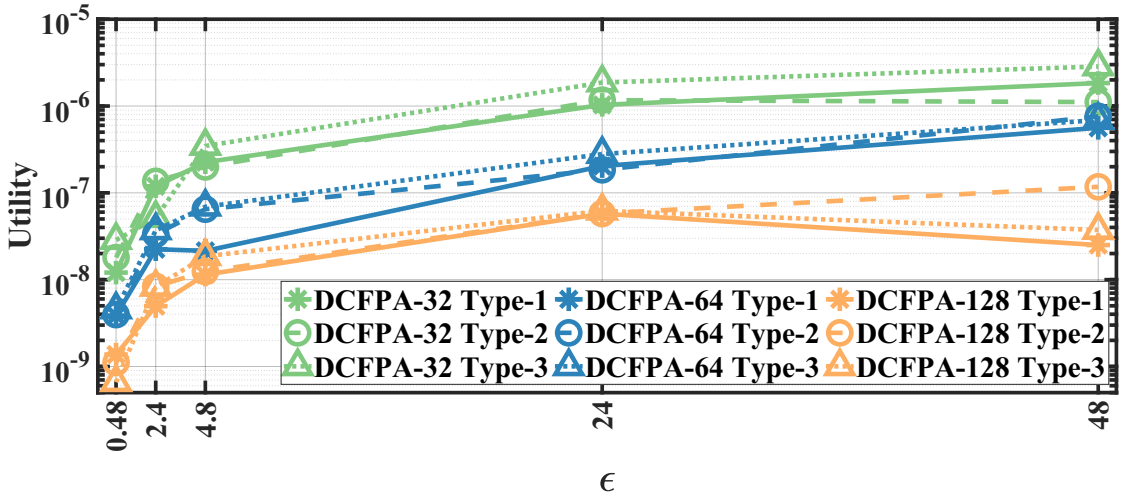


Figure B.12: Utility of the DCFPA for MPIIPrivacEye.

RFs and k-NNs, we use 10 trees and $k = 11$ with a random tie breaker among tied groups, respectively. We normalize the training data to zero mean and unit variance, and apply the same parameters to the test data. Although we do not apply subsampling while generating the differentially private data, which is applied in previous work [143], we use subsampled data for training and testing for document type, gender, and privacy sensitivity classification tasks with window sizes of 10 and 20 for MPIIDPEye and MPIIPrivacEye, respectively, to have a fair comparison and similar amount of data. Apart from the person identification task, all the classifiers are trained and tested in a leave-one-person-out cross-validation setup, which is considered as a more challenging but generic setup. For the person identification task, since it is not possible to carry out the experiments in a leave-one-person-out cross-validation setup, we opt for a similar configuration as in previous work [143] by using the first halves of the signals as training data and the remaining parts as test data. Such setup can be considered as

one of the hypothetical best-case scenarios for an adversary as this simulates some set of prior knowledge for an adversary on participants' visual behaviors. For the person identification task, in order to use similar amount of data with other classification tasks from each signal, we use window sizes of 5 and 10 for MPIIDPEye and MPIIPrivacEye, respectively. For the MPIIDPEye, we evaluate results both with majority voting by summarizing classifications from different time instances for each participant and recording and without majority voting. Privacy sensitivity classification tasks for MPIIPrivacEye are carried only without majority voting since privacy sensitivity of the scene can change at each time step and applying majority voting to such task in our setup is not reasonable.

While classification results cannot be treated directly as the utility, they provide insights into the usability of the differentially private data in practice. We first evaluate document type classification task in the majority voting setting in Table B.1 for MPIIDPEye dataset as it is possible to compare our results with the previous work [143]. As previous results quickly drop to the 0.33 guessing probability in high privacy regions, we significantly outperform them particularly with DCFPA and FPA with the accuracies over 0.60 and 0.85, respectively. In the less private regions towards $\epsilon = 48$, this trend still exists with the CFPA and FPA with accuracy results over 0.7 and 0.85. Chunk-based methods perform slightly worse than the FPA in the document type classifications even though the utility of the FPA is lower. We observe that the reading patterns are hidden easier with chunk-based methods; therefore, document type classification task becomes more challenging. This is especially validated with DCFPA methods using different chunk sizes, as DCFPA-128 outperforms smaller chunk-sized DCFPAs even though the sensitivities are higher. Therefore, we conclude that the differential privacy method should be selected for eye movements depending on the further task which will be applied. The document type classification results without majority voting are provided in the table in S1 Table.

Next, we analyze the gender classification results for MPIIDPEye. All methods are able to hide the gender information in the high privacy regions as it is already challenging to identify it with clean data as accuracies are ≈ 0.7 in previous work [143]. While we obtain similar results compared to previous work for the gender classification task, the CFPA method is able to predict gender information correctly in the less private regions, namely $\epsilon = 48$, as it also has the highest utility values in these regions. The FPA applied to the complete signal and the DCFPA are not able to classify genders accurately. We observe that higher amount of noise that is needed by the FPA and noising the fine-grained "difference" information between eye movement observations with DCFPA are the reasons for hiding the gender information successfully in all privacy regions. Overall, the CFPA provides an optimal equilibrium between gender and document type classification success in the less private regions if gender information is not considered as a feature that should be protected from adversaries. Otherwise, all proposed methods are able to hide gender information from the data in the higher privacy regions as expected. Gender classification results are depicted in Table B.2. Especially in some methods with k-NNs and SVMs, gender classification accuracies are close to zero because of the majority voting and it is validated by the results without majority voting in the table in S2

Table.

Lastly for the MPIIDPEye, we evaluate person identification task using differentially private data. The resulting classification accuracies with majority voting are depicted in Table B.3. By using the FPA, it is possible to identify the participants very accurately, which means that even though the document type classification accuracies of the FPA are higher than the others, a strong adversary can also identify personal ids when this method is used. The same trend also exists in the without majority voting setting, which is reported in the table in S3 Table. The CFPA and DCFPA perform well against person identification attempts in the high privacy regions. However, when the CFPA is used, it is possible to identify personal ids in the less private regions. Overall, for the MPIIDPEye dataset, the DCFPA performs better than the others due to its resistance against person identification and gender classification and relatively high classification accuracies for the document type predictions. We conclude that this is due to the robust decorrelation effect of the DCFPA.

For the MPIIPrivacEye, we report privacy sensitivity classification accuracies using differentially private eye movement features in the Table B.4. The FPA performs worse than our methods. The DCFPA, particularly with the chunk size of 32, outperforms all other methods slightly in the higher privacy regions as it is also the case for the utility results. In the lower privacy regions, the CFPA performs the best with ≈ 0.60 accuracy. Since performance does not drop significantly in the higher chunk sizes, it is reasonable to use higher chunk-sized methods as they decrease the temporal correlations better. While having an accuracy of approximately 0.6 in a binary classification problem does not form the best performance, according to the previous work [28], privacy sensitivity classification using only eye movements with clean data in a person-independent setup only performs marginally higher than 0.60. Therefore, we show that even though we use differentially private data in the most private settings, we obtain similar results to the classification results using clean data. This means that differentially private eye movements can be used along with scene features for detecting privacy sensitive scenes in AR setups.

The results of the person identification task in the MPIIPrivacEye dataset are similar to the results of the MPIIDPEye dataset and the results with majority voting are depicted in Table B.5. Personal identifiers are predicted very accurately when the FPA is used. The CFPA and DCFPA are resistant to person identification attacks in all privacy regions performing around the random guess probability in almost all cases. Similar to the classification results of the MPIIDPEye dataset, the DCFPA method performs the best when utility-privacy trade-off is taken into consideration. The person identification results without majority voting are presented in the table in S4 Table.

B.1.5 Discussion

We compared our differential privacy methods with the standard Laplace mechanism as well as the FPA method, which is proposed for temporally correlated data, by using the MPIIDPEye

and MPIIPrivacEye datasets. The utility results based on the NMSE metric show that due to the reduced sensitivities as a result of the chunking operations, the CFPA and DCFPA perform better than the FPA and standard Laplace mechanism. While larger chunk sizes applied with the CFPA and DCFPA decrease the effects of temporal correlations on the differential privacy mechanisms, they also increase the sensitivities, leading to higher amount of noise addition to the data and worse utility performance. Utility evaluations represent how much differentially private signals diverge from the original signals. Having eye movement feature signals less diverged from the original values by providing the differential privacy would lead better performance in various tasks. While the FPA, CFPA, and DCFPA are appropriate for temporally correlated data, the DCFPA uses the consecutive differences of eye movement feature signals, which are significantly less correlated than the original feature signals, as illustrated in Figures B.4 and B.6. Thus, the DCFPA is less vulnerable to temporal correlations in the differential privacy context.

In addition to utility results, we evaluated document type, gender, and person identification tasks for the MPIIDPEye dataset and privacy sensitivity classification of the observed scene and person identification task for the MPIIPrivacEye dataset and compared our results with the previous works especially in the eye tracking literature. The FPA outperforms the CFPA and DCFPA in document type classification task since the chunks perturb "Z"-type reading patterns. However, this might be a task-specific outcome as the CFPA and DCFPA perform better in terms of utility. In addition, when the FPA is used, personal identifiers can be estimated with high accuracies in both datasets. On the contrary, especially the DCFPA provides decreased probabilities for person identification tasks in the MPIIDPEye, and probabilities close to the random guess probability for the MPIIPrivacEye, which are optimal from a privacy-preservation perspective. We remark that this outcome is also related to decreased temporal correlations. Gender information is successfully hidden in all methods and scene privacy can be predicted to some extent using differentially private eye movement signals. In addition, privacy sensitivity detection results on the MPIIPrivacEye are consistent with the utility results based on the NMSE metric.

Due to the significant reduction of temporal correlations and high utility and relatively accurate classification results in different tasks, the DCFPA is the best performing differential privacy method for eye movement feature signals. In addition, it is not possible to recognize the person accurately from eye movement data when the DCFPA is used. From correlation reduction point of view, in both methods namely, CFPA and DCFPA, when the performances are similar, it is reasonable to use higher chunk sizes as such chunks are less vulnerable to temporal correlations as illustrated in Figures B.3 and B.5. Overall, our methods outperform the state-of-the-art for differential privacy for aggregated eye movement feature signals.

B.1.6 Conclusion

We proposed different methods to achieve differential privacy for eye movement feature signals by correcting, extending, and adapting the FPA method. Since eye movement features are correlated over time and are high dimensional, standard differential privacy methods provide low utility and are vulnerable to inference attacks. Thus, we proposed privacy solutions for temporally correlated eye movement data. Our methods can be easily applied to other biometric human-computer interaction data as well since they are independent of the used data and outperform the state-of-the-art methods in terms of both NMSE and classification accuracy and reduce the correlations significantly. In future work, we will analyze the actual privacy metric ϵ' which takes the data correlations into account and choose k values in a private manner for the centralized differential privacy setting.

Table B.1: Document type classification accuracies in the MPIIDPEye dataset using differentially private eye movement features with majority voting.

	Document type classification accuracies (k-NN|SVM|DT|RF)				
Method	$\epsilon = 0.48$	$\epsilon = 2.4$	$\epsilon = 4.8$	$\epsilon = 24$	$\epsilon = 48$
FPA	0.50|0.63|0.82|**0.87**	0.51|0.63|0.81|**0.87**	0.5|0.61|0.81|**0.87**	0.52|0.63|0.82|**0.87**	0.52|0.64|0.83|**0.88**
CFPA-32	0.39|0.37|0.45|0.44	0.40|0.38|0.45|0.44	0.40|0.44|0.46|0.44	0.58|0.58|0.55|0.60	**0.71**|0.69|0.66|0.66
CFPA-64	0.41|0.36|0.45|0.45	0.40|0.37|0.44|0.45	0.40|0.41|0.44|0.45	0.57|0.59|0.55|0.59	0.70|0.70|0.66|0.66
CFPA-128	0.36|0.33|0.45|0.45	0.36|0.33|0.44|0.44	0.37|0.35|0.44|0.45	0.52|0.56|0.52|0.57	0.69|0.68|0.64|0.66
DCFPA-32	0.51|0.37|0.46|0.44	0.51|0.36|0.47|0.42	0.47|0.35|0.47|0.43	0.49|0.37|0.46|0.44	0.48|0.36|0.47|0.45
DCFPA-64	0.61|0.45|0.43|0.41	0.55|0.35|0.43|0.41	0.56|0.41|0.43|0.41	**0.60**|0.43|0.45|0.42	0.59|0.40|0.44|0.43
DCFPA-128	**0.64**|0.45|0.46|0.48	**0.62**|0.42|0.45|0.46	**0.69**|0.50|0.44|0.46	0.57|0.45|0.45|0.46	0.60|0.42|0.45|0.46

Table B.2: Gender classification accuracies in the MPIIDPEye dataset using differentially private eye movement features with majority voting.

	Gender classification accuracies (k-NN|SVM|DT|RF)				
Method	$\epsilon = 0.48$	$\epsilon = 2.4$	$\epsilon = 4.8$	$\epsilon = 24$	$\epsilon = 48$
FPA	0.44|0.30|0.43|0.38	0.45|0.30|0.41|0.37	0.44|0.28|0.41|0.39	0.43|0.27|0.43|0.38	0.44|0.31|0.42|0.39
CFPA-32	0.04|0.01|0.26|0.24	0.05|0.01|0.27|0.25	0.05|0.02|0.28|0.27	0.36|0.30|0.50|0.45	0.62|0.50|0.67|0.53
CFPA-64	0.08|0.05|0.27|0.26	0.08|0.04|0.28|0.27	0.10|0.06|0.31|0.27	0.38|0.34|0.52|0.47	0.62|0.51|0.68|0.54
CFPA-128	0.18|0.15|0.32|0.30	0.16|0.12|0.31|0.30	0.18|0.10|0.32|0.31	0.36|0.30|0.50|0.46	0.60|0.47|0.68|0.54
DCFPA-32	0.03| ≈ 0|0.22|0.31	0.04| ≈ 0|0.23|0.32	0.04| ≈ 0|0.22|0.32	0.04| ≈ 0|0.23|0.31	0.04| ≈ 0|0.23|0.32
DCFPA-64	0.04| ≈ 0|0.30|0.33	0.04| ≈ 0|0.30|0.34	0.04| ≈ 0|0.30|0.32	0.04| ≈ 0|0.29|0.34	0.03| ≈ 0|0.30|0.34
DCFPA-128	0.09|0.01|0.34|0.35	0.08| ≈ 0|0.32|0.34	0.08|0.01|0.32|0.35	0.07| ≈ 0|0.33|0.34	0.07|0.01|0.34|0.34

Table B.3: Person identification accuracies in the MPIIDPEye dataset using differentially private eye movement features with majority voting.

Method	Person identification accuracies (k-NN|SVM|DT|RF)				
	$\epsilon = 0.48$	$\epsilon = 2.4$	$\epsilon = 4.8$	$\epsilon = 24$	$\epsilon = 48$
FPA	1	1	1	1	1
CFPA-32	0.15|0.08|0.44|0.37	0.13|0.08|0.46|0.39	0.11|0.08|0.48|0.41	0.40|0.11|0.64|0.70	0.72|0.16|0.87|0.92
CFPA-64	0.14|0.08|0.42|0.34	0.13|0.08|0.44|0.37	0.12|0.08|0.45|0.38	0.39|0.11|0.63|0.71	0.70|0.17|0.85|0.92
CFPA-128	0.16|0.05|0.39|0.36	0.15|0.05|0.41|0.36	0.17|0.05|0.43|0.39	0.45|0.07|0.55|0.63	0.70|0.16|0.74|0.88
DCFPA-32	0.06|0.10|0.39|0.37	0.06|0.10|0.39|0.36	0.08|0.10|0.39|0.36	0.10|0.10|0.39|0.37	0.10|0.10|0.40|0.38
DCFPA-64	0.10|0.10|0.33|0.35	0.10|0.10|0.32|0.34	0.10|0.10|0.32|0.33	0.13|0.10|0.31|0.34	0.13|0.10|0.32|0.33
DCFPA-128	0.09|0.05|0.24|0.28	0.09|0.05|0.25|0.27	0.10|0.05|0.23|0.27	0.10|0.06|0.24|0.26	0.10|0.05|0.22|0.25

Table B.4: Privacy sensitivity classification accuracies in the MPIIPrivacEye dataset using differentially private eye movement features.

Method	Privacy sensitivity classification accuracies (k-NN|SVM|DT|RF)				
	$\epsilon = 0.48$	$\epsilon = 2.4$	$\epsilon = 4.8$	$\epsilon = 24$	$\epsilon = 48$
FPA	0.49|0.58|0.51|0.55	0.49|0.58|0.51|0.55	0.49|0.58|0.51|0.55	0.50|0.58|0.51|0.55	0.50|0.59|0.51|0.55
CFPA-32	0.55|0.59|0.52|0.56	0.55|0.58|0.52|0.56	0.55|0.58|0.52|0.56	0.56|0.58|0.53|0.57	0.58|**0.60**|0.54|0.58
CFPA-64	0.55|0.58|0.52|0.56	0.55|0.58|0.52|0.56	0.55|0.58|0.52|0.56	0.56|0.58|0.53|0.57	0.58|0.59|0.54|0.58
CFPA-128	0.55|0.57|0.52|0.56	0.55|0.57|0.52|0.56	0.55|0.57|0.52|0.56	0.56|0.58|0.53|0.57	0.58|0.59|0.54|0.59
DCFPA-32	0.54|**0.59**|0.52|0.56	0.55|**0.59**|0.52|0.56	0.55|**0.59**|0.52|0.56	0.54|**0.59**|0.52|0.56	0.55|0.59|0.52|0.56
DCFPA-64	0.54|0.58|0.52|0.56	0.54|0.58|0.52|0.56	0.54|0.58|0.52|0.56	0.54|0.58|0.52|0.56	0.54|0.58|0.52|0.56
DCFPA-128	0.54|0.57|0.52|0.56	0.54|0.57|0.52|0.56	0.54|0.57|0.52|0.56	0.54|0.57|0.52|0.56	0.54|0.57|0.52|0.56

Table B.5: Person identification classification accuracies in the MPIIPrivacEye dataset using differentially private eye movement features with majority voting.

	Person identification classification accuracies (k-NN|SVM|DT|RF)				
Method	$\epsilon = 0.48$	$\epsilon = 2.4$	$\epsilon = 4.8$	$\epsilon = 24$	$\epsilon = 48$
FPA	1	1	1	1	1
CFPA-32	0.05|0.06|0.07|0.07	0.05|0.06|0.07|0.07	0.06|0.06|0.08|0.07	0.07|0.06|0.09|0.11	0.11|0.06|0.14|0.16
CFPA-64	0.06|0.06|0.06|0.07	0.06|0.06|0.06|0.06	0.06|0.06|0.07|0.07	0.07|0.06|0.09|0.09	0.11|0.06|0.16|0.16
CFPA-128	0.06|0.06|0.07|0.07	0.06|0.06|0.07|0.07	0.06|0.06|0.07|0.08	0.07|0.06|0.09|0.11	0.11|0.06|0.15|0.15
DCFPA-32	0.06|0.05|0.08|0.07	0.06|0.06|0.07|0.08	0.07|0.05|0.08|0.08	0.07|0.05|0.08|0.08	0.07|0.06|0.08|0.08
DCFPA-64	0.06|0.05|0.06|0.06	0.06|0.05|0.06|0.06	0.06|0.05|0.06|0.06	0.05|0.06|0.06|0.06	0.06|0.05|0.06|0.06
DCFPA-128	0.05|0.05|0.06|0.06	0.05|0.05|0.05|0.06	0.06|0.05|0.06|0.06	0.05|0.05|0.05|0.06	0.06|0.05|0.05|0.06

B.1.7 Supporting Information

S1 Table. Document type classification results without majority voting for the MPIIDPEye dataset.

Table B.6: Document type classification accuracies in the MPIIDPEye dataset using differentially private eye movement features without majority voting.

	Document type classification accuracies (k-NN|SVM|DT|RF)				
Method	$\epsilon = 0.48$	$\epsilon = 2.4$	$\epsilon = 4.8$	$\epsilon = 24$	$\epsilon = 48$
FPA	0.46|0.52|0.68|0.73	0.46|0.52|0.67|0.73	0.45|0.51|0.67|0.73	0.46|0.52|0.68|0.73	0.47|0.52|0.68|0.74
CFPA-32	0.34|0.35|0.36|0.38	0.34|0.35|0.36|0.38	0.34|0.36|0.36|0.38	0.39|0.44|0.38|0.42	0.47|0.53|0.44|0.49
CFPA-64	0.34|0.35|0.36|0.38	0.34|0.35|0.36|0.38	0.34|0.36|0.36|0.38	0.39|0.44|0.38|0.42	0.47|0.53|0.44|0.49
CFPA-128	0.34|0.34|0.36|0.39	0.34|0.34|0.36|0.39	0.34|0.34|0.36|0.39	0.38|0.42|0.37|0.42	0.46|0.51|0.43|0.48
DCFPA-32	0.36|0.35|0.36|0.37	0.36|0.34|0.35|0.37	0.35|0.34|0.36|0.37	0.36|0.35|0.35|0.37	0.35|0.34|0.36|0.38
DCFPA-64	0.38|0.37|0.35|0.37	0.37|0.35|0.35|0.37	0.37|0.36|0.35|0.37	0.37|0.36|0.35|0.37	0.37|0.36|0.35|0.37
DCFPA-128	0.40|0.38|0.36|0.38	0.39|0.37|0.35|0.37	0.41|0.39|0.35|0.38	0.38|0.37|0.35|0.38	0.39|0.37|0.35|0.38

S2 Table. Gender classification results without majority voting for the MPIIDPEye dataset.

Table B.7: Gender classification accuracies in the MPIIDPEye dataset using differentially private eye movement features without majority voting.

	Gender classification accuracies (k-NN|SVM|DT|RF)				
Method	$\epsilon = 0.48$	$\epsilon = 2.4$	$\epsilon = 4.8$	$\epsilon = 24$	$\epsilon = 48$
FPA	0.48|0.42|0.48|0.45	0.48|0.42|0.47|0.44	0.47|0.41|0.47|0.45	0.47|0.41|0.48|0.44	0.48|0.43|0.48|0.45
CFPA-32	0.43|0.31|0.44|0.40	0.43|0.31|0.45|0.41	0.43|0.32|0.46|0.41	0.46|0.42|0.49|0.47	0.51|0.47|0.53|0.53
CFPA-64	0.44|0.35|0.45|0.40	0.44|0.35|0.45|0.41	0.44|0.35|0.46|0.42	0.46|0.43|0.49|0.47	0.51|0.48|0.54|0.53
CFPA-128	0.45|0.39|0.46|0.42	0.45|0.38|0.46|0.42	0.45|0.38|0.46|0.42	0.46|0.43|0.49|0.47	0.51|0.47|0.53|0.53
DCFPA-32	0.44|0.27|0.45|0.42	0.44|0.27|0.45|0.42	0.44|0.27|0.45|0.42	0.44|0.27|0.45|0.42	0.44|0.27|0.46|0.42
DCFPA-64	0.44|0.30|0.46|0.43	0.43|0.29|0.46|0.43	0.44|0.30|0.46|0.43	0.43|0.30|0.46|0.43	0.43|0.30|0.46|0.43
DCFPA-128	0.44|0.32|0.46|0.43	0.44|0.32|0.46|0.43	0.44|0.32|0.47|0.43	0.44|0.31|0.46|0.43	0.44|0.32|0.47|0.43

S3 Table. Person identification results without majority voting for the MPIIDPEye dataset.

Table B.8: Person identification accuracies in the MPIIDPEye dataset using differentially private eye movement features without majority voting.

	Person identification accuracies (k-NN|SVM|DT|RF)				
Method	$\epsilon = 0.48$	$\epsilon = 2.4$	$\epsilon = 4.8$	$\epsilon = 24$	$\epsilon = 48$
FPA	1|1|0.98|1	1|1|0.98|1	1|1|0.98|1	1|1|0.97|1	1|1|0.95|1
CFPA-32	0.09|0.11|0.16|0.16	0.08|0.11|0.16|0.17	0.09|0.11|0.17|0.17	0.12|0.15|0.18|0.21	0.18|0.21|0.23|0.27
CFPA-64	0.09|0.11|0.17|0.17	0.09|0.11|0.17|0.17	0.09|0.11|0.17|0.17	0.12|0.15|0.19|0.21	0.17|0.21|0.23|0.27
CFPA-128	0.11|0.13|0.17|0.18	0.11|0.13|0.17|0.18	0.11|0.13|0.17|0.18	0.13|0.16|0.18|0.20	0.18|0.20|0.21|0.25
DCFPA-32	0.09|0.10|0.15|0.16	0.09|0.11|0.14|0.16	0.09|0.11|0.14|0.16	0.09|0.11|0.14|0.16	0.09|0.11|0.15|0.16
DCFPA-64	0.09|0.10|0.13|0.15	0.09|0.10|0.13|0.15	0.09|0.10|0.13|0.15	0.09|0.10|0.13|0.15	0.09|0.10|0.13|0.15
DCFPA-128	0.08|0.09|0.12|0.13	0.08|0.09|0.11|0.13	0.08|0.09|0.11|0.13	0.08|0.09|0.12|0.13	0.08|0.09|0.11|0.13

S4 Table. Person identification results without majority voting for MPIIPrivacEye dataset.

Table B.9: Person identification classification accuracies in the MPIIPrivacEye dataset using differentially private eye movement features without majority voting.

	Person identification classification accuracies (k-NN|SVM|DT|RF)				
Method	$\epsilon = 0.48$	$\epsilon = 2.4$	$\epsilon = 4.8$	$\epsilon = 24$	$\epsilon = 48$
FPA	1|1|0.99|1	1|1|0.99|1	1|1|0.99|1	1|1|0.97|1	1|1|0.94|1
CFPA-32	0.06|0.07|0.06|0.06	0.06|0.07|0.06|0.06	0.06|0.07|0.06|0.06	0.07|0.09|0.07|0.07	0.08|0.10|0.08|0.08
CFPA-64	0.06|0.07|0.06|0.06	0.06|0.07|0.06|0.06	0.06|0.07|0.06|0.06	0.07|0.09|0.07|0.07	0.08|0.10|0.07|0.08
CFPA-128	0.06|0.08|0.06|0.07	0.06|0.08|0.06|0.07	0.06|0.08|0.06|0.07	0.07|0.09|0.07|0.08	0.08|0.11|0.07|0.08
DCFPA-32	0.06|0.07|0.06|0.06	0.06|0.07|0.06|0.06	0.06|0.07|0.06|0.06	0.06|0.07|0.06|0.06	0.06|0.07|0.06|0.06
DCFPA-64	0.06|0.06|0.06|0.06	0.06|0.06|0.06|0.06	0.06|0.06|0.06|0.06	0.06|0.06|0.06|0.06	0.06|0.07|0.06|0.06
DCFPA-128	0.06|0.06|0.06|0.06	0.06|0.06|0.06|0.06	0.06|0.06|0.06|0.06	0.06|0.06|0.06|0.06	0.06|0.06|0.06|0.06

Acknowledgments

O. Günlü thanks Ravi Tandon for his useful suggestions. E. Bozkir thanks Martin Pawelczyk and Mete Akgün for useful discussions.

Author Contributions

Conceptualization: Efe Bozkir, Onur Günlü, Wolfgang Fuhl, Rafael F. Schaefer, Enkelejda Kasneci.
Data curation: Efe Bozkir, Onur Günlü.
Formal analysis: Efe Bozkir, Onur Günlü.
Investigation: Efe Bozkir, Onur Günlü.
Methodology: Efe Bozkir, Onur Günlü.
Software: Efe Bozkir, Onur Günlü.
Supervision: Efe Bozkir, Onur Günlü, Wolfgang Fuhl, Rafael F. Schaefer, Enkelejda Kasneci.
Validation: Efe Bozkir, Onur Günlü.
Visualization: Efe Bozkir, Onur Günlü.
Writing – original draft: Efe Bozkir, Onur Günlü.
Writing – review & editing: Efe Bozkir, Onur Günlü, Wolfgang Fuhl, Rafael F. Schaefer, Enkelejda Kasneci.

Peer Review History

PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0255979.

Data Availability Statement

Relevant data files are provided via following url: https://atreus.informatik.uni-tuebingen.de/bozkir/dp_eye_tracking.

Funding

O. Günlü and R. F. Schaefer are supported by the German Federal Ministry of Education and Research (BMBF) within the national initiative for “Post Shannon Communication (NewCom)” under the Grant 16KIS1004. We acknowledge support by Open Access Publishing Fund of University of Tübingen. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests

The authors have declared that no competing interests exist.

B.2 Privacy Preserving Gaze Estimation using Synthetic Images via a Randomized Encoding Based Framework

B.2.1 Abstract

Eye tracking is handled as one of the key technologies for applications that assess and evaluate human attention, behavior, and biometrics, especially using gaze, pupillary, and blink behaviors. One of the challenges with regard to the social acceptance of eye tracking technology is however the preserving of sensitive and personal information. To tackle this challenge, we employ a privacy-preserving framework based on randomized encoding to train a Support Vector Regression model using synthetic eye images privately to estimate the human gaze. During the computation, none of the parties learn about the data or the result that any other party has. Furthermore, the party that trains the model cannot reconstruct pupil, blinks or visual scanpath. The experimental results show that our privacy-preserving framework is capable of working in real-time, with the same accuracy as compared to non-private version and could be extended to other eye tracking related problems.

B.2.2 Introduction

Recent advances in the fields of Head-Mounted-Display (HMD) technology, computer graphics, augmented reality (AR), and eye tracking enabled numerous novel applications. One of the most natural and non-intrusive ways of interaction with HMDs or smart glasses is achieved by gaze-aware interfaces using eye tracking. However, it is possible to derive a lot of sensitive and personal information from eye tracking data such as intentions, behaviors, or fatigue since eyes are not fully controlled in a conscious way.

It has been shown that cognitive load [318, 241], visual attention [212], stress [289], task identification [319], skill level assessment and expertise [320, 321, 302], human activities [298, 290], biometric information and authentication [131, 133, 132, 27, 322], or personality traits [304] can be obtained using eye tracking data. Since highly sensitive information can be derived from eye tracking data, it is not surprising that HMDs or smart glasses have not been adopted by large communities yet. According to a recent survey [143], people agree to share their eye tracking data only when it is co-owned by a governmental health-agency or is used for research purposes. This indicates that people are hesitant about sharing their eye tracking data in commercial applications. Therefore, there is a likelihood that larger communities could adopt HMDs or smart glasses if privacy-preserving techniques are applied in the eye tracking applications. The reasons why privacy preserving schemes are needed for eye tracking are discussed in [127] extensively. However, until now, there are not many studies in privacy-preserving eye tracking. Recently, a method to detect privacy sensitive everyday situations [28], an approach to degrade iris authentication while keeping the gaze tracking utility in an acceptable accuracy [147], and differential privacy based techniques to protect personal information on heatmaps and eye movements [144, 143] are introduced.

While differential privacy can be applied to eye tracking data for various tasks, it introduces additional noise on the data which causes decrease in the utility [144, 143], and it might lead to less accurate results in computer vision tasks, such as gaze estimation or activity recognition.

In light of the above, function-specific privacy models are required. In this work, we focus on the gaze estimation problem as a proof-of-concept by using synthetic data including eye landmarks and ground truth gaze vectors. However, the same privacy-preserving approach can be extended to any feature-based, eye tracking problem such as intention, fatigue, or activity detection, in HMD or unconstrained setups due to the demonstrated real-time working capabilities. In our study, the gaze estimation task is solved by using Support Vector Regression (SVR) models in a privacy-preserving manner by computing the dot product of eye landmark vectors to obtain the kernel matrix of the SVR for a scenario, where two parties have the eye landmark data, each of which we call *input-party*, and one *function-party* that trains a prediction model on the data of the input-parties. This scenario is relevant when the input-parties use eye tracking data to improve the accuracy of their models and do not share the data due to the privacy concerns. To this end, we utilize a framework employing randomized encoding [178]. In the computation, neither the eye images nor the extracted features are revealed to the function-party directly. Furthermore, the input-parties do not infer the raw eye tracking data or result of the computation. Eye images that are used for training and testing are rendered using UnityEyes [69] synthetically and 36 landmark-based features [74] are used. To the best of our knowledge, this is the first work that applies a privacy-preserving scheme based on function-specific privacy models on an eye tracking problem.

B.2.3 Threat Model

We assume that the input-parties are semi-honest (honest but curious) that are not allowed to deviate from the protocol description while they try to infer some valuable information about other parties' private inputs using their views of the protocol execution. We also assume that the function-party is malicious and the input-parties and the function-party do not collude.

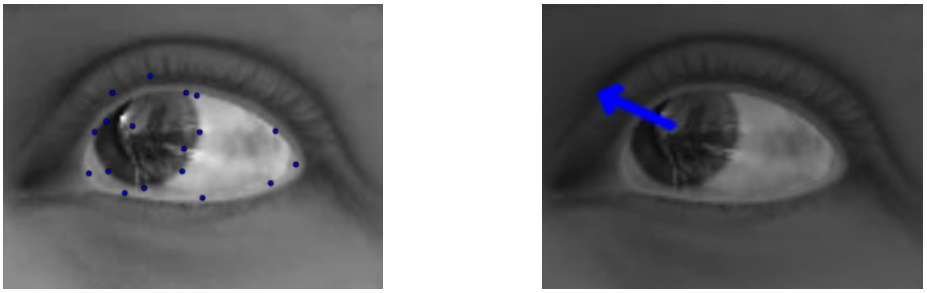
B.2.4 Methodology

In this section, we discuss the data generation, randomized encoding, and privacy-preserving gaze estimation framework.

Data Generation

To train and evaluate the gaze estimator, we generate eye images and gaze vectors. As our work is a proof-of-concept and requires high amount of data, synthetic images from UnityEyes [69], which is based on the Unity3D, are used. *Camera parameters* and *Eye parameters* are chosen as $(0,0,0,0)$ (fixed camera) and $(0,0,30,30)$ (eyeball pose range parameters in degrees), respectively. $20,000$ images are rendered in *Fantastic* quality setting and 512×384 screen

resolution. Then, processing and normalization pipeline from [74] is employed. In the end, we obtain 128×96 sized eye images, 18 eye landmarks including eight iris edge, eight eyelid, one iris center, and one iris-center-eyeball-center vector normalized according to Euclidean distance between eye corners, and gaze vectors using pitch and yaw angles. Final feature vectors consist of 36 elements. Figure B.13 shows an example illustration.



(a) Landmarks. (b) Gaze.

Figure B.13: Eye landmarks and gaze on a synthetic image.

Randomized Encoding

The utilized framework employs randomized encoding (RE) [141, 161] to compute the dot product of the landmark vectors. The dot product is needed to compute kernel matrix of the SVR which is later used for training the gaze estimator and validation of the framework.

In the randomized encoding, the computation of a function $f(x)$ is performed by a randomized function $\hat{f}(x;r)$ where x is the input value, which corresponds to eye landmarks in our setup, and r is the random value. The idea is to encode the original function by using random value(s) such that the combination of the components of the encoding reveals only the output of the original function. In the framework, the computation of the dot product is accomplished by utilizing the decomposable and affine randomized encoding (DARE) of addition and multiplication [142]. The encoding of multiplication is as follows.

Definition 5 (Perfect RE for Multiplication [142])**.** A multiplication function is defined as $f_m(x_1, x_2) = x_1 \cdot x_2$ over a ring R. One can perfectly encode the f_m by employing the DARE $\hat{f_m}(x_1, x_2; r_1, r_2, r_3)$:

$$\hat{f_m}(x_1, x_2; r_1, r_2, r_3) = (x_1 + r_1, x_2 + r_2, r_2 x_1 + r_3, r_1 x_2 + r_1 r_2 - r_3),$$

where r_1, r_2 and r_3 are uniformly chosen random values. The recovery of $f_m(x_1, x_2)$ can be accomplished by computing $c_1 \cdot c_2 - c_3 - c_4$ where $c_1 = x_1 + r_1$, $c_2 = x_2 + r_2$, $c_3 = r_2 x_1 + r_3$ and $c_4 = r_1 x_2 + r_1 r_2 - r_3$. The simulation of $\hat{f_m}$ can be done perfectly by the simulator $\mathsf{Sim}(y; a_1, a_2, a_3) := (a_1, a_2, a_3, a_1 a_2 - y - a_3)$ where a_1, a_2 a_3 are random values.

Framework

To perform the private gaze estimation task in our scenario, we inspire from the framework as in [178] due to its efficiency compared to other approaches in the literature. The framework is proposed to compute the addition or multiplication of the input values of two input-parties in the function-party by utilizing randomized encoding. We utilize the multiplication operation over the eye landmark vectors to compute the dot product of these vectors to obtain kernel matrix of the SVR in a privacy-preserving way.

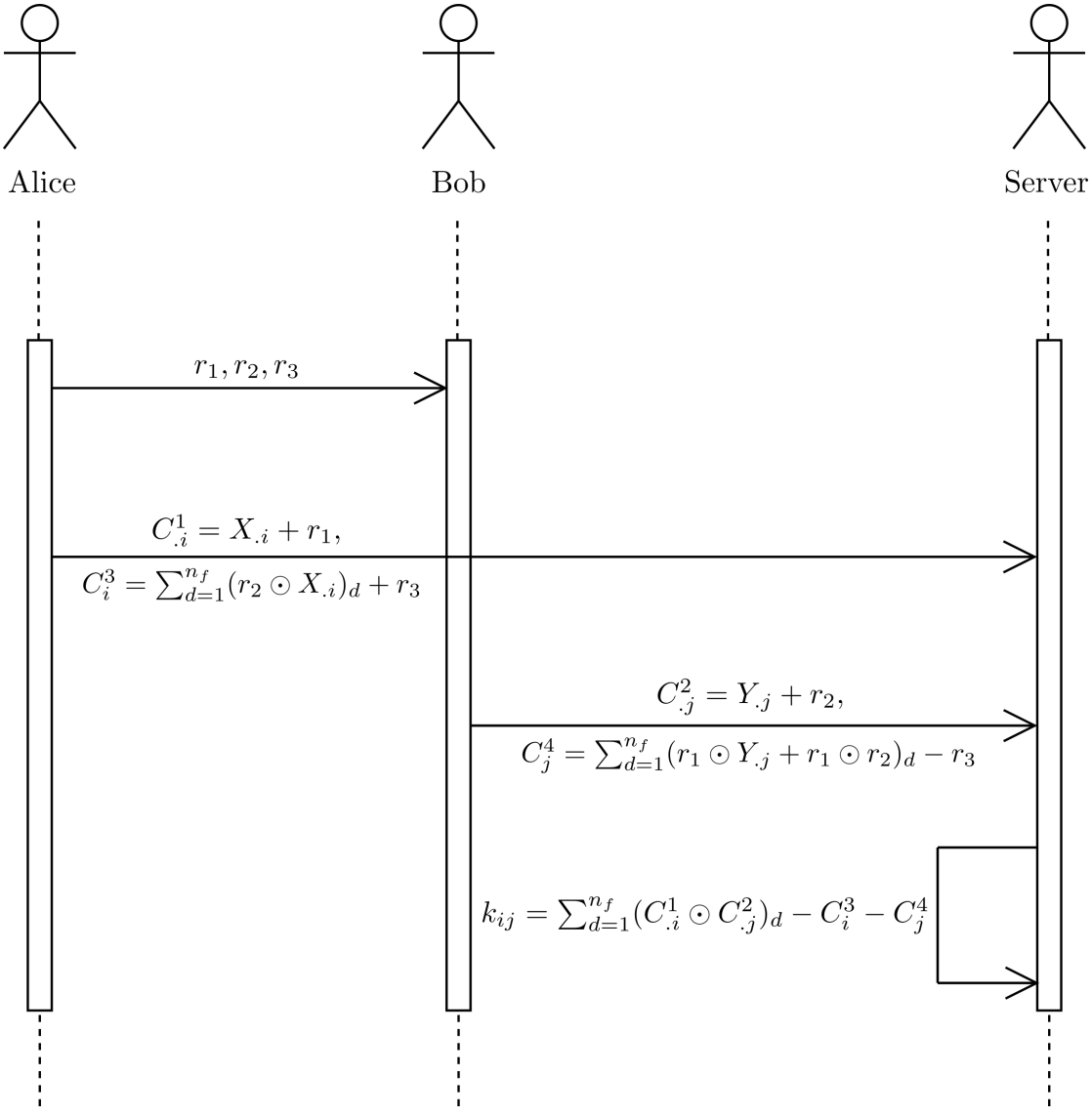


Figure B.14: Overall protocol execution.

We have two input-parties as Alice and Bob, having the eye landmark data as $X \in \mathbb{R}^{n_f \times n_a}$ and $Y \in \mathbb{R}^{n_f \times n_b}$ where n_a and n_b represent the number of samples in Alice and Bob, respectively, and n_f is the number of features. In addition to the input-parties, there exists a server that

trains a model on the data of the input-parties. $A_{.j}$ for any matrix A represents the j-th column of the corresponding matrix and "$\odot$" represents the element-wise multiplication of the vectors. As a first step, Alice creates a uniformly chosen random value $r_3 \in \mathbb{R}$ and two vectors $r_1, r_2 \in \mathbb{R}^{n_f}$ with uniformly chosen random values, which are used to encode the element-wise multiplication of the vectors and shares them with Bob. Afterwards, Bob computes $C^2_{.j} = Y_{.j} + r_2$ and $C^4_j = \sum_{d=1}^{n_f}(r_1 \odot Y_{.j} + r_1 \odot r_2)_d - r_3$, $\forall\, j \in \{1, \cdots, n_b\}$ where $C^2 \in \mathbb{R}^{n_f \times n_b}$ and $C^4 \in \mathbb{R}^{n_b}$. Meanwhile, Alice computes $C^1_{.i} = X_{.i} + r_1$ and $C^3_i = \sum_{d=1}^{n_f}(r_2 \odot X_{.i})_d + r_3$, $\forall\, i \in \{1, \cdots, n_a\}$ where $C^1 \in \mathbb{R}^{n_f \times n_a}$ and $C^3 \in \mathbb{R}^{n_a}$. Input-parties send their share of the encoding to the server with the gram matrix of their samples, which is the dot product among their samples. Then, the server computes the dot product between samples of Alice and Bob to complete the missing part of the gram matrix of all samples. To achieve this, the server computes $k_{ij} = \sum_{d=1}^{n_f}(C^1_{.i} \odot C^2_{.j})_d - C^3_i - C^4_j$, $\forall\, i \in \{1, \cdots, n_a\}$ and $\forall\, j \in \{1, \cdots, n_b\}$ where k_{ij} is the i-th row j-th column entry of the gram matrix between the samples of the input-parties. Once the server has all components of the gram matrix, it constructs the complete gram matrix K by simply concatenating the parts of it. In our solution, Alice and Bob send to the server (C^1, C^3) and (C^2, C^4) tuples, respectively. These components reveal nothing but only the gram matrix of the samples after decoding. Furthermore, the input-parties shuffle their raw data before the computation to avoid the possibility of private information leakage such as the behavior of the person due to the nature of the visual sequence information. The overall flow is summarized in Figure B.14.

After having the complete gram matrix for all samples that Alice and Bob have, the server uses it as a kernel matrix as if it was computed by the linear kernel function on pooled data. Additionally, it is also possible to compute a kernel matrix as if it was computed by the polynomial or radial basis kernel function (RBF) by utilizing the resulting gram matrix. As an example, the calculation of RBF from the gram matrix is as follows.

$$K(x, y) = \exp\left(-\frac{\left\| x \cdot x - 2x \cdot y + y \cdot y \right\|^2}{2\sigma^2}\right),$$

where "$\cdot$" represents the dot product of vectors, which is possible to obtain from the gram matrix, and σ is the parameter utilized to adjust the similarity level. Once the desired kernel matrix is computed, it is possible to train an SVR model by employing the computed kernel matrix to estimate the gaze. In the process of the computation of the dot product, the amount of data transferred among parties is $(n_f n_a + n_f n_b + n_a + n_b + 2n_f) \times d$ bytes where d is the size of one data unit.

B.2.5 Security Analysis

A semi-honest adversary who corrupts any of the input-parties cannot learn anything about the private inputs of the other input-party. During the protocol execution, two vectors of random values and a single random value are sent from Alice to Bob. The views of the input-

parties consist only of vectors with random values. Using these random values, it is not possible for one party to infer something about the other party's private inputs [178].

Theorem 3. A malicious adversary $\mathscr{A}$ corrupting the function-party learns nothing more than the result of gram matrix. It is computationally infeasible for $\mathscr{A}$ to infer any information about the input-parties' data X and Y as long as Perfect RE multiplication is semantically secure (Definition 5).

Proof. We first show the correctness of our solution. We assume $n_f = 2$ and encode the function $f_d(x,y) = x_1y_1 + x_2y_2$ over some finite ring R by the following DARE:

$$\begin{aligned}\hat{f}_d(x,y;r) = (&x_1 + r_{11}, y_1 + r_{12}, x_2 + r_{21}, y_2 + r_{22},\\ &r_{12}x_1 + r_{22}x_2 + r_3,\\ &r_{11}y_1 + r_{11}r_{12} + r_{21}y_2 + r_{21}r_{22} - r_3)\end{aligned}$$

Given an encoding $(c_1,c_2,c_3,c_4,c_5,c_6)$, $f_d(x,y)$ is recovered by computing $c_1c_2 + c_2c_4 + c_5 + c_6$.

By the concatenation lemma in [142], we can divide c_5 and c_6 into n_f shares by using n_f random values instead of a single r_3 value.

$$\begin{aligned}\hat{f}_d(x,y;r) = (&x_1 + r_{11}, y_1 + r_{12}, r_{12}x_1 + r_{13}, r_{11}y_1 + r_{11}r_{12} - r_{13},\\ &x_2 + r_{21}, y_2 + r_{22}, r_{22}x_2 + r_{23}, r_{21}y_2 + r_{21}r_{22} - r_{23})\end{aligned}$$

Given an encoding $(c_1,c_2,c_3,c_4,c_5,c_6,c_7,c_8)$,

$$\begin{aligned}\hat{f_m}(x_1,y_1;r) &= (c_1,c_2,c_3,c_4)\\ \hat{f_m}(x_2,y_2;r) &= (c_5,c_6,c_7,c_8)\end{aligned}$$

By the concatenation lemma in [142], $\hat{f}_d(x,y;r) = (\hat{f_m}(x_1,y_1;r), \hat{f_m}(x_2,y_2;r))$ perfectly encodes the function $f_d(x,y)$ if Perfect RE multiplication is semantically secure.

After showing the correctness, we analyze the security with the simulation paradigm. In the simulation paradigm, there is a simulator who generates the view of a party in the execution. A party's input and output must be given to the simulator to generate its view. Thus, security is formalized by saying that a party's view can be simulatable given its input and output and the parties learn nothing more than what they can derive from their input and prescribed output.

The function-party $\mathscr{F}$ does not have any input and output. A simulator $\mathscr{S}$ can generate the views of incoming messages received by $\mathscr{F}$. $\mathscr{S}$ creates four vectors $C^{1'}$,$C^{2'}$,$C^{3'}$,$C^{4'}$ with uniformly distributed random values using a pseudorandom number generator G'. Finally, $\mathscr{S}$ outputs $\{C^{1'},C^{2'},C^{3'},C^{4'}\}$.

In the execution of the protocol π, $\mathscr{A}$ receives four messages which are masked with uniformly random values generated using a pseudorandom number generator G. The view of $\mathscr{A}$ includes $\{C^1, C^2, C^3, C^4\}$. The distribution over G is statistically close to the distribution over G'. This implies that

$$\{\mathscr{S}(C^{1'}, C^{2'}, C^{3'}, C^{4'})\} \stackrel{c}{\equiv} \{view^{\pi}_{\mathscr{A}}(C^1, C^2, C^3, C^4)\}$$

□

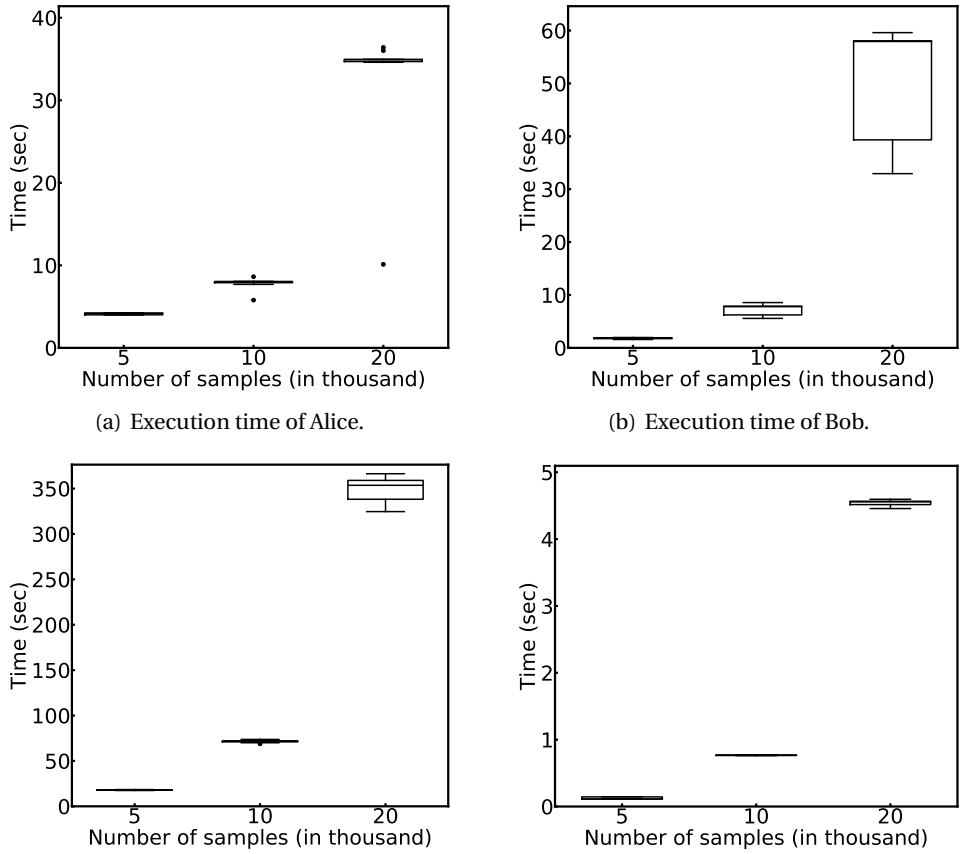
B.2.6 Results

To demonstrate the performance, we conduct experiments on a PC equipped with Intel Core i7-7500U with 2.70 GHz processor and 16 GB memory RAM. We employ varying sizes of eye landmark data, that are 5,000, 10,000 and 20,000 samples of which one-fifth is the test data and we split the data between the input-parties equally. The framework allows us to optimize the parameters of the model in the server without further communicating with the input-parties. Thanks to this, we utilize 5-fold cross-validation to optimize the parameters, which are the similarity adjustment parameter $\gamma \in \{2^{-3}, 2^{-2}, \cdots, 2^4\}$ of the Gaussian RBF kernel, the misclassification penalty parameter $C \in \{2^{-3}, 2^{-2}, \cdots, 2^3\}$, and the tolerance parameter $\epsilon \in \{0.005, 0.01, 0.05, 0.1, 0.5, 1\}$ of SVR. After parameter optimization, we repeat the experiment on varying sizes of eye landmark data with the optimal parameter set 10 times to assess the execution time. To evaluate the gaze estimation results, we employ mean angular error in the same way as in [74]. Table B.10 demonstrates the relationship between the dataset size and the resulting mean angular error. Since no additional noise is introduced during the computation of the kernel matrix, the results from our privacy-preserving framework are the same with the non-private ones. The mean angular errors are lower compared to the state-of-the-art gaze estimation techniques since we use synthetic data and fixed camera position during image rendering.

Table B.10: The mean angular errors for varying dataset sizes.

# of samples	Mean angular error
5k	0.21
10k	0.18
20k	0.17

The amount of time to train and test the models increases as the sample sizes increase since computation requirements get larger. The increment in the dataset size increases the communication cost among parties. The execution times of all parties for 10 runs with the optimal parameters are shown in Figure B.15. We also demonstrate the amount of time to predict the test samples, which corresponds to one-fifth of the total number of samples to emphasize the real-time working capabilities. In the experiment with 20,000 samples,



(a) Execution time of Alice.

(b) Execution time of Bob.

(c) Execution time of Server.

(d) Prediction time of Server.

Figure B.15: The execution time of (a) Alice, (b) Bob and (c) the server are given. We also demonstrate (d) the time required for the prediction of the test samples, which are 20% of the total number of samples in each case.

for instance, we spend ≈ 4.5 seconds to predict $4,000$ test samples, which corresponds to 1.125 ms per sample. When the current sampling frequencies of eye trackers are taken into consideration, it is possible to deploy and use the framework to estimate gaze if an optimized communication between the parties is established.

B.2.7 Conclusion

In this work, we utilized a framework based on randomized encoding to estimate human gaze in a privacy-preserving way and in real-time. Our solution can provide improved gaze estimation if input-parties want to use each other's data for different reasons such as to account for genetic structural differences in the eye region. None of the input-parties has the access to the eye landmark data of the others or the result of the computation in the function party, while the function-party cannot infer anything about the data of the input-parties. Temporal information of the visual scanpath, pupillary, or blinks cannot be reconstructed due to the

shuffling of the data, and lack of sensory information and direct access to the eye landmarks. Our solution works in real-time, hence it could be deployed along with HMDs for different use-cases and extended to similar eye tracking related problems if similar amount of features is used. To the best of our knowledge, this is the first work based on function-specific privacy models in the eye tracking domain. The number of parties is a limitation of our solution. Thus, as future work we will extend our work to a larger number of parties.

C Accessibility of Eye Tracking in VR in Daily Setups

This chapter includes the following publication:

1. **Efe Bozkir**, Shahram Eivazi, Mete Akgün, and Enkelejda Kasneci. Eye tracking data collection protocol for VR for remotely located subjects using blockchain and smart contracts. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR) Work-in-progress papers*, New York, NY, USA, 2020. IEEE. doi: 10.1109/AIVR50618.2020.00083.

C.1 Eye Tracking Data Collection Protocol for VR for Remotely Located Subjects using Blockchain and Smart Contracts

C.1.1 Abstract

Eye tracking data collection in the virtual reality context is typically carried out in laboratory settings, which usually limits the number of participants or consumes at least several months of research time. In addition, under laboratory settings, subjects may not behave naturally due to being recorded in an uncomfortable environment. In this work, we propose a proof-of-concept eye tracking data collection protocol and its implementation to collect eye tracking data from remotely located subjects, particularly for virtual reality using Ethereum blockchain and smart contracts. With the proposed protocol, data collectors can collect high quality eye tracking data from a large number of human subjects with heterogeneous socio-demographic characteristics. The quality and the amount of data can be helpful for various tasks in data-driven human-computer interaction and artificial intelligence.

C.1.2 Introduction

Over past decades, head-mounted display (HMD) technologies have taken advantage of innovations from imaging and eye tracking research to improve image quality and utility of user interfaces. To date, several consumer level HMDs have integrated eye trackers, providing opportunity for researchers to collect eye movement data for user behavior analysis and data-driven interaction.

In the virtual reality (VR) context, it has been shown that eye tracking is helpful for assessing human attention [212], detecting human stress [211], assessing cognitive load [169], predicting human future gaze locations [96], supporting evaluation and diagnosis of diseases [121], motion sickness detection [256], foveated rendering [91, 89], continuous authentication [27], gaze-based interaction [258], training [116], and redirected walking [257]. Many of these tasks are data-driven and require a large quantity of eye tracking data which are usually collected in laboratory settings. Subjects are frequently compensated with some amount of money or gifts for their participation. Two drawbacks of these settings are the lack of heterogeneity in socio-demographic characteristics of data collected subjects and potential for unnatural behaviors of subjects due to the constraints of the laboratory settings. While VR is a unique and controlled environment and requires dedicated hardware such as HMDs, as personal usage of such devices increases, we foresee that it should be possible to collect data from remotely located subjects, i.e., at their homes. Especially in situations such as COVID-19, this possibility could help experimental works continue in a remote setting. Currently, for crowd-sourcing or similar purposes, platforms such as Amazon Mechanical Turk[1] are used. While it is not possible to collect VR data with such platforms, for other types of data collection significant compensations are paid to manage the remote subjects' work. In addition, these

[1] https://www.mturk.com/

third-party platforms store and manage data. In fact, as eye tracking and movement data represent unique information about the subjects, the data manipulation possibility of the third parties should be prevented. Third parties should only act as a bridge between the data collector and the subjects in case there is no direct communication between the parties.

To overcome the disadvantages of the laboratory setting and enable remotely located subject participation in eye tracking experiments in the VR context, we propose a blockchain-based protocol on the Ethereum blockchain using smart contracts, where we use the blockchain for validation of data integrity and smart contract for compensation management. For this study, we focus mainly on collecting eye tracking data in VR environments as many modern HMDs come with integrated eye trackers. This means that subjects do not need any additional effort to integrate any sensor into their setup. It is relatively easier to control environmental configurations in HMDs when compared to other setups such as illumination and light-sources which may affect subject behaviors or eye movement patterns. However, the proposed protocol can also be used in similar setups as long as identical experiment configurations are guaranteed.

While the first prominent usage of the blockchains is Bitcoin [180] and most of the applications are in the financial domain, blockchains also draw attention of the human-computer interaction (HCI), eye tracking, and VR communities. Opportunities and challenges for the HCI and interaction design and the role of HCI community were discussed in [203] and [323], respectively. An augmented reality (AR)-based cryptocurrency wallet was developed in [202] to familiarize users with blockchain wallet services. In addition, GazeCoin is a cryptocurrency for VR/AR which is exchanged between content makers, advertisers, and the users [204]. Apart from the financial use-cases, due to their immutability blockchains are used as notary. Additionally, Ethereum platform brings the smart contract [324] concept to the blockchains [181]. One of the straightforward usages of smart contracts is escrow services. For the remote purchase of goods, buyer and seller parties use the smart contracts without trusting one another and a trusted centralized party during the escrow. The smart contracts that are deployed on the blockchains distribute the money to the parties once buyer and seller parties fulfill their obligations in the remote purchase. In our protocol, we treat recorded eye tracking data as digital good so that compensation distribution is done by the smart contracts. To assure that the recorded data are not altered by the subjects, the hash of the recorded data using white-box cryptography [182] is stored in the blockchain, which enables the blockchain as a notary for data integrity. Our major contributions are as follows.

- A blockchain-based eye tracking data collection protocol for remotely located subjects that can be used for eye tracking experiments in VR, which presents the opportunity to collect data from a various number of subjects.
- Delegation of mutual trust issues for compensation management and integrity of the recorded eye tracking data to smart contracts and blockchains, respectively.
- Elimination of the centralized third parties for compensation management, data collection and manipulation, which is optimal from a privacy perspective.

C.1.3 Preliminary Definitions

As our protocol consists of interdisciplinary work from different domains such as virtual reality, blockchains, and cryptography, we provide some definitions that are used throughout the paper.

Blockchain [180]: An immutable ledger that consists of a chain of blocks that keeps records of transactions, maintained by several machines in a peer-to-peer network. Each block consists of a timestamp, transaction data, and the cryptographic hash of the previous block. As each block consists of the cryptographic hash of the one prior, immutability is automatically preserved unless one party has the majority of the computational power.

Ethereum [181]: Public, open-source, blockchain-based, and smart contract supporting distributed platform.

Ether (ETH) [181]: The cryptocurrency of the Ethereum platform.

Smart Contract [181]: A self-executing, irreversible, and transparent contract between buyer and seller, implemented in the code.

White-box cryptography [179]: "Software protection technology which allows for the application of cryptographic operations without revealing any critical information such as secret keys."

C.1.4 Protocol

In this section, we discuss our protocol and its flow, assumptions, and details of the implementation.

Flow

Our proposed protocol consists of two parties as data collector and subjects. The data collector is responsible for providing the VR application for eye tracking data collection and subjects are tasked with carrying out the experiment and providing the recorded eye tracking data. At the end of a valid experiment, subjects are compensated for their participation. Let us assume that each subject is compensated with X unit of ETH for the valid data recorded from an experiment session. A relevant amount can be set for compensation depending on the experiment.

Figure C.1 shows the overall flow and short descriptions of each step of the protocol. As the **step 1**, subjects fetch the application from the data collector and carry out the experiment. While the content of the stimuli changes depending on the use-case, the VR application validates the eye tracking data quality at the end of each experimental session by using tracking rates or confidence intervals that are provided by the eye tracker. If the recorded eye

tracking data are too noisy, subjects are not supposed to send the data to the data collector, where they are informed by the VR application. This obligation forces the subjects to follow the instructions of the VR experiment, such as eye tracker calibration, carefully while the data collector obtains better quality data in the end. After the validation success, the VR application calculates the hash output of the recorded eye tracking data and saves it. Saving the hash output is required for assessing the data integrity; however, adversarial subjects can easily find out the hashing algorithm using the executable of the VR application on their own devices. Therefore, we opt for a white-box [179, 182] paradigm for calculating the data hash. In the white-box paradigm, the adversary is supposed to have visibility of the inputs, outputs, and other intermediate steps. White-box cryptography achieves protection of confidential information such as secret keys while keeping the application semantically the same. Even if adversaries infer the hash function, due to the lack of secret key, it is not possible to generate a hash output for altered data. Consequently, subjects are obliged to behave honestly, where honest behavior means not altering the recorded data. In the end of the first step, once the recorded data is validated and hash value is saved, the subjects are informed by the VR application that the recorded eye tracking data is reportable.

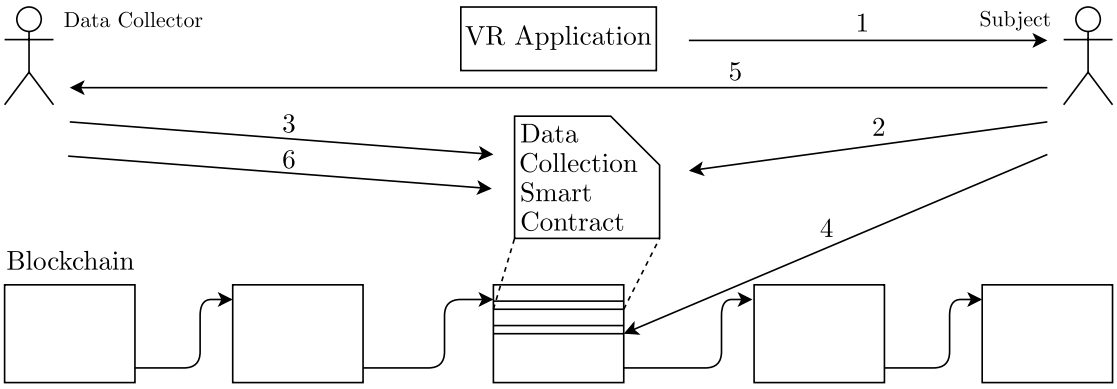


Figure C.1: Blockchain-based protocol and its steps. (1) Subject fetches the application and carries out the experiment. (2) Subject initiates the smart contract. (3) Data collector confirms the contract creation and stakes. (4) Subject stores the recorded data hash in blockchain. (5) Subject transfers the recorded data to the data collector. (6) Data collector confirms the data collection.

As the **step 2**, the subjects initiate the smart contract and stake double the amount of compensation, which is $2X$ ETH for our case. Staking double the amount of compensation that they will obtain from the smart contract forces subjects to act honestly; otherwise, they lose the amount that they stake. As the **step 3**, the data collector confirms the data collection and stakes the same amount as the subject, which is $2X$ ETH to the smart contract. While the compensation is X ETH per experiment, the data collector is supposed to stake double the amount of compensation so that it also becomes an obligation to behave honestly. Otherwise, the doubled amount of compensation will be lost without obtaining the recorded data. As the **step 4**, the subjects store the hash output that is reported by the VR application in the blockchain and, as the **step 5**, they send the recorded data along with the transaction hash of

the transaction for storing the data hash in the blockchain to the data collector. If subjects try to alter the data, the hash in the blockchain and the altered data will not match and it will be discovered by the data collector. As the **step 6**, the data collector obtains the recorded eye tracking data and transaction hash of the data hash and checks whether or not the obtained data and the hash provided by the subjects overlap using the hash function that is implemented in the VR application and secret keys. If the reported data and hash value stored in the blockchain overlap, the data collector confirms the smart contract and that the obtained data are valid. Then, the smart contract automatically distributes $3X$ and X ETH to the subject and the data collector, respectively. In the end, each subject earns X unit of ETH for participation in the experiment, where the data collector obtains the recorded eye tracking data. Due to the immutable nature of blockchains and smart contracts, none of the parties can alter the values in the blockchain and behave as an adversary.

In the protocol, as both parties stake more than the amount they are supposed to spend or earn, they have to act honestly in order to achieve successful data collection and compensation distribution, otherwise data collection is not finalized and parties lose the amount they stake. In particular, the subjects have to stake double the amount of compensation that they will receive whereas the data collectors have to stake double the amount of compensation that they will give. Since the smart contracts are immutable and stored in the blockchain, a third-party application is not needed for compensation distribution or data manipulation, which is useful from a privacy preservation point of view.

Assumptions

We have three main assumptions in our protocol. Firstly, validation of the quality of the recorded eye tracking data is automatically completed by the VR application at the end of each experiment by using metrics such as tracking ratio or confidence levels reported by the eye tracker. Due to poor calibration for eye tracking, removal of the head-mounted display (HMD) in the middle of experiment, or similar reasons, recorded eye tracking data may have an extensive amount of noise level. Instead of cleaning data offline extensively after the experiments, our protocol assumes that data validity is checked at the end of each experiment by the VR application and the application informs the subjects whether the quality of the data is valid and reportable.

Secondly, the recorded eye tracking data is hashed using white-box cryptography and stored at the end of the experiment by the VR application to be stored in the blockchain for validation of the data integrity. In traditional eye tracking experiments, subjects participate in the experiments on the devices that are provided by the data collectors. However, in the remotely located subject participation, subjects run the applications on their own devices. Therefore, they have direct access to the provided application and if any adversarial subject analyzes the binary implementation of an application that does not use white-box paradigm, they can easily infer the used hash function and generate hash output for fake data. On the contrary, when using white-box cryptography, the secret keys are not leaked even if adversaries analyze

the binary implementation. Even if an adversary infers the hash function, a hash output for fake data cannot be generated without secret keys. Therefore, white-box paradigm is used by the VR application. If subjects alter the recorded data or send fake data to the data collector, the generated hash value will not match the recorded data, which leads subjects to lose their staked compensation in the smart contract.

Lastly, as our protocol does not use any centralized third party, a secure direct communication is needed for exchanging the application and the recorded data between the data collector and subjects. In case it is not available, a bridging third party only for communication purposes can be implemented.

Implementation

We select the Ethereum platform for our proof-of-concept due to its public blockchain, relatively higher number of nodes, and status as one of the most mature platforms in the blockchain domain. However, any blockchain-based platform that supports smart contracts can be opted in.

We implement the blockchain related part of the protocol, particularly the steps 2, 3, 4, and 6 discussed in Section C.1.4, using Solidity[2] and a simple purchase smart contract [325] on the Ropsten Testnet of the Ethereum platform. In the beginning of the data collection, both the data collector and the subject hold 1 ETH in their wallets. We select the compensation amount as 0.025 ETH. For the calculation of the hash output of the recorded eye tracking data, we use synthetic data; however, any eye tracker integrated to modern HMDs can be used in a real-world implementation. The hash value of the data is calculated using Keyed-Hashing for Message Authentication (HMAC) [326] and Secure Hash Algorithm3-512 (SHA3-512) [327] as it is possible to have white-box implementation of the HMAC. The calculated hash value is stored in the input data field of a self transaction from the subject. After the protocol execution, the data collector and the subject hold ≈ 0.975 and ≈ 1.025 ETH when the transaction fees are subtracted, respectively. The smart contract, overall procedure, the data collector, and the subject parties are available on the Ropsten Testnet via following link: https://ropsten.etherscan.io/address/0x0e937a4a4618dd8d5a12ec4a9f8fd61d6bfd13e4.

In the above link, there are three transactions in chronological order that correspond to steps 2, 3, and 6 of our protocol. The subject (address starting with $0x89$) and the data collector (address starting with $0x44$) of our implementation are available in the source of the first and the second transactions of the smart contract, respectively. There are three transactions in the subject address. The first and second transactions are for depositing the test ETH and initiating the smart contract, respectively. The third transaction in the subject address is a self transaction and corresponds to step 4 of our protocol. In the "Input Data" field of the self-transaction, the calculated data hash is available.

[2] https://docs.soliditylang.org/

C.1.5 Conclusion and Discussion

We proposed a blockchain-based protocol for collecting eye tracking data in VR from remotely located subjects. As eye tracking experiments are usually conducted in laboratory settings with a limited number of subjects from similar backgrounds in terms of socio-demographic characteristics, it is a challenge to draw generic data-driven conclusions. Due to the laboratory settings, subjects may not behave naturally. While our protocol overcomes the drawbacks of the traditional eye tracking data collection setups without needing a centralized third party for data collection and compensation management, it also creates an opportunity to carry out the data collection anonymously, which is optimal for the privacy of subjects. We focused on the eye tracking data collection in VR setups as validation of the eye tracking data and generation of the controlled environments with VR can be done easily. In addition, current availability of eye tracker integrated HMDs in the consumer market supports our protocol for VR and eye tracking data; however, the proposed protocol may be useful for other types of eye trackers, sensors, or environments as long as identical configurations between subjects can be generated. In contrast to traditional eye tracking experiments, subject consent, additional questionnaire, or similar information should be collected digitally using our protocol. Our protocol may also require an application-level effort to have one-to-one mapping between subjects and experiments.

As future work, we plan to have an end-to-end implementation of our protocol along with a real VR application and HMD-integrated eye tracker. In addition, while transactions are applied anonymously on the public blockchains, it is possible to track them. Recent work on eye tracking, HCI, and VR [264, 143, 263, 262, 176] emphasize the importance of privacy preservation. Combining privacy-preserving methods with our protocol remains as part of future work.

Acknowledgments

E.B. thanks Batuhan Sarıoğlu for useful discussions on blockchains.

Bibliography

[1] Tim Bradshaw. Oculus seeks to boost Rift virtual reality headset with price cut. Online, Financial Times, Mar 2017. URL https://www.ft.com/content/0fb0fa8e-fe20-11e6-96f8-3700c5664d30. Accessed: 2021-09-29.

[2] Grand View Research, Inc. Virtual reality headset market size, share & trends analysis report by end-device (Low-end, high-end), by product type (Standalone, smartphone-enabled), by application (Gaming, education), and segments forecasts, 2021 - 2028. Virtual Reality Headset Market Share Report, 2021-2028, Mar 2021. URL https://www.grandviewresearch.com/industry-analysis/virtual-reality-vr-headset-market.

[3] Google VR. Google cardboard. Online, 2021. URL https://arvr.google.com/cardboard/. Accessed: 2021-09-20.

[4] HTC Corporation. HTC Vive Pro Eye. Online, 2011-2021. URL https://www.vive.com/eu/product/vive-pro-eye/overview/. Accessed: 2021-09-20.

[5] Steven M. LaValle. *Virtual Reality*. Cambridge University Press, Cambridge, United Kingdom, 2020. URL http://lavalle.pl/vr/.

[6] Warner Bros. Entertainment Inc. The Matrix Trilogy. Online, 2021. URL https://www.warnerbros.co.uk/movies/matrix-trilogy. Accessed: 2021-09-23.

[7] Samuel Thomas von Sömmerring. *Über das Organ der Seele*. Nicolovius, Königsberg, 1796. Afterword by Immanuel Kant.

[8] Antonin Artaud. *Le théâtre et son double*. Gallimard, Paris, France, 1938. ISBN 978-2-07-020285-0.

[9] Myron W. Krueger. *Artificial Reality*. Addison-Wesley, Boston, MA, USA, 1983. ISBN 978-0-201-04765-3.

[10] Michael Heim. *The Metaphysics of Virtual Reality*. Oxford University Press, Oxford, United Kingdom, 1994. ISBN 978-0-19-509258-5.

[11] Nicholas C. Burbules. Rethinking the virtual. *E-Learning and Digital Media*, 1(2): 162–183, 2004. doi: 10.2304/elea.2004.1.2.2.

[12] Joe Durbin. NVIDIA estimates VR is 20 years away from resolutions that match the human eye. Online, May 2017. URL https://uploadvr.com/nvidia-estimates-20-years-away-vr-eye-quality-resolution/. Accessed: 2021-09-15.

[13] Tom Alexander Garner, Wendy Powell, and Vaughan Powell. *Everyday Virtual Reality*, pages 1–9. Springer, Cham, Switzerland, 2018. doi: 10.1007/978-3-319-08234-9_259-1.

[14] Adalberto L. Simone, Benjamin Beyers, Svetlana Bialkova, Misha Sra, Jan Gugenheimer, Augusto Esteves, Xuesong Zhang, and Jihae Han. WEVR 2021, 7th Workshop on everyday virtual reality at IEEE VR. Online, Apr 2021. URL https://wevr.adalsimeone.me/. Accessed: 2021-09-15.

[15] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. The experience of presence: Factor analytic insights. *Presence*, 10(3):266–281, 2001. doi: 10.1162/105474601300343603.

[16] Matthew Lombard, Theresa Bolmarcich, and Lisa Weinstein. Measuring presence: The temple presence inventory. In *Proceedings of the 12th Annual International Workshop on Presence*, pages 1–15, Los Angeles, CA, USA, 2009. The International Society for Presence Research. ISBN 978-0-9792217-3-6.

[17] Robert S. Kennedy, Norman E. Lane, Kevin S. Berbaum, and Michael G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology*, 3(3):203–220, 1993. doi: 10.1207/s15327108ijap0303_3.

[18] Sandra G. Hart. Nasa-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, 2006. doi: 10.1177/154193120605000909.

[19] Wolfgang Fuhl, Marc Tonsen, Andreas Bulling, and Enkelejda Kasneci. Pupil detection for head-mounted eye tracking in the wild: an evaluation of the state of the art. *Machine Vision and Applications*, 27(8):1275–1288, 2016. doi: 10.1007/s00138-016-0776-4.

[20] Jackson Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91(2):276–292, 1982. doi: 10.1037/0033-2909.91.2.276.

[21] Moritz Kassner, William Patera, and Andreas 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*, pages 1151–1160, New York, NY, USA, 2014. ACM. doi: 10.1145/2638728.2641695.

[22] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in VR: How do people explore virtual environ-

ments? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1633–1642, 2018. doi: 10.1109/TVCG.2018.2793599.

[23] Brendan David-John, Candace Peacock, Ting Zhang, T. Scott Murdison, Hrvoje Benko, and Tanya R. Jonker. Towards gaze-based prediction of the intent to interact in virtual reality. In *ACM Symposium on Eye Tracking Research and Applications*, pages 2:1–2:7, New York, NY, USA, 2021. ACM. doi: 10.1145/3448018.3458008.

[24] Zhiming Hu, Andreas Bulling, Sheng Li, and Guoping Wang. FixationNet: Forecasting eye fixations in task-oriented virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2681–2690, 2021. doi: 10.1109/TVCG.2021.3067779.

[25] John G. Daugman. High confidence visual recognition of persons by a test of statistical independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11): 1148–1161, 1993. doi: 10.1109/34.244676.

[26] Ajay Kumar and Arun Passi. Comparison and combination of iris matchers for reliable personal authentication. *Pattern Recognition*, 43(3):1016–1026, 2010. doi: 10.1016/j.patcog.2009.08.016.

[27] Yongtuo Zhang, Wen Hu, Weitao Xu, Chun Tung Chou, and Jiankun Hu. Continuous authentication using eye movement response of implicit visual stimuli. *ACM Interactive Mobile Wearable Ubiquitous Technologies*, 1(4):177:1–177:22, 2018. doi: 10.1145/3161410.

[28] Julian Steil, Marion Koelle, Wilko Heuten, Susanne Boll, and Andreas Bulling. PrivacEye: Privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features. In *ACM Symposium on Eye Tracking Research & Applications*, pages 26:1–26:10, New York, NY, USA, 2019. ACM. doi: 10.1145/3314111.3319913.

[29] Ceenu George, Mohamed Khamis, Emanuel Zezschwitz, Marinus Burger, Henri Schmidt, Florian Alt, and Heinrich Hussmann. Seamless and secure VR: Adapting and evaluating established authentication systems for virtual reality. In *Network and Distributed System Security Symposium (NDSS)*, Reston, VA, USA, 2017. The Internet Society. doi: 10.14722/usec.2017.23028.

[30] European Commission. Directorate-General for Justice and Consumers. *The GDPR: New Opportunities, New Obligations : What Every Business Needs to Know about the EU's General Data Protection Regulation*. Publications Office of the European Union, Luxembourg, 2018. ISBN 978-92-79-79430-8.

[31] State of California Department of Justice. California Consumer Privacy Act, 2018. URL https://oag.ca.gov/privacy/ccpa.

[32] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998. doi: 10.1037/0033-2909.124.3.372.

[33] Ecenaz Alemdag and Kursat Cagiltay. A systematic review of eye tracking research on multimedia learning. *Computers & Education*, 125:413–428, 2018. doi: 10.1016/j.compedu.2018.06.023.

[34] Laura A. Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 478–479, New York, NY, USA, 2004. ACM. doi: 10.1145/1008992.1009079.

[35] Oskar Palinko, Andrew L. Kun, Alexander Shyrokov, and Peter Heeman. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 141–144, New York, NY, USA, 2010. ACM. doi: 10.1145/1743666.1743701.

[36] Philip S. Holzman, Leonard R. Proctor, and Dominic W. Hughes. Eye-tracking patterns in schizophrenia. *Science*, 181(4095):179–181, 1973. doi: 10.1126/science.181.4095.179.

[37] Kathy Conklin, Ana Pellicer-Sánchez, and Gareth Carrol. *Eye-tracking: A guide for applied linguistics research*. Cambridge University Press, Cambridge, United Kingdom, 2018. ISBN 978-1-108-41535-4.

[38] Jennifer Romano Bergstrom and Andrew Jonathan Schall. *Eye tracking in user experience design*. Elsevier, Waltham, MA, USA, 2014. ISBN 978-0-12-408138-3.

[39] Maria Laura Mele and Stefano Federici. Gaze and eye-tracking solutions for psychological research. *Cognitive Processing*, 13(1):261–265, 2012. doi: 10.1007/s10339-012-0499-z.

[40] Halszka Jarodzka, Kenneth Holmqvist, and Hans Gruber. Eye tracking in educational science: Theoretical frameworks and research agendas. *Journal of Eye Movement Research*, 10(1):1–18, 2017. doi: 10.16910/jemr.10.1.3.

[41] Unaizah Obaidellah, Mohammed Al Haek, and Peter C.-H. Cheng. A survey on the usage of eye-tracking in computer programming. *ACM Computing Surveys*, 51(1):5:1–5:58, 2018. doi: 10.1145/3145904.

[42] Michel Wedel and Rik Pieters. Eye tracking for visual marketing. *Foundations and Trends in Marketing*, 1(4):231–320, 2008. doi: 10.1561/1700000011.

[43] Junichi Shimizu and George Chernyshov. Eye movement interactions in google cardboard using a low cost EOG setup. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pages 1773–1776, New York, NY, USA, 2016. ACM. doi: 10.1145/2968219.2968274.

[44] Anton Mølbjerg Eskildsen and Dan Witzner Hansen. Analysis of iris obfuscation: Generalising eye information processes for privacy studies in eye tracking. In *ACM Symposium on Eye Tracking Research and Applications*, pages 2:1–2:10, New York, NY, USA, 2021. ACM. doi: 10.1145/3448017.3457385.

[45] Priya Kansal and Sabarinathan Devanathan. EyeNet: Attention based convolutional encoder-decoder network for eye region segmentation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3688–3693, New York, NY, USA, 2019. IEEE. doi: 10.1109/ICCVW.2019.00456.

[46] Aayush K. Chaudhary, Rakshit Kothari, Manoj Acharya, Shusil Dangi, Nitinraj Nair, Reynold Bailey, Christopher Kanan, Gabriel Diaz, and Jeff B. Pelz. RITnet: Real-time semantic segmentation of the eye for gaze tracking. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3698–3702, New York, NY, USA, 2019. IEEE. doi: 10.1109/ICCVW.2019.00568.

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, Switzerland, 2015. Springer. doi: 10.1007/978-3-319-24574-4_28.

[48] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, New York, NY, USA, 2017. IEEE. doi: 10.1109/CVPR.2017.243.

[49] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Eye-MMS: Miniature multi-scale segmentation network of key eye-regions in embedded applications. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3665–3670, New York, NY, USA, 2019. IEEE. doi: 10.1109/ICCVW.2019.00452.

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, New York, NY, USA, 2017. IEEE. doi: 10.1109/ICCV.2017.244.

[51] Wolfgang Fuhl, David Geisler, Wolfgang Rosenstiel, and Enkelejda Kasneci. The applicability of cycle GANs for pupil and eyelid segmentation, data generation and image refinement. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4406–4415, New York, NY, USA, 2019. IEEE. doi: 10.1109/ICCVW.2019.00541.

[52] Rakshit S. Kothari, Aayush K. Chaudhary, Reynold J. Bailey, Jeff B. Pelz, and Gabriel J. Diaz. EllSeg: An ellipse segmentation framework for robust gaze tracking. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2757–2767, 2021. doi: 10.1109/TVCG.2021.3067765.

[53] Jonathan Perry and Amanda S. Fernandez. EyeSeg: Fast and efficient few-shot semantic segmentation. In *Computer Vision – ECCV 2020 Workshops*, pages 570–582, Cham, Switzerland, 2020. Springer. doi: 10.1007/978-3-030-66415-2_37.

[54] Yiru Shen, Oleg Komogortsev, and Sachin S. Talathi. Domain adaptation for eye segmentation. In *Computer Vision – ECCV 2020 Workshops*, pages 555–569, Cham, Switzerland, 2020. Springer. doi: 10.1007/978-3-030-66415-2_36.

[55] Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. D-ID-Net: Two-stage domain and identity learning for identity-preserving image generation from semantic segmentation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3677–3682, New York, NY, USA, 2019. IEEE. doi: 10.1109/ICCVW.2019.00454.

[56] Dongheng Li, David Winfield, and Derrick J. Parkhurst. Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, pages 79–79, New York, NY, USA, 2005. IEEE. doi: 10.1109/CVPR.2005.531.

[57] Lech Świrski and Neil A. Dodgson. A fully-automatic, temporal approach to single camera, glint-free 3D eye model fitting. In *Proceedings of ECEM 2013*, 2013. URL http://www.cl.cam.ac.uk/research/rainbow/projects/eyemodelfit/.

[58] Wolfgang Fuhl, Thomas Kübler, Katrin Sippel, Wolfgang Rosenstiel, and Enkelejda Kasneci. ExCuSe: Robust pupil detection in real-world scenarios. In *Computer Analysis of Images and Patterns*, pages 39–51, Cham, Switzerland, 2015. Springer. doi: 10.1007/978-3-319-23192-1_4.

[59] Amir-Homayoun Javadi, Zahra Hakimi, Morteza Barati, Vincent Walsh, and Lili Tcheang. SET: a pupil detection method using sinusoidal approximation. *Frontiers in Neuroengineering*, 8:4, 2015. doi: 10.3389/fneng.2015.00004.

[60] Wolfgang Fuhl, Thiago C. Santini, Thomas Kübler, and Enkelejda 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*, pages 123–130, New York, NY, USA, 2016. ACM. doi: 10.1145/2857491.2857505.

[61] Wolfgang Fuhl, Thiago Santini, Carsten Reichert, Daniel Claus, Alois Herkommer, Hamed Bahmani, Katharina Rifai, Siegfried Wahl, and Enkelejda Kasneci. Non-intrusive practitioner pupil detection for unmodified microscope oculars. *Computers in Biology and Medicine*, 79:36–44, 2016. doi: 10.1016/j.compbiomed.2016.10.005.

[62] Wolfgang Fuhl, Shahram Eivazi, Benedikt Hosp, Anna Eivazi, Wolfgang Rosenstiel, and Enkelejda 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*, pages 48:1–48:5, New York, NY, USA, 2018. ACM. doi: 10.1145/3204493.3204558.

[63] Wolfgang Fuhl, David Geisler, Thiago Santini, Tobias Appel, Wolfgang Rosenstiel, and Enkelejda 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*, pages 8:1–8:6, New York, NY, USA, 2018. ACM. doi: 10.1145/3204493.3204559.

[64] Thiago Santini, Wolfgang Fuhl, and Enkelejda Kasneci. PuReST: Robust pupil tracking for real-time pervasive eye tracking. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 61:1–61:5, New York, NY, USA, 2018. ACM. doi: 10.1145/3204493.3204578.

[65] Thiago Santini, Diederick C. Niehorster, and Enkelejda 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*, pages 17:1–17:10, New York, NY, USA, 2019. ACM. doi: 10.1145/3314111.3319835.

[66] Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. Learning-by-synthesis for appearance-based 3D gaze estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1821–1828, New York, NY, USA, 2014. IEEE. doi: 10.1109/CVPR.2014.235.

[67] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, New York, NY, USA, 2015. IEEE. doi: 10.1109/CVPR.2015.7299081.

[68] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. A 3d morphable eye region model for gaze estimation. In *Computer Vision – ECCV 2016*, pages 297–313, Cham, Switzerland, 2016. Springer. doi: 10.1007/978-3-319-46448-0_18.

[69] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 131–138, New York, NY, USA, 2016. ACM. doi: 10.1145/2857491.2857492.

[70] Wolfgang Fuhl, Thiago Santini, Gjergji Kasneci, and Enkelejda Kasneci. PupilNet: Convolutional neural networks for robust pupil detection. CoRR, 2016. URL https://arxiv.org/abs/1601.04902v1.

[71] Wolfgang Fuhl, Thiago Santini, Gjergji Kasneci, Wolfgang Rosenstiel, and Enkelejda Kasneci. PupilNet v2.0: Convolutional neural networks for CPU based real time robust pupil detection. CoRR, 2017. URL https://arxiv.org/abs/1711.00112v1.

[72] Kang Wang, Rui Zhao, and Qiang Ji. A hierarchical generative model for eye image synthesis and eye gaze estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 440–448, New York, NY, USA, 2018. IEEE. doi: 10.1109/CVPR.2018.00053.

[73] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *Computer Vision – ECCV 2018*, pages 741–757, Cham, Switzerland, 2018. Springer. doi: 10.1007/978-3-030-01261-8_44.

[74] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 21:1–21:10, New York, NY, USA, 2018. ACM. doi: 10.1145/3204493.3204545.

[75] Joohwan Kim, Michael Stengel, Alexander Majercik, Shalini De Mello, David Dunn, Samuli Laine, Morgan McGuire, and David 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*, pages 550:1–550:12, New York, NY, USA, 2019. ACM. doi: 10.1145/3290605.3300780.

[76] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7314–7324, New York, NY, USA, 2020. IEEE. doi: 10.1109/CVPR42600.2020.00734.

[77] Michael Stengel, Steve Grogorick, Martin Eisemann, Elmar Eisemann, and Marcus A. Magnor. An affordable solution for binocular eye tracking and calibration in head-mounted displays. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pages 15–24, New York, NY, USA, 2015. ACM. doi: 10.1145/2733373.2806265.

[78] Scott W. Greenwald, Luke Loreti, Markus Funk, Ronen Zilberman, and Pattie Maes. Eye gaze tracking with google cardboard using purkinje images. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*, pages 19–22, New York, NY, USA, 2016. ACM. doi: 10.1145/2993369.2993407.

[79] Thomas C Kübler, Enkelejda Kasneci, and Wolfgang Rosenstiel. SubsMatch: Scanpath similarity in dynamic scenes based on subsequence frequencies. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 319–322, New York, NY, USA, 2014. ACM. doi: 10.1145/2578153.2578206.

[80] David Geisler, Nora Castner, Gjergji Kasneci, and Enkelejda Kasneci. A MinHash approach for fast scanpath classification. In *ACM Symposium on Eye Tracking Research and Applications*, pages 4:1–4:9, New York, NY, USA, 2020. ACM. doi: 10.1145/3379155.3391325.

[81] Marcel A. Just and Patricia A. Carpenter. Eye fixations and cognitive processes. *Cognitive Psychology*, 8(4):441–480, 1976. doi: 10.1016/0010-0285(76)90015-3.

[82] Joseph H. Goldberg and Xerxes P. Kotval. Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, 24(6):631–645, 1999. doi: 10.1016/S0169-8141(98)00068-7.

[83] Joseph H. Goldberg, Mark J. Stimson, Marion Lewenstein, Neil Scott, and Anna M. Wichansky. Eye tracking in web search tasks: Design implications. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, pages 51–58, New York, NY, USA, 2002. ACM. doi: 10.1145/507072.507082.

[84] Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, pages 71–78, New York, NY, USA, 2000. ACM. doi: 10.1145/355017.355028.

[85] Ioannis Agtzidis, Mikhail Startsev, and Michael Dorr. 360-Degree video gaze behaviour: A ground-truth data set and a classification algorithm for eye movements. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1007–1015, New York, NY, USA, 2019. ACM. doi: 10.1145/3343031.3350947.

[86] Enkelejda Tafaj, Gjergji Kasneci, Wolfgang Rosenstiel, and Martin Bogdan. Bayesian online clustering of eye movement data. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 285–288, New York, NY, USA, 2012. ACM. doi: 10.1145/2168556.2168617.

[87] Thiago Santini, Wolfgang Fuhl, Thomas Kübler, and Enkelejda Kasneci. Bayesian identification of fixations, saccades, and smooth pursuits. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 163–170, New York, NY, USA, 2016. ACM. doi: 10.1145/2857491.2857512.

[88] Tobias Appel, Peter Gerjets, Stefan Hoffman, Korbinian Moeller, Manuel Ninaus, Christian Scharinger, Natalia Sevcenko, Franz Wortha, and Enkelejda Kasneci. Cross-task and cross-participant classification of cognitive load in an emergency simulation game. *IEEE Transactions on Affective Computing*, pages 1–1, 2021. doi: 10.1109/TAFFC.2021.3098237.

[89] Xiaoxu Meng, Ruofei Du, and Amitabh Varshney. Eye-dominance-guided foveated rendering. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1972–1980, 2020. doi: 10.1109/TVCG.2020.2973442.

[90] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics*, 35(6):179:1–179:12, 2016. doi: 10.1145/2980179.2980246.

[91] Elena Arabadzhiyska, Okan Tarhan Tursun, Karol Myszkowski, Hans-Peter Seidel, and Piotr Didyk. Saccade landing position prediction for gaze-contingent rendering. *ACM Transactions on Graphics*, 36(4):50:1–50:12, 2017. doi: 10.1145/3072959.3073642.

[92] Chih-Fan Hsu, Anthony Chen, Cheng-Hsin Hsu, Chun-Ying Huang, Chin-Laung Lei, and Kuan-Ta Chen. Is foveated rendering perceivable in virtual reality? exploring the efficiency and consistency of quality assessment methods. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 55–63, New York, NY, USA, 2017. ACM. doi: 10.1145/3123266.3123434.

[93] Henry Griffith and Oleg Komogortsev. A shift-based data augmentation strategy for improving saccade landing point prediction. In *ACM Symposium on Eye Tracking Research and Applications*, pages 20:1–20:6, New York, NY, USA, 2020. ACM. doi: 10.1145/3379157.3388935.

[94] Petr Kellnhofer, Piotr Didyk, Karol Myszkowski, Mohamed M. Hefeeda, Hans-Peter Seidel, and Wojciech Matusik. GazeStereo3D: Seamless disparity manipulations. *ACM Transactions on Graphics*, 35(4):68:1–68:13, 2016. doi: 10.1145/2897824.2925866.

[95] Martin Weier, Thorsten Roth, André Hinkenjann, and Philipp Slusallek. Predicting the gaze depth in head-mounted displays using multiple feature regression. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 19:1–19:9, New York, NY, USA, 2018. ACM. doi: 10.1145/3204493.3204547.

[96] Zhiming Hu, Sheng Li, Congyi Zhang, Kangrui Yi, Guoping Wang, and Dinesh Manocha. DGaze: CNN-based gaze prediction in dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1902–1911, 2020. doi: 10.1109/TVCG.2020.2973473.

[97] Zhiming Hu, Congyi Zhang, Sheng Li, Guoping Wang, and Dinesh Manocha. SGaze: A data-driven eye-head coordination model for realtime gaze prediction. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2002–2010, 2019. doi: 10.1109/TVCG.2019.2899187.

[98] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52):16054–16059, 2015. doi: 10.1073/pnas.1510393112.

[99] Matthias Kummerer, Thomas S. A. Wallis, Leon A. Gatys, and Matthias Bethge. Understanding low- and high-level contributions to fixation prediction. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4799–4808, New York, NY, USA, 2017. IEEE. doi: 10.1109/ICCV.2017.513.

[100] Richard Droste, Jianbo Jiao, and J. Alison Noble. Unified image and video saliency modeling. In *Computer Vision – ECCV 2020*, pages 419–435, Cham, Switzerland, 2020. Springer. doi: 10.1007/978-3-030-58558-7_25.

[101] Anh Nguyen, Zhisheng Yan, and Klara Nahrstedt. Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 1190–1198, New York, NY, USA, 2018. ACM. doi: 10.1145/3240508.3240669.

[102] Guangxiao Ma, Shuai Li, Chenglizhao Chen, Aimin Hao, and Hong Qin. Stage-wise salient object detection in 360° omnidirectional image via object-level semantical saliency ranking. *IEEE Transactions on Visualization and Computer Graphics*, 26(12): 3535–3545, 2020. doi: 10.1109/TVCG.2020.3023636.

[103] Thies Pfeiffer and Cem Memili. Model-based real-time visualization of realistic three-dimensional heat maps for mobile eye tracking and eye tracking in virtual reality. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 95–102, New York, NY, USA, 2016. ACM. doi: 10.1145/2857491.2857541.

[104] Teresa Hirzle, Jan Gugenheimer, Florian Geiselhart, Andreas Bulling, and Enrico Rukzio. A design space for gaze interaction on head-mounted displays. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 625:1–625:12, New York, NY, USA, 2019. ACM. doi: 10.1145/3290605.3300855.

[105] Xinyao Ma, Zhaolin Yao, Yijun Wang, Weihua Pei, and Hongda Chen. Combining brain-computer interface and eye tracking for high-speed text entry in virtual reality. In *23rd International Conference on Intelligent User Interfaces*, pages 263–267, New York, NY, USA, 2018. ACM. doi: 10.1145/3172944.3172988.

[106] Vijay Rajanna and John Paulin Hansen. Gaze typing in virtual reality: Impact of keyboard design, selection method, and motion. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 15:1–15:10, New York, NY, USA, 2018. ACM. doi: 10.1145/3204493.3204541.

[107] Xueshi Lu, Difeng Yu, Hai-Ning Liang, Wenge Xu, Yuzheng Chen, Xiang Li, and Khalad Hasan. Exploration of hands-free text entry techniques for virtual reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 344–349, New York, NY, USA, 2020. IEEE. doi: 10.1109/ISMAR50242.2020.00061.

[108] Ludwig Sidenmark and Anders Lundström. Gaze behaviour on interacted objects during hand interaction in virtual reality for eye tracking calibration. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 6:1–6:9, New York, NY, USA, 2019. ACM. doi: 10.1145/3314111.3319815.

[109] Ken Pfeuffer, Lukas Mecke, Sarah Delgado Rodriguez, Mariam Hassib, Hannah Maier, and Florian Alt. Empirical evaluation of gaze-enhanced menus in virtual reality. In *26th ACM Symposium on Virtual Reality Software and Technology*, pages 20:1–20:11, New York, NY, USA, 2020. ACM. doi: 10.1145/3385956.3418962.

[110] Anh Nguyen and Andreas Kunz. Discrete scene rotation during blinks and its effect on redirected walking algorithms. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, pages 29:1–29:10, New York, NY, USA, 2018. ACM. doi: 10.1145/3281505.3281515.

[111] Yiran Zhang, Nicolas Ladeveze, Huyen Nguyen, Cedric Fleury, and Patrick Bourdot. Virtual navigation considering user workspace: Automatic and manual positioning before teleportation. In *26th ACM Symposium on Virtual Reality Software and Technology*, pages 9:1–9:11, New York, NY, USA, 2020. ACM. doi: 10.1145/3385956.3418949.

[112] Lung-Pan Cheng, Eyal Ofek, Christian Holz, Hrvoje Benko, and Andrew D. Wilson. Sparse haptic proxy: Touch feedback in virtual environments using a general passive prop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3718–3728, New York, NY, USA, 2017. ACM. doi: 10.1145/3025453.3025753.

[113] Ashima Keshava, Anete Aumeistere, Krzysztof Izdebski, and Peter Konig. Decoding task from oculomotor behavior in virtual reality. In *ACM Symposium on Eye Tracking Research and Applications*, pages 30:1–30:5, New York, NY, USA, 2020. ACM. doi: 10.1145/3379156.3391338.

[114] Rawan Alghofaili, Michael S Solah, Haikun Huang, Yasuhito Sawahata, Marc Pomplun, and Lap-Fai Yu. Optimizing visual element placement via visual attention analysis. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 464–473, New York, NY, USA, 2019. IEEE. doi: 10.1109/VR.2019.8797816.

[115] Daniel Lange, Tim Claudius Stratmann, Uwe Gruenefeld, and Susanne Boll. HiveFive: Immersion preserving attention guidance in virtual reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, New York, NY, USA, 2020. ACM. doi: 10.1145/3313831.3376803.

[116] Yining Lang, Liang Wei, Fang Xu, Yibiao Zhao, and Lap-Fai Yu. Synthesizing personalized training programs for improving driving habits via virtual reality. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 297–304, New York, NY, USA, 2018. IEEE. doi: 10.1109/VR.2018.8448290.

[117] Enkelejda Kasneci, Gjergji Kasneci, Ulrich Trautwein, Tobias Appel, Maike Tibus, Susanne M. Jaeggi, and Peter Gerjets. Do your eye movements reveal your performance on an IQ test? a study linking eye movements and socio-demographic information to fluid intelligence. PsyArXiv, 2021. URL https://doi.org/10.31234/osf.io/dru93.

[118] Tiffany Luong, Nicolas Martin, Anaïs Raison, Ferran Argelaguet, Jean-Marc Diverrez, and Anatole Lécuyer. Towards real-time recognition of users mental workload using integrated physiological sensors into a VR HMD. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 425–437, New York, NY, USA, 2020. IEEE. doi: 10.1109/ISMAR50242.2020.00068.

[119] Justin C. Wilson, Suku Nair, Sandro Scielzo, and Eric C. Larson. Objective measures of cognitive load using deep multi-modal learning: A use-case in aviation. *Proceedings of the ACM Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):40:1–40:35, 2021. doi: 10.1145/3448111.

[120] Thomas Christian Kübler, Enkelejda Kasneci, and Florentin Vintila. Pupil response as an indicator of hazard perception during simulator driving. *Journal of Eye Movement Research*, 10(4), 2017. doi: 10.16910/jemr.10.4.3.

[121] Jason Orlosky, Yuta Itoh, Maud Ranchet, Kiyoshi Kiyokawa, John Morgan, and Hannes Devos. Emulation of physician tasks in eye-tracked virtual reality for remote diagnosis of neurodegenerative disease. *IEEE Transactions on Visualization and Computer Graphics*, 23(4):1302–1311, 2017. doi: 10.1109/TVCG.2017.2657018.

[122] Benedikt W. Hosp, Florian Schultz, Oliver Höner, and Enkelejda Kasneci. Soccer goalkeeper expertise identification based on eye movements. *PLOS ONE*, 16(5):1–22, 2021. doi: 10.1371/journal.pone.0251070.

[123] Karan Ahuja, Deval Shah, Sujeath Pareddy, Franceska Xhakaj, Amy Ogan, Yuvraj Agarwal, and Chris Harrison. Classroom digital twins with instrumentation-free gaze tracking. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 484:1–484:9, New York, NY, USA, 2021. ACM. doi: 10.1145/3411764.3445711.

[124] Abraham Hani Mhaidli and Florian Schaub. Identifying manipulative advertising techniques in XR through scenario construction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 296:1–296:18, New York, NY, USA, 2021. ACM. doi: 10.1145/3411764.3445253.

[125] Nelson Silva, Tanja Blascheck, Radu Jianu, Nils Rodrigues, Daniel Weiskopf, Martin Raubal, and Tobias Schreck. Eye tracking support for visual analytics systems: Foundations, current applications, and research challenges. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 11:1–11:10, New York, NY, USA, 2019. ACM. doi: 10.1145/3314111.3319919.

[126] Christina Katsini, Yasmeen Abdrabou, George E. Raptis, Mohamed Khamis, and Florian Alt. The role of eye gaze in security and privacy applications: Survey and future HCI research directions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–21, New York, NY, USA, 2020. ACM. doi: 10.1145/3313831.3376840.

[127] Daniel J. Liebling and Sören Preibusch. Privacy considerations for a pervasive eye tracking world. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 1169–1177, New York, NY, USA, 2014. ACM. doi: 10.1145/2638728.2641688.

[128] John Daugman. Iris recognition at airports and border-crossings. In *Encyclopedia of Biometrics*, pages 819–825, Boston, MA, USA, 2009. Springer. doi: 10.1007/978-0-387-73003-5_24.

[129] Mauro Barni, Giulia Droandi, Riccardo Lazzeretti, and Tommaso Pignata. SEMBA: Secure multi-biometric authentication. *IET Biometrics*, 8(6):411–421, 2019. doi: 10.1049/iet-bmt.2018.5138.

[130] Xinxia Song, Zhigang Chen, and Dechao Sun. Iris ciphertext authentication system based on fully homomorphic encryption. *Journal of Information Processing Systems*, 16 (3):599–611, 2020. doi: 10.3745/JIPS.03.0138.

[131] Tomi Kinnunen, Filip Sedlak, and Roman Bednarik. Towards task-independent person authentication using eye movement signals. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 187–190, New York, NY, USA, 2010. ACM. doi: 10.1145/1743666.1743712.

[132] Oleg V. Komogortsev and Corey D. Holland. Biometric authentication via complex oculomotor behavior. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems*, pages 1–8, New York, NY, USA, 2013. IEEE. doi: 10.1109/BTAS.2013.6712725.

[133] Oleg V. Komogortsev, Sampath Jayarathna, Cecilia R. Aragon, and Mechehoul Mahmoud. Biometric identification via an oculomotor plant mathematical model. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 57–60, New York, NY, USA, 2010. ACM. doi: 10.1145/1743666.1743679.

[134] Simon Eberz, Kasper B. Rasmussen, Vincent Lenders, and Ivan Martinovic. Looks like eve: Exposing insider threats using eye movement biometrics. *ACM Transactions on Privacy and Security*, 19(1):1:1–1:31, 2016. doi: 10.1145/2904018.

[135] Huadi Zhu, Wenqiang Jin, Mingyan Xiao, Srinivasan Murali, and Ming Li. BlinKey: A two-factor user authentication method for virtual reality devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):164:1–164:29, 2020. doi: 10.1145/3432217.

[136] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, Germany, 2006. Springer. doi: 10.1007/11681878_14.

[137] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg, Germany, 2006. Springer. doi: 10.1007/11787006_1.

[138] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014. doi: 10.1561/0400000042.

[139] Andrew C. Yao. Protocols for secure computations. In *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*, pages 160–164, New York, NY, USA, 1982. IEEE. doi: 10.1109/SFCS.1982.38.

[140] Andrew Chi-Chih Yao. How to generate and exchange secrets. In *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*, pages 162–167, New York, NY, USA, 1986. IEEE. doi: 10.1109/SFCS.1986.25.

[141] Benny Applebaum, Yuval Ishai, and Eyal Kushilevitz. Cryptography in NCˆ0. *SIAM Journal on Computing*, 36(4):845–888, 2006. doi: 10.1137/S0097539705446950.

[142] Benny Applebaum. Garbled circuits as randomized encodings of functions: a primer. In *Tutorials on the Foundations of Cryptography*, pages 1–44. Springer, Cham, Switzerland, 2017. doi: 10.1007/978-3-319-57048-8_1.

[143] Julian Steil, Inken Hagestedt, Michael Xuelin Huang, and Andreas Bulling. Privacy-aware eye tracking using differential privacy. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 27:1–27:9, New York, NY, USA, 2019. ACM. doi: 10.1145/3314111.3319915.

[144] Ao Liu, Lirong Xia, Andrew Duchowski, Reynold Bailey, Kenneth Holmqvist, and Eakta Jain. Differential privacy for eye-tracking data. In *ACM Symposium on Eye Tracking Research & Applications*, pages 28:1–28:10, New York, NY, USA, 2019. ACM. doi: 10.1145/3314111.3319823.

[145] Tao Zhang, Tianqing Zhu, Renping Liu, and Wanlei Zhou. Correlated data in differential privacy: Definition and analysis. *Concurrency and Computation: Practice and Experience*, page e6015, 2020. doi: 10.1002/cpe.6015.

[146] Jingjie Li, Amrita Roy Chowdhury, Kassem Fawaz, and Younghyun Kim. Kaleido: Real-time privacy control for eye-tracking systems. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1793–1810, Berkeley, CA, USA, 2021. USENIX Association. URL https://www.usenix.org/conference/usenixsecurity21/presentation/li-jingjie.

[147] Brendan John, Sanjeev Koppal, and Eakta Jain. EyeVEIL: Degrading iris authentication in eye tracking headsets. In *ACM Symposium on Eye Tracking Research & Applications*, pages 37:1–37:5, New York, NY, USA, 2019. ACM. doi: 10.1145/3314111.3319816.

[148] Brendan John, Ao Liu, Lirong Xia, Sanjeev Koppal, and Eakta Jain. Let it snow: Adding pixel noise to protect the user's identity. In *ACM Symposium on Eye Tracking Research and Applications*, pages 43:1–43:3, New York, NY, USA, 2020. ACM. doi: 10.1145/3379157.3390512.

[149] Brendan John, Sophie Jörg, Sanjeev Koppal, and Eakta Jain. The security-utility trade-off for iris authentication and eye animation for social virtual avatars. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):1880–1890, 2020. doi: 10.1109/TVCG.2020.2973052.

[150] Aayush Kumar Chaudhary and Jeff B. Pelz. Privacy-preserving eye videos using rubber sheet model. In *ACM Symposium on Eye Tracking Research and Applications*, pages 22:1–22:5, New York, NY, USA, 2020. ACM. doi: 10.1145/3379156.3391375.

[151] Inken Hagestedt, Michael Backes, and Andreas Bulling. Adversarial attacks on classifiers for eye-based user modelling. In *ACM Symposium on Eye Tracking Research and Applications*, pages 44:1–44:3, New York, NY, USA, 2020. ACM. doi: 10.1145/3379157.3390511.

[152] Valve Corporation. Steam. Online, 2021. URL https://store.steampowered.com/. Accessed: 2021-09-20.

[153] Facebook Technologies, LLC. Lifecycle of an Oculus VR App. Online, 2021. URL https://developer.oculus.com/distribute/publish-app-review/. Accessed: 2021-09-20.

[154] Mozilla. hubs moz://a. Online, 2021. URL https://hubs.mozilla.com/. Accessed: 2021-09-20.

[155] Google LLC. YouTube. Online, 2021. URL https://www.youtube.com/. Accessed: 2021-09-20.

[156] Xiao Ma, Megan Cackett, Leslie Park, Eric Chien, and Mor Naaman. Web-based VR experiments powered by the crowd. In *Proceedings of the 2018 World Wide Web Conference*, pages 33–43, Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee. doi: 10.1145/3178876.3186034.

[157] Amazon Mechanical Turk, Inc. Amazon Mechanical Turk. Online, 2005-2018. URL https://www.mturk.com/. Accessed: 2021-09-20.

[158] Radiah Rivu, Ville Mäkelä, Sarah Prange, Sarah Delgado Rodriguez, Robin Piening, Yumeng Zhou, Kay Köhle, Ken Pfeuffer, Yomna Abdelrahman, Matthias Hoppe, Albrecht Schmidt, and Florian Alt. Remote VR studies – a framework for running virtual reality studies remotely via participant-owned HMDs. CoRR, 2021. URL https://arxiv.org/abs/2102.11207v1.

[159] Evangelos Markopoulos, Mika Luimula, Pasi Porramo, Tayfun Pisirici, and Aleksi Kirjonen. Virtual reality (VR) safety education for ship engine training on maintenance and safety (ShipSEVR). In *Advances in Creativity, Innovation, Entrepreneurship and Communication of Design*, pages 60–72, Cham, Switzerland, 2020. Springer. doi: 10.1007/978-3-030-51626-0_7.

[160] Svetlana Bialkova and Dick Ettema. Cycling renaissance: The VR potential in exploring static and moving environment elements. In *2019 IEEE 5th Workshop on Everyday Virtual Reality (WEVR)*, pages 1–6, New York, NY, USA, 2019. IEEE. doi: 10.1109/WEVR.2019.8809586.

[161] Benny Applebaum, Yuval Ishai, and Eyal Kushilevitz. Computationally private randomizing polynomials and their applications. *Computational Complexity*, 15(2):115–162, 2006. doi: 10.1007/s00037-006-0211-8.

[162] Synergy Research Group. The decade's megatrends in numbers – part 1: Cloud goes from 0 to 100 in ten years while enterprise data center spending stagnates. Online, Jan 2020. URL https://www.srgresearch.com/articles/the-decades-megatrends-in-numbers-part-1. Accessed: 2021-07-06.

[163] Patricia Goldberg, Ömer Sümer, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, and Ulrich Trautwein. Attentive or not? Toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educational Psychology Review*, 33(1):27–49, 2021. doi: 10.1007/s10648-019-09514-z. Published online: 2019.

[164] Benjamin Fauth, Wolfgang Wagner, Christiane Bertram, Richard Göllner, Janina Roloff, Oliver Lüdtke, Morgan S. Polikoff, Uta Klusmann, and Ulrich Trautwein. Don't blame the teacher? the need to account for classroom characteristics in evaluations of teaching quality. *Journal of Educational Psychology*, 112(6):1284–1302, 2020. doi: 10.1037/edu0000416.

[165] Abraham Savitzky and Marcel J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964. doi: 10.1021/ac60214a047.

[166] Sebastiaan Mathôt, Jasper Fabius, Elle Van Heusden, and Stefan Van der Stigchel. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods*, 50(1):94–106, 2018. doi: 10.3758/s13428-017-1007-2.

[167] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 143–146, New York, NY, USA, 2011. ACM. doi: 10.1145/1978942.1978963.

[168] Scott D. Roth. Ray casting for modeling solids. *Computer Graphics and Image Processing*, 18(2):109–144, 1982. doi: 10.1016/0146-664X(82)90169-1.

[169] Efe Bozkir, David Geisler, and Enkelejda 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)*, pages 1834–1837, New York, NY, USA, 2019. IEEE. doi: 10.1109/VR.2019.8797758.

[170] Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. Agreeing to cross: How drivers and pedestrians communicate. CoRR, 2017. URL https://arxiv.org/abs/1702.03555v1.

[171] Jun Zhao, Junshan Zhang, and H. Vincent Poor. Dependent differential privacy for correlated data. In *IEEE Globecom Workshops (GC Wkshps)*, pages 1–7, New York, NY, USA, 2017. IEEE. doi: 10.1109/GLOCOMW.2017.8269219.

[172] Yang Cao, Masatoshi Yoshikawa, Yonghui Xiao, and Li Xiong. Quantifying differential privacy in continuous data release under temporal correlations. *IEEE Transactions on Knowledge and Data Engineering*, 31(7):1281–1295, 2019. doi: 10.1109/TKDE.2018.2824328.

[173] Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems*, 39(1):3:1–3:36, 2014. doi: 10.1145/2514689.

[174] Nisarg Raval, Ashwin Machanavajjhala, and Jerry Pan. Olympus: Sensor privacy through utility aware obfuscation. *Proceedings on Privacy Enhancing Technologies*, 2019(1):5–25, 2018. doi: 10.2478/popets-2019-0002.

[175] Vibhor Rastogi and Suman Nath. Differentially private aggregation of distributed time-series with transformation and encryption. In *ACM SIGMOD International Conference on Management of Data*, pages 735–746, New York, NY, USA, 2010. ACM. doi: 10.1145/1807167.1807247.

[176] Efe Bozkir, Onur Günlü, Wolfgang Fuhl, Rafael F. Schaefer, and Enkelejda Kasneci. Differential privacy for eye tracking with temporal correlations. *PLOS ONE*, 16(8):1–22, 2021. doi: 10.1371/journal.pone.0255979. Preprint: https://arxiv.org/abs/2002.08972.

[177] Onur Günlü. Design and analysis of discrete cosine transform based ring oscillator physical unclonable functions. Master's thesis, Technical University of Munich, Munich, Germany, 2013.

[178] Ali Burak Ünal, Mete Akgün, and Nico Pfeifer. A framework with randomized encoding for a fast privacy preserving calculation of non-linear kernels for machine learning applications in precision medicine. In *Cryptology and Network Security*, pages 493–511, Cham, Switzerland, 2019. Springer. doi: 10.1007/978-3-030-31578-8_27.

[179] Brecht Wyseur. White-box cryptography: Hiding keys in software. MISC magazine, Apr 2012. URL https://whiteboxcrypto.com.

[180] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. Online, 2008. URL https://bitcoin.org/bitcoin.pdf. Accessed: 2020-10-29.

[181] Vitalik Buterin. A next-generation smart contract and decentralized application platform. Ethereum, Online, 2014. URL https://translatewhitepaper.com/wp-content/uploads/2021/04/EthereumOrijinal-ETH-English.pdf. Accessed: 2020-09-15.

[182] Alex Biryukov and Aleksei Udovenko. Attacks and countermeasures for white-box designs. In *Advances in Cryptology – ASIACRYPT 2018*, pages 373–402, Cham, Switzerland, 2018. Springer. doi: 10.1007/978-3-030-03329-3_13.

[183] Richard J. Shavelson, Judith J. Hubner, and George C. Stanton. Self-concept: Validation of construct interpretations. *Review of Educational Research*, 46(3):407–441, 1976. doi: 10.3102/00346543046003407.

[184] Marc Pomplun, Tyler Garaas, and Marisa Carrasco. The effects of task difficulty on visual search strategy in virtual 3D displays. *Journal of Vision*, 13(3):24, 2013. doi: 10.1167/13.3.24.

[185] Lawrence Baretto. 'You get the same buzz as racing for real' - Lando Norris on the thrill of sim racing. Online, Apr 2020. URL https://www.formula1.com/en/latest/article.you-get-the-same-buzz-as-racing-for-real-lando-norris-on-the-thrill-of-sim.lFM26RLpKxViWXTYwSz0J.html. Accessed: 2021-08-04.

[186] @LandoNorris. Everyone practices like this on the simulator, right? [Twitter post]. Twitter, Online, Mar 02, 7:33 PM 2020. URL https://twitter.com/LandoNorris/status/1234547426209550336. Accessed: 2021-08-04.

[187] Tobii Pro. What's the formula behind an F1 driver? Eye tracking could uncover the recipe for success. Online, Jun 2016. URL https://www.tobiipro.com/blog/formular-1-eye-tracking/. Accessed: 2021-08-04.

[188] Gregory Kramida. Resolving the vergence-accommodation conflict in head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics*, 22(7):1912–1931, 2016. doi: 10.1109/TVCG.2015.2473855.

[189] Frank D. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *ACM SIGMOD International Conference on Management of Data*, pages 19–30, New York, NY, USA, 2009. ACM. doi: 10.1145/1559845.1559850.

[190] BBC News. Bitcoin: El Salvador makes cryptocurrency legal tender. Online, Jun 2021. URL https://www.bbc.com/news/world-latin-america-57398274. Accessed: 2021-08-03.

[191] Neal Stephenson. *Snow Crash*. Bantam Spectra, New York, NY, USA, 1992. ISBN 0-553-08853-X.

[192] John David N. Dionisio, William G. Burns III, and Richard Gilbert. 3D virtual worlds and the Metaverse: Current status and future possibilities. *ACM Computing Surveys*, 45(3), 2013. doi: 10.1145/2480741.2480751.

[193] Lik-Hang Lee, Tristan Braud, Pengyuan Zhou, Lin Wang, Dianlei Xu, Zijun Lin, Abhishek Kumar, Carlos Bermejo, and Pan Hui. All one needs to know about Metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. CoRR, 2021. URL https://arxiv.org/abs/2110.05352v3.

[194] Christian Braunagel, Wolfgang Rosenstiel, and Enkelejda Kasneci. Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness. *IEEE Intelligent Transportation Systems Magazine*, 9(4):10–22, 2017. doi: 10.1109/MITS.2017.2743165.

[195] Christian Braunagel, Enkelejda Kasneci, Wolfgang Stolzmann, and Wolfgang Rosenstiel. Driver-activity recognition in the context of conditionally autonomous driving. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 1652–1657, New York, NY, USA, 2015. IEEE. doi: 10.1109/ITSC.2015.268.

[196] Grand View Research, Inc. Autonomous vehicle market size, share & trends analysis report by application (Transportation, defense), by region (North & South America, Europe, APAC, MEA), and segment forecasts, 2021 - 2030. Autonomous Vehicle Market Size & Share Report, 2021-2030, Mar 2020. URL https://www.grandviewresearch.com/industry-analysis/autonomous-vehicles-market. Accessed: 2021-09-15.

[197] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, Cambridge, MA, USA, 2016. ISBN 978-0-262-03561-3.

[198] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, New York, NY, USA, 2016. ACM. doi: 10.1145/2976749.2978318.

[199] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. CoRR, 2021. URL https://arxiv.org/abs/2109.12298v3. Privacy in Machine Learning (PriML) workshop, NeurIPS 2021.

[200] Matthew Joseph, Aaron Roth, Jonathan Ullman, and Bo Waggoner. Local differential privacy for evolving data. *Journal of Privacy and Confidentiality*, 10(1):1–29, 2020. doi: 10.29012/jpc.718.

[201] Huajie Chen, Ali Burak Ünal, Mete Akgün, and Nico Pfeifer. Privacy-preserving SVM on outsourced genomic data via secure multi-party computation. In *Proceedings of the Sixth International Workshop on Security and Privacy Analytics*, pages 61–69, New York, NY, USA, 2020. ACM. doi: 10.1145/3375708.3380316.

[202] You-Ping Chen and Ju-Chun Ko. CryptoAR wallet: A blockchain cryptocurrency wallet application that uses augmented reality for on-chain user data display. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 39:1–39:5, New York, NY, USA, 2019. ACM. doi: 10.1145/3338286.3344386.

[203] Marcus Foth. The promise of blockchain technology for interaction design. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, pages 513–517, New York, NY, USA, 2017. ACM. doi: 10.1145/3152771.3156168.

[204] GazeCoin. Gazecoin: A unit of exchange between advertisers, content makers and users based on 'gaze'/eye tracking. Online, Oct 2017. URL https://www.gazecoin.io/s/GazeCoin_WhitePaper.pdf. Accessed: 2020-10-29.

[205] Kevin Sekniqi, Daniel Laine, Stephen Buttolph, and Emin Gün Sirer. Avalanche platform. Online, 2020. URL https://www.avalabs.org/whitepapers. Accessed: 2021-08-05.

[206] Gavin Wood. Polkadot: Vision for a heterogeneous multi-chain framework. Online, 2016. URL https://github.com/w3f/polkadotwhite-paper/raw/master/PolkaDotPaper.pdf. Accessed: 2021-08-05.

[207] Joseph Psotka. Immersive training systems: Virtual reality and education and training. *Instructional Science*, 23(5):405–431, 1995. doi: 10.1007/BF00896880.

[208] Sandra Helsel. Virtual reality and education. *Educational Technology*, 32(5):38–42, 1992. URL https://www.learntechlib.org/p/170758.

[209] Andrea Casu, Lucio Davide Spano, Fabio Sorrentino, and Riccardo Scateni. Riftart: Bringing masterpieces in the classroom through immersive virtual reality. In *Smart Tools and Apps for Graphics - Eurographics Italian Chapter Conference*, pages 77–84, Geneva, Switzerland, 2015. The Eurographics Association. doi: 10.2312/stag.20151294.

[210] Natasha Anne Rappa, Susan Ledger, Timothy Teo, Kok Wai Wong, Brad Power, and Bruce Hilliard. The use of eye tracking technology to explore learning and performance within virtual reality and mixed reality settings: a scoping review. *Interactive Learning Environments*, 0(0):1–13, 2019. doi: 10.1080/10494820.2019.1702560.

[211] Christian Hirt, Marcel Eckard, and Andreas Kunz. Stress generation and non-intrusive measurement in virtual environments using eye tracking. *Journal of Ambient Intelligence and Humanized Computing*, 11(1):1–13, 2020. doi: 10.1007/s12652-020-01845-y.

[212] Efe Bozkir, David Geisler, and Enkelejda Kasneci. Assessment of driver attention during a safety critical situation in VR to generate VR-based training. In *ACM Symposium on Applied Perception 2019*, pages 23:1–23:5, New York, NY, USA, 2019. ACM. doi: 10.1145/3343036.3343138.

[213] Ömer Sümer, Patricia Goldberg, Kathleen Stürmer, Tina Seidel, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Teachers' perception in the classroom. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2315–2324, New York, NY, USA, 2018. IEEE. URL https://arxiv.org/abs/1805.08897v1.

[214] Laura Freina and Michela Ott. A literature review on immersive virtual reality in education: State of the art and perspectives. In *Proceedings of the 11th International Scientific Conference eLearning and Software for Education*, pages 133–141, Bucharest, Romania, 2015. Carol I NDU Publishing House. doi: 10.12753/2066-026X-15-020.

[215] Christine Youngblut. Educational uses of virtual reality technology. Technical report, Institute for Defense Analyses, Alexandria, VA, USA, 1998.

[216] Christian Moro, Zane Štromberga, Athanasios Raikos, and Allan Stirling. The effectiveness of virtual and augmented reality in health sciences and medical anatomy. *Anatomical Sciences Education*, 10(6):549–559, 2017. doi: 10.1002/ase.1696.

[217] Wadee Alhalabi. Virtual reality systems enhance students' achievements in engineering education. *Behaviour & Information Technology*, 35(11):1–7, 2016. doi: 10.1080/0144929X.2016.1212931.

[218] Richard Lamb and Elisabeth A. Etopio. Virtual reality: A tool for preservice science teachers to put theory into practice. *Journal of Science Education and Technology*, 29(4): 573–585, 2020. doi: 10.1007/s10956-020-09837-5.

[219] Ananda Bibek Ray and Suman Deb. Smartphone based virtual reality systems in classroom teaching — a study on the effects of learning outcome. In *2016 IEEE Eighth International Conference on Technology for Education (T4E)*, pages 68–71, New York, NY, USA, 2016. IEEE. doi: 10.1109/T4E.2016.022.

[220] Kun Hung Cheng and Chin Chung Tsai. A case study of immersive virtual field trips in an elementary classroom: Students' learning experience and teacher-student interaction behaviors. *Computers & Education*, 140:103600, 2019. doi: 10.1016/j.compedu.2019.103600.

[221] Kate S. Hone and Ghada R. El Said. Exploring the factors affecting MOOC retention: A survey study. *Computers & Education*, 98:157–168, 2016. doi: 10.1016/j.compedu.2016.03.016.

[222] Ronald B. Marks, Stanley D. Sibley, and J. Ben Arbaugh. A structural equation model of predictors for effective online learning. *Journal of Management Education*, 29(4): 531–563, 2005. doi: 10.1177/1052562904271199.

[223] Elena Olmos-Raya, Janaina Ferreira-Cavalcanti, Manuel Contero, M Concepción Castellanos, Irene Alice Chicchi Giglioli, and Mariano Alcañiz. Mobile virtual reality as an educational platform: A pilot study on the impact of immersion and positive emotion induction in the learning process. *EURASIA Journal of Mathematics, Science and Technology Education*, 14(6):2045–2057, 2018. doi: 10.29333/ejmste/85874.

[224] Jan Herrington, Thomas C. Reeves, and Ron Oliver. Immersive learning technologies: Realism and online authentic learning. *Journal of Computing in Higher Education*, 19 (1):80–99, 2007. doi: 10.1007/BF03033421.

[225] Sharad Sharma, Ruth Agada, and Jeff Ruffin. Virtual reality classroom as an constructivist approach. In *2013 Proceedings of IEEE Southeastcon*, pages 1–5, New York, NY, USA, 2013. IEEE. doi: 10.1109/SECON.2013.6567441.

[226] Meng-Yun Liao, Ching-Ying Sung, Hao-Chuan Wang, and Wen-Chieh Lin. Virtual classmates: Embodying historical learners' messages as learning companions in a VR classroom through comment mapping. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 163–171, New York, NY, USA, 2019. IEEE. doi: 10.1109/VR.2019.8797708.

[227] Adalberto L. Simeone, Marco Speicher, Andreea Molnar, Adriana Wilde, and Florian Daiber. LIVE: The human role in learning in immersive virtual environments. In *Symposium on Spatial User Interaction*, pages 5:1–5:11, New York, NY, USA, 2019. ACM. doi: 10.1145/3357251.3357590.

[228] Tuomas Kantonen, Charles Woodward, and Neil Katz. Mixed reality in virtual world teleconferencing. In *2010 IEEE Virtual Reality Conference (VR)*, pages 179–182, New York, NY, USA, 2010. IEEE. doi: 10.1109/VR.2010.5444792.

[229] Muhammad Sikandar Lal Khan, Haibo Li, and Shafiq Ur Réhman. Tele-immersion: Virtual reality based collaboration. In *International Conference on Human-Computer Interaction*, pages 352–357, Cham, Switzerland, 2016. Springer. doi: 10.1007/978-3-319-40548-3_59.

[230] Dongsik Jo, Ki-Hong Kim, and Gerard Jounghyun Kim. Effects of avatar and background representation forms to co-presence in mixed reality (MR) tele-conference systems. In *SIGGRAPH ASIA 2016 Virtual Reality Meets Physical Reality: Modelling and Simulating Virtual Humans and Environments*, pages 12:1–12:4, New York, NY, USA, 2016. ACM. doi: 10.1145/2992138.2992146.

[231] Gongjin Lan, Ziyun Luo, and Qi Hao. Development of a virtual reality teleconference system using distributed depth sensors. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, pages 975–978, New York, NY, USA, 2016. IEEE. doi: 10.1109/CompComm.2016.7924850.

[232] Jeremy N. Bailenson, Nick Yee, Jim Blascovich, Andrew C. Beall, Nicole Lundblad, and Michael Jin. The use of immersive virtual reality in the learning sciences: Digital transformations of teachers, students, and social context. *Journal of the Learning Sciences*, 17(1):102–141, 2008. doi: 10.1080/10508400701793141.

[233] Friederike Blume, Richard Göllner, Korbinian Moeller, Thomas Dresler, Ann-Christine Ehlis, and Caterina Gawrilow. Do students learn better when seated close to the teacher? a virtual classroom study considering individual levels of inattention and hyperactivity-impulsivity. *Learning and Instruction*, 61:138–147, 2019. doi: 10.1016/j.learninstruc.2018.10.004.

[234] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Jarodzka Halszka, and Joost van de Weijer. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press, Oxford, United Kingdom, 2011. ISBN 978-0-19-969708-3.

[235] Unai Díaz-Orueta, Cristina García-López, Nerea Crespo-Eguílaz, Rocío Sánchez-Carpintero, Gema Climent, and Juan Narbona. AULA virtual reality test as an attention measure: Convergent validity with conners' continuous performance test. *Child Neuropsychology*, 20(3):328–342, 2014. doi: 10.1080/09297049.2013.792332.

[236] Seung-hun Seo, Eunjoo Kim, Peter Mundy, Jiwoong Heo, and Kwanguk Kim. Joint attention virtual classroom: A preliminary study. *Psychiatry Investigation*, 16(4):292–299, 2019. doi: 10.30773/pi.2019.02.08.

[237] Pierre Nolin, Annie Stipanicic, Mylène Henry, Yves Lachapelle, Dany Lussier-Desrochers, Albert S. Rizzo, and Philippe Allain. ClinicaVR: Classroom-CPT: A virtual reality tool for assessing attention and inhibition in children and adolescents. *Computers in Human Behavior*, 59:327–333, 2016. doi: 10.1016/j.chb.2016.02.023.

[238] Aman Mangalmurti, William Kistler, Barrington Quarrie, Wendy Sharp, Susan Persky, and Philip Shaw. Using virtual reality to define the mechanisms linking symptoms with cognitive deficits in attention deficit hyperactivity disorder. *Scientific Reports*, 10(1):529, 2020. doi: 10.1038/s41598-019-56936-4.

[239] Albert A. Rizzo, Todd Bowerly, J. Galen Buckwalter, Dean Klimchuk, Roman Mitura, and Thomas D. Parsons. A virtual reality scenario for all seasons: The virtual classroom. *CNS Spectrums*, 11(1):35–44, 2006. doi: 10.1017/S1092852900024196.

[240] Ricardo Böheim, Tim Urdan, Maximilian Knogler, and Tina Seidel. Student hand-raising as an indicator of behavioral engagement and its role in classroom learning. *Contemporary Educational Psychology*, 62:101894, 2020. doi: 10.1016/j.cedpsych.2020.101894.

[241] Tobias Appel, Christian Scharinger, Peter Gerjets, and Enkelejda Kasneci. Cross-subject workload classification using pupil-related measures. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 4:1–4:8, New York, NY, USA, 2018. ACM. doi: 10.1145/3204493.3204531.

[242] Tobias Appel, Natalia Sevcenko, Franz Wortha, Katerina Tsarava, Korbinian Moeller, Manuel Ninaus, Enkelejda Kasneci, and Peter Gerjets. Predicting cognitive load in an emergency simulation based on behavioral and physiological measures. In *2019 International Conference on Multimodal Interaction*, pages 154–163, New York, NY, USA, 2019. ACM. doi: 10.1145/3340555.3353735.

[243] Jeremy N. Bailenson, Andrew C. Beall, and Jim Blascovich. Gaze and task performance in shared virtual environments. *The Journal of Visualization and Computer Animation*, 13(5):313–320, 2002. doi: 10.1002/vis.297.

[244] Jeremy N. Bailenson, Eyal Aharoni, Andrew C. Beall, Rosanna E. Guadagno, Aleksandar Dimov, and Jim Blascovich. Comparing behavioral and self-report measures of embodied agents' social presence in immersive virtual environments. In *Proceedings of the 7th Annual International Workshop on Presence*, pages 216–223, Valencia, Spain, 2004. The International Society for Presence Research. ISBN 84-9705-649-3.

[245] David Weintrop, Elham Beheshti, Michael Horn, Orton Kai, Kemi Jona, Laura Trouille, and Uri Wilensky. Defining computational thinking for mathematics and science

classrooms. *Journal of Science Education and Technology*, 25(1):127–147, 2016. doi: 10.1007/s10956-015-9581-5.

[246] Shivsevak Negi and Ritayan Mitra. Fixation duration and the learning process: an eye tracking study with subtitled videos. *Journal of Eye Movement Research*, 13(6), 2020. doi: 10.16910/jemr.13.6.1.

[247] Kuei-Pin Chien, Cheng-Yue Tsai, Hsiu-Ling Chen, Wen-Hua Chang, and Sufen Chen. Learning differences and eye fixation patterns in virtual and physical science laboratories. *Computers & Education*, 82:191–201, 2015. doi: 10.1016/j.compedu.2014.11.023.

[248] Joseph T. Coyne, Cyrus Foroughi, and Ciara Sibley. Pupil diameter and performance in a supervisory control task: A measure of effort or individual differences? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61(1):865–869, 2017. doi: 10.1177/1541931213601689.

[249] Richard E. Mayer. Unique contributions of eye-tracking research to the study of learning with graphics. *Learning and Instruction*, 20(2):167–171, 2010. doi: 10.1016/j.learninstruc.2009.02.012.

[250] Jacob Hadnett-Hunter, George Nicolaou, Eamonn O'Neill, and Michael Proulx. The effect of task on visual attention in interactive virtual environments. *ACM Transactions on Applied Perception*, 16(3):17:1–17:17, 2019. doi: 10.1145/3352763.

[251] Sam Kavanagh, Andrew Luxton-Reilly, Burkhard Wuensche, and Beryl Plimmer. A systematic review of virtual reality in education. *Themes in Science and Technology Education*, 10(2):85–119, 2017.

[252] Elizabeth Johnston, Gerald Olivas, Patricia Steele, Cassandra Smith, and Liston Bailey. Exploring pedagogical foundations of existing virtual reality educational applications: A content analysis study. *Journal of Educational Technology Systems*, 46(4):414–439, 2018. doi: 10.1177/0047239517745560.

[253] Albert A. Rizzo, J. Galen Buckwalter, Todd Bowerly, Cheryl Van Der Zaag, Lorie Humphrey, Ulrich Neumann, Clint Chua, Chris Kyriakakis, Andre Van Rooyen, and D. Sisemore. The virtual classroom: A virtual reality environment for the assessment and rehabilitation of attention deficits. *CyberPsychology & Behavior*, 3(3):483–499, 2000. doi: 10.1089/10949310050078940.

[254] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *Computer Vision – ECCV 2020*, pages 365–381, Cham, Switzerland, 2020. Springer. doi: 10.1007/978-3-030-58558-7_22.

[255] Anastasia Schmitz, Andrew MacQuarrie, Simon Julier, Nicola Binetti, and Anthony Steed. Directing versus attracting attention: Exploring the effectiveness of central and

peripheral cues in panoramic videos. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 63–72, New York, NY, USA, 2020. IEEE. doi: 10.1109/VR46266.2020.00024.

[256] Tae Min Lee, Jong-Chul Yoon, and In-Kwon Lee. Motion sickness prediction in stereoscopic videos using 3D convolutional neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1919–1927, 2019. doi: 10.1109/TVCG.2019.2899186.

[257] Eike Langbehn, Frank Steinicke, Markus Lappe, Gregory F. Welch, and Gerd Bruder. In the blink of an eye: Leveraging blink-induced suppression for imperceptible position and orientation redirection in virtual reality. *ACM Transactions on Graphics*, 37(4):66:1–66:11, 2018. doi: 10.1145/3197517.3201335.

[258] Mohamed Khamis, Carl Oechsner, Florian Alt, and Andreas Bulling. VRPursuits: Interaction in virtual reality using smooth pursuit eye movements. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, pages 18:1–18:8, New York, NY, USA, 2018. ACM. doi: 10.1145/3206505.3206522.

[259] Herbert W. Marsh and John W. Parker. Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology*, 47(1):213–231, 1984. doi: 10.1037/0022-3514.47.1.213.

[260] Elizabeth B. Cloude, Daryn A. Dever, Megan D. Wiedbusch, and Roger Azevedo. Quantifying scientific thinking using multichannel data with crystal island: Implications for individualized game-learning analytics. *Frontiers in Education*, 5:217, 2020. doi: 10.3389/feduc.2020.572546.

[261] Megan D. Wiedbusch and Roger Azevedo. Modeling metacomprehension monitoring accuracy with eye gaze on informational content in a multimedia learning environment. In *ACM Symposium on Eye Tracking Research and Applications*, pages 20:1–20:9, New York, NY, USA, 2020. ACM. doi: 10.1145/3379155.3391329.

[262] Ömer Sümer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Automated anonymisation of visual and audio data in classroom studies. In *The Workshops of the Thirty-Forth AAAI Conference on Artificial Intelligence*, pages 1–7, Palo Alto, CA, USA, 2020. AAAI Press. URL https://arxiv.org/abs/2001.05080v1.

[263] Wolfgang Fuhl, Efe Bozkir, and Enkelejda Kasneci. Reinforcement learning for the privacy preservation and manipulation of eye tracking data. In *Artificial Neural Networks and Machine Learning – ICANN 2021*, pages 595–607, Cham, Switzerland, 2021. Springer. doi: 10.1007/978-3-030-86380-7_48. Preprint: https://arxiv.org/abs/2002.06806.

[264] Efe Bozkir, Ali Burak Ünal, Mete Akgün, Enkelejda Kasneci, and Nico Pfeifer. Privacy preserving gaze estimation using synthetic images via a randomized encoding based framework. In *ACM Symposium on Eye Tracking Research and Applications*, pages 21:1–21:5, New York, NY, USA, 2020. ACM. doi: 10.1145/3379156.3391364.

[265] IRTAD. Road safety annual report 2018. ITF/OECD, May 2018. URL https://www.itf-oecd.org/sites/default/files/docs/irtad-road-safety-annual-report-2018_2.pdf.

[266] Vassilis Charissis and Stylianos Papanastasiou. Human–machine collaboration through vehicle head up display interface. *Cognition, Technology & Work*, 12(1):41–50, 2010. doi: 10.1007/s10111-008-0117-0.

[267] Felix Schwarz and Wolfgang Fastenmeier. Augmented reality warnings in vehicles: Effects of modality and specificity on effectiveness. *Accident Analysis & Prevention*, 101: 55–66, 2017. doi: 10.1016/j.aap.2017.01.019.

[268] Cuong Tran, Karlin Bark, and Victor Ng-Thow-Hing. A left-turn driving aid using projected oncoming vehicle paths with augmented reality. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 300–307, New York, NY, USA, 2013. ACM. doi: 10.1145/2516540.2516581.

[269] Michelle L. Rusch, Mark C. Schall, Jr., John D. Lee, Jeffrey D. Dawson, and Matthew Rizzo. Augmented reality cues to assist older drivers with gap estimation for left-turns. *Accident Analysis & Prevention*, 71:210–221, 2014. doi: 10.1016/j.aap.2014.05.020.

[270] Karlin Bark, Cuong Tran, Kikuo Fujimura, and Victor Ng-Thow-Hing. Personal navi: Benefits of an augmented reality navigational aid using a see-thru 3D volumetric HUD. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 1–8, New York, NY, USA, 2014. ACM. doi: 10.13140/2.1.3582.0801.

[271] Chris Dijksterhuis, Arjan Stuiver, Ben Mulder, Karel A. Brookhuis, and Dick de Waard. An adaptive driver support system: User experiences and driving performance in a simulator. *Human Factors*, 54(5):772–785, 2012. doi: 10.1177/0018720811430502.

[272] Wai-Tat Fu, John Gasper, and Seong-Whan Kim. Effects of an in-car augmented reality system on improving safety of younger and older drivers. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 59–66, New York, NY, USA, 2013. IEEE. doi: 10.1109/ISMAR.2013.6671764.

[273] Lutz Lorenz, Philipp Kerschbaum, and Josef Schumann. Designing take over scenarios for automated driving: How does augmented reality support the driver to get back into the loop? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58 (1):1681–1685, 2014. doi: 10.1177/1541931214581351.

[274] Sabine Langlois and Boussaad Soualmi. Augmented reality versus classical hud to take over from automated driving: An aid to smooth reactions and to anticipate maneuvers. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1571–1578, New York, NY, USA, 2016. IEEE. doi: 10.1109/ITSC.2016.7795767.

[275] Michelle L. Rusch, Mark C. Schall Jr., Patrick Gavin, John D. Lee, Jeffrey D. Dawson, Shaun Vecera, and Matthew Rizzo. Directing driver attention with augmented reality

cues. *Transportation Research Part F: Traffic Psychology and Behaviour*, 16:127–137, 2013. doi: 10.1016/j.trf.2012.08.007.

[276] Laura Pomarjanschi, Michael Dorr, and Erhardt Barth. Gaze guidance reduces the number of collisions with pedestrians in a driving simulator. *ACM Transactions on Interactive Intelligent Systems*, 1(2):8:1–8:14, 2012. doi: 10.1145/2070719.2070721.

[277] Minh T. Phan, Indira Thouvenin, and Vincent Frémont. Enhancing the driver awareness of pedestrian using augmented reality cues. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1298–1304, New York, NY, USA, 2016. IEEE. doi: 10.1109/ITSC.2016.7795724.

[278] Hyungil Kim, Joseph L. Gabbard, Alexandre M. Anon, and Teruhisa Misu. Driver behavior and performance with augmented reality pedestrian collision warning: An outdoor user study. *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1515–1524, 2018. doi: 10.1109/TVCG.2018.2793680.

[279] Anuj Kumar Pradhan, Kim R. Hammel, Rosa DeRamus, Alexander Pollatsek, David A. Noyce, and Donald L. Fisher. Using eye movements to evaluate effects of driver age on risk perception in a driving simulator. *Human Factors*, 47(4):840–852, 2005. doi: 10.1518/001872005775570961.

[280] Daniel L. Roenker, Gayla M. Cissell, Karlene K. Ball, Virginia G. Wadley, and Jerri D. Edwards. Speed-of-processing and driving simulator training result in improved driving performance. *Human Factors*, 45(2):218–233, 2003. doi: 10.1518/hfes.45.2.218.27241.

[281] Donald L. Fisher, Anuj K. Pradhan, Alexander Pollatsek, and Michael A. Knodler, Jr. Empirical evaluation of hazard anticipation behaviors in the field and on driving simulator using eye tracker. *Transportation Research Record*, 2018(1):80–86, 2007. doi: 10.3141/2018-11.

[282] Ganesh Pai Mangalore, Yalda Ebadi, Siby Samuel, Michael A. Knodler, and Donald L. Fisher. The promise of virtual reality headsets: Can they be used to measure accurately drivers' hazard anticipation performance? *Transportation Research Record: Journal of the Transportation Research Board*, 2673(10):455–464, 2019. doi: 10.1177/0361198119847612.

[283] Uijong Ju, June Kang, and Christian Wallraven. You or me? personality traits predict sacrificial decisions in an accident situation. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1898–1907, 2019. doi: 10.1109/TVCG.2019.2899227.

[284] Unity3D. Unity3D colliders overview. Unity Technologies, Online, 2019. URL https://docs.unity3d.com/Manual/CollidersOverview.html. Accessed: 2019-07-04.

[285] Cathy Cavanaugh. Augmented reality gaming in education for engaged learning. In *Gaming and Simulations: Concepts, Methodologies, Tools and Applications*, pages 45–56. IGI Global, Hershey, PA, USA, 2011. doi: 10.4018/9781609601959.ch103.

[286] Pega Zarjam, Julien Epps, and Fang Chen. Characterizing working memory load using EEG delta activity. In *2011 19th European Signal Processing Conference*, pages 1554–1558, New York, NY, USA, 2011. IEEE. URL https://ieeexplore.ieee.org/document/7074062.

[287] Carina Walter, Wolfgang Rosenstiel, Martin Bogdan, Peter Gerjets, and Martin Spüler. Online EEG-based workload adaptation of an arithmetic learning environment. *Frontiers in Human Neuroscience*, 11:286, 2017. doi: 10.3389/fnhum.2017.00286.

[288] M. Sazzad Hussain, Rafael A. Calvo, and Fang Chen. Automatic cognitive load detection from face, physiology, task performance and fusion during affective interference. *Interacting with Computers*, 26(3):256–268, 2014. doi: 10.1093/iwc/iwt032.

[289] Thomas C Kübler, Enkelejda Kasneci, Wolfgang Rosenstiel, Ulrich Schiefer, Katja Nagel, and Elena Papageorgiou. Stress-indicators and exploratory gaze for the analysis of hazard perception in patients with visual field loss. *Transportation Research Part F: Traffic Psychology and Behaviour*, 24:231–243, 2014. doi: 10.1016/j.trf.2014.04.016.

[290] Christian Braunagel, David Geisler, Wolfgang Rosenstiel, and Enkelejda Kasneci. Online recognition of driver-activity based on visual scanpath classification. *IEEE Intelligent Transportation Systems Magazine*, 9(4):23–36, 2017. doi: 10.1109/MITS.2017.2743171.

[291] Hyungil Kim, Xuefang Wu, Joseph L. Gabbard, and Nicholas F. Polys. Exploring head-up augmented reality interfaces for crash warning systems. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 224–227, New York, NY, USA, 2013. ACM. doi: 10.1145/2516540.2516566.

[292] Panos Konstantopoulos, Peter Chapman, and David Crundall. Driver's visual attention as a function of driving experience and visibility. using a driving simulator to explore drivers' eye movements in day, night and rain driving. *Accident Analysis & Prevention*, 42(3):827–834, 2010. doi: 10.1016/j.aap.2009.09.022.

[293] Johan Engström, Gustav Markkula, Trent Victor, and Natasha Merat. Effects of cognitive load on driving performance: The cognitive control hypothesis. *Human Factors*, 59(5): 734–764, 2017. doi: 10.1177/0018720817690639.

[294] Yutaka Yoshida, Hayato Ohwada, Fumio Mizoguchi, and Hirotoshi Iwasaki. Classifying cognitive load and driving situation with machine learning. *International Journal of Machine Learning and Computing*, 4(3):210–215, 2014. doi: 10.7763/IJMLC.2014.V4.414.

[295] Catherine Gabaude, Bruno Baracat, Christophe Jallais, Marion Bonniaud, and Alexandra Fort. Cognitive load measurement while driving. in: Human factors: a view from an integrative perspective. In *Proceedings HFES Europe Chapter Conference Toulouse*, pages 67–80, Liverpool, United Kingdom, 2012. Human Factors and Ergonomics Society Europe Chapter. ISBN 978-0-945289-44-9.

[296] Fumio Mizoguchi, Hayato Ohwada, Hiroyuki Nishiyama, and Hirotoshi Iwasaki. Identifying driver's cognitive load using inductive logic programming. In *Inductive Logic*

Programming, pages 166–177, Berlin, Heidelberg, Germany, 2013. Springer. doi: 10.1007/978-3-642-38812-5_12.

[297] Lex Fridman, Bryan Reimer, Bruce Mehler, and William T. Freeman. Cognitive load estimation in the wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 652:1–652:9, New York, NY, USA, 2018. ACM. doi: 10.1145/3173574.3174226.

[298] Julian Steil and Andreas Bulling. Discovery of everyday human activities from long-term visual behaviour using topic models. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 75–85, New York, NY, USA, 2015. ACM. doi: 10.1145/2750858.2807520.

[299] Shoya Ishimaru, Kai Kunze, Koichi Kise, Jens Weppner, Andreas Dengel, Paul Lukowicz, and Andreas Bulling. In the blink of an eye: Combining head motion and eye blink frequency for activity recognition with google glass. In *ACM Augmented Human International Conference*, pages 15:1–15:4, New York, NY, USA, 2014. ACM. doi: 10.1145/2582051.2582066.

[300] Krzysztof Krejtz, Andrew T. Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLOS ONE*, 13(9):1–23, 2018. doi: 10.1371/journal.pone.0203629.

[301] Yasunori Yamada and Masatomo Kobayashi. Detecting mental fatigue from eye-tracking data gathered while watching video: Evaluation in younger and older adults. *Artificial Intelligence in Medicine*, 91:39–48, 2018. doi: 10.1016/j.artmed.2018.06.005.

[302] Nora Castner, Enkelejda Kasneci, Thomas Kübler, Katharina Scheiter, Juliane Richter, Thérése Eder, Fabian Hüttig, and Constanze Keutel. Scanpath comparison in medical image reading skills of dental students: Distinguishing stages of expertise development. In *ACM Symposium on Eye Tracking Research & Applications*, pages 39:1–39:9, New York, NY, USA, 2018. ACM. doi: 10.1145/3204493.3204550.

[303] Peter M. van Leeuwen, Stefan de Groot, Riender Happee, and Joost C. F. de Winter. Differences between racing and non-racing drivers: A simulator study using eye-tracking. *PLOS ONE*, 12(11):1–19, 2017. doi: 10.1371/journal.pone.0186871.

[304] Shlomo Berkovsky, Ronnie Taib, Irena Koprinska, Eileen Wang, Yucheng Zeng, Jingjie Li, and Sabina Kleitman. Detecting personality traits using eye-tracking data. In *ACM Conference on Human Factors in Computing Systems*, pages 221:1–221:12, New York, NY, USA, 2019. ACM. doi: 10.1145/3290605.3300451.

[305] Yosef Razin and Karen Feigh. Learning to predict intent from gaze during robotic hand-eye coordination. In *AAAI Conference on Artificial Intelligence*, pages 4596–4602, Palo Alto, CA, USA, 2017. AAAI Press. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewPaper/14270.

[306] Molly B. Ungrady, Maurice Flurie, Bonnie M. Zuckerman, Daniel Mirman, and Jamie Reilly. Naming and knowing revisited: Eyetracking correlates of anomia in progressive aphasia. *Frontiers in Human Neuroscience*, 13:354, 2019. doi: 10.3389/fnhum.2019.00354.

[307] Gerardo Fernández, Facundo Manes, Luis Politi, David Orozco, Marcela Schumacher, Liliana Castro, Osvaldo Agamennoni, and Nora Rotstein. Patients with mild alzheimer's disease fail when using their working memory: Evidence from the eye tracking technique. *Journal of Alzheimer's Disease*, 50(3):827–838, 2016. doi: 10.3233/JAD-150265. Accepted in 2015.

[308] Onur Günlü. *Key Agreement with Physical Unclonable Functions and Biometric Identifiers*. PhD thesis, Technical University of Munich, Munich, Germany, 2018. published by Dr. Hut Verlag in Feb. 2019. ISBN 978-3-8439-3912-6.

[309] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125, New York, NY, USA, 2008. IEEE. doi: 10.1109/SP.2008.33.

[310] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 1054–1067, New York, NY, USA, 2014. ACM. doi: 10.1145/2660267.2660348.

[311] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *International Conference on Neural Information Processing Systems*, pages 3574–3583, Red Hook, NY, USA, 2017. Curran Associates Inc. URL https://proceedings.neurips.cc/paper/2017/file/253614bbac999b38b5b60cae531c4969-Paper.pdf.

[312] Onur Günlü and Onurcan İşcan. DCT based ring oscillator physical unclonable functions. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8198–8201, New York, NY, USA, 2014. IEEE. doi: 10.1109/ICASSP.2014.6855199.

[313] Onur Günlü, Tasnad Kernetzky, Onurcan İşcan, Vladimir Sidorenko, Gerhard Kramer, and Rafael F. Schaefer. Secure and reliable key agreement with physical unclonable functions. *Entropy*, 20(5), 2018. doi: 10.3390/e20050340.

[314] Brendan David-John, Diane Hosfelt, Kevin Butler, and Eakta Jain. A privacy-preserving approach to streaming eye-tracking data. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2555–2565, 2021. doi: 10.1109/TVCG.2021.3067787.

[315] Yang Cao, Masatoshi Yoshikawa, Yonghui Xiao, and Li Xiong. Quantifying differential privacy under temporal correlations. In *IEEE International Conference on Data Engineering*, pages 821–832, New York, NY, USA, 2017. IEEE. doi: 10.1109/ICDE.2017.132.

[316] Georgios Kellaris and Stavros Papadopoulos. Practical differential privacy via grouping and smoothing. *VLDB*, 6(5):301–312, 2013. doi: 10.14778/2535573.2488337.

[317] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *IEEE International Conference on Computer Vision*, pages 3686–3695, New York, NY, USA, 2017. IEEE.

[318] Siyuan Chen and Julien Epps. Using task-induced pupil diameter and blink rate to infer cognitive load. *Human–Computer Interaction*, 29(4):390–413, 2014. doi: 10.1080/07370024.2014.892428.

[319] Ali Borji and Laurent Itti. Defending yarbus: Eye movements reveal observers' task. *Journal of Vision*, 14(3):29:1–29:22, 2014. doi: 10.1167/14.3.29.

[320] Yan Liu, Pei-Yun Hsueh, Jennifer Lai, Mirweis Sangin, Marc-Antoine Nussli, and Pierre Dillenbourg. Who is the expert? analyzing gaze data to predict expertise level in collaborative applications. In *2009 IEEE International Conference on Multimedia and Expo*, pages 898–901, New York, NY, USA, 2009. IEEE. doi: 10.1109/ICME.2009.5202640.

[321] Shahram Eivazi, Ahmad Hafez, Wolfgang Fuhl, Hoorieh Afkari, Enkelejda Kasneci, Martin Lehecka, and Roman Bednarik. Optimal eye movement strategies: a comparison of neurosurgeons gaze patterns when using a surgical microscope. *Acta Neurochirurgica*, 159(6):959–966, 2017. doi: 10.1007/s00701-017-3185-1.

[322] Yasmeen Abdrabou, Mohamed Khamis, Rana Mohamed Eisa, Sherif Ismail, and Amrl Elmougy. Just gaze and wave: Exploring the use of gaze and gestures for shoulder-surfing resilient authentication. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 29:1–29:10, New York, NY, USA, 2019. ACM. doi: 10.1145/3314111.3319837.

[323] Chris Elsden, Arthi Manohar, Jo Briggs, Mike Harding, Chris Speed, and John Vines. Making sense of blockchain applications: A typology for HCI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 458:1–458:14, New York, NY, USA, 2018. ACM. doi: 10.1145/3173574.3174032.

[324] Nick Szabo. Formalizing and securing relationships on public networks. *First Monday*, 2(9), 1997. doi: 10.5210/fm.v2i9.548.

[325] Jackson Ng. Escrow service as a smart contract: The execution. Online, May 2018. URL https://jacksonng.org/escrow-service-smart-contract-execution. Accessed: 2020-09-01.

[326] Hugo Krawczyk, Mihir Bellare, and Ran Canetti. HMAC: Keyed-hashing for message authentication. Internet RFC2104, Feb 1997. URL https://doi.org/10.17487/RFC2104.

[327] Morris J. Dworkin. SHA-3 standard: Permutation-based hash and extendable-output functions. Federal Inf. Process. Stds., NIST FIPS - 202, Aug 2015. URL https://doi.org/10.6028/NIST.FIPS.202.