Spring 2024

Bringing "Virtual" to "Reality": Enhancing Security and Usability on VR System and Applications

Huadi Zhu
The University of Texas at Arlington

BRINGING “VIRTUAL” TO “REALITY”: ENHANCING SECURITY AND USABILITY ON VR SYSTEM AND APPLICATIONS

by

HUADI ZHU

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2024

To my family, with deepest gratitude and love,

especially to my parents,
whose boundless motivation and exemplary guidance have shaped me into the person I am today,

to my beloved wife,
whose unwavering love, support, and encouragement sustained me through the challenges of this journey,

and to my unborn child,
whose forthcoming arrival fills me with hope and purpose.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my Ph.D. supervisor, Prof. Ming Li, as my foremost acknowledgment. Her exceptional guidance, tremendous support, and invaluable insights have been instrumental in shaping both my academic journey and personal growth. Prof. Li's profound knowledge and extensive experience have served as a beacon leading me through the challenges of my career. The privilege of learning from her is a treasured gift that I will carry with me.

I extend my heartfelt appreciation to the members of my dissertation committee, Prof. Yonghe Liu, Prof. Jia Rao, and Prof. Dajiang Zhu, for their constructive feedback and suggestions that significantly contributed to the refinement of my research and for all the assistance in different stages of my Ph.D. study.

I am thankful to a remarkable group of friends and colleagues who offered unconditional help and mental support, and the family-like environment we created together. Special gratitude goes to Srinivasan Murali, Chaowei Wang, Youngtak Cho, Mingyan Xiao, Wenqiang Jin, and Tianhao Li, for their steadfast companionship and contributions. Our shared experiences, countless discussions, and cherished memories have enriched this endeavor beyond measure.

I would also like to thank the University of Texas at Arlington for providing the resources and facilities essential for conducting my research. Additionally, I am grateful to the staff and faculty members who have offered assistance whenever needed.

Lastly, but most importantly, I would like to thank my parents, my wife, and other family members, whose love, encouragement, and understanding have provided the strength and motivation needed to sail through challenges in academia and life. Thank you for be-

ing my pillars and cornerstones and for instilling in me the values of perseverance and resilience. This achievement is as much yours as it is mine.

April 18, 2024

ABSTRACT

BRINGING "VIRTUAL" TO "REALITY": ENHANCING SECURITY AND USABILITY ON VR SYSTEM AND APPLICATIONS

HUADI ZHU, Ph.D. Computer Science

The University of Texas at Arlington, 2024

Supervising Professor: Dr. Ming Li

With the rapid advancements in computer science, electronics, optics, and related fields, virtual reality (VR) gradually penetrates into our daily lives, and is predicted to become a core technology in the near future. Despite their potentials, however, existing designs and solutions for VR applications remain at the infant stage, introducing limited usability and efficiency for real-world users. Besides, the increasing prevalence of VR presents new security and privacy threats due to the vast amount of information stored in or accessible through VR devices. To bridge this gap, we exploit and combine techniques from computer science and human biology, as well as other related domains, to enhance security, usability, and efficiency of these novel applications.

In the dissertation, we investigate the emerging security and privacy threats of existing VR systems, propose novel mitigation schemes, and develop new techniques to improve user experience in emerging VR applications. Our contributions are mainly threefold. First, we introduce novel user authentication schemes on VR via a secure and convenient visual channel. Specifically, we leverage the customized blink patterns and the biometric pupil variations to identify legitimate users, which is deployable for commercial VR devices. We

further enhance this work by exploiting the phenomenon of auditory-pupillary response and introducing an effort-free biometric authentication scheme for VR devices, which outperforms all state-of-the-art solutions. Second, we propose to harness users' ocular behaviors to enable accurate quality of experience (QoE) assessment for 360-degree videos, by modeling these cues into a graph and applying graph learning techniques to extract hidden information in predicting the QoE score. Third, we build a novel video recommender system for VR users leveraging additional insights from users' physiological signals to learn their preferences and interests and make corresponding recommendations.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iv

ABSTRACT vi

1. INTRODUCTION 1

2. BLINKEY: A TWO-FACTOR USER AUTHENTICATION METHOD FOR VIRTUAL REALITY DEVICES 4

2.1 Introduction 4

2.1.1 Motivation 4

2.1.2 Proposed Methodology 6

2.2 Related Work 8

2.2.1 Knowledge-based Authentication 8

2.2.2 Biometric Authentication 9

2.2.3 Rhythm-Based Authentication 11

2.3 Threat Model 13

2.4 BlinKey Characterization and Features 14

2.4.1 Definition of BlinKey 14

2.4.2 Feature Selection 14

2.5 System Design 17

2.5.1 System Overview 17

2.5.2 Start and End Detection 19

2.5.3 Pre-processing 20

2.5.4 Feature Extraction 23

2.5.5 Classification 24

2.6 Performance Evaluation 31
2.6.1 Prototype Implementation & Experiment Setup 31
2.6.2 Robustness Against Attacks 33
2.6.3 Usability 40
2.6.4 Survey Results 43
2.7 Discussions 43
2.8 Conclusions 47
3. SOUNDLOCK: A NOVEL USER AUTHENTICATION SCHEME FOR VR DEVICES USING AUDITORY-PUPILLARY RESPONSE 48
3.1 Introduction 48
3.2 Preliminaries 52
3.2.1 Background on Auditory-Pupillary Response 52
3.2.2 Measurement Study 54
3.2.3 Problem Statement 56
3.3 Basic Scheme Design 57
3.3.1 Preprocessing 57
3.3.2 Feature Extraction and Selection 58
3.3.3 Classification 62
3.4 Advanced Scheme with Multi-stimuli 63
3.5 Security Analysis 67
3.5.1 Robustness Against Attacks 67
3.5.2 Entropy Analysis 68
3.6 Experiment Methodology 70
3.6.1 Experiment Setup 70
3.6.2 Experiment Design 70
3.6.3 Recruitment and Ethical Aspects 72

3.7 Results . . . 74
3.7.1 Pilot Study–Classifier Selection . . . 74
3.7.2 In-field Study–System Performance . . . 76
3.7.3 Performance Under Various Scenarios . . . 78
3.7.4 User Study . . . 81
3.8 Related Work . . . 84
3.9 Limitations and Future Work . . . 87
3.10 Conclusion . . . 89
4. EYEQOE: A NOVEL QoE ASSESSMENT MODEL FOR 360-DEGREE VIDEOS USING OCULAR BEHAVIORS . . . 94
4.1 Introduction . . . 94
4.2 Related Work . . . 99
4.3 Background . . . 101
4.4 Measurement Study . . . 103
4.5 Modeling Ocular Behaviors into Graphs . . . 107
4.6 EyeQoE . . . 109
4.6.1 A Basic GCN-based QoE Assessment Model . . . 109
4.6.2 Dealing with Subjects and Visual Stimuli Heterogeneity . . . 112
4.6.3 Dealing with Unseen Videos . . . 114
4.6.4 Piecing All Together . . . 117
4.7 Evaluation . . . 117
4.7.1 Settings . . . 117
4.7.2 Overall Performance . . . 120
4.7.3 Micro Benchmarks . . . 126
4.8 Discussion and Future Work . . . 131
4.9 Conclusion . . . 134

5. *Phyre*: A NOVEL VIDEO RECOMMENDER SYSTEM FOR VIRTUAL REALITY USING PHYSIOLOGICAL SIGNALS . . . 135
5.1 Introduction . . . 135
5.1.1 Background . . . 135
5.1.2 Challenges . . . 136
5.1.3 Our Solution: *Phyre* . . . 137
5.2 Related Work . . . 139
5.3 Measurement Study . . . 141
5.4 System Overview . . . 143
5.5 Graph-based Recommendation . . . 144
5.5.1 Graph Construction . . . 144
5.5.2 GCN-based Video Recommendation . . . 146
5.6 Cross-Modality Cross-Context Domain Adaptation . . . 150
5.7 Energy-efficient Adaptive Encoding . . . 154
5.8 Evaluation . . . 155
5.8.1 Evaluation Setup . . . 155
5.8.2 Overall Performance . . . 158
5.8.3 Impact of Base Models and Public Datasets . . . 159
5.8.4 Ablation Study . . . 160
5.8.5 Robustness Against Impact Factors . . . 161
5.8.6 System Overhead . . . 162
5.8.7 Micro Benchmarks . . . 163
5.9 Discussion and Future Work . . . 163
5.10 Conclusion . . . 164
6. CONCLUSION AND FUTURE WORK . . . 166
REFERENCES . . . 168

CHAPTER 1

INTRODUCTION

The rapid development of virtual reality (VR) has been seen in the past few years with a consistently growing popularity. According to a recent report [217], the VR market is around $28 billion in 2022; by 2030, the number is forecast to reach over $87 billion with a constant annual growth rate of 15%. With the capability of providing an immersive and interactive experience, VR has revolutionized gaming and entertainment and permeated a variety of applications, including e-commerce, education, healthcare, and military [261]. For example, retailers can bridge physical and online stores via VR to provide an immersive shopping experience for customers [168]; medical practitioners may communicate with patients in a VR environment for remote diagnosis [177]; military actions can be simulated and practiced in a virtual battlefield [159]. In the above applications, tremendous amounts of sensitive data are collected, processed, and stored on VR devices, such as customers' credit card information, patients' health status, and military secrets. Adversarial access to VR devices would cause data breaches and other critical consequences. Therefore, implementing user authentication mechanisms in VR is a crucial step in resisting unauthorized access. On the other hand, novel VR applications such as video streaming and recommendation suffer poor usability. Novel solutions tailored for the VR context are in dire need in improving user experience with VR.

In this dissertation, we intend to enhance the security and usability of VR technologies by exploiting human biosignals, aiming to bring the "virtual" VR techniques into the tangible "reality" of everyone's daily lives. The rest of the dissertation is organized as follows.

Chapter 2 investigates security threats against existing user authentication schemes on VR, and presents the design, implementation, and evaluation of a two-factor user authentication scheme, *BlinKey*, for VR devices that are equipped with an eye tracker. A user's secret passcode is a set of recorded rhythms when he/she blinks, together with the unique pupil size variation pattern. We call this passcode as a blinkey, which can be jointly characterized by knowledge-based and biometric features. To examine the performances, *BlinKey* is implemented on an HTC Vive Pro with a Pupil Labs eye tracker. Through extensive experimental evaluations with 52 participants, we show that our scheme can achieve the average EER as low as 4.0% with only 6 training samples. Besides, it is robust against various types of attacks. *BlinKey* also exhibits satisfactory usability in terms of login attempts, memorability, and impact of user motions. We also carry out questionnaire-based pre-/post-studies. The survey result indicates that *BlinKey* is well accepted as a user authentication scheme for VR devices.

Chapter 3 further proposes SoundLock, a novel user authentication scheme for VR devices using auditory-pupillary response as biometrics. During authentication, auditory stimuli are presented to the user via the VR headset. The corresponding pupillary response is captured by the integrated eye tracker. User's legitimacy is then determined by comparing the response with the template generated during the enrollment stage. To strike a balance between security and usability in the scheme design, an optimization problem is formulated. Due to its non-linearity, a two-stage heuristic algorithm is proposed to solve it efficiently. The solution provides necessary guidance for selecting effective auditory stimuli and determining their corresponding lengths. We demonstrate through extensive in-field experiments that SoundLock outperforms state-of-the-art biometric solutions with FAR (FRR) as low as 0.76% (0.91%) and is well received among participants in the user study.

In Chapter 4, we develop EyeQoE, a novel method that models eye-based cues into graphs and develop a GCN-based classifier to produce QoE assessment by extracting intrinsic features from graph-structured data. We further exploit the Siamese network to eliminate the impact from subjects and visual stimuli heterogeneity. A domain adaptation scheme named MADA is also devised to generalize our model to a vast range of unseen 360-degree videos. Extensive tests are carried out with our collected dataset. Results show that EyeQoE achieves the best prediction accuracy at 92.9%, which outperforms state-of-the-art approaches. As another contribution of this work, we have publicized our dataset on `https://github.com/MobiSec-CSE-UTA/EyeQoE_Dataset.git`.

In Chapter 5, we introduce *Phyre*, a video recommender system tailored for VR. Our approach leverages viewers' physiological responses as they engage with VR videos to infer their preferences and thus make future recommendations. We integrate these new physiological user-video interaction measures into the mainstream recommendation framework and renovate the graph learning-based paradigm to accommodate the new changes. The recommender system is further empowered with a novel domain adaptation approach named CMCCDA to address the data scarcity problem for model training. We also develop an energy-efficient adaptive encoding scheme to reduce the energy consumption on the VR device. We collect a physiological dataset for video recommendation in VR and demonstrate through extensive evaluation that *Phyre* significantly outperforms state-of-the-art schemes by up to 68.0% in recommendation precision and up to 28.8% in ranking quality.

Finally, Chapter 6 concludes this dissertation and discusses future research work.

CHAPTER 2

BLINKEY: A TWO-FACTOR USER AUTHENTICATION METHOD FOR VIRTUAL REALITY DEVICES

2.1 Introduction

2.1.1 Motivation

Virtual Reality (VR) is an immersive technology that allows users to experience a virtual world with a head mount device. The rapid development of VR has been seen in the past few years with a consistently growing popularity. According to [83], 20.8 million people in the US used VR headset in 2019. This number is forecast to grow to 28.1 million by 2021. Statistics also show that the worldwide shipment of VR devices has grown over 60% in the past two years [73]. By 2025, the value of the VR market is expected to reach USD 87.97 billion, from USD 11.52 billion in 2019 [120]. While VR is traditionally used for recreational purposes, it is now rapidly permeating a variety of mission-critical applications ranging from e-business [11, 50, 258], healthcare [61, 222, 273], social networking [97, 119, 283], manufacturing [36, 89, 181], military training [142, 243, 271], and education [9, 22, 101].

In these applications, VR devices store their users' personal information, such as emails, photos, videos, and browsing history, as well as their online login accounts and passwords. Recently, online shopping and in-app purchases have emerged as important e-commerce opportunities for VR. For example, eBay launched a VR department store, where users can shop around in a virtual environment and make transactions online [33]. VR is also deemed as the future of social media interactions. In March 2020, Facebook started beta testing for its new VR social network "Horizons" where users engage with

news content, share information, and entertain themselves in the virtual world by logging into Horizons using their accounts and passwords [191]. In the above scenarios, as the process of inputting data to current VR systems tends to be tedious, users may store their account and credit card information for auto-login and in-app purchase [194]. As a result, such practices may result in the security breach and even financial loss if the device is accidentally left unattended to people with ill purposes, including close friends and roommates. Therefore, the employment of user authentication mechanisms is crucial for VR devices. Only the owner or authorized users are able to unlock the device, while outliers are prohibited from access.

Unfortunately, user authentication on VR devices is yet far from well investigated. Current solutions, including password, digital PIN, and drawing pattern, mostly follow conventional approaches for general personal devices. However, these schemes have been proved vulnerable to shoulder-surfing attacks [82, 99, 105], as how password/PIN/pattern entered in VR device leaves little leverage to obfuscate the secret entry process. If the adversary is aware of the virtual digit board layout, it can easily decode hand movements to infer PIN inputs. The inference is even easier for the pattern-based authentication since the attacker only needs to track the hand movement trajectory without exquisite knowledge of the virtual board input design. Moreover, because a user's view is completely blocked from the physical world by the headset, it renders the user challenging to be aware of the presence of shoulder-surfing attackers.

To resist shoulder-surfing attacks, a shuffled keyboard has been proposed [14, 211]; the system adopts a new randomly generated keyboard layout each time a user intends to enter the credential. While leaving the key inference almost impossible, it sacrifices the authentication usability. Extra effort is incurred to the user in searching for keys on a shuffled keyboard. Recently, some novel user authentication methods for VR devices have been introduced. A couple of them focus on the improvement of the explicit knowledge-

based authentication schemes, such as 3D password [98, 305] and spatial targets [94, 125]. These methods provide more robust authentication by implementing more complicated secret codes. However, they do not improve usability, if not further worsening it. For example, in [305], users are required to remember and enter a complicated 3D drawing pattern for authentication, which results in longer authentication time and a higher error rate. Some existing efforts employ the implicit biometrics to defend against shoulder-surfing attacks [21, 145, 185, 204]. Nonetheless, using biometrics alone suffer from irrevocability, which renders replay attacks a severe threat if even a single user's biometric sample is acquired by an attacker [87]. There are also some prior works on two-factor authentication [17, 34, 155]. So far, the existing solutions either rely on highly advanced equipment, such as a customized sensory headset with a number of electrodes to capture the brain signals[155], which is not readily available on current VR devices, or introduce heavy cognitive load that has users to perform complex and tedious authentication tasks.

2.1.2 Proposed Methodology

In this paper we propose *BlinKey*, a practical two-factor authentication scheme for VR devices that are equipped with eye trackers. Users authenticate themselves by blinking eyes following certain rhythm only known by themselves. It is a new passcode-style authentication. Rather than numbers, letters, or characters, users choose different beats/rhythms when blinking. Basically, a blinkey[1] can be easily created by the user, for example, by extracting some beats from his/her favorite songs or jingles. The knowledge-based feature of a blinkey is characterized by the timing of its blinks, which can be recorded by the eye tracker together with the system clock. Additionally, a blinkey is also characterized by its biometric features. We observe that how human pupils adapt to light after blinks, more

[1] In this paper, we utilize the Italian font *BlinKey* to represent the authentication scheme, while the regular font blinkey as the password itself.

specifically, the variation of pupil size, is unique for each person. As a blinkey is composed of multiple blinks, we then treat the pupil size variation, captured by the eye tracker, between blinks as a biological marker. Incorporating the above knowledge-based and biometric features, *BlinKey* serves as two-factor authentication to determine whether a user is legitimate or not.

BlinKey can be an ideal solution for user authentication on VR devices. First, it can effectively resist shoulder-surfing attacks. Unlike conventional PIN/password/pattern authentication, which requires users to hold the controller to enter credentials, *BlinKey* is simply performed by user blinking eyes. As the visual sight is blocked by the headset, it is impossible for the adversary to observe the passcode entry process. Second, it is convenient to perform. *BlinKey* is a hand-free authentication without imposing effort-demanding tasks. Third, as it involves both explicit knowledge and implicit biometric features, it is robust against attacks, such as guessing attacks and shoulder-surfing attacks. Although *BlinKey* only works for VR devices that are equipped with eye trackers, they are not a small population. To our knowledge, many VR headsets, such as HTC Vive Pro Eye [116], FOVE 0 [93], Pico Neo series [2], and Varjo VR-1 [270], are all in this category. These devices can therefore provide eye blinks and pupil size variations to the authentication unit on the device. We would like to note that integrating eye-tracking technology is a trend of VR headsets [223, 247], as it significantly improves user experience. For example, it helps VR headsets to simulate depth of field and focus, providing a more realistic and natural visual experience. *BlinKey*, as another user authentication scheme, can be employed for accessing both stand-alone devices and online accounts. This is also the case for many other user authentication schemes. For example, fingerprint-based authentication is widely adopted not only by a broad set of personal devices but also by some online services, such as online banking [86, 193].

BlinKey is composed of two phases. In the enrollment phase, users are asked to create their own blinkeys and enter them multiple times for the training purpose. During the login phase, the user simply enters the previously enrolled blinkey to unlock the device. If it matches the training samples, the user is authorized; otherwise, the access request is denied. To investigate the performance of *BlinKey*, we recruit 52 volunteers and collect 1306 blinkey samples from them. Classification accuracy is studied concerning different parameter settings. Based on the result, we implement our scheme on a commercial VR device with the parameter values that optimize the authentication performance. Another 43 participants are recruited. Multiple in-field experiments are conducted to evaluate the system performance in terms of attack resistance, time consumption, login attempts, the impact of user motions, and memorability, which outperform state-of-the-art solutions.

2.2 Related Work

2.2.1 Knowledge-based Authentication

In recent years, how to authenticate users in VR devices has been increasingly explored in both computing and security research communities. George et al. carried out user study for the direct transfer of well-established user authentication concepts, including PIN and pattern lock, into VR [99]. Due to their vulnerability to shoulder-surfing attacks, a shuffled keyboard is proposed [14, 211]. Users enter their credentials on a virtual keyboard with a randomly generated layout each time. Yu et al. then develop a 3D pattern lock that creates an additional entropy for user's secret credentials [305]. Funk et at. [94] developed a graphical authentication mechanism based on gaze-tracking, called LookUnlock. The passcode consists of a set of virtual objects that a user's gaze focuses on in the correct sequence. A similar idea is adopted by [98, 125]. These schemes produce rather limited key space. For a passcode constructed by selecting 4 objects in a sequence from a total of 9 objects, the key

space is merely $P(9,4) = 3,024$, even smaller than that of the 4-digit PIN[2]. Moreover, it is not an easy task to remember the correct sequence of 4 objects. For example, according to the result [98], their 7-day recall rate is 74.1%. As shown in Section 2.6.3.3, this value is 89.6% for *BlinKey*. Mathis et al. proposed RubikAuth [170], where users select digits from a virtual 3D cube manipulated with a handheld controller. Following a similar idea, RubikBiom [169] further takes into account user behavioral biometric features such as hand movement when entering credentials from the virtual 3D cube. With the introduction of an additional layer of protection, RubikBiom is more robust against guessing attacks and shoulder-surfing attacks. As noted by the authors, both schemes require two-handed interactions which are inconvenient for users with motor disabilities. *BlinKey* is free from such a restriction for allowing users to enter their authentication credentials with eye blinks. Alsulaiman and Saddik [17] propose a 3D password that combines textual passwords and the user's behavior biometrics for entering the password.

2.2.2 Biometric Authentication

Unlike knowledge-based authentication, which is based on "what you know", biometric authentication leverages "who you are" by looking into the unique biometrics that people are naturally born with. It has gained preference in certain situations due to its robustness against guessing attacks and shoulder-surfing attacks.

Gesture biometrics: Prior works [145, 185, 204] extract user's distinctive biometric features from their head/hand/body movements for user authentication. These schemes require users to turn the head, bend the body in different directions, or throw/catch particular virtual objects. The involved actions may be awkward to perform especially in public places.

[2]We will discuss in Section 2.7 that *BlinKey* offers the key space orders of magnitude higher than conventional passcodes, such as digit-PIN and password, of the same length.

Table 2.1: Comparison among different user authentication approaches on VR devices. (*) The work [99] discusses both PIN and pattern lock for VR. ●: method fulfills criterion. ○: method does not fulfill criterion. –: not enough information.

Scheme	Key space	Hand-free	Sensor-free	Accuracy	Security	Auth speed	Memorability
PIN [99]*	⋆	○	●	⋆⋆⋆	⋆	Short	⋆⋆⋆
Pattern lock [99]	⋆	○	●	⋆⋆⋆	⋆	Short	⋆⋆⋆
Shuffled keyboard [14]	⋆	○	●	⋆⋆⋆	⋆⋆	⋆⋆	⋆⋆⋆
LookUnlock [94]	⋆	●	●	⋆⋆⋆	⋆⋆	⋆⋆	⋆
3D Pattern [305]	⋆⋆	○	●	⋆⋆⋆	⋆⋆	Short	⋆
RubikAuth [170]	⋆⋆	○	●	⋆⋆⋆	⋆⋆	Short	⋆
Hand gesture[145]	⋆⋆	○	●	⋆	⋆⋆	Short	-
Brain biometrics [155]	⋆⋆⋆	●	○	⋆⋆	⋆⋆	⋆⋆	-
Head movement [185]	⋆⋆	●	●	⋆	⋆⋆	Long	-
SkullConduct [234]	⋆⋆	●	○	⋆	⋆⋆	Short	-
Eye movements [240]	⋆⋆	●	○	⋆⋆	⋆⋆	Short	-
3D Password [17]	⋆⋆⋆	○	●	⋆⋆	⋆⋆⋆	⋆⋆	⋆⋆⋆
RubikBiom [169]	⋆⋆⋆	○	●	⋆⋆	⋆⋆⋆	Short	⋆
BlinKey (this work)	⋆⋆⋆	●	●	⋆⋆	⋆⋆⋆	⋆⋆	⋆⋆

Gaze biometrics: Gaze tracking has recently been explored for user authentication. Existing solutions either examine the position or the content that a user is looking at or eye movement. The former is based on the hypotheses that each user's gaze behaves uniquely when watching the screen [45, 220, 221]. These schemes rely on a large number of data samples to extract sufficient features for accurate authentication. As a result, they typically take more than one minute to authenticate a user. The second class of gaze-based authentication leverages the uniqueness of eye movement to fingerprint each user [28, 80, 114, 115, 137]. Relevant features include eye movement velocity and saccade latency. As pointed out by [311], these solutions suffer from irrevocability, which is in fact a common pitfall for many pure biometric-based authentication schemes. To address this issue, [240, 311] introduce the idea of random stimuli. As a result, the biometric features observed in each authentication trial become dynamic, leaving reply attacks in-

feasible. Nonetheless, they only work with precise eye movement tracking. For example, [240] requires a sampling rate of up to 500 Hz and tracking error within 0.4°, which cannot be met by current add-on eye trackers for VR devices.

Other biometrics. Schneegass et al. [234] present SkullConduct, a biometric system that uses bone conduction of sound through the user's skull for user identification. A microphone is used to capture the skull vibration. Recently, Lin et al. [155] utilized responsive brainwaves when a user is presented with visual stimuli for authentication. Sophisticated electrodes should be integrated into VR headsets to capture the human brainwave.

2.2.3 Rhythm-Based Authentication

Only a few rhythm-based authentication schemes have been proposed so far. Wobbrock [287] developed an authentication system for single-key devices called "TapSongs", which enables user authentication on a single "binary" sensor (e.g., button) by matching the rhythm of tap down/up events to a jingle timing model created by the user. A group authentication scheme, Thumprint [68], was proposed by Das et al., using the rhythm of a secret knock to authenticate a group of users, while each user's expression of the secret is discernible. Chen [54] built a two-factor rhythm-based authentication scheme for multi-touch mobile devices. Recently, Hutchins et al. [118] developed a rhythm-based authentication scheme for wearable devices equipped with a touching sensor. TapMeIn [188] is another authentication method for smartwatches. On top of the secret tapping rhythm, it jointly considers biometric features, such as pressure and finger size of tapping. All these features are captured by smartwatch's touching screen/sensors that are missing from current VR devices. Thus, TapMeIn is inapplicable to our case. Observing that how human pupils adapt to light after blinks, more specifically, the variation of pupil size, is unique for each person, we then treat it as a biological marker. Together with the user's blinking rhythm, they are both captured by the eye tracker and serve as secret credentials for user authentication.

Summary. Table 3.9 provides a comprehensive comparison between some representative user authentication schemes for VR and *BlinKey*. The existing schemes are categorized into three groups, knowledge-based authentication (light gray), biometric-based authentication (medium gray), and two-factor authentication (dark gray). The comparison is made from the aspects of security (including "key space" and "security") and usability (including "hand-free", "extra sensors", "accuracy", "login time", and "memorability").

The salient advantage of knowledge-based authentication is mainly on its usability, with the highest accuracy and the lowest login time among the three groups. As user's passcodes are mostly entered by hand controllers, no extra sensor is needed. Nonetheless, these schemes have been criticized for their security, for example, vulnerable to shoulder-surfing and/or statistic attacks. This issue is partially resolved by some biometric-based authentication schemes. First, biometric features can barely be eavesdropped. Second, user's unique biometrics introduce a much larger key space. On the other hand, due to the hardware restriction, the explorable biometrics from VR devices are still limited so far. Some approaches require users to perform body/head/hand movement that can be readily captured by VR devices; some others rely on extra sensors, e.g., EMG and ECG sensors, to extract biometric features. There also have been a couple of two-factor authentication schemes that combine regular knowledge-based passcodes and user biometrics. Most of them exhibit better security performances than the other two. However, due to the involvement of behavior biometric features, which are dynamic even from the same user, their accuracy is degraded a little bit than knowledge-based schemes. Apparently, *BlinKey* belongs to the third group. Compared with [17, 169], it is entered hand free and thus friendly for users with motor disabilities. Moreover, as discussed in Section 7, rhythmic patterns produce a significantly larger key space than conventional PIN/password/pattern lock; so is *BlinKey* compared with [17, 169].

2.3 Threat Model

The adversary's goal is to impersonate the legitimate user and successfully authenticate itself to the VR device. This work pertains to the discussion of the following commonly seen attacks.

- *Zero-effort attack.* The adversary does not have any side information of the enrolled blinkey and tries to get authenticated by random guessing. It is also referred to as *guessing attack* in some other literature.
- *Statistical attack.* The adversary has access to a large volume of blinkeys and is aware of the set of features utilized by the scheme. It performs statistical analysis over the dataset and derives probability distribution over each feature. Then, the adversary forges synthetic blinkeys following the acquired distributions.
- *Shoulder-surfing attack.* The adversary is able to observe the authentication process while the victim is entering a blinkey. Then it mimics the legitimate user by repeating what it has observed.
- *Credential-aware attack.* This attack is even more powerful than the shoulder-surfing attack. We assume that the adversary has the full information of the legitimate user's secret blinking rhythm. The only difference from the shoulder-surfing attack is that the latter acquires blinking rhythm via visual observation.

We also make the following assumptions throughout the paper. The adversary cannot compromise the VR device or its connected server to access the user's blinkeys; otherwise, it renders secure user authentication design impossible. Due to the similar reason, the connection between the VR device and the server is also deemed secure.

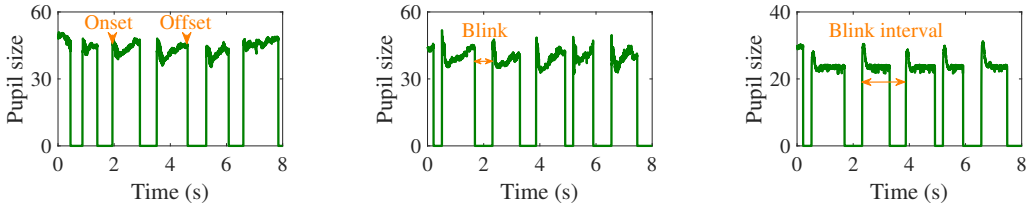
2.4 BlinKey Characterization and Features

2.4.1 Definition of BlinKey

A blinkey is composed of time instances stamped by the system clock when a user blinks in a self-designed rhythm, together with variations of pupil size exhibited between consecutive blinks. Both information can be recorded by the eye tracker. Figure 2.1 gives three exemplary blinkeys. Blink onset/offset indicates the moment that eyes are open/closed. An eye is deemed closed when its pupil size is measured zero and open otherwise. The length of a blinkey is simply the number of blinks it contains. For example, all the three blinkeys in Figure 2.1 are of length 6. We observe that the pupil size is not fixed between blinks. It experiences some fluctuations in the following procedures. When the eye is open, the eye tracker quickly captures the pupil's instantaneous size, which is at a large value. Then the pupil quickly adapts to ambient light by adjusting its diameter. After a short period, around dozens to hundreds of milliseconds, the pupil returns to a relatively stable status with micro-fluctuations. More importantly, we find that such a pupil's adaptation pattern varies across people. Figure 2.1(b) and 2.1(c) demonstrate the same blinking rhythm performed by two users. While the rhythm is almost identical, the way how pupil size changes is clearly distinct between two trials. This is due to the pupil dilation/constriction that is controlled by the iris muscles with a biologically unique pattern [228]. Moreover, we also notice in Figure 2.1 that the pupil size variation pattern is consistent from the same user. Based on the above observation, we thus treat the pupil size variation between blinks as an additional dimension of features that fingerprint individuals.

2.4.2 Feature Selection

Since a blinkey consists of both knowledge-based features ("something you know") and biometric features ("something you are"), we are interested in identifying suitable feature set for user authentication.



(a) Blinkey 1 (b) Blinkey 2 (c) Blinkey 3

Figure 2.1: Three exemplary blinkeys.

2.4.2.1 Knowledge-based Features

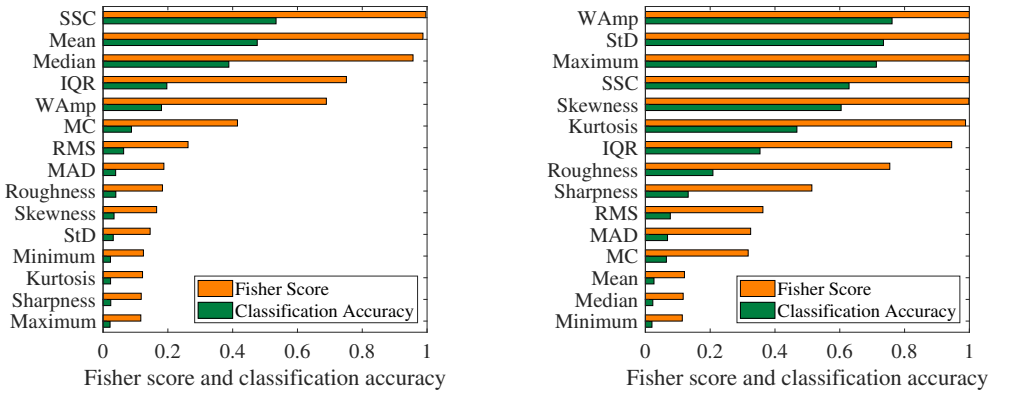
The knowledge-based features are the blink rhythm designed by the legitimate user. We mainly focus on the following three features *blink time instance*, *blink interval*, and *relative interval*.

- **Blink time instance.** The blink rhythm can be uniquely identified by a set of blink onsets and blink offsets, indexed by their timestamps, which are represented by two vectors $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \cdots, \alpha_n\}$ and $\boldsymbol{\beta} = \{\beta_1, \beta_2, \cdots, \beta_n\}$. Here, α_i and β_i are the timestamps for the i^{th} blink onset and offset, respectively, and n is the blinkey length, i.e., the number of blinks contained. For analysis consistency, we index the first blink onset as 0, $\alpha_1 = 0$. In other words, we deem the starting point of a blinkey as the moment when a user opens her eyes for the first time to perform her blinkey.
- **Blink intervals.** To characterize a blinkey's rhythm, we further extract the inter-onset intervals of a blinkey, defined as the time duration between two adjacent blink onsets, as shown in Figure 2.1: $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \gamma_3, ..., \gamma_{n-1}\}$, where $\gamma_i = \alpha_{i+1} - \alpha_i$.
- **Relative intervals.** In actual scenarios, users' input speed may be influenced by their moods or other factors. Thus the time instance for each blink and their intervals may be different even for the same user entering a same blinkey. To take this into account, we introduce another feature, relative interval, which is defined as the ratio of a blink interval to its previous one: $\boldsymbol{\eta} = \{\eta_1, \eta_2, \eta_3, ..., \eta_{n-2}\}$, where $\eta_i = \frac{\gamma_{i+1}}{\gamma_i}$.

2.4.2.2 Biometric Features

As discussed above, the pupil size variation of each user can be treated as her biometric identifier. We now investigate the proper set of features to extract for authentication.

- **Fourier coefficients.** From the perspective of frequency analysis, the pupil size variation consists of components under different frequencies. To extract this information, we then apply the fast Fourier transform (FFT) over time-domain samples. The Fourier coefficient associated with each frequency component then serves as part of biometric features, $\boldsymbol{\phi} = \{\phi_1, \phi_2, \cdots, \phi_m\}$, where ϕ_i ($i \in [1, m]$) is the mean Fourier coefficient of the i^{th} frequency component. The larger coefficient of a higher frequency component a user produces, the more agile her pupils adapt to luminance. Computation and parameter setting details regarding Fourier coefficient extraction will be discussed in Section 2.5.3.
- **Statistical features.** In addition to Fourier coefficients, we further explore a few statistical features in both time and frequency domains that have been widely adopted in characterizing signals [18, 146, 164]. A set of candidate statistical features include, `Maximum`, `Minimum`, `Mean`, `Median`, `Root Mean Square (RMS)`, `Standard Deviation (StD)`, `Mean Absolute Deviation (MAD)`, `Kurtosis`, `Skewness`, `Interquartile Range (IQR)`, `Roughness`, `Sharpness`, `Mean Crossing (MC)`, `Willison Amplitude (WAmp)`, `Slope Sign Change (SSC)`, in time (T) and frequency (F) domains.
  Since not all of them play essential roles in our task, it is necessary to filter out non-critical ones. For this purpose, we calculate the *Fisher score* for each above feature. As one of the most commonly used supervised feature selection methods [158, 244], the Fisher score takes the inter-class variance and the in-class variance over the values of a given feature and computes their ratio. A higher ratio indicates that the distances between classes are much larger than those within the same class.



(a) Statistical features in time domain (b) Statistical features in frequency domain

Figure 2.2: Fisher score and classification accuracy for statistical features.

Classification accuracy, an accuracy indicator of feeding each feature alone into the classifier, shows how well these features work for the classifier individually. Hence, we compute both the Fisher score and the classification accuracy for each statistical feature. Their results are shown in Figure 2.2(a) and Figure 2.2(b), respectively, in a descending order of a combination of both metrics. To facilitate the discussion, the Fisher score is normalized. We thus pick the top-ten best features, i.e., with the highest combined values, to constitute the statistical feature set $\boldsymbol{s} = \{\text{WAmp}_F, \text{StD}_F, \text{Maximum}_F, \text{SSC}_F, \text{Skewness}_F, \text{SSC}_T, \text{Mean}_T, \text{Kurtosis}_F, \text{Median}_T, \text{IQR}_F\}$. The result is shown in Table 2.6.

The entire feature set to characterize a blinkey is then written as $\boldsymbol{f} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\phi}, \boldsymbol{s}\}$.

2.5 System Design

2.5.1 System Overview

Figure 2.3 shows an overview of the *BlinKey* system. It involves registration (or called enrollment) phase and login (or called testing) phase. For either phase, the workflow of data processing is summarized as follows. Authentication is turned on when a user

Table 2.2: The selected statistical features.

Feature	Definition	Fisher	Accuracy
WAmp_F	The count of significant changes in frequency	0.9999	0.7614
StD_F	The extent of deviation in frequency	0.9998	0.7348
Maximum_F	The maximum amplitude in frequency	0.9999	0.7129
SSC_F	The count of slope sign changes in frequency	0.9992	0.6287
Skewness_F	The distortion in frequency	0.9984	0.6042
SSC_T	The count of slope sign changes in time	0.9962	0.5339
Mean_T	The average amplitude in time	0.9873	0.4758
Kurtosis_F	The sharpness of the peak in frequency	0.9882	0.4677
Median_T	The value that divides the signal in half in time	0.9567	0.3882
IQR_F	The 1st quartile subtracted from the 3rd in frequency	0.9454	0.3538

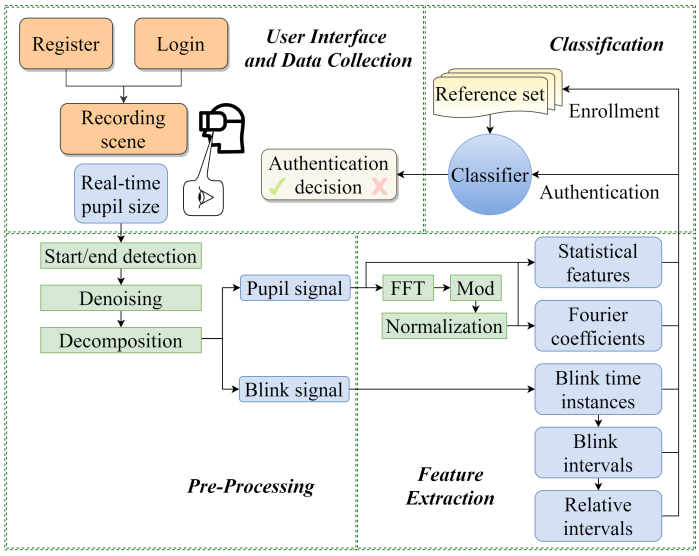


Figure 2.3: System overview.

awakens the screen, opens an app, or triggers a purchase interface. In a pop-up virtual scene, the user is asked to blink in a self-designed pattern as an input blinkey. Once the authentication procedure is activated, the eye tracker keeps recording the user's real-time pupil size signals and transmit them to the server. The signal first passes the start/end detection module so as to segment the entire blinkey. The raw signal is then denoised and decomposed. Its outputs, including *blinking rhythm* and pupil size variations, are then

fed into the feature extraction module to distill knowledge-based and biometric features. Finally, the classifier decides whether the given blinkey is legitimate or not.

2.5.2 Start and End Detection

A challenge of our approach lies in how to detect a blinkey, more specifically, identify its start and end points. This task is easy for authentication on regular personal devices, such as smartphones and tablets. For the case of pattern lock, the moment that a finger touches/leaves the screen is simply the start/end point of one trial. These moments can be accurately recognized by touching sensors embedded in the screen. For the case of password-based authentication, the end of one entry is explicitly indicated by tapping the enter/return key. Unfortunately, such hardware is unavailable at VR devices. One viable solution is to create a virtual enter/return key. However, it may incur extra effort for a user to interact with the virtual screen via a controller. Alternatively, we propose to have a user to indicate the start/end of a blinkey for closing the eyes a while, as shown in Figure 2.4. In this way, the moment that the user opens eyes for the first time after the long blink is treated as the start of the blinkey. Similarly, the moment that the user closes eyes right before the long blink is treated as the end of the blinkey.

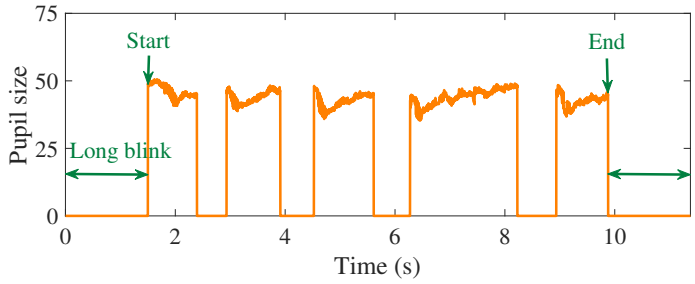


Figure 2.4: Illustration of start/end of a blinkey and the long blink.

The remaining question is to decide the duration for the long blink. Recall that a single blink is determined by the forceful closing of the eyelid. The system should be capable of differentiating between a long blink and a blink belonging to a blinkey or a spontaneous blink. We start by analyzing the statistics of spontaneous blinks based on the 434 blink samples collected from 22 volunteers. Its statistical distribution is plotted in Figure 2.5. We find that the duration of spontaneous blinks ranges from 0.09 to 0.26 second with its mean as 0.12 second, and the 95th percentile as 0.18 second. Our discovery concurs with the result of UCL Researcher [37], stating that the duration of a spontaneous blink is on average 0.1 - 0.15 second, as well as the result of Harvard Database of Useful Biological Numbers [195], stating that the duration of spontaneous blinks mainly ranges from 0.1 to 0.4 second.

We further investigate the statistics of voluntary blinks of blinkeys. Its distribution is derived based on another phase of data collection, where we acquired 1306 blinkey samples from 52 volunteers. The details of this data collection phase are provided in 3.3.3. We observe in Figure 2.5 that the 95th percentile exists at 1.95 seconds. Based on the statistical analysis, we set the duration of the long blink as 2.5 seconds. A longer duration will sacrifice the usability of authentication, while a shorter value renders the detection error-prone. As a note, a user does not have to estimate the exact 2.5 seconds before performing a blinkey, as long as the waiting duration is no less than the threshold. This requirement is easy to meet.

2.5.3 Pre-processing

The objective of this component is twofold, to filter out noise in the raw signal and to decompose the signal into ingredients that contain knowledge-based and biometric features separately.

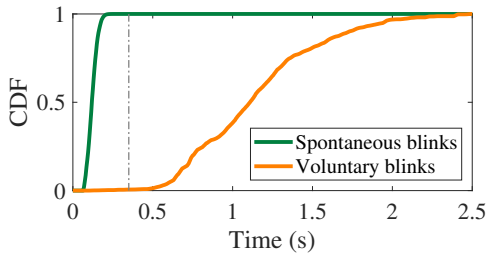


Figure 2.5: Statistical distribution of duration for spontaneous blinks and voluntary blinks.

2.5.3.1 Denoising

As shown in Figure 2.6, the raw signal is mainly composed of three components: voluntary blinks, spontaneous blinks, and the pupil adaptive variations between blinks. The useful information includes voluntary blinks and pupil adaptations. Spontaneous blinks are conducted in the pre-motor brain stem and happen without conscious efforts, like breathing and digestion. It helps to spread the tear to all parts of the eyes and helps to keep them moist [285]. They are done involuntary and distinct from the voluntary blinks in a blinkey. As the involvement of spontaneous blinks brings the noise to the feature extraction and thus authentication accuracy, the goal of this phase is to eliminate spontaneous blinks from the raw signal.

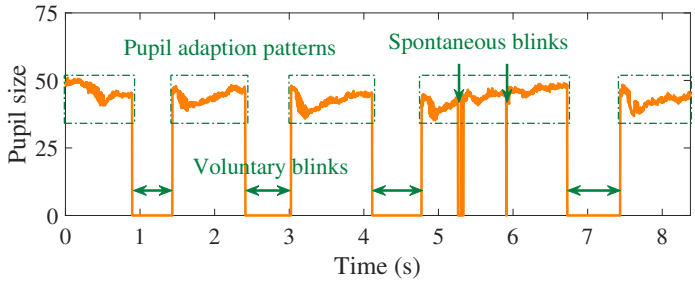


Figure 2.6: The raw signal of a blinkey mainly consisting of voluntary blinks, pupil adaptations, and spontaneous blinks.

As shown in Figure 2.5, the statistical analysis indicates that the time duration of spontaneous blinks and voluntary blinks is clearly distinct from each other. It is noticed that the former mostly falls within the range from 0.09 to 0.26 second, while the latter is between 0.45 and 2.5 seconds. Motivated by this observation, we thus set a detection threshold at 0.35 second. For a blink whose duration is beyond this value, it is treated as a voluntary one; otherwise, it is a spontaneous one, which is eliminated from the raw signal. Meanwhile, it is infeasible to directly set their associated pupil size to 0's, as it will pollute the blinkey's features. Instead, we apply the *spline interpolation* [286]. As a common interpolation technique, it estimates missing data using a mathematical function that minimizes overall surface curvature. In our case, pupil sizes of spontaneous blinks are treated as the missing data and interpolated accordingly. In this way, we eliminate spontaneous blinks from the signal while preserving the blinkey features.

2.5.3.2 Decomposition

The goal of decomposition is to extract from the denoised signal user's *blinking rhythm* and *segments*, which carry knowledge-based features and biometric features of a blinkey, respectively. The decomposition facilitates the feature extraction next. As shown in Figure 2.7, a *segment* is simply the set of non-zero pupil size values between two consecutive blinks. A segment reflects the user's pupil variations after each voluntary blink. The decomposition is done by detecting all onsets and offsets in a blinkey. Since humans perform eyelid opening and closure rapidly, it leads to sharp rises and drops in the observed pupil size. Therefore, the detection of onsets and offsets can be accomplished via simple edge detection algorithms.

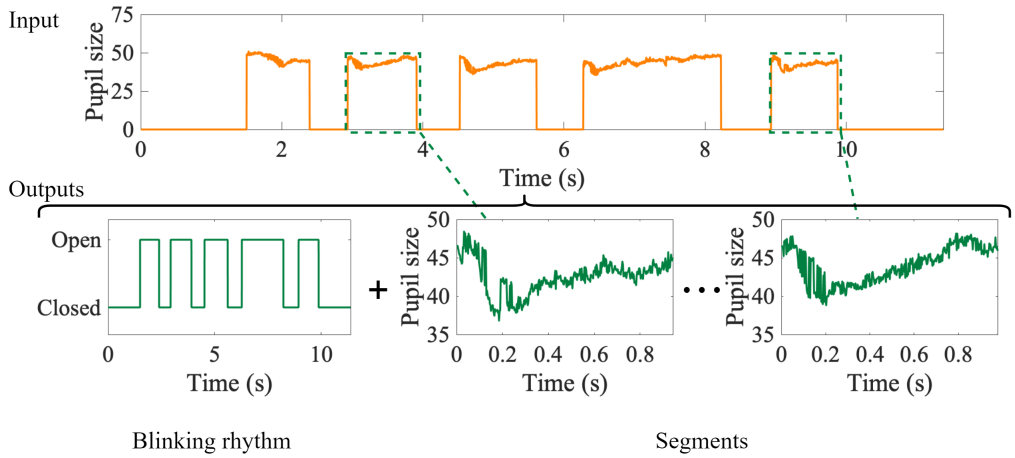


Figure 2.7: Illustration of decomposition. The input (top) is denoised signals and the output (bottom) is *blinking rhythm* and *segments*.

2.5.4 Feature Extraction

Once *blinking rhythm* and *segments* are ready, we are ready to extract from them desired features.

Knowledge-based features can be directly derived from the *blinking rhythm*. Specifically, we first obtain the time instances of onsets and offsets, i.e., $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Then, the blink interval set $\boldsymbol{\gamma}$ and relative interval set $\boldsymbol{\eta}$ are calculated following their definitions in Section 2.4.2.1.

Biometric features involved in our scheme are classified into time-domain features and frequency-domain features. For the former, they include SSC, Mean, Median, etc. They can be computed based on time-series samples from one blinkey entry following definitions of these metrics. For the latter, they include WAmp, StD, Maximum, etc. As the first step, we employ FFT to decompose time-domain samples into their constituent frequencies. The frequency-domain representation can decompose complicated pupil size variations into periodic components that time-domain analysis cannot realize. FFT is applied over each segment. Before that, we first employ zero-padding to ensure that each

segment has the same length of 1024 data points. The reason for choosing 1024 is twofold. First, FFT works most efficiently for a signal with length a power of 2 since it recursively folds the size at each step. Second, we observe from the 1306 collected samples that all segments last within 5 seconds. Given the sampling rate as 200 Hz in our system, every segment is sampled into 1000 data points the maximum. Based on the above discussion, we pad the segment into 1024 data points. Once the Fourier coefficients are derived for each segment, we take their average over all segments for each frequency component. It then produces the Fourier coefficient feature $\boldsymbol{\phi}$. Frequency-domain statistical features are computed following a similar method for time-domain statistical features.

2.5.5 Classification

Once features of a blinkey are extracted following previous steps, the remaining task is to apply classification methods for user authentication, i.e., to discriminate the legitimate user and imposters. Two common classification methods are considered, one-class Support Vector Machine (SVM) and K-Nearest Neighbors (k-NN). To determine which one best serves our system, we conduct comprehensive evaluations based on our dataset consisting of 1306 blinkeys from 52 volunteers. These volunteers are all college students, including 36 females and 16 males. The classification performances are examined through the following metrics.

- False Rejection Rate (FRR). The probability that a legitimate user is rejected by the system. It is calculated as the ratio of the number of a legitimate user's incorrect authentications to the total number of attempts.
- False Acceptance Rate (FAR). The probability that an impostor is given access, computed as the ratio of the number of an impostor's authentication attempts that are accepted by the system to the total number of attempts.

- Equal Error Rate (EER). The point at which FRR and FAR are equal, by adjusting parameter values.

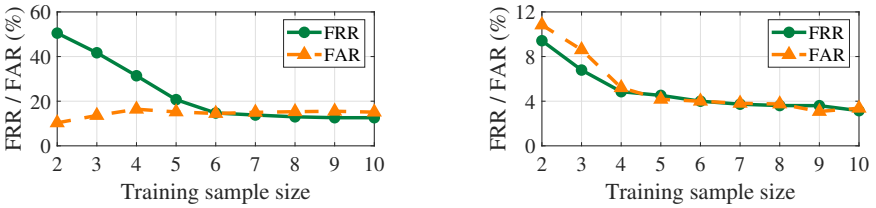
Note that FRR reflects the user convenience in our system; a lower FRR implies that a legitimate user can successfully unlock the VR device at a higher probability. FAR reflects the security aspect; a lower FAR implies that the imposter will be denied at a higher probability. It is worth noting that two blinkeys are deemed different with different lengths. For example, if the legitimate blinkey has a length of 6, then any testing input with a different length will be rejected immediately. Hence, in the following we only focus on the classification over blinkeys of the same length.

To investigate the performance of *BlinKey*, we performed two user studies. In phase I, the objective is to collect blinkeys created by different users so as to carry out statistical analysis and classification model selection as discussed here. In phase II, a prototype of *BlinKey* is built. We then conduct a series of in-field experiments to evaluate the security and utility of our system which will be covered in the next section.

In the phase-I user study, a specialized app is developed and implemented on the VR system to facilitate the data collection. A total of 52 volunteers are recruited. They are all college students aged from 18 to 35. Among them, there are 36 females and 16 males. Their demographic details are provided in Table 2.3. Before the data collection, they are explained how *BlinKey* works. Each volunteer is asked to design several different blinking patterns. For each pattern, a video of the volunteer's pupils is recorded by an eye tracker in the VR headset. Afterwards, they are shown the detected pupil size signal and are asked to manually mark the voluntary blinks from the spontaneous blinks according to their self-designed patterns. In total, we obtain 1306 samples containing 7,528 voluntary blinks and 3,673 spontaneous blinks. The collected dataset is used to derive statistics of blinkeys. Besides, we also aim to identify suitable parameters for the classifier.

Table 2.3: Demographics of volunteers in the phase-I study.

Gender	No.	Age range	No.	Eye color	No.	Eye wear type	No.
Female	36	18-23	19	Black	29	None	25
Male	16	24-29	28	Brown	14	Colorless glasses	21
		30-35	5	Hazel	8	Colorless contact lenses	4
				Green	1	Colored contact lenses	2



(a) One-class SVM (b) One-class k-NN

Figure 2.8: FRR and FAR under different training sample sizes. (a) One-class SVM. (b) One-class k-NN.

2.5.5.1 Support Vector Machine

One-class SVM has been successfully applied to a number of classification problems. It generalizes the idea of finding an optimal hyper-plane in high-dimensional space to perform classification. Compared to other classification methods, it has advantages in implementation simplicity and efficiency in dealing with high-dimensional, non-linear datasets. Here, one-class SVM is implemented with the Radial Basis Function (RBF) kernel.

The number of training samples is an important indicator of classification performances. We tune the value from 2 to 10 and evaluate its impact. Figure 2.8(a) shows the authentication accuracy with respect to the training sample size. We observe that the FRR is as high as 50.5% with only 2 training samples. It drops quickly to 14.6% under 6 training samples. It mildly decreases to 12.6% as the training sample size grows to 10. The FAR grows from 10.3% with 2 training samples and keeps relatively stable around 15.0% as the training sample size increases to 10. The minimum EER 14.6% is achieved with 6 training samples.

We further evaluate the performance of SVM with respect to the kernel coefficients γ and ν in Figure 2.9. Here, γ is the standard deviation of the kernel function. It influences the decision boundary qualitatively. As γ grows, FAR increases while FRR decreases, which means both legitimate users and impostors are more likely to get authenticated. In fact, for a larger γ, the decision criteria tend to be relaxed to avoid the hazard of overfitting. For a smaller γ, the decision boundary tends to be strict and sharp. In contrast to the former situation, it tends to overfit. The parameter ν is an upper bound on the fraction of margin errors and a lower bound of the fraction of support vectors relative to the total number of training samples. For example, ν=0.01 means that at most 1% of the training samples are misclassified (at the cost of a small margin, though) and at least 1% of the training samples are support vectors. Hence, as shown in Figure 2.9, a larger ν leads to a lower FAR but at the cost of a higher FRR. Combining the results above, EER reaches its lowest point at 14.6% when training sample size, γ, and ν are set to 6, 0.018, and 0.028, respectively. Hence, one-class SVM produces unsatisfactory authentication accuracy in our system.

2.5.5.2 K-Nearest Neighbors

Another classification method under consideration is k-NN. It measures the similarity between the testing sample and training samples. The similarity is represented by the *Manhattan distance*. If the score is below the threshold, the testing sample is considered a legitimate input; otherwise, it is an outlier.

We first examine the classification accuracy with respect to the training sample size. As shown in Figure 2.8(b), both FRR and FAR decreases with a larger training sample size. The detection accuracy improvement becomes insignificant, with 6 or more training samples. To balance between accuracy and usability, we use 6 samples to train the model. Comparing between Figure 2.8(a) and Figure 2.8(b), we find k-NN produces a much lower

FRR (%)	$\gamma \cdot 10^{-2}$																			
$\nu \cdot 10^{-2}$	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4
1.2	17.7	20.5	17.6	16.5	15.2	15.1	14.9	14.7	13.3	12.9	12.5	11.4	11.1	10.6	9.72	8.74	8.53	8.62	8.35	8.38
1.4	18.1	20.7	18.1	16.7	15.6	15.9	15.4	15.1	13.5	13.1	12.7	11.7	11.2	11.1	10.2	9.36	9.27	8.88	8.84	8.22
1.6	20.9	21.2	18.5	17.2	16.1	16.1	15.6	15.3	13.5	13.3	13.3	11.6	11.4	11.1	10.8	9.60	9.48	9.19	9.33	8.73
1.8	21.8	21.2	18.7	17.6	16.3	16.7	15.8	15.6	13.8	13.5	13.5	11.9	11.6	11.2	10.9	9.79	10.0	9.47	9.31	8.91
2	22.3	21.5	18.9	17.8	16.9	16.9	15.7	15.5	13.9	13.8	13.7	12.1	11.3	11.8	11.2	10.3	10.2	9.67	9.51	9.33
2.2	22.7	21.7	19.4	18.1	17.3	17.1	16.0	16.0	14.2	14.0	13.7	12.4	11.6	12.0	11.4	10.5	10.4	9.87	9.71	9.53
2.4	23.1	22.1	19.6	18.3	17.5	17.3	16.2	16.0	14.1	14.1	13.9	12.8	12.5	12.2	11.6	10.7	10.6	10.1	9.91	9.73
2.6	23.5	22.3	19.8	18.5	17.9	17.6	16.5	16.3	14.5	14.2	14.1	13.0	12.7	12.5	11.8	10.9	10.8	10.3	10.1	9.76
2.8	25.4	22.7	20.0	19.0	18.1	17.8	16.7	16.2	**14.6**	14.3	14.3	13.4	12.9	12.7	12.0	11.1	11.0	10.5	10.3	10.0
3.0	26.2	23.0	20.2	19.2	18.3	18.1	16.8	16.4	14.9	14.5	14.5	13.6	13.1	12.8	11.9	11.2	11.2	10.7	10.5	10.2

FAR (%)	$\gamma \cdot 10^{-2}$																			
$\nu \cdot 10^{-2}$	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4
1.2	13.5	12.4	13.7	15.2	16.6	16.9	16.6	17.9	17.9	18.6	19.4	19.6	20.2	21.5	22.2	22.5	23.6	26.3	36.1	38.1
1.4	11.7	11.8	13.5	15.7	16.4	16.5	16.4	17.7	15.7	18.5	17.3	19.3	20.0	21.3	20.0	21.3	23.4	25.2	33.3	35.1
1.6	12.2	11.6	13.3	15.5	16.2	16.2	16.2	18.0	15.5	17.6	15.6	19.1	20.0	21.1	20.0	21.1	22.8	25.0	32.1	33.7
1.8	11.8	11.4	13.1	15.3	15.8	15.5	15.8	15.3	15.6	15.9	15.7	18.9	19.8	20.9	19.8	20.9	22.6	24.8	31.9	32.3
2	10.1	11.2	12.9	15.1	14.7	14.9	14.7	15.1	15.8	15.1	15.5	19.1	19.6	20.7	19.6	20.7	22.4	24.4	30.5	30.6
2.2	9.70	11.0	12.7	14.9	14.5	14.9	14.5	14.9	15.2	15.5	15.3	18.5	19.4	20.5	19.4	20.5	22.4	24.2	28.8	29.2
2.4	8.46	10.8	12.5	14.0	14.3	14.5	14.3	14.7	15.0	14.7	15.1	17.4	21.4	20.3	21.4	20.3	22.3	25.1	27.4	27.8
2.6	9.30	10.6	12.3	13.8	14.1	14.4	14.1	14.5	14.8	14.8	14.9	17.2	19.4	20.2	19.4	20.2	22.1	24.4	26.0	26.4
2.8	9.10	10.4	12.1	14.1	13.9	14.3	13.9	14.3	**14.6**	14.6	14.7	17.4	19.2	19.8	19.2	19.8	21.5	23.2	25.2	26.5
3.0	8.90	10.2	11.9	13.9	13.7	13.8	13.7	14.1	14.4	14.2	14.5	17.2	19.0	19.1	19.0	19.1	21.3	23.0	24.0	25.1

Figure 2.9: FRR and FAR with respect to γ and ν under one-class SVM.

error rate. Given 6 training samples, EER of SVM and k-NN is 14.6% and 4.0%, respectively. The latter is less than 1/3 of the former.

We then investigate the impact of two critical parameters, k, the number of neighbors to select, and α, the threshold from the Manhattan distance matrix. A larger k indicates that more neighbors are taken into the calculation of the classification score. A larger α means a testing sample is more likely to be accepted legitimate. The results demonstrated in Figure 2.10 meet our expectations. A larger α, i.e., a loose detection rule, results in lower FRR but a higher FAR. As we increase the value of k, the classification becomes more stable due to majority voting/averaging, and thus, is more likely to make more accurate detection. Nonetheless, as k is beyond a certain value, we will witness an increasing number of errors as the value of k is pushed too far. As shown in Figure 2.10, the lowest EER exists at 4.0% with $k = 3$ and $\alpha = 1.0$.

FRR (%)		α															
		0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
k	1	39.4	22.8	9.21	4.55	2.61	0.92	0.56	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	49.3	28.9	17.8	12.4	5.29	3.81	2.42	0.93	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	3	73.3	49.3	26.5	15.0	7.70	**4.00**	3.86	2.59	1.48	0.74	0.56	0.19	0.19	0.00	0.00	0.00
	4	88.6	65.7	47.7	38.1	22.8	15.0	8.84	4.55	2.42	0.93	0.56	0.56	0.19	0.19	0.19	0.19
	5	100	97.0	54.9	40.0	26.1	16.2	17.2	6.03	3.81	1.87	1.68	1.48	0.56	0.56	0.56	0.56
	6	100	100	98.8	97.0	67.1	49.3	37.0	22.8	9.21	5.29	4.55	3.07	2.61	2.22	0.93	0.56

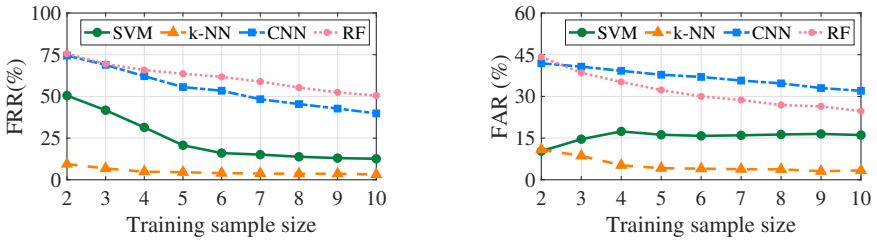
FAR (%)		α															
		0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
k	1	0.00	0.75	0.42	3.90	8.82	12.8	14.7	19.6	30.0	40.3	44.5	47.6	50.5	54.8	55.5	57.3
	2	0.00	0.00	0.75	1.50	3.25	5.48	9.40	12.8	14.5	17.2	24.8	35.8	41.5	27.0	48.6	51.6
	3	0.00	0.00	2.00	0.75	1.92	**4.00**	6.03	7.54	10.5	12.6	12.8	14.5	16.0	26.4	35.8	40.3
	4	0.00	0.00	0.00	0.00	0.75	1.92	4.03	6.20	8.14	10.4	13.2	13.8	15.9	23.5	27.6	32.6
	5	0.00	0.00	0.00	0.00	0.75	1.50	3.00	3.87	8.00	10.1	12.2	12.4	13.5	18.1	23.1	26.8
	6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.75	1.50	2.00	3.25	3.87	5.56	10.4	12.8	13.9

Figure 2.10: FRR and FAR with respect to α and k under one-class k-NN.

2.5.5.3 Other Classifiers

We further examine the classification accuracy of convolutional neural networks (CNN) and random forests (RF) in the latest version. Specifically, one-class CNN and one-class RF are considered. The former is based on CNN for one-class classification problems. Its idea is to use a zero centered Gaussian noise in the latent space as the pseudo-negative class and train the convolutional network using the cross-entropy loss to learn a good representation and the decision boundary for a given class [200]. CNN has been widely applied to computationally complex classification tasks, such as image defect detection [308] and face verification [199]. One-class RF is a method based on a random forest algorithm and an original outlier generation procedure that makes use of classifier ensemble randomization principles [74]. The basic idea is to to use some randomization principles of ensemble learning methods to sub-sample the number of features and the number of training target instances to make possible the generation of outliers from the computation perspective, and to make use of the information given by the target samples to adapt accordingly the outlier distribution. Compared to CNN, it is faster to perform and requires fewer data samples.

As shown in Figure 2.11, given the same training sample size, k-NN achieves the lowest FRR and FAR among the four classifiers, while CNN and RF exhibit the worst performance. This is because the latter two generally require a large dataset to properly train their models. An empirical implication indicates that it typically takes at least 5,000 samples to train CNN with 10 or more layers and hundreds of neurons for satisfying accuracy in applications like image classification. Similarly, the training sample size is around 500 to train RF for relatively good performance in a classification problem. On the other hand, only 6 samples are needed for k-NN to obtain EER as low as 4.0%. It indicates that k-NN attains a promising authentication accuracy with much fewer training samples, especially compared with CNN and RF. Besides, with simple structures, k-NN and SVM consume fewer computation resources for training and testing than the other two. Thus, they are deployable to a wide spectrum of VR devices with heterogeneous resource capacities.



(a) FRR (b) FAR

Figure 2.11: Authentication accuracy comparison among four classifiers.

To sum up, k-NN outperforms SVM, CNN, and RF in terms of classification accuracy, given the same training sample size in our case. More importantly, our design acquires limited training samples, as few as 6. Hence, the enrollment of a blinkey can be performed efficiently. Besides, our approach also outperforms [185, 311], two recently proposed user

authentication schemes for VR devices, in terms of authentication accuracy. For [185], its EER is 7.4%. For [311], its EER is 6.9%. Both are higher than ours.

2.6 Performance Evaluation

2.6.1 Prototype Implementation & Experiment Setup

As a proof-of-concept implementation, we develop the prototype of *BlinKey* on an HTC Vive Pro head-mount device, connected to a local server[3] running SteamVR to support the VR environment. We install a Pupil Labs eye tracker in the VR device to record the real-time pupil size. The sampling rate is set to 200 Hz, i.e., pupil size samples are collected every 5 milliseconds. The collected data are fed into the server through ZeroMQ application program interface (API). All the functions, such as start/end detection, pre-processing, feature extraction, and classification, are implemented in Unity, a cross-platform engine for VR games. As observed in Section 3.3.3, k-NN yields better accuracy than SVM in our system. Hence, we implement the former as the classifier in our prototype. The training sample size is set to 6, which means a user is asked to enter her blinkey 6 times in the enrollment phase. We set the parameters k as 3 and α as 1.0, since the k-NN demonstrates the best authentication accuracy with this setting. For comparison purposes, we also implement the basic PIN and pattern lock authentication schemes on the same VR device. Their corresponding passcodes are entered using controllers paired with the device.

To evaluate the security and usability of *BlinKey*, another 43 participants are recruited to conduct experiments. Among them, 13 volunteers also participated in the prior data col-

[3]The local server is a typical arrangement for the tethered VR headset, which our prototype device HTC Vive Pro belongs to. The local server is not a required element for *BlinKey*. Although our prototype makes use of the local server to do classification, the computation load is pretty light. Instead of any resource-demanding classification models, such as neural networks, *BlinKey* employs the light-weight k-NN. Thus, the computation can be practically supported on standalone VR devices with on-board computing units.

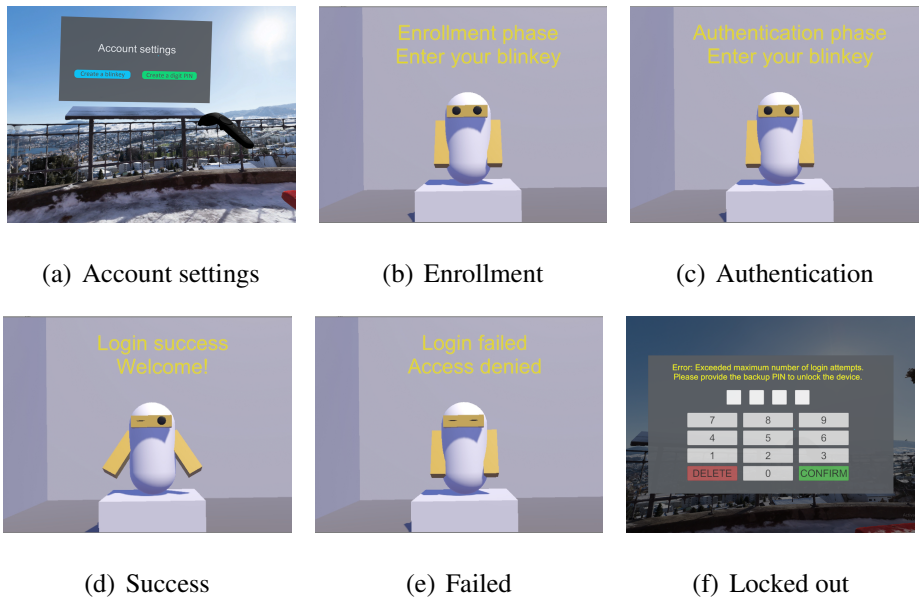
Table 2.4: Distribution of volunteer information.

Gender	No.	Age	No.	Eye color	No.	Eye wear type	No.	Experience	No.
Female	16	18-23	15	Black	20	None	23	None	25
Male	27	24-29	26	Brown	18	Glasses	17	Limited	14
		30-35	2	Hazel	3	Colorless lenses	2	Proficient	4
				Blue	2	Colored lenses	1		

lection session. The distribution of the participants' information is shown in Table 2.4. At the beginning of the experiment, the basic idea of *BlinKey* is explained to the participants. They are then trained on how to correctly enter a blinkey. Thereafter, they are asked to create their own blinkeys.

Screenshots of the user interface (UI) of our prototype are shown in Figure 2.12. UI is implemented in a virtual scene in Unity and displayed in the VR headset to guide users for enrollment and authentication. For the blinkey enrollment, we follow the basic steps of how an iPhone enrolls a user's fingerprints. Specifically, when legitimate users boot their new VR devices for the first time, they are guided to the process of account setting. As one of the steps, users are prompted to enroll their blinkeys (see Figure 2.12(a)). Users are asked to enter their blinkeys repeatedly until 6 valid samples have been collected (see Figure 2.12(b)). If a user tends to enroll another blinkey, the user is first required to provide the existing blinkey correctly. Then the rest steps similarly follow the ones for the initial account setup. The authentication is automatically triggered as a user puts on the VR headset, initiates an online purchase, or tries to log into her Internet account. A dialog box pops up, asking the user to enter her valid blinkey (as shown in Figure 2.12(c)). Based on the input, the classifier decides whether this entry is from the legitimate user: if yes, the access is granted (see Figure 2.12(d)); otherwise, the access is denied with an error message shown on the screen (see Figure 2.12(e)). If denied, a user can re-enter her blinkey until

reaching the maximum number of attempts allowed, say 5. Then, the account is temporarily locked, and the recovery process is invoked (see Figure 2.12(f)).



(a) Account settings (b) Enrollment (c) Authentication

(d) Success (e) Failed (f) Locked out

Figure 2.12: Screenshots of UI for *BlinKey*.

2.6.2 Robustness Against Attacks

The adversary's goal is to impersonate a legitimate user and successfully get authenticated to the VR device. We assume that the adversary has physical access to the device. In practice, such physical access can be gained in ways such as a thief stealing a device, finders finding a lost device, and a roommate temporarily accessing a device when the owner is taking a shower. In the experiment, we consider the following types of attacks: *zero-effort attacks*, *statistical attacks*, *shoulder-surfing attacks*, and *credential-aware attacks*.

Table 2.5: Success rate of zero-effort attacks under different blinkey lengths.

Blinkey length	3	4	5	6	7	8	9	10
FAR (%)	8.1	4.4	3.4	1.9	0	0	0	0

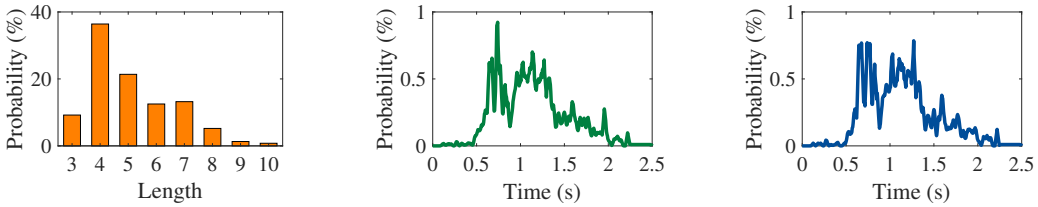
2.6.2.1 Zero-effort Attacks

Zero-effort attacks may be the most common type of attacks against an authentication system, where the attacker guesses the secret or tries the authentication procedure without much knowledge of the legitimate password. In our case, each volunteer (attacker) is asked to randomly pick blinkeys without any prior knowledge of the legitimate one and tries to pass the authentication by chance. Up to five authentication attempts can be made. An attack is considered to succeed if any one of them passes the authentication.

Table 2.5 shows the success rate of zero-effort attacks, which is directly the FAR of our mechanism. Among 1306 collected blinkeys, all of them have the length between 3 and 10. Hence, we conduct tests over blinkey with their lengths falling within this range. Clearly, the blinkey length plays a critical role in the success rate of zero-effort attacks. The longer a blinkey is, the less possible it can be compromised by an adversary. Particularly, if the length is 7 or longer, the success rate drops to zero. Therefore, in the practical implementation of *BlinKey*, the system can impose a hard constraint over a valid blinkey's minimum length, say 7, to defeat zero-effort attacks.

2.6.2.2 Statistical Attacks

This type of attack assumes that the adversary has access to a abroad set of user's blinkeys. This type of attackers employ knowledge obtained from the statistics of a group of blinkeys as hints to generate authentication attempts. The basic approach is to estimate the feature distribution and then use the most probable feature values to generate the forgery.



(a) Blinkey length (b) Blink duration (c) Opening-phase duration

Figure 2.13: Probability distributions of knowledge-based features.

In the experiment, we use the 1306 collected blinkey samples and produce a set of forgery blinkeys as follows. We first randomly select a length following the probability distribution of all blinkey lengths, as illustrated in Figure 2.13(a). Then we randomly choose values for each eye blink and open following their probability distributions derived from our dataset. Figure 2.13(b) and Figure 2.13(c) depict these two distributions. Finally, a set of 150 forgery blinkeys is generated in this process.

Table 2.6: Success rate of statistical attacks under different blinkey lengths.

Blinkey length	3	4	5	6	7	8	9	10
FAR (%) of statistical attacks	5.2	6.4	2.8	2.4	1.8	0	0	0

An attacker is randomly assigned multiple forgery blinkeys and tries to get authenticated by repeating them. Hence, attackers use their own pupils and thus biometric features to launch the attack. Table 2.6 shows the success rate, i.e., FAR, of statistical attacks of *BlinKey*. The attacker's success rate drops to 0 for blinkeys when their lengths reach 8. Notably, statistic analysis does not grant the attacker much privilege over zero-effort attacks.

We further the variation pattern of *BlinKey*, specifically, the rhythm pattern distribution of blinkeys (without considering the biometric features) based on our dataset. Its purpose is to examine if users tend to choose similar blinking rhythms which would render

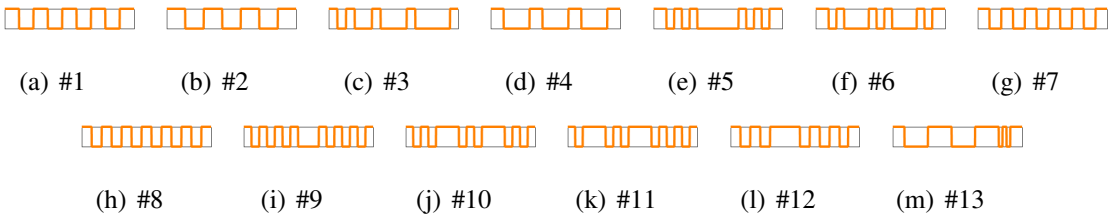
the scheme vulnerable to statistical attacks. As shown in Table 2.7, we list the top-13 most frequently used blinkeys by analyzing 1306 valid enrollments in the dataset. 11 of them are the same, indexed as #1 blinkey, with their frequency calculated as 2.1%. Besides, there are also duplicates for #2–#10 blinkeys, with their occurrence frequencies as 1.5%, 0.9%, 0.8%, 0.6%, 0.6%, 0.4%, 0.4%, 0.4%, and 0.4%, respectively.

Table 2.7: Frequency of blinkeys from collected dataset.

Index	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Frequency	2.1%	1.5%	0.9%	0.8%	0.6%	0.6%	0.4%	0.4%	0.4%	0.4%

It implies that users are less likely to choose the same blinking pattern. Therefore, attackers can barely obtain useful information from the statistic analysis over a set of blinkeys. We acknowledge that our dataset is limited in its size, with only 1306 blinkeys. Still, our analysis partially reflects the blinking pattern distribution in practice. Compared with regular digit-PIN and password, a blinkey can be characterized by a more rich set of features, including tapping time instances, tapping intervals, relative intervals, and even pupil size variations. All these factors make *BlinKey* robust against statistical attacks.

We further visualize in Figure 2.14 the most frequently adopted blinkey patterns that are presented in Table 2.7. As shown, the patterns that exhibit uniform rhythms (#1, #2, #4, #7, and #8) or symmetric rhythms (#5, #6, #9, and #10) are more likely to be adopted. Such a phenomenon is also observed in PINs; the commonly picked PINs include 000000, 010101, etc., which share similar properties above. Note that Blinkey is a two-factor user authentication that also involves biometric features. Hence, it effectively avoids PIN and password pitfalls caused by popular credential selections.



(a) #1 (b) #2 (c) #3 (d) #4 (e) #5 (f) #6 (g) #7

(h) #8 (i) #9 (j) #10 (k) #11 (l) #12 (m) #13

Figure 2.14: Visualization of the top-13 frequently selected blinkey patterns.

2.6.2.3 Credential-Aware Attacks

A credential-aware attack is when the adversary has the full knowledge of the blinking rhythm of a blinkey. Therefore, it can extract all the knowledge-based features, including blink time instances, blink intervals, and relative intervals. To launch this type of attack, we provide the attacker all the above-mentioned information regarding victim blinkeys. As discussed in statistical attacks, it is unlikely for the adversary to reproduce the legitimate user's biometrics. Likewise, to launch credential-aware attacks against PIN and pattern lock, adversaries are informed with details of victim PINs and drawing patterns. Based on this information, the attacker tries to gain access to the system. Table 2.8 compares the success rate against three types of authentication schemes. While PIN and pattern are compromised, *BlinKey* effectively resists the attack. This is because *BlinKey* also involves biometric features, which are hard to mimic, in addition to credentials. Meanwhile, we also notice that the leakage of credentials does provide attackers advantage in compromising the system. For instance, given the length of 7, the attacker's success rate is 0 under zero-effort attacks, while it increases to 14.2% under credential-aware attacks. This result indicates that biometric features alone, i.e., pupil size variations, cannot deliver satisfactory security performance. Luckily, the success rate against *BlinKey* is merely 4.4% when the length is 10. Therefore, one viable solution to defend credential-aware attacks is to adopt a longer blinkey. As a note, the length of a pattern lock is defined by the number of points a user draws through. For instance, the length of a "Z" pattern (1-2-3-5-7-8-9) is 7.

Table 2.8: Success rate of credential-aware attacks on *BlinKey*, PIN, and pattern lock.

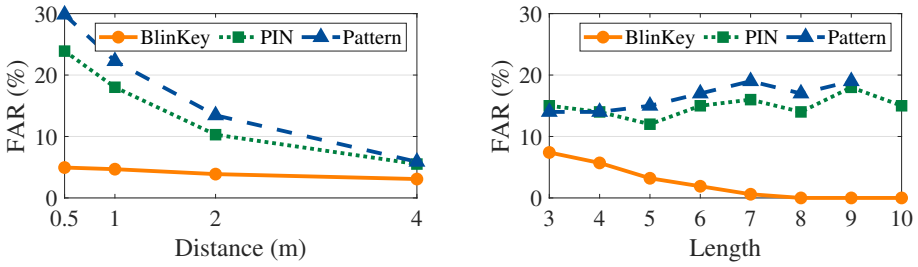
Length	3	4	5	6	7	8	9	10
BlinKey	16.6%	25.4%	19.7%	15.5%	14.2%	10.6%	7.9%	4.4%
PIN	100%	100%	97.1%	100%	100%	100%	100%	100%
Pattern lock	100%	100%	100%	100%	100%	99.3%	97.1%	96.8

2.6.2.4 Shoulder-Surfing Attacks

Shoulder-surfing attacks are another general type of attacks against an authentication system, in which the adversary obtains authentication information via visual observation. It is more severe towards PIN/password/pattern authentication on VR devices than regular personal devices. Because the victim's vision is blocked by the headset, they are unaware of the surrounding environment, including the presence of shoulder-surfing attackers. We randomly pick 22 out of 43 participants involved in the phase-II user study and group them into 11 pairs. Each of them was told to replay his/her partner's passcode. Firstly, one user of the pair acts as an attacker, the other as a legitimate user, and then the roles are exchanged. During the experiment, the legitimate user repeats the same passcode for three times with a pause in between. Then, the attacker watches the entire process and tries to reproduce it. Every attacker makes three access attempts. The attacker is considered a success in a shoulder surfing if any one of the five trials passes the authentication.

Figure 2.15(a) plots the FAR, i.e., attacker's success rate, of *BlinKey*, PIN, and pattern lock with respect to its distance to the legitimate user. When the distance is 0.5 m, the success rate toward PIN and pattern lock is 23.9% and 29.8%, respectively, while that toward *BlinKey* is merely 4.9%. This is intuitive, as a shorter distance enables the attacker to have a closer observation over the legitimate user's login. Thus, it has a better chance to correctly replay the knowledge-based secret. On the other hand, it is hard, if not impossible, for the attacker to observe the user's eyes in a VR headset. Besides, as *BlinKey* involves

biometric features, it is extremely challenging for an attacker to repeat such information. It also explains why FAR keeps almost unchanged as the distance gets longer. Figure 2.15(b) shows the success rate of should-surfing attacks with respect to the length of blinkey, PIN, and drawing pattern. Again, *BlinKey* has the best performance among the three. When the length is 8, FAR of *BlinKey* is 0, i.e., no adversary successfully launches shoulder-surfing attacks, while the value for PIN and pattern lock is 14.1% and 17.0%, respectively. Interestingly, unlike *BlinKey*, PIN and pattern lock become more vulnerable to shoulder-surfing attacks with a larger length. One possible explanation is that a longer key provides the attacker more information about the relative button positions to better infer the keypad structure.



(a) Impact of the victim-attacker distance (b) Impact of key length

Figure 2.15: Success rate of shoulder-surfing attackers against *BlinKey*, PIN, and pattern lock.

The phenomenon that the success rate of shoulder-surfing attacks is non-zero is attributed to two reasons. First, while it is hard to launch the shoulder-surfing attack, the attacker can still guess the secret, i.e., zero-effort attack, even without much insight. As shown in Table 2.5 in the paper, its success rate is 8.1% when a blinkey has a length of only 3. Second, while k-NN exhibits promising authentication accuracy, it is imperfect. As shown in Figure 2.10, the lowest EER (where FAR=FRR) exists at 4.0%. It indicates that there is still certain possibility that an illegitimate blinkey is wrongly classified as a

legitimate one. On the other hand, a close observation over an user's login process does provide the attacker some marginal advantage. For example, a couple of volunteers tend to nod their heads subconsciously following the same rhythm as they blink. This advantage diminishes quickly as the attack-victim distance increases.

2.6.3 Usability

Apart from security, usability is another critical criterion to evaluate a user authentication scheme. We measure the usability of *BlinKey* from aspects of time consumption, legitimate recognition, memorability, and impact of user motions.

2.6.3.1 Time Consumption

We examine the enrollment time and login time needed for *BlinKey*. Specifically, the former refers to the total duration required to enroll all samples to train the classifier, while the latter is the total duration for a user to enter a test blinkey and for the system to make an authentication decision. The distributions of enrollment time and login time are depicted in Figure 2.16(a) and Figure 2.16(b), separately. We observe that the enrollment time of *BlinKey* ranges from 40.8 to 63.5 seconds. Its average, median, and 90-th percentile are 49.5 seconds, 42.9 seconds, and 61.1 seconds, respectively. The login time spans from 7.3 to 11.7 seconds, with its average, median, and 90-th percentile as 9.6 seconds, 8.9 seconds, and 11.2 seconds, respectively. Therefore, the most time-consuming part is the enrollment phase. Luckily, the enrollment only needs to be performed once for a user. Hence, its time consumption is still reasonably practical. The authentication time of our scheme is shorter than many existing solutions, such as [45, 221]. It takes 17 and 60 seconds to authenticate a user in [45] and [221], respectively. Besides, as shown in Table 2.4, only 4 out of 43 volunteers had the experience of performing authentication in a VR device before. This factor partially accounts for the time overhead in our result. We thus optimistically project

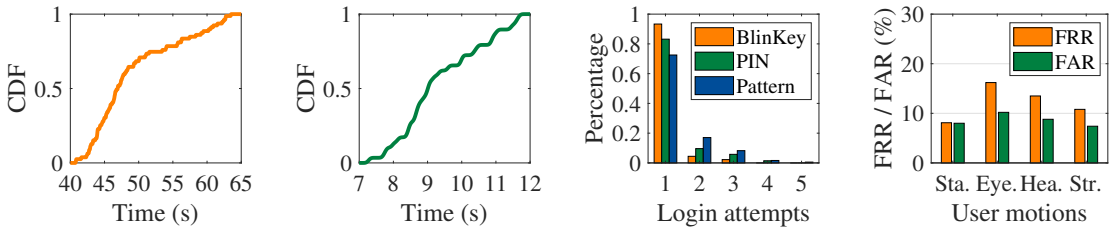
that as users get more familiar with *BlinKey*, the enrollment and login time should be further reduced.

The blinks indicating the start and end of a blinkey have been taken into account for the measurement of both the enrollment time and login time in the evaluation. Specifically, 5 seconds out of the login duration (with the 90-th percentile as 11.2 seconds) are attributed to this overhead. As our future work, we plan to propose efficient approach to indicate the start/end of a blinkey with reduced overhead.

2.6.3.2 Login Attempts

This metric measures how many login attempts a legitimate user needs to unlock the device. A fewer number of attempts are desirable for an authentication scheme with high usability.

93.3% of blinkeys can be successfully authenticated in the first attempt, while this value for PIN and pattern lock is 83.2% and 72.5%, respectively. This is because users make mistakes more often in selecting the correct key or drawing the correct line on a virtual keyboard with the controllers. In contrast, the entering of blinkeys is performed by blinking eyes without interacting with the controller. It only takes 1.09 attempts on average for a legitimate user to get authenticated in *BlinKey*.



(a) Distribution of enrollment time (b) Distribution of login time (c) Login attempts (d) Impact of user motions

Figure 2.16: Evaluation of usability of *BlinKey*.

Table 2.9: The recall rate after a period of time.

Duration between stage-I and -II	No. participants	No. successes	Success rate
7 days	29	26	89.6%
14 days	15	12	80.0%

2.6.3.3 Memorability

Memorability demonstrates how well a secret key can be remembered by its owner, especially after a long period. To evaluate the memorability of *BlinKey*, we designed two follow-up experiments. The participants are invited to perform their *blinkey* after 7 days, and 14 days and test if they can successfully get authenticated. Out of the 45 volunteers who joined in the first-stage experiment, 29 and 15 of them participated in the two second-stage experiments, respectively. As shown in Table 2.9, 26 out of 29 volunteers are able to recall their blinkeys successfully after 7 days and 12 out of 15 volunteers are able to recall their blinkeys after 14 days. While the memorability performance of *BlinKey* is far from perfect, we would like to note that most of the volunteers may not have the chance to practice their blinkeys during 7 days, unlike regular passwords or digit-PINs that are entered to personal devices multiple times a day. We believe the performance will be enhanced with more frequent practices.

2.6.3.4 Impact of User Motions

In practical scenarios, users are not always sitting statically while entering a blinkey. Rather, they may be rotating their eyes, moving their heads, or even walking. An ideal system should be capable of handling these situations. In the experiment, we investigate whether user motions impact the performance of *BlinKey*. Four different types of motions are considered, sitting, rotating eyes, moving head, and strolling. We observe in Figure 2.16(d) that the best accuracy is achieved when the user is sitting, with its FRR at 8.1%

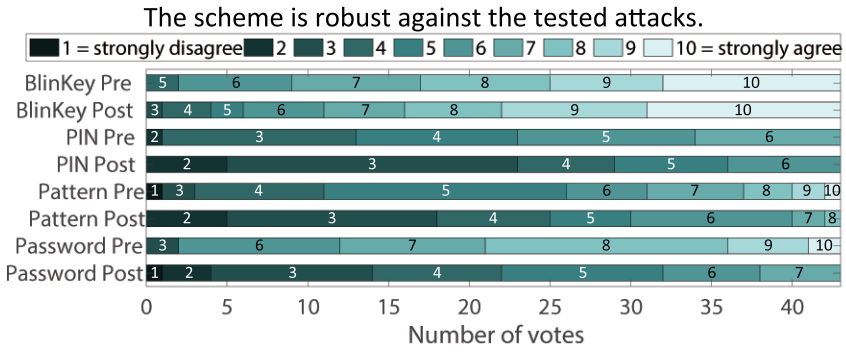
and the FAR at 8.0%. The lowest accuracy is observed when the user is rotating eyes, with the corresponding FRR at 16.9% and FAR at 10.2%. This is because eye movement prevents the eye tracker to accurately estimate real-time pupil size. Nonetheless, neither head movement nor strolling causes significant performance degradation. Besides, we also observe that FAR is relatively stable across all motion status. It means the authentication security is not deteriorated much by motions. Based on the above observation, users will be recommended to enter blinkeys by looking into the virtual screen to prevent significant eyeball movement. There will be no restriction on their body movement, though.

2.6.4 Survey Results

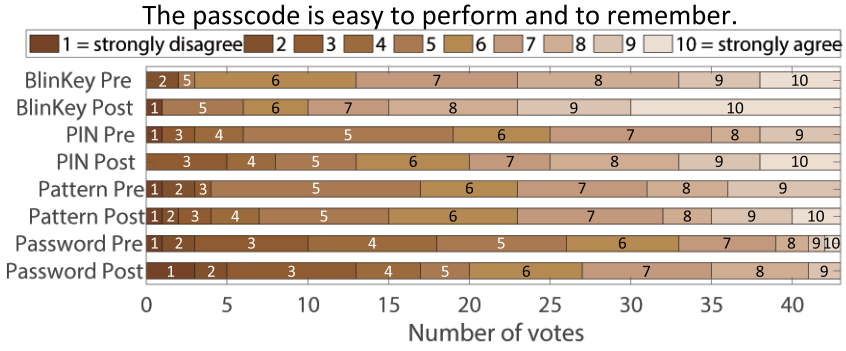
In addition to the experiments, we further evaluate *BlinKey* via survey. The pre-survey was conducted after the introduction of the basic idea of *BlinKey* and before the experiment, while the post-survey is conducted after all experiments. Volunteers are asked to rate *BlinKey* from the perspectives of security and usability and compare them with commonly used methods on mobile devices, including PIN, password, and pattern lock. Questions include 1) Is it safe against attacks being tested? 2) Is it easy to perform and remember? On a 10-point Likert scale (1 = strongly disagree; 10 = strongly agree), participants pick a point that they deem proper. Survey results are shown in Figure 2.17. Most volunteers agree that *BlinKey* is better than the other three listed authentication methods in both aspects. It is worth mentioning that many participants rate *BlinKey* a higher score in the post-study than in the pre-study, which suggests that our scheme outperforms user's expectations.

2.7 Discussions

Raw size of *BlinKey* space. *BlinKey* is a two-factor authentication, a combination of the rhythm passcode and human biometrics, i.e., variations of pupil size. Since the



(a) Security



(b) Usability

Figure 2.17: Pre-/Post-study survey results regarding security and usability.

variability brought by biometric features is hard to quantify, we would like to discuss the key space of *BlinKey* merely taking into account the variability introduced by blinking rhythms. Thus, the real key space of *BlinKey* should be no less than this value.

BlinKey adopts a similar design of the rhythm passcode as a prior work [118]. We thus revise the theoretical result of [118] and derive the key space of *BlinKey*.

Theorem 1. *(Revised from Theorem 5.1 of [118].) The size of* BlinKey*'s key space is*

$$|\Pi| = \sum_{l=1}^{L_{\max}} \binom{\frac{T_{\max}}{\sigma} - (\frac{\tau_b}{\sigma} - 1) \times l - (\frac{\tau_s}{\sigma} - 1) \times (l-1)}{2l-1},$$

where $L_{\max}$, $T_{\max}$, σ, τ_b and τ_s stand for the maximum blinkey length, corresponding maximum time duration, the system clock unit, minimum value of an onset-offset duration and minimum value of a offset-onset duration, respectively.

For an illustration purpose, we let $\sigma = 5$ ms, which is the time unit for Pupil Labs eye tracker's system clock. According to the statistic analysis over our collected dataset, we set the rest parameters as $T_{\max} = 12$ s, $\tau_s = 0.15$ s and $\tau_b = 0.10$ s. Thus, when the blinkey length is 6, the corresponding space size is about 10^{23}. As a reference, the key space for a regular PIN with 6 digits is 10^6. The above theorem is derived without considering pupil size variation. With the introduction of an additional dimension of entropy, the key space of *BlinKey* should be further enlarged.

Practical design. Our design grants the user some error tolerance–when a legitimate user fails to authenticate, she can re-enter her blinkey until the maximum number of attempts is reached. In this case, the user is temporarily locked out, and the recovery process is invoked (see Figure 2.12(f)). Here are two classic recovery methods widely adopted by other user authentication schemes. 1) Provide an alternative way to authenticate users; when a legitimate user fails to authenticate herself with her blinkey, she can still unlock the device by entering a valid passcode or digit-PIN. 2) Have a remote server to send a recovery code to the user's previously authorized email address; the user retrieves the code by accessing the email and unlocks the device by entering the code. These two approaches are deemed robust against attacks.

When an adversary tends to enroll himself in the device, he needs to first enter a valid blinkey, which has been created by the legitimate user earlier, to unlock the device. Otherwise, there is no way for the adversary to enroll himself. This idea has been adopted in many personal devices, such as smartphones and PCs. There is also an exception that the victim VR device has not been secured with any user authentication scheme. In this case, the adversary can directly set up his account associated with his blinkey in the device. To

address this issue, a conventional solution is to enforce the user to enroll her authentication credentials, i.e., blinkey here, during initial account setup.

Impact of environment. User's pupil size is affected by their biophysical status, such as mood, energy level, whether drinking alcohol, illness, etc. Consequently, these factors would impair the authentication accuracy of *BlinKey*. One viable solution is to further deploy a second-option user authentication method, such as digit-PIN or password. Once a legitimate user's input cannot be recognized by the system by any chance, including the above-mentioned situations, she can always unlock the device by a valid digit-PIN. Such an idea has been adopted by current fingerprint-/facial recognition-based user authentication on smartphones. While the brightness of the display does affect pupil size when blinking, it does not necessarily impact the performance of our scheme. As shown in Figure 2.12(c), the screen displays the same image with the same brightness/color/content during the login process. Thus, it eliminates the impact from the display.

Reduce login overhead. Under the current design, the login duration of *BlinKey* spans from 7.3 to 11.7 seconds, with its average, median, and 90-th percentile as 9.6 seconds, 8.9 seconds, and 11.2 seconds, respectively. While this overhead is reasonably practical, it is still longer than conventional PIN and password. The most significant portion of the overhead is attributed to the blinks indicating the start and end of a blinkey, i.e., 5 seconds according to the setting. As our future work, we plan to propose efficient approach to indicate the start/end of a blinkey with reduced overhead. Besides, as shown in Table 2.4, only 4 out of 43 volunteers had the experience of performing authentication in a VR device before. This factor partially accounts for the long time overhead in our result. The login time would be further reduced as users get more familiar with authenticating themselves via *BlinKey* in VR.

2.8 Conclusions

As VR devices are increasingly weaved into our everyday life, providing security to the data acquired by or accessed through these devices becomes critically important. In this study, we develop a two-factor user authentication mechanism, named *BlinKey*, which employs the user-designed blinking rhythm and unique biometrics exhibited in pupil size variation to fingerprint legitimate users. Compared to prior work, our solution delivers secure authentication, incurs low cognitive overhead, and offers great convenience. Through an extensive evaluation that involves 52 volunteers, we observe that the average EER is as low as 4.0% with only 6 training samples. The proposed *BlinKey* is also implemented on an HTC Vive Pro with a Pupil Labs eye tracker. We further measure its security by testing robustness against various types of attackers, and its utility, from aspects of time consumption, login attempts, the impact of user motions, and memorability. We observe that *BlinKey* requires relatively long enrollment time (median: 42.9 seconds). One reason is that many participants have limited experience in authenticating themselves on VR devices. This is likely to be alleviated as users practice it multiple times daily after scheme implementation. Besides, as enrollment is only executed once for each blinkey, the long enrollment time will not incur noticeable overhead from a long-term view. In conclusion, we believe *BlinKey* is a practical authentication method applicable to current VR devices.

CHAPTER 3

SOUNDLOCK: A NOVEL USER AUTHENTICATION SCHEME FOR VR DEVICES USING AUDITORY-PUPILLARY RESPONSE

3.1 Introduction

Motivation. The rapid development of virtual reality (VR) has been seen in the past few years with a consistently growing popularity. According to a recent report [217], the VR market is around $28 billion in 2022; by 2030, the number is forecast to reach over $87 billion with a constant annual growth rate of 15%. With the capability of providing an immersive and interactive experience, VR has revolutionized gaming and entertainment and permeated a variety of applications, including e-commerce, education, healthcare, and military [261]. For example, retailers can bridge physical and online stores via VR to provide an immersive shopping experience for customers [168]; medical practitioners may communicate with patients in a VR environment for remote diagnosis [177]; military actions can be simulated and practiced in a virtual battlefield [159]. In the above applications, tremendous amounts of sensitive data are collected, processed, and stored on VR devices, such as customers' credit card information, patients' health status, and military secrets. Adversarial access to VR devices would cause data breaches and other critical consequences. Therefore, implementing user authentication mechanisms in VR is a crucial step in resisting unauthorized access.

However, user authentication on VR devices is still at the infant stage. Current solutions, including passwords, digital PINs, and pattern locks, mostly follow conventional approaches for general personal devices. Users have to use some external hand controllers to enter the credentials. They have been criticized for the usability deficit: It takes users

substantial effort to select correct keys from the virtual keyboard using the controller [248]. What's worse, they are shown to be vulnerable to shoulder-surfing attacks. As the user enters her credential, the hand movement leaves a trajectory that can be easily mapped to the entered secrets with the keyboard layout [99, 248, 314]. Per the statistics from prior work [99], the success rate of shoulder-surfing attacks towards PINs and drawing patterns in VR is as high as 18%.

To address the above issues, great efforts have been devoted to exploring practical alternatives. Existing approaches can be generally categorized into the following classes: *knowledge-based methods* [94, 98, 99, 170, 305], *physiological biometrics* [19, 55, 155, 234], *behavioral biometrics* [162, 185, 204, 237, 311], *token-based methods* [48], and a mixture of above [169, 297, 314]. Among them, physiological biometrics attract the most attention due to its high usability and authentication accuracy. Nonetheless, its wide deployment is still faced with several challenges. First, to access the user's biometrics, such as electroencephalogram (EEG), electrocardiogram (ECG), electromyography (EMG), and iris patterns, dedicated and costly sensors are needed. These sensors are mostly unavailable in current VR headsets. While iris scans have been deployed on HoloLens 2, a high-end augmented reality (AR) device costing at least $3,500, they are less likely to integrate into an even broader set of medium-/low-end terminals with much lower budgets. Second, most physiological biometrics are irrevocable. Once a biometric credential is compromised or stolen, it cannot be reset. This property is also called cancelability.

Our approach. In this paper, we propose to leverage a new kind of biometric, *auditory-pupillary response*, for user authentication on VR headsets. By presenting users with auditory stimuli, the pupil's reaction, in the form of size changes, is universally observable among human beings [27, 113, 171, 186, 269]. The auditory-pupillary response is an autonomic reflex that dilates or constricts the pupil, mediated by the sympathetic and parasympathetic nervous systems, which are both parts of the autonomous nervous system.

The biological uniqueness in the complex neural pathways and structure of iris muscles present particular features that make it possible to explore auditory-pupillary responses for user identification. As validated in our preliminary study (see Section 4.4), inter-subject pupillary responses exhibit distinguishable patterns under the same stimulus, whereas intra-subject pupillary responses are consistent in multiple trials. These observations motivate us to develop SoundLock, a novel reflex physiological biometric authentication method for VR devices based on the auditory-pupillary response. During authentication, carefully designed auditory stimuli are rendered to the user via the VR device's audio channel. The corresponding pupillary response is captured by the eye tracker integrated into the device. The user's legitimacy is then determined by comparing the response with the template generated during the enrollment stage.

Compared with conventional authentication methods for VR, such as passwords, digital PINs, and drawing patterns, our scheme has the following prominent advantages. First, its usability has been greatly enhanced as it significantly reduces user effort for credential entry. A user's biometric, i.e., the auditory-pupillary response, is automatically gathered by the device. The entire process is hand-free and relieves users from memory burdens. Second, since the user's eyes are completely blocked by the VR headset, it is impossible for an adversary to gain visual observation of the authentication process to launch shoulder-surfing. Meanwhile, SoundLock, as a new kind of reflex physiological biometric for VR, outperforms existing static biometric [45, 145, 162, 185, 220] in the following aspects: First, auditory-pupillary responses are revocable. In the case of having one pupillary response stolen or counterfeited, a new credential can be easily generated by changing its associated stimulus. Second, SoundLock can be implemented on many mainstream VR headsets, e.g., HTC VIVE Pro Eye, Pico Neo series, Varjo VR-3, and Fove VR [2, 93, 270, 274], which are already equipped with eye trackers. It is well accepted that incorporating eye-

tracking technology is a trend in VR to assist in simulating depth of field and focus and providing users a more realistic and natural visual experience [60, 130, 246].

Despite these attractive properties, the design of SoundLock is faced with several non-trivial challenges. First, while pupillary response exhibits prominent inter-subject distinguishability, identifying essential features out of raw pupil size measurement for accurate user authentication is not an easy task. No prior research has been conducted on this topic. We thoroughly investigate 60 features, including morphological features that are pupillary response-specific and general statistical features, and narrow them down to 20 that best represent the uniqueness of each individual. We validate through a comprehensive evaluation that the selected features effectively produce high authentication accuracy. Second, to enlarge the credential pool, we adopt multiple auditory stimuli. However, the multi-stimuli prolong the authentication time and thus impair usability. To mitigate this issue, we model the problem into an optimization problem that maximizes the authentication accuracy while satisfying a hard constraint on the authentication time (see Section 3.4). It aims to balance security and usability. Realizing that it is challenging to directly solve the problem optimally owing to its non-linearity, we devise a two-stage heuristic algorithm to find the approximate solution efficiently. Lastly, like other biometrics, the auditory-pupillary response may exhibit variations over time. As a result, its authentication performance may degrade over a long time span. To deal with this issue, we adopt an adaptive biometric strategy to consistently update the classification model with the coming of new samples.

To evaluate the performance of SoundLock, we implement it on a VR device and carry out extensive experiments involving 44 participants. It achieves an F1-score of 0.984, FAR of 0.76%, and FRR of 0.91%, outperforming state-of-the-art solutions. Besides, our scheme can be performed within a practical authentication time of 7 s. SoundLock also demonstrates satisfactory consistency under various testing conditions. Finally, the user study manifests that our scheme is well received among the participants; especially, 72.7%

of them are willing to adopt SoundLock as the authentication scheme on their (future) VR devices.

To summarize, the contributions of this paper include:

- We investigate a new kind of reflex physiological biometric, auditory-pupillary response, for user authentication on VR devices. We validate its feasibility through a measurement study.
- To model the response for user authentication, we investigate a set of morphological and statistical features, which are proven effective in producing high authentication accuracy.
- To strike a balance between security and usability in the design, we formulate an optimization problem. A two-stage heuristic algorithm is proposed to efficiently solve the problem with an approximate solution.
- We perform extensive in-field experiments to evaluate SoundLock. Results demonstrate that the proposed scheme outperforms state-of-the-art biometric authentication solutions and is well received among participants in the user study.

3.2 Preliminaries

3.2.1 Background on Auditory-Pupillary Response

The pupil size has been proven sensitive to a wide variety of auditory stimuli [27, 113, 171, 186, 269]. Figure 3.1 exhibits pupil size, measured in pixels, changes as a subject is presented with an auditory stimulus, a white noise that starts at 1 s and stops at 5 s. This sample is randomly selected from our collected dataset. Measures from only one eye are collected since pupillary responses in both eyes have been confirmed to be consensual [140]. The presentation of an auditory stimulus results in a multi-phasic pupillary response. The initial phasic response is evoked with transient pupil dilation shortly after the stimulus

onset, followed by a constriction. This process is followed by the second round of, and sometimes more rounds of, dilations and constrictions with attenuated amplitudes. After the stimulus offset, the pupil gradually returns to its baseline, i.e., the pupil size under the no-stimulus condition, accompanied by minor fluctuations [210].

Physiologically, the pupillary response is controlled by two muscles: the *iris radial muscle* (IRM) increasing the pupil size and the *iris sphincter muscle* (ISM) reducing the pupil size [38]. The balance between the sympathetic and parasympathetic nervous systems determines pupil size. The underlying mechanisms are complex; the relative contribution of the two systems depends on a variety of factors, such as stimulus characteristics and cognitive activities. Pupil dilation is controlled by the IRM. IRM consists of fibers that are oriented radially and connect the exterior of the iris with the interior. When IRM contracts, it pulls the interior of the iris outward, thus increasing the size of the pupil. Upon perception of auditory stimuli, psycho-sensory arousals are first triggered at the hypothalamus and the locus coeruleus. The activities on the hypothalamus and the locus coeruleus reflect arousals and project to the intermediolateral column of the spinal cord. The arousals finally reach IRM via a complicated network of nerves and cause contraction. In contrast, pupil constriction is controlled by ISM, which encircles the pupil like a cord that reduces pupil size when it contracts. As shown in Figure 3.1, the pupil constricts once it dilates to a large extent. This process operates through the opposite action of pupil dilation. ISM is directed through the parasympathetic pathway. The activated Edinger-Westphal nucleus transmits information via the oculomotor nerve to the ciliary ganglion, which is located behind the eyeball. The information is further sent via the short ciliary nerve to innervate the ISM to contract. In short, the pupil dynamics observed under auditory stimuli are a joint effect delivered by IRM, ISM, and their corresponding neural pathways [163, 171, 186, 255, 277, 278].

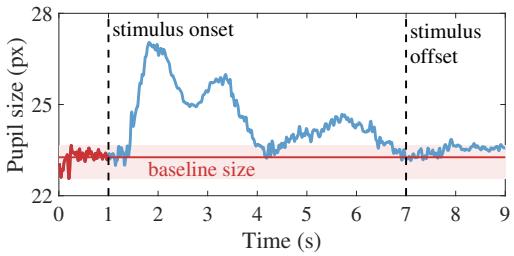


Figure 3.1: Pupillary response to auditory white noise.

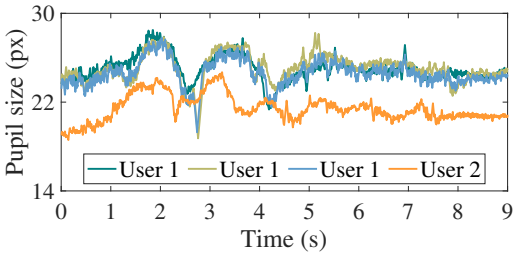


Figure 3.2: Intra-/Inter-subject pupillary response.

3.2.2 Measurement Study

While the phenomenon of auditory-pupillary response is well recognized, whether it can be exploited for user authentication remains unclear. Our measurement study intends to answer this question by carrying out extensive experiments. A total of 32 subjects are invited. They listen to auditory stimuli of different types via the HTC VIVE Pro VR headset. A total of 20 stimuli are adopted, including white noise, monotones, prompt sounds, natural sounds, and human voices. They have been widely adopted in prior works on auditory-pupillary response [27, 113, 154, 171, 186]. Each auditory stimulus is a 6-second audio track. Subjects' pupillary responses are captured by a Pupil Labs eye tracker that is integrated into the headset. To facilitate the data collection, a specialized app is built using Unity, a cross-platform engine for VR development. To avoid impact from visual stimuli, participants are exposed to a dark VR environment, i.e., no image is displayed. The above process is repeated 20 times for each participant. The following analysis is conducted based on the collected 12,800 samples, i.e., time-resolved pupil size sequences.

Intra- and inter-subject pupillary response. Figure 3.2 shows pupillary responses from four trials under the same stimulus. Three of them are collected from the same subject. The three intra-subject responses exhibit similar patterns, although they are from different trials. It indicates that pupillary response is relatively consistent for the same user. Meanwhile, inter-subject responses exhibit distinguishable patterns. To better quan-

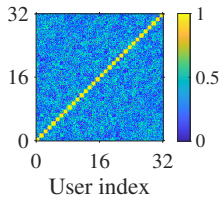


Figure 3.3: Confusion matrix of PCC among 32 subjects.

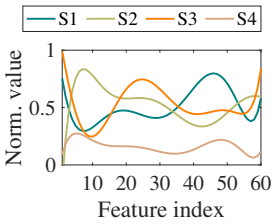


Figure 3.4: Normalized values of 60 extracted features.

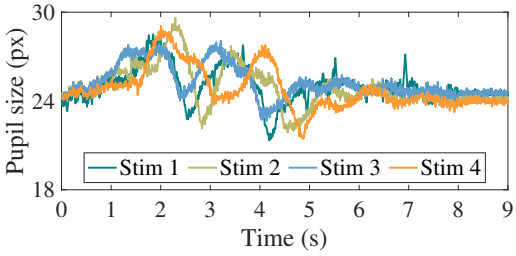


Figure 3.5: Pupillary response under different stimuli.

tify the intra-/inter-subject response relationship, Figure 3.3 further plots the confusion matrix (160×160) of pupillary responses among the 32 participants in response to one stimulus. 5 samples are randomly selected from each participant. The Pearson correlation coefficient (PCC) is adopted. The PCC values on the diagonal line ($\mu = 0.91, \sigma = 0.02$) are significantly higher than those off the line ($\mu = 0.36, \sigma = 0.14$). It implies that individuals exhibit diverse pupillary responses when presented with the same auditory stimulus, while those from the same subject are consistent.

Pupillary response under various stimuli. We then play a variety of auditory stimuli to the subject. It is observed in Figure 3.5 that the corresponding pupillary responses vary across the stimuli. We further extract 60 features out of the raw measures. Figure 3.4 depicts their normalized values. Polynomial regression is applied for better illustration. The feature vectors are distinguishable with respect to various stimuli. Intuitively, it is possible to generate a large number of credentials for a user from her pupillary responses by applying various auditory stimuli. More importantly, these credentials can be easily revoked: In the case of having one pupillary response stolen, a new credential can be generated by changing its associated stimulus, which is called *cancelability* [212]. In contrast, this property does not exist in conventional biometrics, such as fingerprints, irises, and faces, which are static to human beings. Once their credentials are damaged or counterfeited, the user cannot cancel the pre-stored credentials or reset them.

Summary. Our findings are encouraging. First, given the same auditory stimulus, intra-subject pupillary responses exhibit consistent patterns in multiple trials, while inter-subject pupillary responses are distinguishable. This property lays the foundation for our idea that utilizes auditory-pupillary response as a new kind of biometric for user authentication. Second, the responses are diverse with respect to various stimuli. It thus motivates us to employ a sequence of stimuli to enlarge the pupillary response-based credential pool. More importantly, the property that the induced credential is stimuli-dependent offers the potential to design a cancelable biometric. An in-use pupil credential can be revoked and updated by simply applying new auditory stimuli. Lastly, we observe in the measurement that the pupil demonstrates a stable behavior in response to auditory stimuli: It first dilates with the stimulus onset and then constricts, followed by a couple of more rounds of dilation-constriction until the stimulus offset. The transitional changes in the pupil size generate consecutive waveforms bearing rich information for authentication. We will investigate in Section 3.3.2 how to extract essential features.

3.2.3 Problem Statement

System model. We consider a general user authentication scenario on VR devices, where a user has to provide a correct credential to log in. We assume that the headset is equipped with an eye tracker for pupil detection and pupil size measurement. The proposed authentication scheme is composed of two stages. In the enrollment stage, the headset plays carefully designed audio stimuli and records users' corresponding pupillary responses. A set of relevant features are extracted upon which a classification model is trained and optimized. In the login stage, a user is presented with the same stimuli. The collected pupillary response is compared with the enrolled ones to determine the user's legitimacy.

Many mainstream VR headsets are equipped with eye trackers nowadays, such as Meta Quest Pro, HTC VIVE Pro Eye, PlayStation VR2, Pico Neo series, Varjo VR-3,

and Fove VR [2, 93, 274], Varjo. The list continues to grow. It is well recognized that eye tracking benefits VR in the following aspects: a) delivering a higher-quality graphics experience through foveated rendering, b) improving wearing comfort by automatically adapting the device to the user via calculating the user's inter-pupillary distance, and c) enhancing the interactions among virtual avatars to better reflect the user's visual attention. It is well accepted that incorporating the eye-tracking technology is a trend in the future development of VR [130, 246].

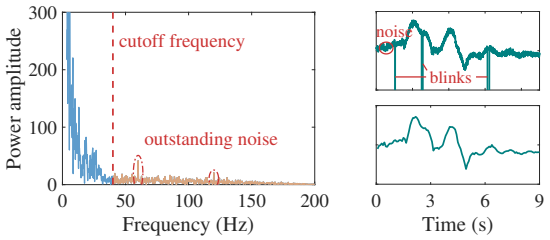
Adversary model. The adversary's goal is to impersonate the legitimate user and log into the VR headset. The adversary is assumed to have physical control of the headset and sufficient time to perform the attack. For example, the VR device is lost or stolen. We primarily consider the impersonation attack [166] throughout this work. The adversary intends to use its own biometric credential, i.e., pupillary response, under the auditory stimuli to get authenticated. Other common attacks will be discussed in Section 3.5.1.

3.3 Basic Scheme Design

We start by introducing a basic scheme that renders a single auditory stimulus. It consists of three main components: preprocessing, feature extraction and selection, and classification. Upon the acquisition of a pupillary response, it is first preprocessed for signal cleaning. Then a set of response-specific features are extracted as well as selected. In the enrollment stage, these features are used to train the classifier; in the login stage, they are fed into the trained classifier for authentication.

3.3.1 Preprocessing

The pupillary response is acquired by an embedded eye tracker sampling at 200 Hz. Figure 3.6(b) (top) plots the raw measurements, which are mixed with noise and zero-readings. This component aims to eliminate them and extract useful information from



(a) Frequency domain (b) Time domain

Figure 3.6: Preprocessing pupillary response.

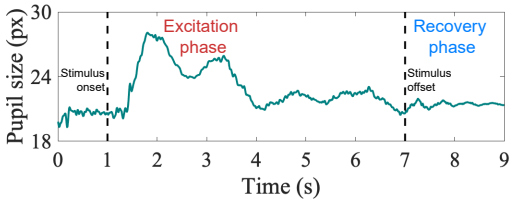
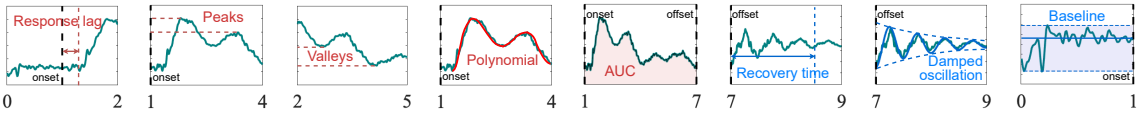


Figure 3.7: Illustration of excitation and recovery phase in a pupillary response.



(a) Lag of response (b) Peak magnitudes (c) Valley magnitudes (d) Polynomial coeff. (e) AUC (f) Recovery time (g) Damped oscillation (h) Baseline size

Figure 3.8: Representative morphological features extracted from pupillary response.

the raw measurement. The background noise is mainly caused by internal and external electromagnetic radiations (e.g., VR display refreshing, power line emanation, and their harmonics) that primarily exist above 50 Hz. In opposition, the frequency components of pupil size variations mainly reside at the lower end of the frequency band, as shown in Figure 3.6(a). Hence, we apply a low-pass filter with a cutting frequency of 40 Hz to eliminate the above-mentioned noise. The intermittent zero-readings exist in the measurement due to spontaneous blinks. We apply the classic interpolation method to smooth the pupillary response signals. Figure 3.6(b) (bottom) plots the pupillary response after preprocessing.

3.3.2 Feature Extraction and Selection

We extract two types of features from the processed pupillary response: morphological features and statistical features. The former is features specifically proposed to outline the morphology of the auditory-pupillary response patterns; they reveal the intrinsic geo-

metrical characteristics in the multi-phasic signals. The latter is provides a more general description of the signal statistics. As demonstrated in Figure 3.7, a pupillary response can be divided into two phases: *excitation phase* and *recovery phase*. In the following, we provide details of extracting the candidate morphological features from both phases.

Excitation phase. It is between the stimulus onset and the stimulus offset. In this phase, the pupil is provoked by the stimulus and experiences transitional dilations and constrictions.

- *Response lag.* It is defined as the latency between the stimulus onset and the moment the pupil reacts to it, as shown in Figure 3.8(a). Prior studies show that this value is mostly determined by the neural pathways while less affected by mechanical properties of the iris muscles [171, 278]. Differences in response latency among individuals have been reported [25, 30, 90, 259, 268, 302]. In general, senior people tend to exhibit longer response lag [259, 268].
- *Peak/Valley magnitudes.* Upon the stimulus onset, the pupil size increases as the pupil dilates and reaches a large extent. Thereafter, the pupil size decreases as it constricts. Multi-round dilations and constrictions generate a series of waveforms. The corresponding peak (valley) magnitudes then serve as the features as shown in Figure 3.8(b) (Figure 3.8(c)). A classic peak detection technique [42] is applied to identify peaks and valleys in the response waveforms.
- *Dilation/Constriction rates.* Apart from the peak and valley magnitudes in the response waveforms, we are also interested in the dilation/constriction rates. They are manipulated by a complex mechanism involving the iris muscles and many components along the neural pathways such as the nerve fibers in the intermediolateral column, the super cervical ganglion, and the ciliary nerves [163, 171, 255]; these rates reflect the biological heterogeneity in the human nervous systems and iris mus-

cles. The dilation rate is calculated as the pupil size change in one dilation divided by the associated time duration. The definition of the constriction rate follows similarly.

- *Polynomial coefficients.* n-degree polynomials are applied to approximate the response waveforms during the excitation phase. We mainly focus on the first two waveforms as the rest tend to attenuate mixed with more noise. n is set to 4 empirically. Figure 3.8(d) depicts derived approximate polynomials; they align well with the ground truth. The corresponding coefficients in the polynomials are treated as a subset of features.
- *Area under the curve (AUC).* It is the area of the response curve during the excitation phase, as illustrated in Figure 3.8(e). In general, the AUC tends to be larger when a user is more agile with the auditory stimulus. AUC is derived by taking the integral of the pupillary response over time.

Recovery phase. It starts from stimulus offset and lasts until the response cutoff.

- *Recovery time.* It is the time the pupil takes to return to its baseline. As depicted in Figure 3.8(f), it denotes the time interval between the stimulus offset and when the pupil stabilizes with negligible deviations from its baseline.
- *Damped oscillation.* With the stimulus offset, the pupil size gradually returns to its baseline, accompanied by oscillatory behavior, as illustrated in Figure 3.8(g). We propose to approximately characterize this pattern using a classic damped sine wave model: $y(t) = Ae^{-\lambda t}\cos(\omega t - \phi) + C$ [108]. The function parameters, A, λ, ω, and ϕ, are taken as a subset of features.
- *Pupillary unrest index (PUI).* Human eyes exhibit continuous pupil size fluctuations, known as pupillary unrest [126, 187, 235]. Although its origins are complex, this phenomenon is mediated by fluctuating inhibitory activity within the parasympathetic Edinger Westphal nucleus, possibly driven indirectly by the locus coeruleus [128, 215, 241]. The pupillary unrest index (PUI) has been proposed in prior work to

characterize the pupillary unrest behavior [160]. It is defined as cumulative changes in the average pupil size in consecutive observation windows. We thus adopt PUI as part of the features.

- *Baseline size.* The pupil baseline size, depicted in Figure 3.8(h), has been well recognized as a user-specific biometric trait [46, 189]. It is the eye's natural status when no external stimulus is applied. In this work, several baseline-related parameters are considered, including the average size, maximum, minimum, standard deviation, and interquartile range. The baseline can be estimated once the pupil is recovered from the excitation status or before stimulus onset.

Aside from the above-mentioned morphological features, we also take into account general statistical features of pupil size from both phases, such as average, variance, median, skewness, and kurtosis. Since these statistical features have been widely adopted in signal characterization [18, 189, 314], we do not expand the discussion here. Table 3.1 lists all the 60 candidate features introduced in Section 3.3.2, including their names, categories, phases, and notations. They are sorted by the normalized Fisher score described below.

Feature selection. This step selects from the candidate features the most relevant ones for user authentication. The refined feature set helps to reduce the computation complexity and avoid model overfitting. To this end, we calculate the Fisher score for each feature, which is defined as the ratio between the feature's inter-class and intra-class variances; a higher ratio indicates a more significant role in contributing to classification accuracy. All candidate features are sorted according to their normalized Fisher scores in Figure 3.9. Finally, the top 20 features are selected to feed into the classification model. These selected features include morphological features such as the dilation rates, the peak magnitudes, and the second valley magnitude; the only selected statistical feature is the average pupil size. The reason that morphological features rank relatively higher is probably that

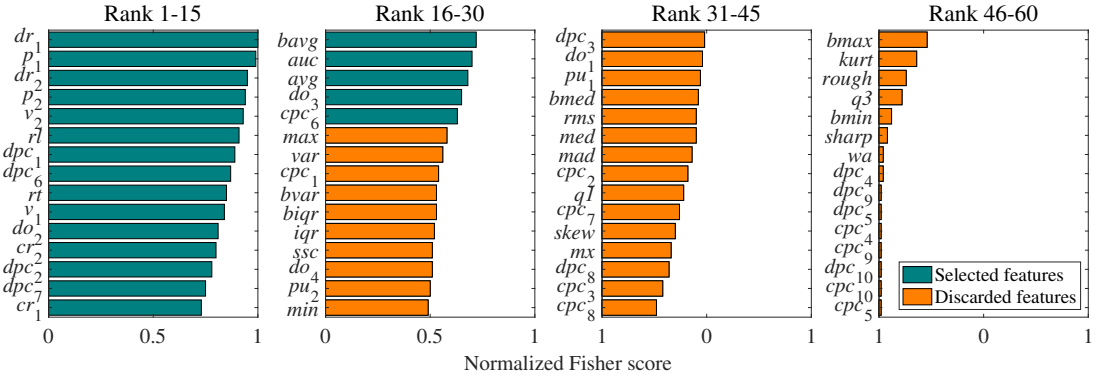


Figure 3.9: All candidate features sorted by their normalized Fisher scores.

they precisely characterize the dynamics in pupillary response, whereas statistical features that are more generic and abstract.

3.3.3 Classification

The remaining task is to apply a classification method over the selected features for user authentication, i.e., to discriminate between the legitimate user and imposters. Two types of classifiers are adopted and evaluated in this work: one-class classifiers and binary classifiers. The former is trained only with samples from the class of interest, i.e., the enrolled legitimate user. The latter is trained on an explicitly labeled dataset of both classes, i.e., the legitimate user's samples and imposters' samples. The following representative machine learning models are employed. *k-nearest neighbor (k-NN):* It measures the similarity between testing samples and training samples and makes the decision by comparing the similarity with a threshold. It has been proven effective especially in cases with small training datasets. *Support vector machines (SVM):* Its main idea is to find a hyperplane in a multi-dimensional space that distinctly separates data points from different classes. Aside from k-NN and SVM, other common classification methods, including *logistic regression (LR)*, *Gaussian Naive Bayes (GNB)*, and *random forest (RF)*, are also evaluated in this work.

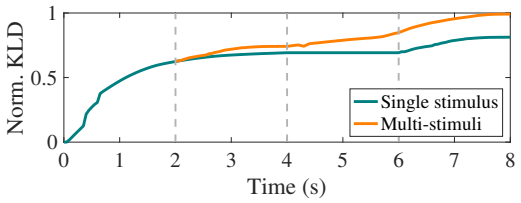


Figure 3.10: Normalized KLD with respect to time: single stimulus vs multi-stimuli.

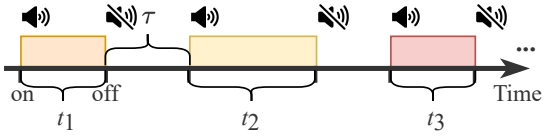


Figure 3.11: The time sequence of multiple stimuli.

3.4 Advanced Scheme with Multi-stimuli

The basic scheme utilizes one auditory stimulus. Inspired by the strong password selection criteria, e.g., more characters and a mixture of numbers, letters, and special characters, we propose to present the user multi-stimuli to enhance the response feature diversity. Specifically, a series of stimuli are played sequentially. Then, all the responses are concatenated and serve as the user's credentials. While the idea is simple, a critical issue is to decide the duration of each stimulus, as a too-long overall duration would impair the usability.

To facilitate the discussion, we adopt the metric Kullback-Leibler divergence (KLD) [144]. It is an indicator of similarity between two probability distributions $P(x)$ and $Q(x)$

$$D_{KL} = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right). \tag{3.1}$$

We let $P(x)$ be the feature distribution of the enrolled user, and $Q(x)$ be that of the reference users, i.e., all the other users from the dataset. X stands for the feature space, and $x \in X$ denotes any possible interval of a feature value. Here KLD represents the distinguishability of the enrolled user from all other users. The larger the value is, the more distinguishable the user is, and the more accurately it can be identified. We further formulate KLD as a function of time. After the stimulus onset, more features are extracted from the measurement as time proceeds. For instance, features in the excitation phase are first derived, followed by features from the recovery phase. Figure 3.10 shows the normalized

KLD with respect to time. We first plot the KLD under a single stimulus (the cyan curve labeled as "Single stimulus" in Figure 3.10). The stimulus starts at $t = 0$. KLD first rises quickly as t is between 0 s and 2 s. Its growth slows down as t passes 2 s. This implies the marginal benefit diminishes for involving more features under the same stimulus. We also show the KLD of employing two stimuli (the orange curve labeled as "Multi-stimuli" in Figure 3.10). The first stimulus starts at $t = 0$ and stops at $t = 2$ s; then, the stimulus is off for 2 s, after which the second stimulus emerges from $t = 4$ s. The KLD experiences another significant increase shortly after the presence of the second stimulus. We make the following observations from Figure 3.10. First, the features do not contribute equally in terms of user classification. The features identified earlier tend to play a more significant role than the ones identified later. Second, the involvement of multiple stimuli introduces more diversity in the pupillary response features and thus benefits classification accuracy.

Problem formulation. In the following, we discuss how to design auditory stimuli. An optimization problem is formulated, where the objective is to maximize the overall KLD in the corresponding pupillary response while keeping the entire authentication time within a practical threshold T_0, which sets a hard constraint on the authentication time. Formally, the optimization problem is expressed as follows

$$\begin{aligned} \max \quad & D_{KL}(P||Q) \\ \text{s.t.} \quad & \sum_{i=1}^{N} (t_i + \tau) \times m_i \leq T_0 \\ & m_i \in \{0, 1\} \end{aligned} \tag{3.2}$$

We aim to select a couple of (e.g., 2-4) auditory stimuli from the pool of size N, i.e., N different audio tracks. The binary variable m_i equals 1 if the i-th stimulus is picked and 0 otherwise. The stimuli selection is necessary as the pupil reacts differently toward various stimuli, as evidenced by our measurement study. Some stimuli are more effective in eliciting distinct patterns in pupillary responses than others. The variable t_i stands for

the duration of the i-th stimulus. τ is a constant representing the interval duration between two adjacent stimuli, which manages the tradeoff between accuracy (the possibility that the pupil has returned to its baseline at the next stimulus onset) and usability (reasonable authentication time). After closely inspecting our collected data, we set it to 2 s empirically with 1.5% outliers. $\sum_{i=1}^{N}(t_i+\tau)\times m_i$ is thus the authentication time. The variables in the above-mentioned optimization problem include t_i's and m_i's, $i \in [1,N]$. Note that the problem formulation is user-specific, because the feature distribution in each individual's pupillary response is diverse. Correspondingly, the solutions of t_i's and m_i's are different across users; that is, each user is associated with a diverse optimum stimuli set and its duration. The problem formulation and calculation are performed during the enrollment stage.

A heuristic algorithm. The objective function and constraint of the above optimization problem are both non-linear. Besides, the two variable sets $\boldsymbol{m}$ and $\boldsymbol{t}$ are linked to each other. Hence, it is impractical to optimally solve it directly. In the following, we propose a heuristic algorithm to find the approximate solution with high computational efficiency. The algorithm is composed of two stages, each fixing the value of $\boldsymbol{m}$ and $\boldsymbol{t}$, respectively. The algorithm takes P_i ($i \in [1,N]$) and Q as inputs, where P_i is the user's feature distribution in the pupillary response under stimulus i and Q is the feature distribution of all reference users. In the first stage, we rank the KLD of each stimulus and select K candidate stimuli that generate the highest KLD. Here K is calculated as $\lceil \frac{T_0}{\tau} \rceil$. It represents the maximum number of stimuli that can be accommodated within T_0. Recall that τ is the interval duration between two adjacent stimuli. In the second stage, we exhaustively search for the maximum KLD among 2^K-1 possible stimuli combinations. To this end, we calculate KLD for each stimuli combination. Since m_i's are fixed under each combination as a result of the first stage, the original optimization problem is significantly simplified with t_i's as the only variables. Now the remaining question is how to solve the simplified optimization problem.

Algorithm 1: Two-stage heuristic algorithm

input : P_i ($i \in [1,N]$) and Q

output: Solution of $\boldsymbol{m}$ and $\boldsymbol{t}$

Calculate KLD for each stimulus i;

$K = \lceil \frac{T_0}{\tau} \rceil$;

Select top-K stimuli with highest KLD;

for $j = 1$ **to** $2^K - 1$ **do**

    Formulate the simplified (3.2) given the j-th stimuli combination;

    Solve it via the *approximate gradient descent* algorithm;

Pick the stimuli combination with the highest KLD;

The corresponding $\boldsymbol{m}$ and $\boldsymbol{t}$ serve as the final solution.

For this, we employ the *approximate gradient descent* (AGD) algorithm [23, 167, 275]. It is an iterative method and useful especially when the derivative is hard to derive directly as in our case. The AGD algorithm finds an approximate solution for t_i's.

Dealing with long-term biometric changes. Like other biometrics, the auditory-pupillary response may exhibit variations over time [205, 225]. As a result, it can make the template acquired during the enrollment stage poorly representative of newly collected data samples, leading to degraded authentication performances. This phenomenon is known as *template aging* [124]. Many strategies have been developed to account for this issue [207, 213, 225]. Their main idea is to consistently update the classification model with new samples. In this work, we follow the existing approach to tackle the possible biometric pattern changes in the pupillary response. The core idea is to retrain the classification model with new samples from successful authentication trials. Our key steps are summarized as follows. 1) The system maintains a training dataset (reference set) of a fixed size after initial enrollment. The optimum training size is determined by the classifier, which is

investigated in Section 3.7.1. Like traditional passwords, this training dataset is securely stored in the device. 2) When a new authentication sample arrives, it is labeled legitimate if the authentication is successful. 3) The dataset is updated with new samples in a first-in-first-out manner: these new samples are added into the reference set while the same number of outdated samples is discarded in the meantime. 4) The classification model is retrained over the updated dataset each several days or even more frequently, depending on the authentication frequency of the user. Since lightweight classifiers are employed in the proposed authentication scheme, the corresponding computation overhead of training is minimal. Note that there are even more sophisticated adaptive mechanisms (e.g., [161, 174, 206]). We plan to integrate them into our design in future studies.

3.5 Security Analysis

3.5.1 Robustness Against Attacks

We primarily consider the impersonation attack throughout this work. The adversary intends to use its biometric credential, i.e., pupillary response, under the auditory stimuli to get authenticated. To launch the attack, the attacker is assumed to have physical access to the victim's VR headset. It happens, for example, when the device is lost/stolen or temporarily possessed by the victim's roommate. Our evaluation results show that the success rate of such attacks is merely 0.76% on average. The robustness of SoundLock against the impersonation attack will be presented with details in Section 3.7.2.

Like other biometric methods, adversaries can also attack SoundLock via the *replay attack*, where the adversary injects a previously recorded sample of the pupillary response. Such an attack is extremely difficult to perform in our case. As the user's eyes are fully covered by the VR headset, it is impossible to record the target's pupillary response externally. On the other hand, it is possible for the adversary to access the victim's pupillary response

samples via, say, pre-installing malware to the headset. Luckily, our scheme adopts the challenge-response authentication framework. With the interactive property, the attacker should know the auditory stimuli in advance to output the timely and correct response from the list of pre-recorded samples. It renders the attack very difficult to execute. Moreover, we argue that the device would be faced with an even more severe situation, if malware is pre-installed with access to the on-device authentication database.

Recent studies have also shown the feasibility of fabricating fake fingers and faces to bypass biometric authentication [26, 76, 289]. They are considered as a special kind of *mimicry attacks*. This attack is almost impossible to execute in our case, as the fabricated eyeball should be able to react to specific auditory stimuli. The pupil changes are subtle, smooth, dynamic, and unique to each individual. It is of great challenge, if not impossible, to build a mechanical device to mimic pupil dilation and constriction precisely. We are aware of some bionic eyes, which are essentially miniature cameras with necessary HCIs to optic nerves. Still, there is no "pupil" in bionic eyes. Besides, it costs around $150,000, which is extremely costly to deploy [276].

It is also possible that the auditory-pupillary response may be leaked, say, because of using a malicious (or compromised) device. Luckily, this new kind of biometric is revocable. In the case of having one pupillary response stolen or counterfeited, a new credential can be easily generated by changing its associated stimulus. It is also one of the prominent advantages of adopting auditory-pupillary response over other conventional biometrics for authentication.

3.5.2 Entropy Analysis

Entropy has been widely adopted to evaluate the security strength of authentication methods such as passwords [279] and PINs [280]. It is a measure of uncertainty in a random variable [63]. The classic entropy of a variable x with the distribution $P(x)$ is

defined as $H = -\sum_{x \in X} P(x) \log P(x)$. In the context of biometric systems, however, the classic entropy overlooks intra-user variability by assuming each user has fixed biometric features and overestimates biometric information [253]. To tackle this issue, some prior works adopt an alternative metric *relative entropy* to measure the security of a biometric system [12, 253, 301]. We thus consider this metric too. Relative entropy is defined as the decrease in uncertainty about a person's identity due to a set of biometric features measurements [12]. It is measured under the framework of KLD, $K = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right)$, where $P(x)$, $Q(x)$, and X represent the feature distribution of the target user, that of the reference set, and the feature space. It quantifies how much a single user's biometric feature distributions diverge from those of the population. It is noteworthy that the dataset plays an important role in the entropy computation as it defines the feature distributions $P(x)$ and $Q(x)$. According to the samples and their feature distributions collected in our dataset, K is calculated as 81 bits on average. Table 3.2 shows the relative entropy of SoundLock, keystroke, iris, fingerprint, and face, and the classic entropy of password and PIN. We can tell from the equations of these two kinds of entropy that classic entropy is an upper bound of relative entropy. In other words, the latter is a more conservative measure of authentication system security than the former [253]. The result shows that even the relative entropy of SoundLock (81 bits) largely exceeds those of passwords and PINs. SoundLock ranks second among all methods. It implies that dynamic pupillary response bears high uncertainty in the biometric information across individuals. It thus serves as a promising biometric for user identification. While the iris is associated with the highest relative entropy, the iris scanner is prohibitively costly to equip to a wide spectrum of VR devices.

3.6 Experiment Methodology

3.6.1 Experiment Setup

Apparatus. We perform all experiments using an HTC VIVE Pro VR headset tethered to an Exxact TensorEX 1x Intel Core X-Series processor workstation. A Pupil Labs eye tracker is integrated into the VR headset. All virtual scenes and the prototype of SoundLock are implemented using Unity, a cross-platform engine for VR development, and scripted in C# and Python. The prototype is developed to render stimuli and capture the pupillary response (i.e., time-series pupil size) through the eye tracker's API. It includes functions of enrollment, optimization, authentication, and device lock/recovery.

Experiment setup. Before the experiment, participants receive an introduction to the concept of SoundLock as well as experimental instructions. After providing informed consent to take part in the study, they are asked to fill out a pre-study questionnaire based on the introduction to evaluate the expected usability of SoundLock. Then, participants are instructed to put on the VR headset. A student researcher assists in adjusting the device to ensure the wearing comfort and the correctness of eye tracker readings. Throughout the entire experiment, the lab environment is kept quiet by default. Next, the participant's pupillary response is recorded while performing the authentication tasks. Task details are presented in Section 3.6.2. There is a short break between authentication tasks. After the experiment, participants are asked to fill out a post-study questionnaire to evaluate the perceived usability of SoundLock through the tasks.

To facilitate evaluation, we adopt several commonly used metrics: false acceptance rate (FAR), false rejection rate (FRR), equal error rate (EER), F1-score, and area under the ROC curve (AUC).

3.6.2 Experiment Design

The entire experiment consists of two phases: a pilot study and an in-field study.

Pilot study. The purpose of the pilot study is a) to collect preliminary data for the measurement study (see Section 4.4), b) to select from the candidate classifiers the one with the best overall performance, and c) to fix the classifier's training size and hyperparameters. In the pilot study, each participant is asked to listen to a set of 20 auditory stimuli consecutively. Their corresponding pupillary response is recorded. The auditory stimuli include white noise, monotones, prompt sounds, natural sounds, and human voices. Each stimulus is a 6-second audio track. Each stimulus is repeated 20 times for all participants. With the collected dataset, we carefully tune the training size and hyperparameters of each candidate classifier proposed in Section 3.3.3. Then, we compare all candidate classifiers and select the one with the best performance. Results will be discussed in Section 3.7.1.

- *Enrollment*: Each participant is presented with a set of 20 auditory stimuli, with each stimulus 5 times. Their auditory-pupillary responses are recorded. All the samples are used to train the classifier as well as to determine the user-specific stimuli via the algorithm introduced in Section 3.4. In this way, each participant's biometric credential is enrolled.
- *Authentication*: In this task, the user-specific stimuli sequence is presented to the participant. Access is granted if the newly recorded pupillary response is classified as a legitimate one. A participant has three chances to pass the authentication. It is deemed successful if the biometric credential is verified in at least one in three trials.
- *Impersonation attack*: In this task, participants are asked to perform impersonation attacks. The attacker intends to use its own biometric credential, i.e., pupillary response, under the auditory stimuli to get authenticated. Specifically, each participant is randomly assigned with three other participants' biometrics to mimic. The attacker is presented with the victim's customized stimuli. The attack is deemed successful if the attacker gets authenticated in any one of three consecutive trials.

- Participants are asked to repeat the authentication task in a few follow-up sessions to examine the scheme performance under various conditions (see Section 3.7.3). Specifically, to investigate the impact of user motion, participants are asked to perform authentication under four types of motions: static (baseline), eye movement, head rotation, and body stretch. To evaluate the SoundLock performance across different time of day, a series of sessions are scheduled for the same group of people from 10 AM to 6 PM, with a 2-hour interval in between. To examine the impact of visual fatigue, authentication tasks are also conducted as participants are exposed to the VR environments for different time duration. We further carry out a longitudinal study. Participants are re-invited 7 days and 14 days after the main session to repeat the tasks. The purpose is to show if auditory-pupillary response as a biometric credential would change over time.

Attendance and time consumption. A total of 32 participants completed the pilot study. The average time spent is around 60 min, including 50 min for displaying all auditory stimuli samples and data recording with 10 min overhead. In the in-field study, 44 participants completed the main session, which consists of the enrollment, authentication, and impersonation attack tasks. They all participated in the impact of the user motion and the visual fatigue sessions right after the main session. The above sessions take around 50 min including necessary overhead, such as Q&A and reading/signing the consent form. 25 of them participated in the impact of time session. 28 and 18 of them completed the 7-day and 14-day longitudinal study, respectively. A user study is conducted; it consists of a pre-study and a post-study before and after the main session, respectively.

3.6.3 Recruitment and Ethical Aspects

Participant recruitment and demographics. The participants are recruited and informed through emails, social media postings (departmental Facebook website), and verbal

communications. When a participant shows interest in participating in our study, we provide him/her a screening questionnaire asking about age, gender, race, and hearing and visual abilities. We screen participants with no hearing and visual impairments (corrected hearing ability with hearing aids and corrected visual ability with glasses and contact lenses will be allowed). Efforts have been made to recruit a diversified population based on age, gender, and race. After that, the participants are officially invited and asked to schedule a time and date with the researchers for the study. A total of 44 participants are recruited. They are all college students, faculty, and staff, aged between 17 and 40. Their demographic information is given in Table 5.1. Each phase takes around 1 hour on average. Participants are compensated at a rate of $10 per hour.

Ethical aspects. The participants are provided with the Informed Consent document before the study. The document provides a detailed description of the study's procedure, benefits/risks, intentions, compensation, possible risks/discomforts, and rights. In order to make sure that participants are aware of the study procedures, the research team reads the summarized and important contents of the consent document at the beginning of each experiment and answers any questions the participant may have. The consent document is signed in person when the participants are in the lab. Subjects have the option to decide whether to participate in the experiments or not. During the experiment, they are free to take a break or quit at any time without penalty. They can ask any questions related to this research. The research team signs a confidentiality agreement with the participants regarding the protection of their biometric data, which are anonymized and securely stored, and will only be used for the purpose of this research. The entire study is IRB-approved.

3.7 Results

3.7.1 Pilot Study–Classifier Selection

In the pilot study, the objective is to examine the candidate classification models and select the one that fits our scenario the best. The results will be used in the prototype development.

Classification model comparison. We implement different classification models as discussed in Section 3.3.3, namely k-NN, OC-SVM, B-SVM, LR, GNB, and RF. 10-fold cross-validation is performed with the collected dataset. Specifically, the dataset is randomly split into two subsets, a training set and a testing set. Then, the classifier is trained and tested, with each user iteratively regarded as legitimate and the rest being imposters. This process is repeated 10 folds to prevent overfitting. We plot FAR and FRR in Figure 3.12 by tuning the hyperparameters of the classification models.

For k-NN, it measures the distance between the testing sample and k training samples and compares it to a threshold α: if the distance is below α, the testing sample is deemed legitimate; otherwise, it is adversarial. Therefore, a larger α implies a looser detection rule that more likely considers an input sample legitimate and vice versa. By controlling the hyperparameter α, i.e., the distance threshold, we obtain the EER of k-NN equal to 1.5% at $\alpha = 1.0$ (see Figure 3.12(a)).

For SVM, its idea is to find an optimal hyper-plane in high-dimensional space to perform classification. We adopt the radial basis function (RBF) kernel, a popular kernelized function, to transform the non-linear data to higher dimensions. A critical hyperparameter for the RBF kernel is γ, the standard deviation of the kernel function that defines the decision boundary qualitatively; a larger γ indicates a more relaxed decision criterion to avoid the hazard of overfitting, resulting in a higher possibility that the input is accepted; a smaller γ implies a strict and sharp decision boundary. Figure 3.12(b) (3.12(c)) illustrates

the FAR and FRR of the OC-SVM (B-SVM) as γ changes, with other parameters optimized. We find the lowest EER for OC-SVM as 3.4% at $\gamma = 6.3 \times 10^{-3}$. For B-SVM, the lowest EER is 4.3%, obtained at $\gamma = 3.2 \times 10^{-3}$.

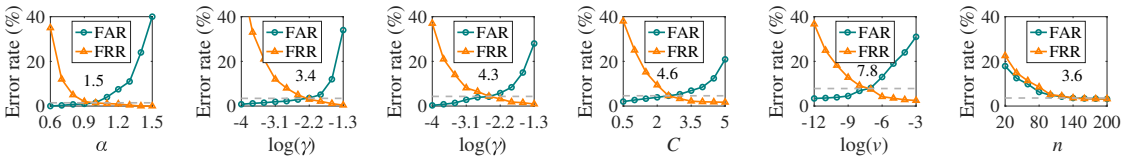
Similarly, for LR, which uses a logistic function to model the dependent variable to generate a classification output, an essential hyperparameter is C, the inverse of regularization strength; a larger C corresponds to less regularization and vice versa. As depicted in Figure 3.12(d), the lowest EER of LR is obtained as 4.6% by tuning C to be 2.5.

As a widely adopted probabilistic machine learning algorithm, GNB works by calculating each data point and assigning the point to the higher class probability that it belongs to. An important hyperparameter is the variance smoothing v, which indicates the portion of the largest variance of all features added to variances for calculation stability. By setting $v = 10^{-7}$, we obtain the lowest EER of GNB as 7.8%, as shown in Figure 3.12(e).

RF consists of many decision trees and uses bagging and feature randomness when building each tree to create an uncorrelated forest of trees whose prediction by committee is the most accurate. An important hyperparameter is n, the number of trees. A larger n leads to more accurate predictions at the cost of higher computation time and power consumption. We plot in Figure 3.12(f) the FAR and FRR curves as a function of the n. We find the EER converges to 3.6% as n approaches 140.

Table 3.4 compares all the classification models in terms of EER, F1-score, and AUC. Among them, k-NN produces the optimal FAR-FRR tradeoff with the lowest EER of 1.5% as well as the highest F1-score (0.983) and AUC (0.996). Its superior performance is primarily due to its robustness with respect to the data size. Compared with other models that generally require a large training dataset, k-NN better fits our scenario, where only a limited number of training samples (around 5) are collected.

Training data size. Figure 3.13 shows the EER with respect to the training data size, i.e., the number of enrolled samples. Given the same training data size, k-NN achieves the



(a) k-NN (b) OC-SVM (c) B-SVM (d) LR (e) GNB (f) RF

Figure 3.12: FAR, FRR, and EER of each classification model.

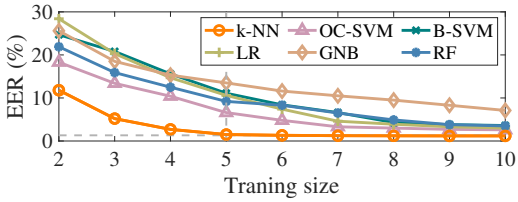
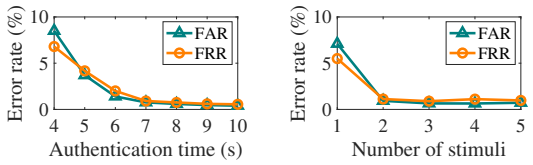


Figure 3.13: Impact of training data size on EER.



(a) (b)

Figure 3.14: Authentication performance.

lowest EER among the six classifiers, while GNB exhibits the worst performance. This is because the latter relies on the assumption that each class follows a Gaussian distribution. A larger dataset is thus needed to properly model the distribution. Empirical study indicates that it typically takes at least tens to hundreds of samples, depending on the task, to deliver a satisfying performance [20, 183, 197]. In contrast, only 5 samples are needed for k-NN to obtain EER as low as 1.5%. It indicates that k-NN attains a promising authentication accuracy with much fewer training samples.

To sum up, k-NN outperforms the other five models in classification accuracy, given the same training data size in our case. More importantly, it takes as few as 5 samples to sufficiently train the classifier. Hence, the enrollment stage can be performed efficiently.

3.7.2 In-field Study–System Performance

As a proof-of-concept implementation, we develop the prototype of SoundLock. Motivated by the results from the pilot study, we implement k-NN as the classifier and fix its

hyperparameters as discussed. A total of five training samples are collected from each participant in the enrollment stage. A series of in-field tests are conducted to evaluate the system's performance.

Authentication accuracy vs. authentication time. We first examine the authentication accuracy of SoundLock with respect to authentication time in Figure 3.14(a). Authentication time is defined as the span from stimulus onset until the response cutoff. In other words, it includes the time to present stimuli and the time for the pupil to react. Both FAR and FRR drop given a longer authentication time. This is because more features are extracted and thus enhance the classification accuracy. We also notice that the benefit of a longer duration becomes marginal if it is beyond 7 s, with the average FAR and FRR as low as 0.76% and 0.91%, respectively. Figure 3.14(b) depicts the authentication accuracy by adopting different numbers of stimuli; the error rate decreases with more stimuli presented. It complies with the result in Figure 3.10–more stimuli enhance the distinguishability of the target user. Based on these observations, we adopt the multi-stimuli scheme and set the authentication duration threshold T_0 as 7 s in the optimization formulation to strike a balance between security and usability. Table 3.5 compares the authentication time between SoundLock and existing works. Classic PINs and drawing patterns generally require a shorter time according to evaluation results from [99]. However, it demands relatively high motor skills for users to quickly enter these credentials in VR. They have been criticized as unfriendly to the elderly population and new users. Besides, relying on visual cues may hinder their usage for people with visual impairments. Among biometric schemes, SoundLock exhibits reasonable authentication time. Note that all these schemes need extra sensors, such as an EEG, to acquire the biometric signals.

Authentication accuracy comparison with state-of-the-arts. We further compare overall authentication accuracy between SoundLock and state-of-the-art solutions. Table 3.5 shows that SoundLock almost achieves the best performance among all biometric meth-

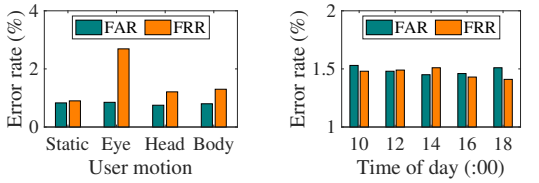
ods in terms of FAR (0.76%), FRR (0.91%), and F1-score (0.984). Besides, it outperforms PIN and drawing pattern in FRR. It means a legitimate user gets denied by PIN or drawing pattern at a higher chance. This is because these two methods require necessary motor skills to perform especially on VR terminals. Errors would occur during credential entry when controllers are not operated properly.

3.7.3 Performance Under Various Scenarios

In practical scenarios, a user may perform the authentication under different conditions, such as motion, time of day, and exposure time to VR environments. It is critical to evaluate if SoundLock is susceptible to these conditions.

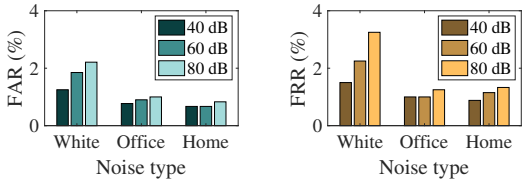
Impact of user motion. Since a user may make some movements during the authentication, it is important to show that the proposed scheme is motion-insensitive. In the experiments, participants are asked to perform four types of motions: sitting (static), eye movement, head rotation, and body stretch. The corresponding authentication accuracy is depicted in Figure 3.15(a). We find that the best performance is achieved at the static status with averaged FAR = 0.83% and FRR = 0.90%. Eye movement is associated with the highest FRR. This is because it introduces errors in the eye tracker calibrating the pupil size. Still, the authentication accuracy is practically acceptable with FAR = 0.85% and FRR = 2.69%. Based on the above results, users would be recommended to minimize their eye movement for the login duration. Other moving actions such as head rotation and body stretch also marginally increase the FRR, possibly due to the slight displacement of the eye tracker. Nevertheless, the increase is negligible; besides, the FAR remains consistent among various types of user motions ($0.80 \pm 0.05\%$), which suggests that user motions would not impact the security property of SoundLock.

Impact of noisy environments. We evaluate the impact of ambient noise on the performance of SoundLock. Three kinds of noises have been considered: white noise,



(a) Impact of user motion (b) Impact of time of day

Figure 3.15: Performance under various conditions.



(a) FAR (b) FRR

Figure 3.16: Performance under various noise types and levels.

office noise, and home noise. In particular, the white noise is synthesized with all the audible frequencies at the same intensity. The office noise is composed of people chatting, typing, phone ringing, computer fans, paperwork, etc. The home noise is a mixture of air conditioning, laundry, door locking, repairing, and TV sounds. All these soundtracks are downloaded from Mixkit [176]. In the experiments, the sounds are played as background noises by a pair of external speakers connecting to a second PC in the lab. We thus simulate the VR usage scenarios in generic, office, and home environments, respectively. Results are shown in Figure 3.16. We find that the performance, FAR and FRR, degrades slightly as the sound level increases from 40 dB to 80 dB. Note that sound levels are in the 40-80 dB range in most offices and homes [62]. It indicates that the ambient noise does influence the pupillary response. On the other hand, the influence is limited. Take home noise as an example. FAR = 0.67% and FRR = 0.88% as the sound level is 40 dB. They become 0.83% and 1.33%, respectively, if the noise is at 80 dB. It may be attributed to the fact that the stimulus audio is played via the VR headset, which is much closer to the user's ears than the noise sources. The former is thus dominant in the perceived sound.

Impact of time of day. We further examine if SoundLock is subject to the time of the day it is performed. A series of tests are scheduled over the same group of participants from 10 AM until 6 PM, with a 2-hour interval in between. We find in Figure 3.15(b) that the performance is relatively stable throughout the day. To quantify the statistical difference

in the FAR and FRR across different time of the day, we employ the *Kruskal-Wallis test* [143]. The test result indicates there is no significant difference on both FAR ($\chi^2 = 2.56$, $p > 0.05$) and FRR ($\chi^2 = 6.05$, $p > 0.05$) with respect to the time of the day.

Impact of exposure time to VR environments. It is well known that wearing VR for long periods can cause visual fatigue and motion sickness due to vergence-accommodation conflict [49]. It is therefore interesting to evaluate the performance of SoundLock with respect to users' exposure time to VR environments. In the experiment, each participant is asked to stay in the immersive environment for various periods of time, i.e., 10, 20, or 30 min, before performing the authentication. A participant can freely quit the experiment whenever they report discomfort or at any time they desire. In particular, users can choose to watch VR videos, play VR games, or browse online via the device. Table 3.6 summarizes the results. We observe that both FAR and FRR slightly increase under a long exposure time. The increase of FRR is relatively more prominent, by 0.74% from 0 min to 30 min. Conversely, FAR only sees a minor increase of 0.16% over time. This indicates the security of SoundLock is not influenced much, since incorrectly accepted adversarial authentications are limited; however, there is a moderately increasing chance that a legitimate user is wrongly classified. It indicates that pupillary response drifts slightly as the user is exposed to the VR environment for a while.

Longitudinal study. To investigate the long-term performance of SoundLock, participants are invited to attend two follow-up sessions, 7 days and 14 days after the main session, to repeat the authentication process. 28 and 18 participated in the two follow-up sessions, respectively. The adaptation strategy introduced in Section 3.4 is adopted. For comparison, we also test in the last session the performance of SoundLock without adaptation. The result is summarized in Table 3.7. The error rate increases as time proceeds without adaptation, with FAR (FRR) rising from 0.79% (0.91%) to 8.89% (5.56%), after a 14-day duration. It implies that the biometrics, i.e., the auditory-pupillary response, drifts

slowly over time. In comparison, the long-term performance becomes stable with the integration of our adaptation strategy. Specifically, the FAR (FRR) is 2.22% (1.48%), which merely exhibits a performance change of +1.46% (+0.57%). It suggests that our approach effectively deals with the temporal variation in pupillary response. Note that participants do not perform authentication in between sessions. We optimistically expect an even better long-term performance when SoundLock is under daily usage as the adaptation can be executed more frequently.

3.7.4 User Study

The goal of the user study is to evaluate the usability of SoundLock from participants' subjective perceptions.

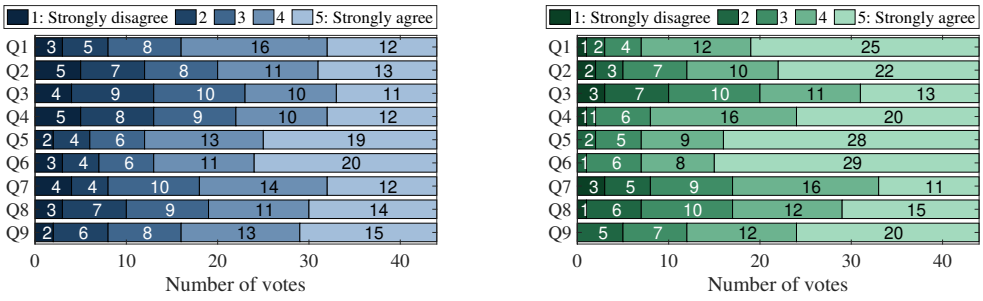
Design. The study consists of a pre-study and a post-study, conducted before and after the main session of the experiment, respectively. To investigate the impact of the in-field experiments on user perception, the same questionnaire is used in both studies. In part-I of the questionnaire, all participants are asked to provide their perception of SoundLock by responding to 9 questions on a 5-point Likert scale (with 1 = strongly disagree and 5 = strongly agree). These questions cover multiple aspects of security and usability. Table 3.8 lists all the questions. Part-II of the questionnaire includes three open-ended questions regarding overall experience *"What's your overall experience with SoundLock?"*, concerns *"Do you have any concerns or did you notice any potential issues of SoundLock?"*, and suggestions *"Do you have any suggestions to improve SoundLock in the future?"*.

Results. All 44 participants respond to the questions. Figure 3.17(b) displays the distribution of answers to part-I questions in post-study. In general, participants express their satisfaction with SoundLock, especially in Q1 ($\mu = 4.32, \sigma = 0.97$, median $= 5$), Q2 ($\mu = 4.07, \sigma = 1.16$, median $= 4$), Q4 ($\mu = 4.20, \sigma = 0.92$, median $= 4$), Q5 ($\mu = 4.43, \sigma = 0.86$, median $= 5$), and Q6 ($\mu = 4.48, \sigma = 0.81$, median $= 5$). The least rated

one is Q3 ($\mu = 3.55, \sigma = 1.25$, median $= 4$). As reported by several participants in the open-ended questions, this is caused by a couple of audio tracks in the stimuli pool. After close examination, listening discomfort is observed in audio tracks with bursting and high-pitch sound.

We then compare the survey results between the pre-study and the post-study using the Student's t-test [250], to investigate whether there is a significant statistical difference between the two studies. According to the test result, the most significant difference between the two studies lies in Q1 ($t(86) = 2.06, p = 0.021 < 0.05$), Q4 ($t(86) = 1.87, p = 0.032 < 0.05$), Q5 ($t(86) = 1.73, p = 0.044 < 0.05$), and Q6 ($t(86) = 1.70, p = 0.046 < 0.05$). Q7 has the least significant inter-study difference ($t(86) = 0.05, p = 0.480$). In general, the average rating is higher in the post-study than the pre-study for all questions; even Q7 sees a slight improvement ($\mu_{pre} = 3.59$ vs. $\mu_{post} = 3.61$). These results indicate that SoundLock exceeds users' expectations after they have real experience with it. Q9 reflects the user's overall attitude towards SoundLock for real-world usage. The result for Q9 in the pre-study (post-study) is $\mu_{pre} = 3.75$, $\sigma_{pre} = 1.20$ ($\mu_{post} = 4.07$, $\sigma_{post} = 1.04$). Meanwhile, 63.6% (72.7%) users report a score larger than 3, i.e., agree or strongly agree, in the pre-study (post-study). This means that most users are willing to adopt SoundLock as the authentication method for VR devices. To summarize, SoundLock is well perceived by users, primarily due to its high security (Q1), ease to use (Q4), ease of learning (Q5), and low cognitive load (Q6).

Subjective feedback. A total of 24 participants respond to the open questions in part-II of the questionnaire. 13 participants leave feedback on the overall experience of SoundLock. Among them, 4 deem the authentication process in SoundLock to be fun, e.g., "*It was a fun experience!*" (P9). 3 appreciate the idea and logic behind SoundLock, e.g., "*The idea of using pupil for authentication is smart.*" (P35). 5 participants report satisfac-



(a) Pre-study results

(b) Post-study results

Figure 3.17: Participants' subjective response distributions.

tory usability of SoundLock, e.g., "*I don't need to do anything and the authentication is automatically done.*" (P40).

Questions are raised by 9 participants. 3 of them question the robustness against consanguinity, e.g., "*Will twins or siblings be able to hack into each other's profile?*" (P35). This question is mainly due to the observation that identical twins or even siblings tend to share certain similar biometrics. Since there are no twins or siblings in our hired participants, we are unable to answer the question. We plan to investigate this as part of our future work with an extended group of subjects. 2 participants express privacy concerns, e.g., "*Will my pupillary response be used to infer what I'm thinking?*" (P38). Since the auditory-pupillary response is a reflex biometric, the pupillary response is stimulus-dependent. Basically, it reflects how human eyes react to an audio sound rather than the user's cortical processing, i.e., mental behavior. So far, we are unaware of any existing results on this topic. 3 participants mention some discomfort in listening to a couple of auditory stimuli with bursting and high-pitch sound. As a solution, we plan to investigate an even larger auditory stimuli pool in our future work. Volunteers will be invited to listen to and rate those stimuli. Any unpleasant ones will be discarded.

10 participants provide their comments for potential improvement. Among them, 3 suggest lowering the sound volume. It is worth mentioning that participants exhibit different tolerance of sound volume. 3 participants suggest combining with other forms of stimuli, such as colors, images, and videos. 4 propose to generalize SoundLock to the AR platform and other terminals, e.g., "*I think the system can be extended to smartphones, which will prove a valuable addition. The speaker can emit a sound and the eye image can be captured by the camera.*" (P1). Many of the comments are valid and inspire us with potential future work.

3.8 Related Work

User authentication on VR. While password and PIN serve as the most popular authentication mechanisms on VR devices [190], they have been criticized for these usability deficits: it takes users substantial effort to select correct letters/digits/characters from the virtual keyboard using the virtual laser extended from their controllers. Moreover, such authentication schemes have been proven highly vulnerable to shoulder-surfing attacks. Due to the occlusion of the VR headset, the user is unaware of the surroundings, rendering it easier for an adversary to acquire the entered credential through observation [82]. To address these issues, both industry and academia have been actively searching for practical alternatives. The existing methods can be broadly categorized into five classes: *knowledge-based methods* [10, 94, 98, 170, 305], *physiological biometrics* [19, 55, 155, 234], *behavioral biometrics* [162, 185, 204, 237, 311], *token-based methods* [48], and a mixture of them [169, 297, 314]. A recently published SoK paper provides an extensive discussion on this topic [248]. Please refer to it for more details. According to its discussion, physiological biometrics seem to outshine their peers so far due to their high usability and accuracy. Nevertheless, they bear at least two limitations for broad deployment. First, to capture users'

biometric information, such as electroencephalogram (EEG), electrooculography (EOG), electrical muscle stimulation (EMS), and iris patterns, sophisticated sensors are required. For example, iris scan has been deployed for user authentication on HoloLens 2 [175], a high-end AR device costing at least $3,500. Due to its high price, the iris scanner is less likely to equip to general VR devices[1] in the near future [260]. Besides, biometrics are unique to an individual. Once such credentials are damaged or counterfeited, the user cannot cancel the pre-stored credentials or reset them with different biometric input. This property is also called *cancelability*. In contrast, our approach exhibits the following advantages. First, it is free from additional high-end sensing devices; instead, it only needs an eye tracker, which has been integrated into many prevalent commercial VR headsets. It is well accepted that incorporating eye-tracking technology is a trend in VR to assist in simulating depth of field and focus and providing users with a more realistic and natural visual experience [60]. Second, auditory-pupillary responses are cancelable. In the case of having one pupillary response stolen, a new credential can be generated by changing its associated stimuli.

Table 3.9 provides a comprehensive comparison among some representative user authentication schemes for VR. The existing schemes are categorized into knowledge-based authentication (white), physiological biometric authentication (light gray), behavioral biometric authentication (medium gray), and multi-factor authentication (dark gray). All schemes are compared from multiple aspects of usability and security.

Pupillary biometrics for user authentication. The idea of exploiting pupillometry for user authentication has been around for a decade [35, 46, 189]. Most efforts have been devoted to enhancing authentication accuracy. For example, Bednarik et al. [28] proposed combining pupillary biometrics with eye movements for user authentication. A similar idea is adopted in [79]. However, implementing these schemes is faced with several practical

[1]The price of Meta Quest 2, the most popular VR device so far, ranges from $299 to $399.

challenges: eye movements and pupillary behaviors are task-dependent and light-sensitive. To overcome these limitations, researchers proposed leveraging pupillary light reflex (PLR) for user authentication [66, 294, 295]. PLR is an involuntary reaction of the human eyes to an external light stimulus: as a user is presented with lights of various combinations of chromas and intensities, her pupils will constrict and dilate accordingly. Typically, to elicit prominent and acute changes in pupil size (to create distinguishable features for classification), users are presented with strong light stimuli. It may lead to snow-blindness and flash-blindness effects [39, 92, 245]. Performing it on a daily basis could potentially bring severe health issues, e.g., temporary or even permanent vision impairment. Alternatively, SoundLock avoids the above concern as it employs auditory stimuli.

Challenge-response protocols for biometric authentication. Challenge-response has recently emerged as a popular authentication protocol and is frequently combined with biometrics for user identification. It leverages a user's physiological response to a given stimulus, i.e., challenge, injected by the interactive device. The assumption is that each user's response to a given challenge is unique. Examples of challenge-response biometrics include: the palm's/finger's response to vibrations [152, 158], EEG response to visual stimuli [19, 155, 288], or muscle response to electrical stimulation [55]. For example, Velody [152] makes use of the unique and nonlinear hand-surface vibration response for user identification. Similarly, VibWrite [158] enables user authentication via finger inputs on ubiquitous surfaces through physical vibration. It is implemented using a pair of vibration motors and a receiver that can be attached to any surface. Lin et al. [155] proposed a psychophysiological authentication protocol using carefully designed visual stimuli to acquire brain response signals. A similar idea is adopted in [19, 288]. Compared to conventional biometric authentication, the credentials created under challenge-response protocols are revocable–once a credential is counterfeited, it is convenient to reset it. Nonetheless, all the above schemes either rely on sophisticated sensors for response data acquisition or

require actuators for challenge generation (e.g., motor vibrator), which do not exist in VR headsets. Hence, they are inapplicable here. Recently, reflexive eye behaviors in response to visual stimuli [240] have been exploited for user authentication. Their stimulus consists of presenting a single red dot on a dark screen that changes position multiple times. Then reflexive saccades are triggered; the distinctive gaze path is treated as the unique signature. This scheme requires explicit action, i.e., eye movement, from the user. Instead, SoundLock elicits users' involuntary pupil size changes in response to auditory stimuli with bare cognitive effort.

3.9 Limitations and Future Work

In this section, we discuss several limitations of this work and present our future research directions.

Enrollment time. SoundLock is associated with a relatively long enrollment time. Under the current design, it ranges between 800 to 820 seconds. This is because SoundLock collects user's pupillary responses to the entire stimuli pool which consists of dozens of audio clips in the enrollment stage. The user-specific optimization is applied to find the best stimuli sequence for an individual. Note that the enrollment is only performed once for each user. To further reduce it, we can replace the current online user-specific optimization with offline optimization on the population scale, that is, an optimal stimuli sequence is derived for a large population group. In this way, only one stimuli sequence is rendered in the enrollment stage rather than the entire pool. The enrollment time would be substantially reduced accordingly. If a user's credential is counterfeited, a new stimuli sequence should be requested. As another possible approach, rather than presenting a user with the whole stimuli pool, we can reasonably present a subtset. We will carefully select the stimuli that

generate the highest entropy among users. Besides, analysis is necessary to evaluate its impact on authentication accuracy.

Multi-modality stimuli. SoundLock only makes use of auditory stimuli. In fact, visual stimuli, such as lights, images, and moving objects, would also evoke pupillary response. In our future work, we plan to investigate biometric authentication methods exploiting multi-modality stimuli. Hopefully, it would introduce new feature dimensions and thus further enhance the system entropy. There are several research questions deserving thorough investigation. First, how to combine visual and auditory stimuli? There are at least two strategies, to display the two kinds of stimuli sequentially or concurrently. Different strategies would lead to distinctive pupillary response patterns (and thus entropy) and time efficiency. Second, under the new design, a new set of prominent and reliable features should be extracted from the raw data to optimize the accuracy. Third, the user-specific stimuli optimization will be revisited to balance security and usability with multi-modality stimuli.

Scalability. SoundLock has been tested among 44 subjects. In our future work, we plan to find out whether the proposed biometric works for a larger and more diverse population. Besides, the current benchmarking of system entropy is based on the dataset collected so far. With extended participation, the calculation result would reflect the ground truth better. Besides, SoundLock is only prototyped and evaluated on one kind of VR model (HTC VIVE Pro) and has been exclusively focused on the VR platform. Next, we plan to evaluate SoundLock on a broader set of VR headsets and examine the impact of device heterogeneity. Additionally, we will also examine the feasibility of generalizing our idea to other platforms, such as AR terminals and smartphones.

3.10 Conclusion

In this paper, we present SoundLock, a novel user authentication scheme designed for VR devices. SoundLock recognizes legitimate users by extracting carefully designed features from pupil size changes in response to auditory stimuli. We first introduce a basic scheme using a single stimulus, followed by an advanced scheme with multi-stimuli. A proof-of-concept prototype of SoundLock is implemented on a VIVE Pro VR headset. Extensive in-field experiments are performed involving 44 participants. Results show that SoundLock offers high authentication accuracy, which outperforms state-of-the-art biometric authentication solutions for VR. SoundLock also exhibits consistent performances under various testing conditions. Our user study reveals that SoundLock is well received; 72.7% of the participants are willing to adopt SoundLock as the authentication mechanism on their (future) VR devices.

Table 3.1: List of all the 60 candidate features.

Index	Feature name	Category	Phase	Notation
1	Response lag	Morphological	Excitation	rl
2-3	Peak magnitudes	Morphological	Excitation	p_{1-2}
4-5	Valley magnitude	Morphological	Excitation	v_{1-2}
6-7	Dilation rates	Morphological	Excitation	dr_{1-2}
8-9	Constriction rates	Morphological	Excitation	cr_{1-2}
10-19	Dilation polynomial coefficients	Morphological	Excitation	dpc_{1-10}
20-29	Constriction polynomial coefficients	Morphological	Excitation	cpc_{1-10}
30	Area under curve	Morphological	Excitation	auc
31	Recovery time	Morphological	Recovery	rt
32-35	Damped oscillation	Morphological	Recovery	do_{1-4}
36-37	Pupillary unrest	Morphological	Recovery	pu_{1-2}
38	Baseline average	Morphological	Recovery	$bavg$
39	Baseline maximum	Morphological	Recovery	$bmax$
40	Baseline minimum	Morphological	Recovery	$bmin$
41	Baseline variance	Morphological	Recovery	$bvar$
42	Baseline median.	Morphological	Recovery	$bmed$
43	Baseline interquartile range	Morphological	Recovery	$biqr$
44	Average	Statistical	-	avg
45	Maximum	Statistical	-	max
46	Minimum	Statistical	-	min
47	Variance	Statistical	-	var
48	Median	Statistical	-	med
49	Root mean square	Statistical	-	rms
50	Skewness	Statistical	-	$skew$
51	Kurtosis	Statistical	-	$kurt$
52	Roughness	Statistical	-	$rough$
53	Sharpness	Statistical	-	$sharp$
54	First quartile	Statistical	-	$q1$
55	Third quartile	Statistical	-	$q3$
56	Interquartile range	Statistical	-	iqr
57	Mean absolute deviation	Statistical	-	mad
58	Slope sign change	Statistical	-	ssc
59	Mean crossing	Statistical	-	mx
60	Willison amplitude	Statistical	-	wa

Table 3.2: Entropy of various authentication methods.

Work	Authentication method	Entropy (bits)
Wang et al. [279]	Password	$20-23$
Wang et al. [280]	PIN (4-digit[1], 6-digit[2])	$8.41^{[1]}$, $13.21^{[2]}$
Sae-Bae et al. [227]	Keystroke	$3.48-4.62$
Youmaran et al. [301]	Iris	$278-288$
Takahashi et al. [256]	Fingerprint	18.6
Adler et al. [12]	Face	$37.0-55.6$
SoundLock (this work)	Pupillometry	81

Table 3.3: Participants' demographics.

Gender	#	%	Age	#	%	Iris color	#	%
Female	16	37	$\leq$18	4	9	Brown	34	77
Male	27	61	19-24	24	55	Hazel	6	14
Other	1	2	25-30	12	27	Blue	2	5
			31-36	3	7	Green	1	2
			$\geq$37	1	2	Other	1	2
Eye wear type	#	%	VR usage	#	%	VR auth experience	#	%
None	28	64	Frequent	5	11	Proficient	3	7
Glasses	13	29	Occasional	8	18	Limited	5	11
Contact lenses	3	7	Rare	13	30	None	36	82
			Never	18	41			

Table 3.4: Performance comparison among different classification models.

Classification type	Model	EER (%)	F1-score	AUC
One-class	k-NN	**1.5**	**0.983**	**0.996**
	OC-SVM	3.4	0.956	0.989
Binary	B-SVM	4.3	0.935	0.986
	LR	4.6	0.929	0.990
	GNB	7.8	0.909	0.956
	RF	3.9	0.944	0.984

Table 3.5: Performance comparison with state-of-the-art schemes. *Values are obtained from [99].

Approach	FAR (%)	FRR (%)	F1-score	Auth time
PIN*	-	>1.14	-	2.54-2.95
Drawing pattern*	-	>5.19	-	2.82-3.87
OcuLock [162]	3.55	3.55	0.983	≤10
SkullConduct [234]	6.90	6.90	-	≤23
Brain Password [155]	2.50	2.50	0.955	≈4.80
ElectricAuth [55]	0.83	2.00	-	≈**1.30**
SoundLock (this work)	**0.76**	**0.91**	**0.984**	≤7

Table 3.6: Performance under different exposure time to VR environments.

Exposure time	FAR (%)	FRR (%)
0 (baseline)	0.76	0.91
10 min	0.81	1.11
20 min	0.88	1.54
30 min	0.92	1.65

Table 3.7: Longitudinal study results. *Without the adaptation strategy.

Time span	FAR (%)	FRR (%)
0 (baseline)	0.76	0.91
7 days	1.19	2.14
14 days	2.22	1.48
14 days*	8.89	5.56

Table 3.8: Part-I questions.

	Question
Q1	SoundLock is a secure authentication scheme.
Q2	The authentication result is accurate.
Q3	There is no discomfort using SoundLock.
Q4	SoundLock is easy to use.
Q5	SoundLock is easy to learn.
Q6	SoundLock does not introduce much cognitive load.
Q7	The login time is acceptable.
Q8	SoundLock can be used on a daily basis.
Q9	I am willing to use SoundLock on my (future) VR device.

Table 3.9: Comparison among different user authentication approaches for VR. ●: method fulfills criterion. ◐: method quasi-fulfills criterion. ○: method does not fulfill criterion. –: not enough information. Att 1-4: replay attack, shoulder-surfing attack, impersonation attack, guessing attack.

Scheme	Sensor free	Hand free	Auth speed	Accuracy	Revocability	Vs att 1	Vs att 2	Vs att 3	Vs att 4
PIN	●	○	⋆⋆⋆	⋆⋆	●	○	○	–	○
Drawing pattern	●	○	⋆⋆⋆	⋆⋆	●	○	○	–	○
3D pattern [305]	●	○	⋆	–	●	○	●	–	○
CueVR [10]	●	○	⋆⋆	⋆	●	○	●	–	○
LookUnlock [94]	●	●	⋆	–	●	○	◐	–	○
RoomLock [98]	●	○	⋆⋆	⋆⋆	●	○	◐	–	○
RubikAuth [170]	●	○	⋆⋆⋆	⋆⋆⋆	●	○	●	–	○
SkullConduct [234]	○	●	⋆	⋆⋆	○	○	●	●	●
Brain Password [155]	○	●	⋆⋆⋆	⋆⋆⋆	●	●	●	●	●
Arias et al. [19]	○	●	⋆⋆	⋆	●	●	●	●	●
ElectricAuth [55]	○	○	⋆⋆⋆	⋆⋆⋆	●	●	●	●	●
SoundLock [316]	●	●	⋆⋆	⋆⋆⋆	●	●	●	●	●
GaitLock [237]	●	●	⋆⋆⋆	⋆⋆⋆	○	○	◐	●	●
OcuLock [162]	○	●	⋆	⋆⋆⋆	○	●	●	●	●
Kupin et al. [145]	○	○	⋆⋆⋆	⋆⋆	○	–	–	●	●
Mustafa et al. [185]	●	●	–	⋆⋆	○	–	●	●	●
Pfeuffer et al. [204]	●	◐	–	⋆	○	○	○	●	●
Zhang et al. [311]	●	●	⋆⋆⋆	⋆⋆	–	–	●	●	●
GlassGesture [297]	●	●	–	⋆⋆⋆	●	–	●	●	●
RubikBiom [169]	●	○	⋆⋆⋆	⋆⋆	●	–	●	●	●
BlinKey [314]	●	●	⋆	⋆⋆⋆	●	–	●	●	●

CHAPTER 4

EYEQOE: A NOVEL QoE ASSESSMENT MODEL FOR 360-DEGREE VIDEOS USING OCULAR BEHAVIORS

4.1 Introduction

Motivation. With the development of Virtual Reality (VR) technologies, 360-degree videos, also referred to as omnidirectional or VR videos, have seen a revolutionary rise over the last decade. As a novel type of multimedia, 360-degree videos provide an immersive and interactive watching experience by rendering spherical frames covering all directions around the viewer, attracting great interest from customers, researchers, and industry. In the meantime, these videos are mostly rendered in high resolutions to maintain fair visual quality. Given the limited network bandwidth, the network and service providers have to strike a balance between resource consumption and service quality for 360-degree video streaming. Hence, it is of essential importance for them to get an in-depth understanding of the user's experience and take necessary adaptive actions in service management. As a critical evaluation indicator, quality of experience (QoE), defined by ITU-T [123] as a measure of the acceptability of an application or service perceived subjectively by end-users, has been widely adopted. In current multimedia services, user's QoE is mainly obtained by asking people to measure their perceived quality via surveys or self-reports. However, such procedures are time-consuming and may be annoying for the users.

To address this issue, extensive prior efforts have been devoted to developing QoE assessment models that map a diverse spectrum of impact factors, such as underlying network conditions and video qualities, to a QoE score given a specific multimedia service type. In this way, it avoids bothering users with questions to collect opinions and feed-

back. QoE assessment is automatically carried out with significantly reduced human labor efforts. Nonetheless, this topic in the context of 360-degree videos in VR environments is yet far from well investigated. One mainstream of existing approaches can be classified as *video-centric models* [52, 67, 88, 239, 251, 252, 290, 303, 306, 313]. QoE is derived by analyzing distortions of videos displayed under various video quality assessment (VQA) metrics. These models are criticized for overlooking subjective factors. More recently, [150, 151, 291] integrate human visual attention to their QoE models. Their basis is that viewers mostly focus on objects of interest in a scene. Thus, distortions on different parts should impose a nonuniform impact on QoE estimation. These works then assign weights in accordance with the viewer's visual attention in aggregating pixel-wise distortions. The above ideas are inherited from QoE modeling of conventional 2D videos and thus incapable of capturing unique characteristics of 360-degree videos. As pointed out by prior studies [104, 134, 266, 317], subject feelings, such as cybersickness, immersiveness, and fatigue, are of essential importance in determining their perceived QoE of watching 360-degree videos, in addition to the well-recognized factors such as video quality. Hence, a QoE model that effectively harnesses all the above factors is in dire need for service management of 360-degree video streaming.

Recently, ocular behaviors, such as eye gaze, fixations, saccades, pupillometry, and blinks, have emerged as a new sensing modality to measure human perceptions. For example, eye blinking rates are reported to increase as the evolvement of visual fatigue [133, 267]. Strong correlations are also observed between visual fatigue and saccade peak velocity, saccade duration, and fixation duration [307]. Eye-based sensing has extended the current multimedia applications and services with an additional perceptive dimension and opened up grand opportunities to enhance service provisioning. For instance, Tesla is starting to use the camera above the rear-view mirror in some car models to help make sure people pay attention to the road while using Autopilot [196]. In the meantime, eye trackers

have been embedded into many prevalent commercial VR headsets [2, 93, 264, 270, 274] to assist in simulating depth of field and focus, providing a more realistic and natural visual experience. It is widely accepted that incorporating eye-tracking technology is a trend of VR headsets [60].

Our approach. Based on these observations, we propose to leverage ocular behaviors captured by eye trackers in VR headsets to model and predict viewer's perceived QoE in watching 360-degree videos. We call the novel prediction model EyeQoE. As presented in our measurement study (Section 4.4), strong correlations are broadly found between ocular behaviors and various impact factors of QoE for 360-degree videos, including the objective ones (e.g., video quality) and subjective ones (e.g., cybersickness, immersiveness, and fatigue). EyeQoE treats the behaviors as indicators of the viewer's perceived experience and aims to bridge these two. It takes the observed behaviors as inputs and produces a corresponding QoE score. In a holistic view, our model is superior to the state-of-the-art approaches from two aspects. First, it takes into account human feelings during QoE assessment, which are largely overlooked by prior works. Second, most prior works endeavor to exhaustively enumerate and include all impact factors in QoE modeling, which are impractical to implement in real-world scenarios. Alternatively, EyeQoE merely utilizes ocular behaviors to reflect the viewer's perceived QoE as a whole. Extensive experiment results show that it outperforms representative prior works in terms of prediction accuracy.

Despite the attractive sense of exploiting ocular behaviors for 360-degree video QoE assessment, enabling it involves several non-trivial challenges. First, ocular behaviors are affected by external visual stimuli [29, 121, 226] and biologically distinct across human subjects [95, 240]. For instance, a subject's gazing patterns tend to be more static when focusing on a tree than tracking a flying bird in a scene [31, 148, 173]. As human eyes have unique physical characteristics (e.g., sizes, biophysical structures, etc.), ocular behaviors may vary among individuals even watching the same video. Thus, EyeQoE needs to cope

with variations introduced by subjects and visual stimuli heterogeneity. Second, because of the intrinsic diversity of visual stimuli, i.e., video clips, the QoE assessment model, once trained over existing videos, may be hard to generalize to unseen videos. To deal with this challenge, a naive approach is to gather as many annotated training samples as possible. It means to cover videos of all kinds, which would lead to considerable overhead.

The proposed EyeQoE is inspired by some advanced techniques in deep neural networks. We first organize observed ocular behaviors into a basic graph, where fixations and saccades are its nodes and edges, respectively. They are connected in chronological order. The constructed graph preserves the visual patterns of the raw data in the temporal domain through modeling the local pairwise relation between adjacent fixations and saccades. We notice that high correlations also exist among fixations associated with the same object of interest in the scene, though they may be separated in the timeline. We thus extend the basic graph by adding additional edges between nodes of high similarity to preserve the content-dependent features. To facilitate learning over graph-structured data, the core of EyeQoE adopts a graph convolution network (GCN) based classifier. GCN is a superior network to produce useful feature representations of nodes and edges from graphs. In this work, it runs over every fixation and saccade and aggregates their layer-wise representation with those of its neighbors. The useful features accumulate and propagate throughout the entire graph as the convolution evolves. The output of the GCN classifier is a QoE score of the given video clip.

To tackle the challenge of subjects and visual stimuli heterogeneity, we enhance the GCN classifier by applying a Siamese network framework with devised training sample selection strategies. The idea of the Siamese network is to employ a pair of substructures with the same GCN and weights. The selected pair of samples are passed through the two substructures separately. The distance metric between two outputs is computed and guides the updates of both substructures. The designed structure, together with the training pro-

cess, allow the model to tolerate inconsistency in ocular behaviors caused by heterogeneous subjects and visual stimuli. To accommodate unseen videos, we formulate our problem as *domain adaptation*. We first categorize all 360-degree videos into various types according to their *colorfulness*, *luminance*, and *motion*. Datasets associated with existing and unseen videos are treated as the source domain and the target domain, respectively. Hence, our problem involves multiple source domains. We then propose a multi-source adversarial domain adaptation (MADA) network based on the classic domain adaptation network [96] that is originally designed for single-source-domain scenarios.

The discussion of this work pertains to PC-tethered VR[1] (e.g., HTC VIVE, Oculus Rift, MS MR) and powerful standalone VR, both with the necessary computing capacity to carry out online inference and domain adaptation. The QoE model is first trained offline, say, at servers or cloud, and then transferred to VR devices, while the prediction is carried out in an online manner.

We highlight our contributions of this paper as follows:

- We introduce EyeQoE, a novel QoE assessment model for 360-degree videos using ocular behaviors. We then construct the behaviors into a graph that preserves both features in the temporal domain and content dependency.
- We develop a GCN-based classifier to facilitate learning over graphs. The classifier is then combined with a Siamese network to deal with subjects and visual stimuli heterogeneity. MADA is further proposed to easily adapt our model to unseen videos.
- We build our own dataset via a three-month data collection campaign. 50 volunteers and 5 student workers get involved. To our knowledge, it would be the first data source of annotated ocular behaviors for 360-degree video QoE assessment.

[1]Tethered VR means that the headset is physically connected to a computer by cables, such as HDMI and/or USB.

- We carry out extensive tests to evaluate EyeQoE based on our dataset. Results indicate that EyeQoE achieves the best prediction accuracy of 92.9%.

The rest of this paper is organized as follows. Section 4.2 reviews prior works related to our topic. Section 4.3 introduces some necessary background of using ocular behaviors for QoE assessment. A measurement study that validates the feasibility of our idea is presented in Section 4.4. The novel graph modeling of ocular behaviors is introduced in Section 4.5 followed by Section 4.6 that provides design details of EyeQoE. We evaluate EyeQoE in Section 4.7. A discussion over the limitations of EyeQoE is provided in Section 4.8. We conclude the paper in Section 4.9.

4.2 Related Work

Video-centric Models. Like conventional videos, some existing QoE assessment models for 360-degree videos directly analyze the displayed videos. QoE is derived by comparing distortions of the displayed video with its original version. This kind of approach is called video quality assessment (VQA). For 360-degree videos, new VQA metrics have been investigated [52, 252, 290, 303, 306, 313]. For example, built upon peak-signal-to-noise ratio (PSNR), a commonly adopted VQA metric for traditional videos, Yu *et al.* [303] modified it into sphere PSNR (S-PSNR) by further considering the impact of the so-called *sphere-to-plane mappings*. Basically, pixels would be distorted when projected from a two-dimensional plane to spherical surface. Sun *et al.* [252] took into account the projection distortion in their VQA metric by multiplying a weight to each pixel that reflects the relation between the sphere and the plane. In the above works, the calculation of VQA metrics is in need of the reference 360-degree videos, i.e., the original version without distortion. Unfortunately, this assumption is impractical in most real-world video streaming scenarios. To overcome the limitation, QoE assessment models with no refer-

ence videos have been developed [67, 88, 239, 251]. VQA metrics are directly derived from the features of impaired videos or network parameters, e.g., bandwidth, packet loss, and latency. Nonetheless, video-centric models are criticized for overlooking viewer's perceptive feelings during QoE assessment, such as immersiveness [104, 317] and cybersickness [131, 134]. As validated through many prior works [15, 184, 265], viewer's subjective experience of watching videos does not necessarily comply with their displayed qualities in many cases.

Visual attention enhanced models. Recently, some works start introducing human factors to QoE assessment of 360-degree videos. In an immersive environment, people cannot see the whole scene from a single viewpoint. Instead, they usually look around and focus on what attracts them. Hence, distortions on different parts of the projection sphere impose a nonuniform impact on QoE. With the basis of the traditional PSNR metric, Xu *et al.* [291] assigned weights on the pixel-wise distortion in calculating the PSNR according to the distribution of the viewer's visual attention. A similar idea is adopted by VQA-OV [150]. Visual attention is generated by tracking the viewer's head and eye movements via the embedded inertial sensors and eye tracker in a VR headset. In [151], they further constructed the subject's field of view (FoV) and saliency map to guide VQA assessment. The strategy of using visual attention or saliency map to boost the video-centric QoE models has also been adopted in the context of conventional videos [84, 147, 157, 291]. As a note, all the above works are still in need of reference videos to calculate pixel-wise distortions. Although these works utilize visual information in their models, it is essentially subject's visual attention. Instead, our work exploits physiological features in viewer's ocular behaviors to infer her satisfaction in watching 360-degree videos. Therefore, our problem formulation and the corresponding inference technique are totally different.

Among the prior works, [209] is the closest to ours. It combines facial expression and gaze direction for traditional video QoE assessment. Our work differs in two main as-

pects. First, we target 360-degree videos in VR environments while they are for traditional videos. Second, our work addresses critical challenges in data-driven QoE modeling, such as subjects and visual stimuli heterogeneity and adaptation to unseen videos. These issues are overlooked in [209].

Some other works investigate the feasibility of leveraging human behavior related data, such as heart rate, facial expression, electrodermal activity (EDA), and electroencephalogram (EEG), to evaluate QoE on various VR applications, including assistive technique systems [229], speech and language assessment applications [132], and general-purpose applications [47, 81]. None of them is designed for 360-degree videos. Besides, to our knowledge, no existing commercial VR headset nowadays is equipped with necessary sensors to acquire these human behavior data.

4.3 Background

Ocular behaviors as indicators of human perceptions. A connection between the ocular behaviors and human perceptions has been accepted for a decade [91, 192, 307]. Such behaviors include eye gaze, fixations, saccades, pupillometry, and various forms of eye opening and closure events. In neurophysiological literature, it is demonstrated that pupils are unconsciously regulated by autonomic nervous system stimulation, which is known to produce responsive output under numerous emotional states. Hess [112] reported behavior changes in subjects who view image stimuli that cause different pupil sizes; images with dilated pupils are deemed more attractive than those with constricted pupils. Eye blinks and gaze behaviors are treated as crucial indicators for visual fatigue, defined as eye-strain or asthenopia, which can be caused by both two-dimensional and stereoscopic moving images [91]. Studies show that eye blinking rates increase due to a prolonged period of time working in front of video display terminals. The exacerbated drying of the ocular

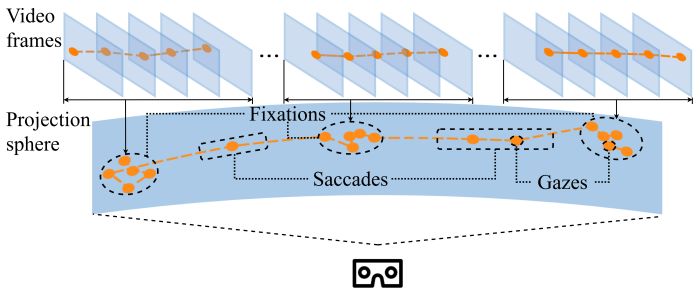


Figure 4.1: Relations of gazes, fixations, and saccades when a viewer is watching a 360-degree video.

surface causes subjects to blink more frequently to lubricate the surface of the cornea and conjunctiva [133, 267]. Prior works also demonstrate strong correlations in visual fatigue versus saccade peak velocity, saccade duration, and fixation duration [307]. Specifically, saccadic oculometrics, saccade peak velocity, and saccade duration significantly decrease as working time progresses, whereas the duration of medium-length fixations increases with fatigue development. All these findings motivate us to exploit ocular behaviors to infer human perceived QoE toward 360-degree videos.

Gazes, fixations, and saccades. Saccades are rapid stepwise movements of both eyes in the same direction that typically last 10-100 ms, depending on the distance covered [78]. They are used to shift the gaze to another location. In contrast to saccades, fixations are relatively focused, low-velocity eye movements with a typical duration of 100-400 ms and are used to stabilize the retina over a stationary object of interest. A visual gaze is the instantaneous visual point landing on the stimulus. A fixation consists of multiple time-series gazes concentrated around the same viewpoint. As shown in Figure 4.1, as a subject watches a 360-degree video, her fixations move over the projection sphere in accordance with the object of interest. Each fixation is associated with a series of frames that typically display a similar scene, in which the location of objects of interest is basically unchanged.

4.4 Measurement Study

While the correlation between ocular behaviors and human perceptions is well recognized, whether the former can serve as an indicator for 360-degree video QoE is unclear. Our measurement study intends to answer this question by carrying out extensive experiments. A total number of 10 subjects are invited to watch 360-degree videos of different qualities via the HTC Vive headset. Each video is of 25 seconds duration. Subjects' ocular behaviors are captured by a Pupil Labs eye tracker that is integrated into the headset. We then examine how they are influenced by various well- recognized impact factors of 360-degree video QoE, including video quality, cybersickness, immersiveness, and fatigue.

Observation 1: Ocular behaviors are impacted by video quality. Figure 4.2 exhibits the impact of video resolutions to ocular behaviors. Figure 4.2(e)-4.2(g) show coordinates of time-series gazes from one fixation with the image resolution of 4K, 1080p, and 720p, respectively. In these figures, the origin is the fixation center and the x-/y-coordinate of each gaze is its horizontal/vertical distance to the center. For fair comparison, we extract the fixations on the same object across the three videos. We find that gazes are more focused when the video is in a higher resolution. This phenomenon is further validated through Figure 4.2(i)-4.2(l) where the probabilistic distribution of gaze distance-to-center (GDC) is displayed. GDC mainly concentrates on the lower end of the x-axis, mostly lower than 0.03 for 4K videos. It becomes scattered as the resolution decreases. We have a similar observation over the gaze velocity in Figure 4.2(m)-4.2(p); eye movements within a fixation tend to slow down when watching a high-quality video, whereas they become faster as the quality is degraded.

Apart from the spatial distortion, we also explore the impact of the video's temporal distortion with stalling events in the video. Figure 4.3(a) shows the GDC of each observed gaze as time proceeds. There are three surges in GDC at the 5th, 12th, and 19th second, which are exactly time instances of the stalling events. It implies that visual attention

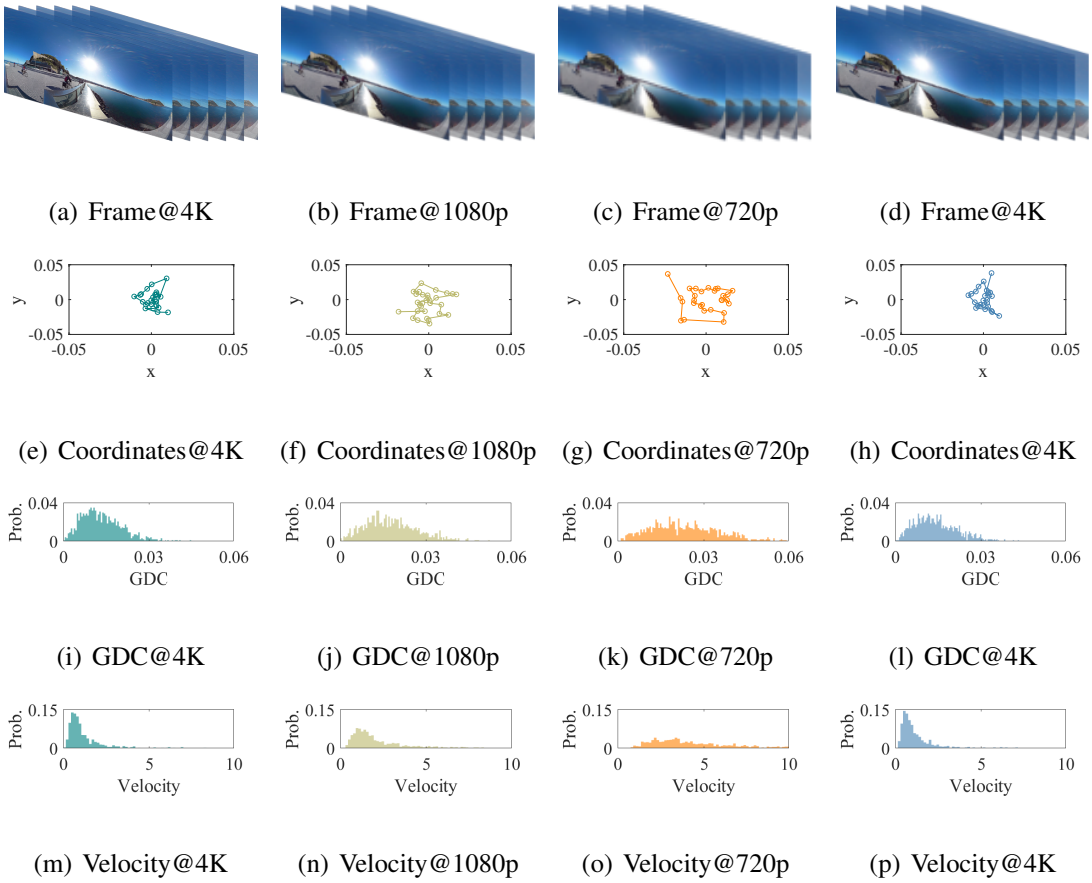
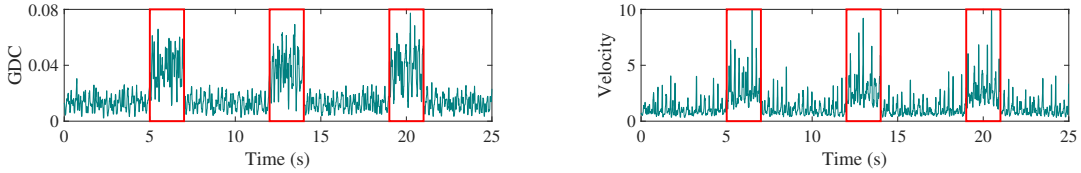


Figure 4.2: (a)-(d): 360-degree videos in resolutions of 4K, 1080p, 720p, and 4K (same subject in a second trial). (e)-(h): Normalized coordinates of gazes in one fixation. (i)-(l): Distribution of GDC. (m)-(p): Distribution of gaze velocity.

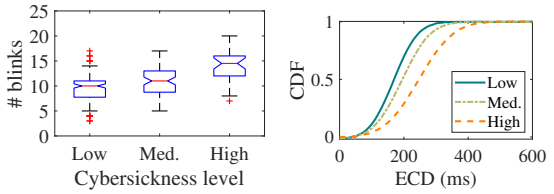
becomes less focused as stalling occurs. As indicated in Figure 4.3(b), gaze velocity also experiences significant increases as the video freezes.

Observation 2: Ocular behaviors are impacted by subjective factors. As verified in prior studies [104, 134, 266, 317], aside from the video quality, 360-degree video QoE is also influenced by subjective factors, namely cybersickness, fatigue, and immersiveness. Cybersickness, or motion sickness, refers to the subject's feeling of sickness, dizziness, nausea, etc., caused by, for example, the physical device, the VR environment, video contents, and the subject's physical status. Fatigue describes the subject's tiresome and is



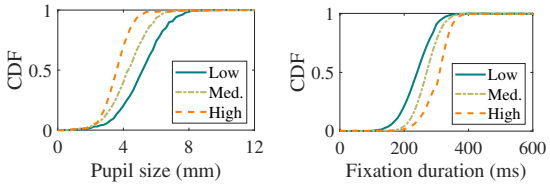
(a) GDC (b) Gaze velocity

Figure 4.3: The impact of video stalling on GCN and gaze velocity.



(a) Number of blinks (b) ECD

Figure 4.4: Impact of cybersickness. (a) The number of blinks in 25 seconds. (b) The CDF of ECD.

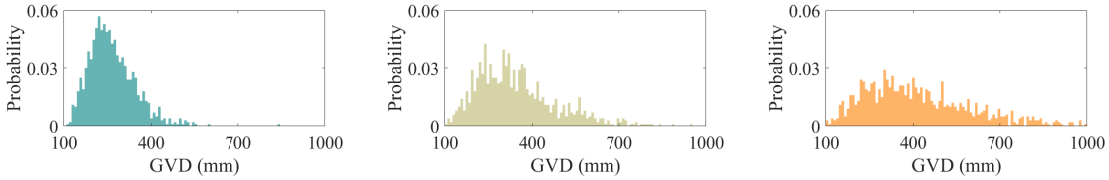


(a) Pupil size (b) Fixation duration

Figure 4.5: Impact of fatigue. (a) The CDF of pupil sizes. (b) The CDF of fixation duration.

mainly impacted by the time duration of watching videos. Immersiveness reflects the subject's perception of being physically present in the VR environment. In the measurement study, subjects are asked to rate their feelings towards cybersickness, fatigue, and immersiveness on a 3-point scale indicating low, medium, and high, respectively, after watching each video.

A correlation is observed between cybersickness and the subject's blink events. Figure 4.4(a) shows the number of blinks that a subject performs in watching a 360-degree video of 25 seconds under three cybersickness levels. Subjects tend to exhibit a higher blink rate when experiencing cybersickness. It may be due to more intense eye-strain symptoms, which leads to higher frequent blinks. Meanwhile, the eye closure duration (ECD) in each blink increases with higher perceived cybersickness, as presented in Figure 4.4(b). Our measurement study also reveals viewer's oculomotor and pupillary behaviors as potential



(a) Low GVD (b) Medium GVD (c) High GVD

Figure 4.6: Impact of immersiveness on viewer's GVD.

indicators of her fatigue. As shown in Figure 4.5, higher perceived fatigue is associated with shrunk pupil sizes and longer fixation durations. A similar finding in contexts other than VR is reported in prior studies [180, 307].

Lastly, we present the impact of immersiveness. Figure 4.6 shows the probabilistic distribution of gaze vergence distance (GVD) subject to various immersiveness levels. Specifically, GVD is defined as the distance between the viewer's eyes and the focused object on display. For low immersiveness, the GVD is more concentrated, while it becomes scattered under higher perceived immersion. It indicates that a viewer's visual attention follows objects of interest that may cover a wide range on the sphere under good immersiveness; it tends to stay in the center of the scene as the perception becomes less satisfactory.

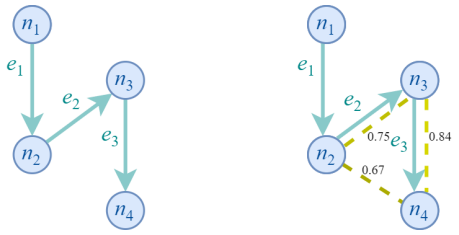
Observation 3: Eye-based patterns are consistent in multiple trials. In the measurement study, we play the same video of the same quality a couple of times to the same subjects and analyze changes in their ocular behaviors. Two trials, as indicated in the first and fourth column of Figure 4.2, are randomly selected. It is observed that ocular patterns, including but not limited to, the spatial distribution of gazes, GDC, and gaze velocity, are quite similar to each other. This observation implies that our QoE assessment model, once well trained on existing ocular behaviors, can be reused over time.

Summary. Our measurement study lays the necessary foundation for the idea of leveraging ocular behaviors to infer the subject's QoE in watching 360-degree videos. The findings are encouraging. First of all, we verify the hypothesis that there are strong correlations between viewers' ocular behaviors and their perceived experience in watching 360-degree videos. Second, ocular behaviors can effectively reflect both objective (e.g., video quality) and subjective (e.g., cybersickness, immersiveness, and fatigue) impact factors of perceived QoE in VR. This property can be achieved neither by the existing video-centric models [52, 88, 252, 290, 303, 306, 313] nor the visual attention enhanced models [150, 151, 291]. Nonetheless, how to perform an accurate QoE assessment based on collected ocular behaviors is a non-trivial task, which is also the focus of Section 4.5 and 4.6 next.

4.5 Modeling Ocular Behaviors into Graphs

The "node-edge" structure of subject's ocular behavior data shown in Figure 4.1 motivates us to transform them into graphs. In the following, we first introduce a basic version that only captures the temporal structure of ocular behaviors, followed by a comprehensive version that further explores content dependencies out of the behaviors.

A basic version. Consider a time series of gazes captured by a VR headset. They form N fixations ($\boldsymbol{N}$) and thus $N-1$ saccades ($\boldsymbol{E}$). The corresponding basic graph is of N nodes and $N-1$ edges. We denote the graph as $G = \{\boldsymbol{N}, \boldsymbol{E}\}$, where $\boldsymbol{N} = \{n_1, ..., n_N\}$ and $\boldsymbol{E} = \{e_1, ..., e_{N-1}\}$. Each saccade $e_k \in \boldsymbol{E}$ links two fixations $n_k,\ n_{k+1} \in \boldsymbol{N}$. Saccades are directional as they present the chronological order from a fixation to its successor in the temporal domain. As depicted in Figure 4.1, inside each fixation, there are many gazes. Typically, these gazes reflect the subject's visual attention to the same object of interest in the same scene. Their time-stamped coordinates then serve as part of attributes of the



A	n_1	e_1	n_2	e_2	n_3	e_3	n_4
n_1	-	1	0	0	0	0	0
e_1	0	-	1	0	0	0	0
n_2	0	0	-	1	0.75	0	0.67
e_2	0	0	0	-	1	0	0
n_3	0	0	0.75	0	-	1	0.84
e_3	0	0	0	0	0	-	1
n_4	0	0	0.67	0	0.84	0	-

(a) Basic graph (b) Comprehensive graph (c) Adjacency matrix

Figure 4.7: (a) An illustration of a basic graph, where circles and arrows denote fixations and saccades, respectively. (b) An illustration of a comprehensive graph. Dashed lines represent newly added edges. (c) The adjacency matrix corresponding to the comprehensive graph.

fixation (i.e., node). Additionally, correlations are observed between the subject's pupillary and oculomotor behaviors and perceived video quality as elaborated in Section 4.4. Hence, time-series pupil sizes and time instances of eyelids open/close events (i.e., blink onsets/offsets) are also treated as part of fixation's attributes. For a saccade, its attributes are similar to those of fixations, including coordinates of gazes of that saccade and time-series pupillary and oculomotor features described above.

A comprehensive version. The basic graph only captures local pairwise relationships between the temporally adjacent fixations and saccades; a fixation (saccade) is connected to two adjacent saccades (fixations). Relationships in other domains remain unexplored. In practice, two fixations, even not directly connected by a saccade, may share high similarities in their attributes. We find in our measurement studies that these similar fixations are typically associated with the same object in a video. For example, ocular behaviors when a viewer focusing on a tree are distinct from tracking a flying bird [31, 148, 173]. Thus, we develop a comprehensive graph that preserves both the temporal and content-dependent information in the collected raw data. The comprehensive version creates additional edges between fixations of high similarities on the basic graph. As shown

in Figure 4.7(b), 3 new edges (indicated by bidirectional dashed lines) are added. Note that new edges do not have any attribute.

Now the remaining question is how to determine the "similarity" of two given fixations. In this work, we employ the *cosine similarity*, a common measure of similarity between two non-zero vectors. Specifically, the similarity score between two fixations n_i and n_j is calculated as $\theta(n_i, n_j) = (n_i \cdot n_j)/(||n_i||\,||n_j||)$. For expression simplicity, here we use the node index n_i to represent its attribute vector. Given a pre-defined threshold θ_0, an edge is added between n_i and n_j if $\theta(n_i, n_j) > \theta_0$. $\theta(n_i, n_j)$ is then treated as the weight of the new edge. The comprehensive graph is thus a weighted graph. It is possible that attributes of fixations and saccades are of unequal size. To facilitate the learning graphs with unequal attribute size, we employ an *encoding process* that transforms arbitrary-length attributes into fixed-length vectors before passing them into the learning model [254].

4.6 EyeQoE

In the following, we first present a basic QoE assessment model that learns from the graph-structured ocular behaviors. We realize that the intrinsic heterogeneity of human visual behaviors and the impact of diverse video contents introduce variations to the learning process. In addition, the assessment model, trained on existing video samples, may not be readily applicable to new unseen videos. Thus, the basic model is further extended to deal with these issues.

4.6.1 A Basic GCN-based QoE Assessment Model

We propose to use GCN neural networks to solve our learning-on-graph problem. GCN is capable of extracting the representation of non-Euclidean graphs using a "convolutional" (neighbor-weight-sharing) kernel [310]. Like other neural networks, a GCN model consists of several layers of neurons; in each layer, higher-level features are ex-

tracted from the input and passed onto the next layer. A GCN model can be designed to classify nodes, subgraphs, or even entire graphs. Aside from GCN, graph neural network (GNN) [71, 153, 230] is another feasible model in handling non-Euclidean characteristics of the complex structure of graphs. We pick the former over the latter due to its efficiency in running backpropagation over time.

Construction of adjacency matrix. We formulate our problem as a graph classification problem, where the classifier takes the comprehensive graph (generated in Section 4.5) as the input and outputs a QoE score on the scale of 1-5. The input consists of an *attribute matrix* and an *adjacency matrix*. Specifically, an attribute matrix is denoted as $X \in \mathbb{R}^{(2N-1)\times D}$, where $2N-1$ comes from N fixations and $N-1$ saccades, and D is the dimension of their attributes after encoding. Each row is the encoded attributes from a fixation/saccade. An adjacency matrix is denoted as $A \in \mathbb{R}^{(2N-1)\times(2N-1)}$, where each row and column corresponds to a fixation or a saccade. The entries of the matrix indicate whether pairs of elements are adjacent or not in the graph. Take Figure 4.7 as an illustration. Since n_1 is linked to e_1, then $A_{1,2} = 1$. On the other hand, $A_{2,1} = 0$ as e_1 is a directional edge. Assume $\theta_0 = 0.5$. For two fixations n_2 and n_4, their corresponding matrix entries are given by their similarity score: $A_{3,7} = A_{7,3} = \theta(n_2, n_4) = 0.67$ as $\theta(n_2, n_4) > \theta_0$. Denote by v_i a node or an edge, the instantiation rule of the adjacency matrix is summarized as

$$A_{i,j} \in A = \begin{cases} 1 & \text{if } v_j \text{ is the successor of } v_i \text{ in the basic graph,} \\ \theta(v_i, v_j) & \text{if } v_i, v_j \in \boldsymbol{N} \text{ and } \theta(v_i, v_j) > \theta_0, \\ 0 & \text{otherwise.} \end{cases}$$

GCN-based model. Our GCN classifier consists of four convolutional layers followed by a max pooling layer [138]; each layer in this classifier can be written as a non-linear function

$$H^{l+1} = f(H^l, A)$$

where $H^l \in \mathbb{R}^{(2N-1)\times D}$ is the matrix of activations in the lth layer with $H^0 = X$. The model is specified by the $f(\cdot,\cdot)$ function of each layer. We adopt the propagation rule introduced in [138]

$$f(H^l, A) = \rho(\hat{\blacksquare}^{-1}\hat{A}H^l W^l) \tag{4.1}$$

where $\hat{A} = A + I$ with I being the identity matrix. $\hat{\blacksquare}$ is the diagonal node dimension matrix of $\hat{A}$, and $W^l \in \mathbb{R}^{(2N-1)\times D}$ is the weight matrix for the l-th layer. ρ is an activation function, e.g., a ReLU $\rho(x) = \max(0,x)$.

The "convolution" operation in Equation (4.1) is designed in a way such that a "one-hop" filter runs over every fixation and saccade and aggregates its layer-wise representation with those of its neighbors. Specifically, for each fixation, the filter adds to it the representations of all other fixations, weighted by their similarity scores, and the representation of its neighboring saccade. For each saccade, since it only has one predecessor fixation as its neighbor, it is only updated by taking the representation of that fixation. Then, the aggregated representation is normalized by dividing with the dimension of the representations. One can incorporate higher-order neighborhoods information by stacking multiple GCN layers. Then features are aggregated and propagated iteratively along with the graph. In the final step, the output from the last layer is passed through a max-pooling layer to generate the classification result z as the estimated QoE score for the given 360-degree video.

We adopt the mean squared error as the loss function:

$$\mathscr{L}_G = \frac{1}{N}\sum_{i=0}^{N}(y_i - z_i)^2 \tag{4.2}$$

where y_i and z_i denote the ground-truth label and the model prediction of the ith sample, respectively, and N stands for the number of training samples.

Table 4.1: Training sample selection rule. ✓: the pair is selected; ✗: the pair is not.

		Same subject	Different subjects
Same label	Same video	✗	✓
	Different videos	✓	✓
Different labels	Same video	✓	✗
	Different videos	✗	✗

4.6.2 Dealing with Subjects and Visual Stimuli Heterogeneity

In practice, the training dataset, i.e., labeled ocular behaviors, is obtained from a group of subjects for watching various 360-degree videos. In addition to objective and subjective impact factors of perceived QoE (as discussed in Section 4.4), the subjects and visual stimuli heterogeneity also affects ocular behaviors. As a result, it introduces an additional dimension of uncertainty to the learning process.

Compared with video quality, QoE should be much less relevant to the video content. It means that two videos are expected to produce similar QoE scores given the same quality and other subjective impact factors (e.g., cybersickness, fatigue, immersiveness, etc.), regardless of the contents displayed. In the meantime, video contents highly affect ocular behaviors, the features considered by EyeQoE for QoE assessment. For example, eyes move faster when watching high-motion scenes than the stationary ones. As one of our contributions, this work aims to eliminate the impact of video contents to QoE assessment, as called *visual stimuli heterogeneity*. To alleviate impacts from both subjects and visual stimuli heterogeneity, we modify the basic GCN-based QoE assessment model by applying the Siamese network [57]. Its idea is to employ a pair of substructures with the same architecture and weights. It passes a pair of input data through the two substructures separately, computes the distance metric between the outputs, and updates both substructures simultaneously.

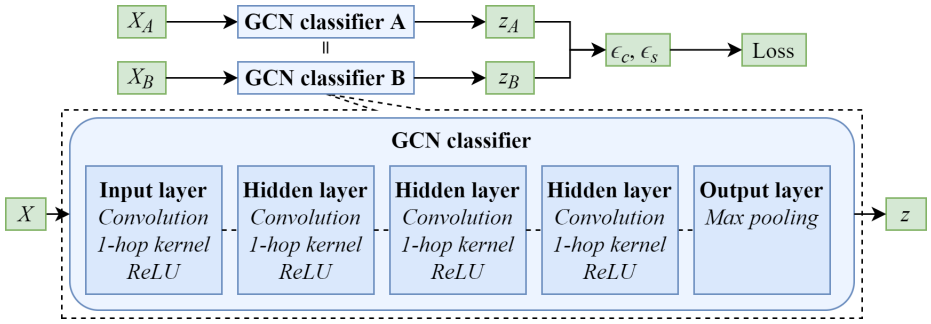


Figure 4.8: Top: the architecture of the Siamese network. Bottom: the GCN classifier model.

The modified model is shown in Figure 4.8. It is composed of two identical GCN classifiers introduced in Section 4.6.1. X_A and X_B stand for the pair of training samples for the two classifiers, respectively. Sample pairs are carefully selected following a scheme as outlined in Table 4.1. Each pair of samples is classified into one of the four categories based on their subjects, video contents, and labels. If their labels are the same, we select the pairs from different subjects and/or video contents. In this way, the model can learn to tolerate differences in ocular behaviors caused by heterogeneous subjects and video contents, i.e., visual stimuli. In contrast, if their labels are different, we select the pairs from the same subjects and video contents; the model then learns to distinguish samples of similar patterns associated with different labels (i.e., QoE scores). The selected sample pairs are passed through the two twin models separately. We then calculate the distance between two outputs. The loss function of the Siamese network is defined as

$$\mathscr{L}_S = \sum_{i=1}^{N} \left(\alpha \left(\eta \varepsilon_c^2 + (1-\eta)(4-\varepsilon_c)^2 \right) + (1-\alpha) \left(\eta \varepsilon_s^2 + (1-\eta)(4-\varepsilon_s)^2 \right) \right) \tag{4.3}$$

where N is the number of sample pairs. ε_c and $\varepsilon_s \in [0,4]$ denote the distances between model outputs of a sample pair concerning visual stimuli and subjects, respectively. η is a binary value indicating whether labels of a sample pair are the same ($\eta = 1$) or different ($\eta = 0$). α is a factor to balance the weight between ε_c and ε_s.

Combining (4.2) and (4.3), the final loss function is expressed as

$$\mathscr{L} = \mathscr{L}_G + \mathscr{L}_S + \lambda \cdot ||w||_2^2 \tag{4.4}$$

where $\lambda \cdot ||w||_2^2$ serves as a regularization term. In the training process, the final loss $\mathscr{L}$ is fed back into the network to update the weights.

4.6.3 Dealing with Unseen Videos

As discussed, the characteristics of the video scenery being displayed also impact the viewer's ocular behaviors. Hence, the QoE assessment model, trained over existing video clips, may not be readily scalable to an even broader set of unseen videos, especially of different characteristics. A conventional approach is to gather as many annotated samples as possible to train the model. In our case, it requires covering videos of all kinds, which would incur prohibitively expensive overhead in data collection. Alternatively, we propose to employ *domain adaptation* [214]. Under this framework, existing videos and new videos are treated as the *source domain* and the *target domain*, respectively. The domain adaptation technique aims to fine-tune parameters of models trained in the source domain to adapt to new circumstances in the target domain. While this technique has been widely adopted in the context of computer vision [65], sentiment analysis [203, 257], and action recognition [56, 182], whether it is effective in 360-degree video QoE assessment is unexplored yet.

Video type categorization. To facilitate the employment of domain adaptation, we first categorize all 360-degree videos in various types[2] according to their *colorfulness*, *luminance*, and *motion*. Existing methods are available to obtain the above information by

[2]In this work, we assume that each 360-degree video clip is of one type without significant scene changes. For longer videos in multiple scenes, they can be divided into multiple segments, each in one scene. We then apply our model to each segment sequentially. The final QoE can be calculated as the aggregated QoEs of all segments.

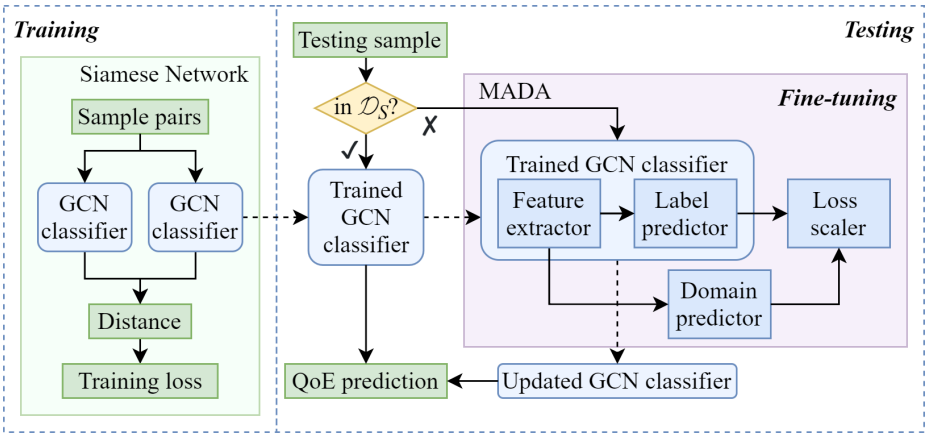


Figure 4.9: The overall architecture of EyeQoE.

inspecting *I-frames* and *P-frames* in videos. As these computations do not involve any sophisticated operations, they can be accomplished within dozens of milliseconds in a computer with moderate settings. Assume that the entire video space is divided into κ types. κ plays an important role in the performance of EyeQoE. We will examine its value selection in Section 4.7.3.

Domain Adaptation Each video type is treated as a domain. Assume that the training videos cover K ($K < \kappa$) domains $\mathscr{D}_S = \{\mathscr{D}_S^1, ..., \mathscr{D}_S^K\}$. The target domain that an unseen video falls into is denoted as $\mathscr{D}_T$. We propose a multi-source adversarial domain adaptation (MADA) network. It is inspired by the classic domain adaptation network introduced in [96] but further extended to scenarios of multiple source domains as in this work. As a note, the classic domain adaptation network is originally designed to deal with single-source-domain scenarios and thus not readily applicable here.

The architecture of MADA is illustrated in Figure 4.9. To fine-tune the trained GCN classifier, MADA takes as inputs the samples from a specific target domain $\mathscr{D}_T$ and a set of samples from each source domain $\mathscr{D}_S^k$ ($k \in [1, K]$). MADA is constructed based on the GCN classifier with four main modules: feature extractor, label predictor, domain predictor, and loss scaler. The feature extractor, together with the label predictor, assemble the same

components of the GCN classifier introduced above. Specifically, the feature extractor is comprised of the first four convolutional layers of GCN. The label predictor is simply the output layer, i.e., the max-pooling layer (Figure 4.8). Given any graph presentation X and A, the above two modules generate a prediction label z. The domain predictor works in an adversarial way. With the high-dimensional features as the input, it aims to decide if the given graph presentation belongs to a source domain or a target domain. Ideally, the domain predictor, once properly trained, cannot distinguish between them. It indicates that our model's inference performances over existing videos and unseen videos are almost the same. The loss scaler computes the loss of label prediction and domain prediction and aggregates them into the final loss value $\mathscr{L}_{tot}$

$$\mathscr{L}_{tot} = \sum_{k=1}^{K} \left(\frac{\gamma^k}{n^k} \sum_{i=1}^{n^k} \mathscr{L}_y^{k,i} \right) - \lambda \left(\sum_{k=1}^{K} \left(\frac{\gamma^k}{n^k} \sum_{i=1}^{n^k} \mathscr{L}_d^{k,i} \right) + \frac{1}{n^{K+1}} \sum_{i=1}^{n^{K+1}} \mathscr{L}_d^{K+1,i} \right).$$

Here $\mathscr{L}_y^{k,i}$ and $\mathscr{L}_d^{k,i}$ ($k \in [1,K]$) stand for the label prediction loss and the domain prediction loss over a sample from source domain $\mathscr{D}_S^k$. $\mathscr{L}_d^{K+1,i}$ stands for the domain prediction loss over a sample from the target domain $\mathscr{D}_T$. n^k and n^{K+1} represent the number of samples of $\mathscr{D}_S^k$ and $\mathscr{D}_T$, respectively. K is the total number of source domains. λ is a parameter that controls the balance between label prediction loss and domain prediction loss. γ^k stands for the similarity between $\mathscr{D}_S^k$ and $\mathscr{D}_T$. It is calculated as the *cosine similarity* between content metrics of videos from these two domains. The content metrics include colorfulness, luminance, and motion as mentioned above. Basically, two domains that share a higher similarity in their videos tend to exhibit similar prediction performance through a trained model. Hence, the prediction loss of each source domain contributes to the total loss with a different weight determined by γ^k: A source domain of a larger γ^k has a more prominent impact.

In the MADA network, the GCN classifier is initiated with parameters derived from the offline training phase, whereas parameters of the domain predictor are set as random

values. MADA is triggered with the arrival of an unseen video out of the source domains. The trained GCN classifier is then fine-tuned through multiple rounds of iterations, where backpropagation is performed and all weights are updated through the gradient descent algorithm. We will examine in Section 4.7.3 with details regarding the efficiency of the fine-tuning process.

4.6.4 Piecing All Together

Figure 4.9 outlines the overall architecture of EyeQoE. The core component is a GCN classifier designed to infer the QoE score given the subject's graph-structured ocular behaviors. To handle the issue of subjects and visual stimuli heterogeneity, we enhance our GCN classifier with a Siamese network which consists of two identical GCN classifiers. Sample pairs are carefully selected and used to train the classifier. Almeida-Pineda algorithm [198], a gradient-based optimization method, is adopted. In the testing stage, given a new sample, EyeQoE first examines if it belongs to any of the source domains. If yes, it indicates that the corresponding video type has been covered during training. Hence, the trained GCN classifier is applied directly for QoE inference. Otherwise, the video is deemed from the target domain. Then our proposed MADA is applied to fine-tune the GCN classifier with the new sample. Finally, the QoE is derived by feeding the sample into the updated classifier.

4.7 Evaluation

4.7.1 Settings

Experiment setup. We implement EyeQoE on a PC running Windows 10 operating system. It is equipped with an Intel Core i7-7820X processor and GeForce RTX 2080 graphic cards. An HTC Vive Pro VR headset is used to provide the VR environment and render videos to subjects. A Pupil Labs eye tracker is embedded inside the VR headset to

capture subjects' eye movements. The VR headset is connected to the PC via a USB cable. EyeQoE is implemented using the Keras 2.3.0 library built on top of the TensorFlow 2.0 framework. The Adam optimizer [136] is employed for optimizing the training process.

Dataset. All source videos are downloaded from two major platforms of 360-degree videos, YouTube and Vimeo. The original version is of 4K resolution and 25 fps frame rate. The videos cover a wide range of genres, such as nature, sports, and city view. To facilitate the experiment, each video is of a 25-second duration without significant scene changes. Each source video is subject to two types of distortions, including resolution degradation and stalling. For the former, we use the JM reference implementation of the H.264 scalable video codec (SVC) to compress the 4K original videos into lower resolutions such as 2K, 1080p and 720p. For the latter, we add freeze frames to simulate stalling in three different versions: 8 stalls each lasting 1 second, 4 stalls each lasting 2 seconds, and 2 stalls each lasting 4 seconds.

We have listed the source videos used for main evaluation in Table 4.2. These videos cover 5 different semantic types, namely Film & Animation (FA), Entertainment (En), Travel & Events (TE), Sports (S), and People & Blogs (PB), as indicated by the source websites. These videos span over 10 different categories, i.e., domains, out of the 27-domain space. These domains are separated by colorfulness, luminance, and motion based on the video content.

A data collection campaign is conducted over three months. 50 subjects are recruited. They are from a university in the United States, most of them are international students from multiple different countries and nations. Table 4.3 summarizes the demographic information of the participants. The diversity is observed in the gender, age, eye color, eye wear, and VR experience. They are asked to wear a VR headset to watch 360-degree videos of different qualities and give a score from 1 to 5 that best describes their experience after watching each video. Original, uncompressed reference videos are randomly placed

Table 4.2: Summary of the video set.

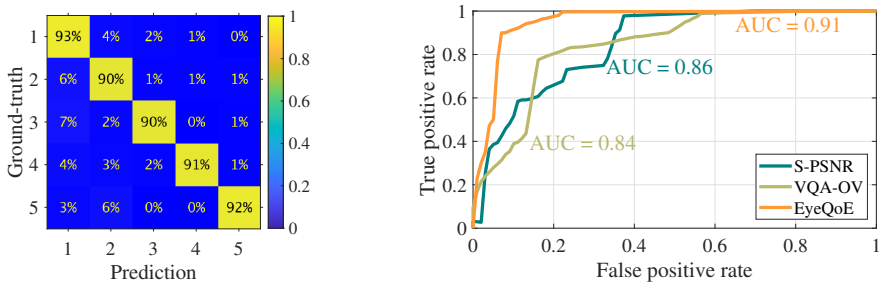
Video	Category	Color.	Lumin.	Motion	Projection	Source
Bar	PB	⋆⋆⋆	⋆	⋆⋆	ERP	Vimeo
Boat	FA	⋆⋆⋆	⋆⋆	⋆⋆	ERP	Vimeo
Bunnies	FA	⋆⋆⋆	⋆⋆	⋆	EAC	YouTube
City	TE	⋆⋆⋆	⋆⋆⋆	⋆	ERP	Vimeo
Dance	En	⋆⋆	⋆⋆⋆	⋆⋆	EAC	YouTube
Girl	PB	⋆⋆	⋆	⋆⋆⋆	ERP	Vimeo
Lions	En	⋆⋆⋆	⋆⋆⋆	⋆⋆	EAC	YouTube
Ski	S	⋆	⋆⋆⋆	⋆	ERP	Vimeo
Snowmobile	S	⋆	⋆⋆⋆	⋆⋆⋆	ERP	Vimeo
Waterfall	En	⋆⋆⋆	⋆	⋆	ERP	Vimeo

Table 4.3: Participant demographic information.

Gender	#	Age	#	Eye color	#	Eye wear type	#	Experience	#
Female	21	18-23	26	Brown	33	None	19	No	34
Male	28	24-29	13	Blue	6	Glasses	22	Yes	16
N/A	1	30-35	9	Hazel	3	Colorless lenses	7		
		$\geq$36	2	Other	8	Colored lenses	2		

amongst the set of videos shown, although the subjects are unaware of their presence. The score that subjects give these references is representative of the bias that the subject carries. By subtracting the reference video scores from those for the distorted videos, the biases are compensated for yielding differential scores for each distorted video. We divide the data collection into two separate sessions, each lasting no more than one hour, to avoid the discomforts caused by watching the immersive videos too long. The interval between two sessions is at least 24 hours. We further implement a UI via Unity, the most widely used VR development platform, to facilitate the data collection.

We did a literature review over the existing open-sourced datasets. As none of them meets our need, we decided to collect our own dataset. We have now publicized it on `https://github.com/MobiSec-CSE-UTA/EyeQoE_Dataset.git`.



(a) Confusion matrix (b) ROC curve

Figure 4.10: Overall performances of EyeQoE and the comparison with two existing QoE models. (a) Confusion matrix of EyeQoE's predictions. (b) ROC curves of different approaches.

4.7.2 Overall Performance

Figure 4.10(a) exhibits the confusion matrix of EyeQoE's prediction results. Rows represent the ground truth from 1 to 5, whereas columns represent the prediction results. Values on the diagonal are the success rate, i.e., the percentage of predicted results that EyeQoE gets right. The result is promising as the success rate is above 90% for all QoE values. Besides, we observe that EyeQoE achieves slightly better performance when predicting low and high QoE scores (1 and 5). It may be attributed to the fact that users generally perform well in distinguishing between the best- and worst-quality videos, while the boundaries for the medium ones tend to be vague in labeling.

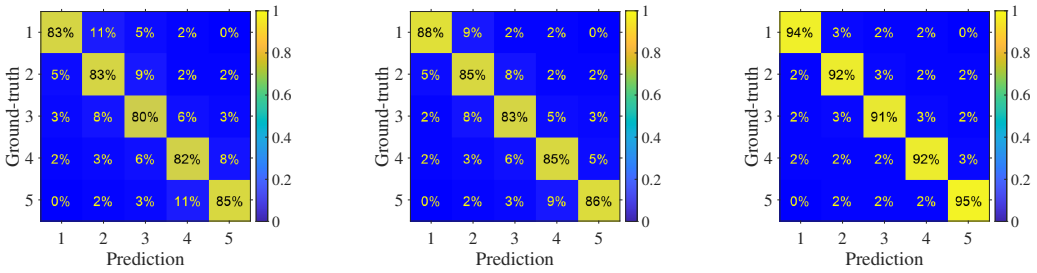
Comparison with state-of-the-art. We compare the performance of EyeQoE with two state-of-the-art solutions for 360-video QoE assessment: S-PSNR [303] and VQA-OV [150]. S-PSNR is a video-centric model; it is built upon the classic PSNR model but further takes into account the pixel distortion issue in projection. VQA-OV belongs to the human factor incorporated model; its main idea is to assign weights on the pixel-wise distortion in calculating the PSNR, where the weights reflect the subject's visual attention on the video.

The ROC curve for each model is depicted in Figure 4.10(b). It is a classic metric to see how a model balances between true positives and false positives. Ideally, the model is expected to have a steep ROC curve to deliver an accurate inference. Clearly, EyeQoE outperforms the other two with the largest AUC (area under the curve) of 0.91. In comparison, those for S-PSNR and VQA-OV are merely 0.86 and 0.84, respectively. S-PSNR and VQA-OV fail to counter critical factors, such as cybersickness, immersiveness, and fatigue, in QoE assessment. In contrast, rather than exhaustively enumerating and considering all possible impact factors for QoE assessment, EyeQoE leverages ocular behaviors as an indicator to reveal the subject's perceived QoE.

Advantage of GCN-based model in QoE assessment. We further compare the accuracy performance between the GCN + Siamese network and prior works, S-PSNR and VQA-OV. Particularly, the GCN + Siamese network is an ablation version of EyeQoE by removing the domain adaptation component. Since none of the above models include domain adaptation, the performance should demonstrate the superiority of our GCN-based design. Figure 4.11 shows the confusion matrices produced by each approach. Apparently, GCN + Siamese yields the best performance among the three. Its diagonal line has larger values, meaning more accurate assessments are produced. For all QoE values, GCN + Siamese maintains a success rate above 91%, whereas the S-PSNR and VQA-OV acquire much lower success rates, ranging from 80% to 88%.

The reasons that the proposed model outperforms S-PSNR and VQA-OV can be summarized as follows. First, our method leverages ocular behaviors, which are neglected by state-of-the-art designs; these behaviors offer valuable information of a user's QoE as validated in Section 4.4. Second, by applying GCN on graphs formed by fixations and saccades, we are able to exploit the temporal dependency and content dependency of the ocular behaviors by inspecting temporal adjacent and similar activities. The "node-link" structure of the irregular non-Euclidean graphs implies that only graph learning techniques

are suitable to explore these dependencies. Third, the Siamese network used during training automatically extracts the most relevant features and eliminates subjects and visual stimuli heterogeneity.

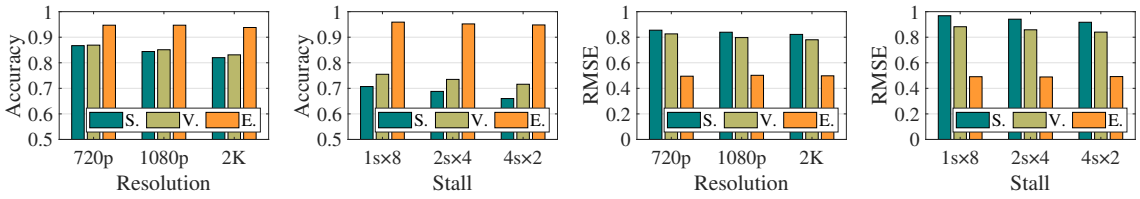


(a) S-PSNR (b) VQA-OV (c) GCN + Siamese only

Figure 4.11: Advantage of GCN-based model - confusion matrices.

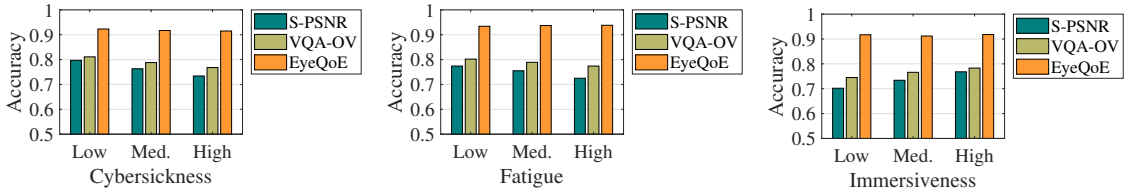
Performance over videos of different distortions. We further investigate the efficacy of EyeQoE over 360-degree videos of various distortions in Figure 4.12. Two kinds of distortions are examined, resolution and stalls. Figure 4.12(a) shows the assessment accuracy by varying the resolution from 720p to 2K. The accuracy of EyeQoE is all above 0.928. Besides, the performance variance under different settings is almost unnoticeable. This is the same case in Figure 4.12(b)-4.12(d). Hence, EyeQoE delivers consistent performance for videos of various distortions. EyeQoE outperforms the other two schemes in all cases, especially the stalling distortion. Recall that S-PSNR and VQA-OV measure video QoE through pixel distortions and are thus incapable of reflecting video quality degradation caused by stalling events.

Impact of subjective factors. Now we evaluate EyeQoE's performance subject to cybersickness, fatigue, and immersiveness. Results are illustrated in Figure 4.13. EyeQoE exhibits high accuracy across various conditions. It implies that ocular behaviors serve as effective indicators of viewer's perceived QoE. Besides, EyeQoE outperforms the other



(a) Accuracy vs resolutions (b) Accuracy vs stalls (c) RMSE vs resolutions (d) RMSE vs stalls

Figure 4.12: Impact of distortion types on prediction performances. S: S-PSNR; V: VQA-OV; E: EyeQoE.



(a) Accuracy vs cybersickness (b) Accuracy vs fatigue (c) Accuracy vs immersiveness

Figure 4.13: Impact of cybersickness, fatigue, and immersiveness on prediction performances.

two models, S-PSNR and VQA-OV, by a clear margin. As discussed, neither S-PSNR nor VQA-OV considers the above subjective factors in QoE modeling. It also explains why their performances become even worse under a high level of cybersickness, fatigue, and immersiveness.

Handling longer videos. EyeQoE is designed in the following way to accommodate longer videos. First, if a video contains multiple scenes, it is divided into several segments, each having one scene. In this way, we obtain S segments of the target video. Then, the subject's ocular behaviors during each segment are structured as one graph and fed into the trained model. The QoE for that segment is thus derived. To aggregate the QoE's from S segments, previous works apply either uniform averaging (e.g., [265]) or weighted

Table 4.4: Performance on long videos.

Video	Duration (min)	Scene rate	Seg. count	Accuracy	RMSE
City view	4:00	10.00	8	0.88	0.77
Coaster	5:32	0.72	12	0.93	0.39
Crime scene	22:24	0.76	48	0.90	0.32
Haydee	2:01	2.98	6	0.90	0.79
Viking village	2:09	0.47	4	0.95	0.45

averaging (e.g., [77, 293]). EyeQoE follows the latter one, where the overall QoE of the video is a weighted average of the QoE for each segment as follows:

$$Q_{total} = \frac{\sum_{i=1}^{S} w(i)Q_i}{\sum_{i=1}^{S} w(i)} \tag{4.5}$$

where Q_i is the QoE output for the i-th segment and $w(\cdot)$ stands for the weight determined by the segment duration and the subject's memory factor. The rationale behind the second design is that a subject's perceived experience over segments rendered later contributes more to the overall QoE [24, 77]. In this way, the temporal dependencies are preserved within each segment.

To further evaluate EyeQoE's performance on longer videos with frequent scene changes, we use 5 long 360-degree videos from YouTube. Table 4.4 lists the duration and scene rate (number of scenes per minute) of these videos as well as the corresponding performance of EyeQoE. We observe that the video duration does not affect much on EyeQoE's performance. However, as the scene rate increases, the overall performance experiences slight degradation with lower accuracy and higher RMSE. Since a long video is divided into multiple segments each with one scene, a higher scene rate thus leads to segments with shorter duration. Hence, the number of features extracted would be reduced, which in turn affects the performance of EyeQoE.

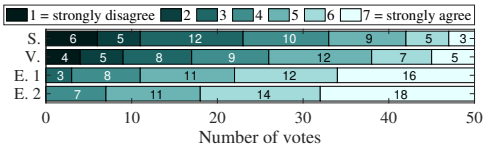
Performance over different video categories. To determine whether EyeQoE achieves consistent performance on different video categories, we look into our experimental results

Table 4.5: EyeQoE's performance on different video categories.

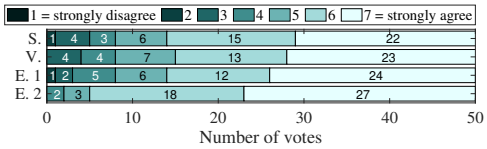
Domain	HHL	HML	LHL	...	LLH	MLM	HLH
Accuracy	0.94	0.93	0.93	...	0.92	0.91	0.91
RMSE	0.49	0.50	0.50	...	0.54	0.57	0.57

and carry out a comprehensive analysis. We list EyeQoE's performance on 3 best and 3 worst domains in Table 4.5. Specifically, each domain is represented by three letters indicating the level of colorfulness, luminance, and motion, respectively (L = low, M = medium, H = high). Overall, the performance remains similar among different categories (i.e., domains in our work). This indicates that EyeQoE does not show significant bias on domains. In the meantime, we observe that EyeQoE performs slightly better on videos with higher luminance and lower motion. This domain largely correlates with nature and sightseeing videos with more static scenes. On the opposite, videos with lower luminance and higher motion, such as gaming and action scenes, tend to produce slightly lower QoE assessment accuracy. One possible reason is that subject's gazes may tend to become less focused under frequent scene changes and a dark view. Consequently, it would slightly impact the ocular behavior features. Still, the accuracy is as high as 0.91 which is satisfactory.

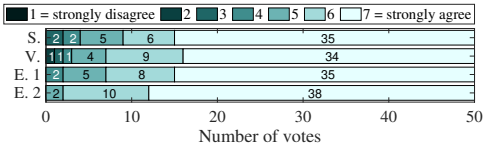
Subjective Survey In the survey, subjects are asked to rate EyeQoE from different perspectives on a 7-point Likert scale (1 = strongly disagree; 7 = strongly agree) as how much they agree with the following statements: S1. The result is accurate and meets my perceived QoE score; S2. I feel physically comfortable during the experiment, without dizziness or sore eyes caused by this model; S3. This model does not interrupt or distract my experience of the video watching; S4. I would like to have this model to rate QoE scores for me for practical use. Survey results are shown in Figure 4.14. Most subjects rated 5 or higher scores for all statements for EyeQoE before and after the experiment. This suggests that EyeQoE is well perceived by users. Particularly, more than half of the



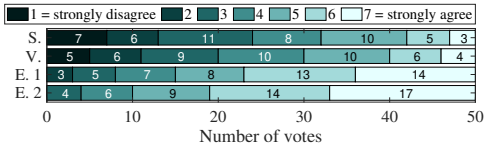
(a) S1. The result is accurate and meets my perceived QoE score



(b) S2. I feel physically comfortable without dizziness or sore eyes caused by this model



(c) S3. This model does not interrupt or distract my experience of the video watching



(d) S4. I would like to have this model to rate QoE scores for me for practical use

Figure 4.14: Survey results (S. = S-PSNR, V. = VQA-OV, E. 1 = EyeQoE before the experiment. E. 2 = EyeQoE after the experiment).

subjects rated the highest score for S2 and S3, indicating that EyeQoE is comfortable to use and does not interfere with normal sessions. Compared to state-of-the-art approaches, EyeQoE receives much higher scores for S1 and S4, suggesting that EyeQoE more accurately reflects subjects' perceived QoE and is preferred for practical use. It is also worth noting that many subjects rate EyeQoE a higher score after the experiment than before it, which demonstrates that EyeQoE outperforms user's expectations.

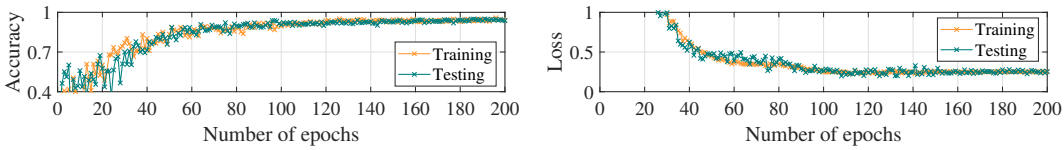
4.7.3 Micro Benchmarks

Impact of the training ratio. The impact of the training ratio on the performances of EyeQoE is analyzed. As presented in Table 4.6, the performance is enhanced steadily as the size of the training dataset increases. It indicates that EyeQoE has robust data scalability. Meanwhile, the performance improvement becomes marginal as the ratio surpasses 60%.

Impact of training epochs. To determine whether the model has been trained properly, we monitor the training process in Figure 4.15. Figure 4.15(a) shows the accuracy

Table 4.6: EyeQoE's performance regarding the training ratio.

Training ratio (%)	10	20	30	40	50	60	70	80
Accuracy	0.71	0.83	0.85	0.86	0.89	0.93	0.93	0.93
RMSE	1.07	0.90	0.77	0.61	0.60	0.51	0.52	0.50



(a) Accuracy vs number of epochs

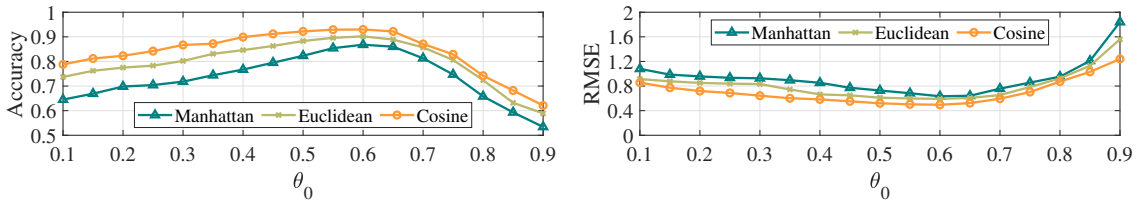
(b) Loss vs number of epochs

Figure 4.15: Training and testing performance over different number of epochs.

with respect to the number of epochs. Note that one epoch is when an entire training dataset is passed both forward and backward through the model once. The accuracy quickly increases to 0.90 and becomes converged after around 80 epochs. Figure 4.15(b) plots the loss value, another indicator of whether the model is properly trained. It is considered as the "price" paid for assessment inaccuracy. As shown, loss tends to be stable after 100 epochs. Combining the results above, it is sufficient to set 100 epochs for training in our case.

Impact of graph construction metrics. Now we evaluate the performance of EyeQoE given different graph construction metrics. To construct a comprehensive graph, similarity is computed between any two fixations to decide if an edge is added. We employ three different similarity metrics: Manhattan similarity, Euclidean similarity, and cosine similarity. They are classic metrics widely adopted for graph modeling [41]. We also examine the impact of threshold θ_0. Recall that an edge is added if $\theta > \theta_0$. As shown in Figure 4.16, cosine similarity leads to the best overall performance among the three similarity metrics. We also find that EyeQoE achieves its best performance with accuracy = 0.93 and RMSE =

0.50 at $\theta_0 = 0.6$. Basically, a too-large value of θ_0 would fail to exploit content-dependency between fixations, while a too-small value would introduce unnecessary noise to learning.



(a) Accuracy vs θ_0 (b) RMSE vs θ_0

Figure 4.16: Impact of similarity metrics and θ_0 on prediction performances.

Performance of domain adaptation. Next we evaluate the performance of EyeQoE on domain adaptation. The fine-tuning process is executed via the proposed MADA with the arrival of an unseen video. The impact of domain space κ is examined. Recall that κ represents the total number of domains, i.e., video types under consideration. In the experiment, three values are adopted $\kappa \in \{8, 27, 64\}$. They are derived by dividing the space of video content metrics, i.e., colorfulness/luminance/motion, into 2, 3, and 4 levels, respectively ($\{8, 27, 64\} = \{2^3, 3^3, 4^3\}$). In the setting, n_T is equal to 0, 5, 10, and 15. Particularly, $n_T = 0$ means the trained model (over existing samples) is directly applied to an unseen video, while $n_T = 5$ means 5 samples in the target domain are used to fine-tune the model. For each n_T value, the same number of samples are randomly picked from each source domain to form the inputs alongside the target domain samples. For comparison, we also test the prediction accuracy on source domains, denoted as D_S in Figure 5.7. This means that the new video belongs to a source domain, and the trained GCN classifier is directly applied without using MADA.

As demonstrated in Figure 5.7, the best overall performance is achieved when $\kappa = 27$ among the three values. In general, a too coarse categorization, i.e., small κ, would fail to

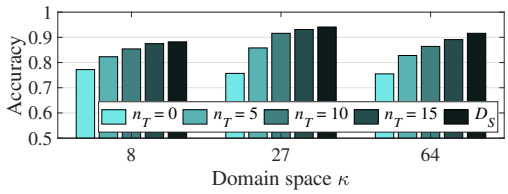
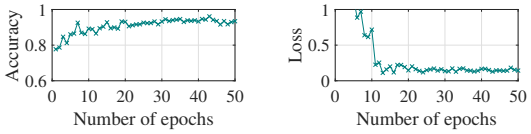
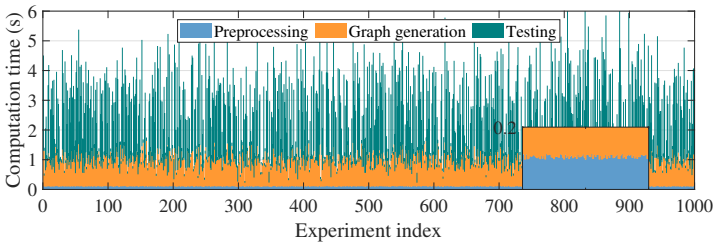


Figure 4.17: Performance of domain adaptation with different configurations.

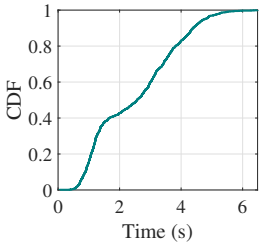


(a) Accuracy vs epochs (b) Loss vs epochs

Figure 4.18: Convergence of MADA with respect to training epochs.



(a) Stacked computation time



(b) CDF of total computation time

Figure 4.19: Computation latency of QoE prediction.

capture the uniqueness of each domain. On the other hand, too fine-grained categorization, i.e., a large κ, would reduce the number of samples in each domain and thus result in overfitting. Both cases affect the test accuracy. We also investigate the impact of n_T. A larger n_T is found to produce higher accuracy, since more samples allow the model to fine-tune its parameters in more rounds to better adapt to the target domain. Meanwhile, it also implies more videos from the same target domain to collect. Fortunately, the accuracy already reaches 0.92 with $n_T = 10$. We thus claim that EyeQoE can deliver satisfactory prediction performance for unforeseen videos within 10 samples of the same type. Figure 4.18 shows the fine-tuning process with respect to the number of epochs. Both the accuracy and the loss value become stable after about 20 epochs. It indicates that the domain adaptation can quickly converge.

Computation latency. We now examine the computation latency of QoE prediction over one video. All the operations include the preprocessing of ocular behaviors, graph generation, and testing (including MADA for domain adaptation). Figure 4.19(a) gives the stacked computation latency of each operation. Among the three, testing incurs the largest overhead, about 1.51 s on average. It is due to the fine-tuning for domain adaptation. The average latency for preprocessing and graph generation is 0.11 s and 0.88 s, respectively. Figure 4.19(b) further illustrates the CDF of the total computation latency of one QoE prediction. The average value is 2.5 s, with 90% of measurements lower than 4.2 s. It indicates that a subject's QoE score can be derived shortly, in a couple of seconds, after a 360-degree video is finished displaying. This duration is comparable to that from the prevalent QoE collection solution, in which users are asked to provide QoE scores manually; yet, EyeQoE is executed automatically without human involvement.

Impact of subject-dependent features on QoE assessment. Different subjects may be impacted in various ways. To investigate the significance and distinction of impact factors, we correlate several objective and subjective factors with the QoE scores from the collected data. Specifically, objective factors such as video resolution and stalling events are directly derived from the preprocessed videos, whereas subjective factors, including cybersickness, fatigue, and immersiveness, are collected during the experiments by confirming with the subjects about their corresponding subjective feelings. Figure 5.14 demonstrates the result, from which we make the following observations. First, among all the listed impact factors, stalling events and cybersickness are the most critical factors, as different cybersickness levels result in the most distinct QoE distributions, and that low QoE scores are induced whenever stalling events occur. Second, QoE scores highly concentrate with different levels of stalling events. For example, 88% of videos with 8 stalls are rated with QoE as 1; the variance of QoE scores is $\sigma^2 = 0.10$. Similarly, 64% and 76% of videos with 4 and 2 stalls are rated with QoE as 2 and 3 ($\sigma^2 = 0.34$ and 0.25), respectively. This means

that with the same levels of stalling events, more subjects perceive similar QoE, which indicates that this factor brings a common significance across various subjects. In contrast, immersiveness results in a relatively even distribution of QoE scores, suggesting that this factor is distinct across different users.

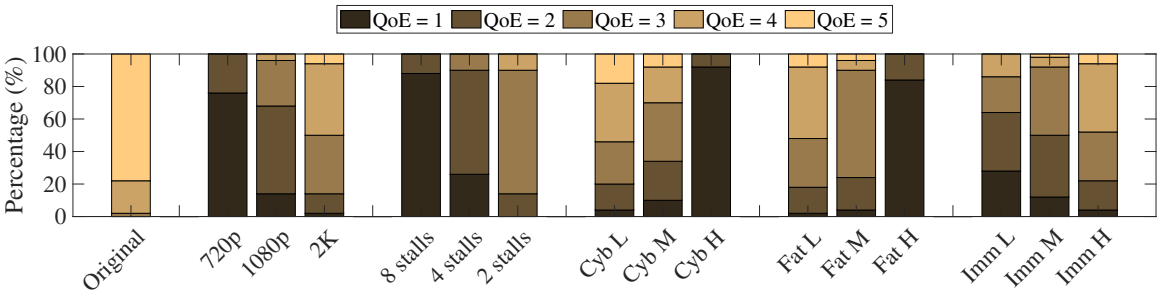


Figure 4.20: Correlation between impact factors and QoE.

4.8 Discussion and Future Work

In this section, we discuss several limitations of this work and present our future research directions.

Extra cost introduced by domain adaptation. Domain adaptation is activated only for unseen videos; that is, the process will be bypassed when the type of videos that are covered in the training process. Hence, no extra training cost is incurred. For unseen videos, domain adaptation does cause certain training cost. To quantify it, we have evaluated the time consumption for domain adaptation in the experiment. Figure 4.19(a) presents the stacked computation time of EyeQoE's all major processes, including preprocessing of ocular behaviors, graph generation, and testing. Specifically, testing is conducted over both seen and unseen videos. The latter includes the domain adaptation operation. We observe that the testing time ranges between 0.1 s and 5.1 s, among which larger values tend to associate with unseen videos due to the domain adaptation.

In current multimedia services, user's QoE is mainly obtained by asking people to rate their perceived quality via surveys or self-reports. However, such procedures are inconvenient and may even be annoying for the users. EyeQoE intends to automate the entire process by constructing a QoE assessment model. User's perceived QoE would be generated and collected automatically. In this sense, timing is not the main consideration of our design. Still, according to the above result, QoE assessment for unseen videos (including domain adaptation) can be done within 5.1 s, which is satisfactory for real-world implementation. Of course, it would be even more desirable if the latency can be further shortened. We plan to investigate this possibility in our future work.

Enhancing the prediction accuracy of EyeQoE. This work demonstrates the feasibility of using ocular behaviors for QoE assessment. While the overall accuracy performance is satisfactory, there is still room for improvement. To this end, we plan to pursue two potential directions. The first one is to combine EyeQoE with traditional objective quality of service (QoS) metrics such as bandwidth, latency, video quality, etc. Specifically, we will integrate the QoS metrics as new dimensions alongside the ocular behaviors as the inputs of our QoE model. The graph modeling will be revised accordingly with the introduction of additional inputs. The selection of QoS metrics will be carefully determined. They should be practical to collect at VR terminals and play positively in enhancing EyeQoE's accuracy. In the other direction, we intend to combine EyeQoE with other existing QoE models for 360-degree videos. The hypothesis is that QoE models capturing a greater diversity of potentially informative features might improve the overall model robustness when included. We plan to apply *ensemble methods* [208, 224] over multiple representative QoE models and EyeQoE to derive the aggregated prediction results. Comparison will be made with each single model over the prediction accuracy.

Reducing QoE prediction latency for unseen videos. Under the current design, online QoE predictions over unseen videos are executed at the level of seconds. The la-

tency is mainly caused by the graph generation and the domain adaptation process, i.e., fine-tuning the trained QoE model. While this value is practically acceptable for pure QoE collection, it would be too large to support real-time QoE-aware service management, which can benefit applications such as adaptive 360-degree video streaming [296]. Essentially, service providers can timely adjust streaming strategies, such as resolutions, rendering speed, and scheduling priority, in accordance with the viewer's QoE estimated in real-time. As our future work, we plan to investigate the feasibility of forecasting viewer's perceived QoE a short period ahead of time, which then better tolerates the prediction latency. There is an important observation that viewer's subjective feelings typically do not change suddenly. For instance, one's cybersickness and fatigue are gradually accumulated as prolonged exposure in a VR environment. Such temporal dependencies can be exploited for QoE forecasting.

Other approaches for adaptation to unseen videos. A critical challenge of this work is to adapt the QoE model, trained by existing video clips, to unseen videos. Aside from domain adaptation as adopted here, another interesting future direction is to leverage few-shot learning [53, 282]. We frame the challenge as a few-shot learning problem, that is: how to train the GCN classifier such that it can quickly adapt to an unseen video after a few learning iterations with a small number of annotated samples from the same category (Section 4.6.3) that the unseen video belongs to. Few-shot learning is promising in classifying new data when only a few training samples with supervised information are available and has been successfully applied in language processing [298], text classification [304], and image classification [51].

4.9 Conclusion

In this paper, we present EyeQoE, a novel QoE prediction model for 360-degree videos using subjects' ocular behaviors. To extract useful features from the behaviors, we propose a novel method that models them into graphs and then build a GCN-based classifier to learn over graphs. Our design also involves the Siamese network that deals with learning uncertainty caused by subjects and visual stimuli heterogeneity. A domain adaptation scheme named MADA is further proposed to ensure the efficacy of EyeQoE on unseen videos. A 3-month data collection campaign is carried out to build our own visual-based QoE assessment dataset. Our comprehensive evaluation shows that EyeQoE advances the literature by a suite of new capabilities. First, its best accuracy performance is 92.9% which beats other state-of-the-art models. Second, EyeQoE is capable of capturing various impact factors, such as video stalls and viewer's subjective feelings (e.g., cybersickness, immersiveness, and fatigue), in QoE prediction, while they are largely overlooked in prior models. Moreover, all the online operations of EyeQoE can be efficiently performed with 90-percentile computation latency within 4.2 seconds.

ACKNOWLEDGMENTS

We sincerely thank the anonymous reviewers for their insightful comments and suggestions. We are also grateful to NSF (CNS-1943509) for partially funding this research.

CHAPTER 5

Phyre: A NOVEL VIDEO RECOMMENDER SYSTEM FOR VIRTUAL REALITY USING PHYSIOLOGICAL SIGNALS

5.1 Introduction

5.1.1 Background

In recent years, the integration of virtual reality (VR) technology has revolutionized individuals' encounters with digital content, casting a notable impact on video consumption. Watching videos in VR offers an immersive and 3-degrees-of-freedom interactive experience, allowing users to enter a three-dimensional virtual environment. As advancements in VR technology have made such experiences more accessible, a substantial growth in video consumption in VR can be expected in the near future.

As one of the core tasks in video services, video recommendation plays an essential role in providing accurate suggestions for watchers, enabling them to navigate through the overwhelming content and efficiently discover videos that truly capture their interest. With the continuously rising popularity of VR videos, the need for effective video recommendations in this context becomes increasingly crucial. However, to our knowledge, *there is no video recommender system specially tailored for VR users*. Currently, video recommendation schemes in VR are directly borrowed from existing frameworks adopted by traditional platforms such as YouTube; these frameworks are used for conventional computing terminals such as PCs and smartphones. In comparison, recommending videos in VR presents the following uniqueness and potential opportunities. First, unlike traditional 2D videos, VR videos bring to its viewers' unique perceptive feelings, such as cybersickness, immersiveness, and presence [104, 134, 135, 266, 317]. These unique attributes potentially intro-

duce a new set of factors influencing viewer preferences in VR settings. Second, many VR headsets nowadays are equipped with a variety of onboard sensors, from IMU to eye trackers. They can capture novel types of user-video interactions, providing valuable insights that can be utilized to enhance video recommendation in VR contexts.

Recently, physiological data has emerged as a new sensing modality to measure user preference during video watching [58, 59, 72, 110, 129, 149, 178, 242, 300]. For example, Christoforou *et al.* [58] employed eye-tracking data to quantify the impact of narrative-based video stimuli to the preferences of large audiences. Lee *et al.* [149] studied the link between users' head movement data and their preference on VR videos. This evidence motivates us to leverage users' physiological responses when engaging with videos to infer their preferences and exploit such information to make future video recommendations. As an initial effort in this research topic, we start by examining two commonly accessible physiological measures from VR headsets: eye gaze[1] and head rotation. Through an extensive measurement study, we validate that these two measures can serve as effective indicators of whether a user enjoys watching a video. Encouraged by this promising finding, we propose incorporating them into VR video recommendation frameworks to enhance recommendation precision.

5.1.2 Challenges

Despite the appeal of this concept, its implementation poses the following non-trivial challenges.

$\mathscr{C}_1$: **Coping with new user-video interaction metrics.** First, with new kinds of user-video interaction metrics, a new data structure and a novel learning model are needed

[1]This covers a range of mainstream COTS VR models, including VIVE Focus 3 [7], VIVE Pro 2 [3], Meta Quest Pro [6], PlayStation VR2 [5], Pico Neo series [2], Varjo VR-3 [1], Fove [4], and Vision Pro [8].

to effectively extract prominent features from the complex raw readings for video recommendation.

$\mathscr{C}_2$: **Lack of training datasets.** Second, to train the recommender model properly, it is essential to acquire a sizable and diverse labeled physiological dataset. This typically involves data from thousands of users and videos, encompassing up to a million interactions. As an initial effort to utilize physiological data to refine video recommendations, we face the challenge of a lack of existing annotated datasets. Consequently, assembling a comprehensive dataset of a meaningful magnitude and scope presents a significant hurdle.

$\mathscr{C}_3$: **Energy consumption overhead.** Lastly, continuously uploading the new user-video interaction metrics to a server where the recommendation is performed is energy-consuming and may quickly deplete the battery of standalone VR headsets. Hence, how to achieve energy efficiency for VR terminals is another critical aspect to consider in *Phyre*.

5.1.3 Our Solution: *Phyre*

In this paper, we propose *Phyre*, a novel video recommender system for VR enhanced by physiological signals. It aims to exploit the correlation between physiological measures and user-video preferences to enhance VR video recommendation. During a video session, the user's physiological responses, gaze and head movement particularly, are collected by the built-in sensors of the VR device; these signals are uploaded to the server to infer the user's preference. A recommender system on the server takes physiological signals in all user-video interactions as the input, extracts their intrinsic and collaborative information, and makes recommendations accordingly.

To tackle the challenges ($\mathscr{C}_1$-$\mathscr{C}_3$), we make the following technical contributions. To address $\mathscr{C}_1$, we propose to formulate users, videos, and their interactions as a graph, where physiological signals are modeled as node and edge embeddings. Graph convolutional network (GCN), a classic graph learning model, is applied over the graph for feature

extraction. To accommodate the new property of the constructed graph, we renovate the conventional message passing function in the convolutional layers of the GCN, improving its capability to learn from physiological signals and extract collaborative information.

To solve $\mathscr{C}_2$, we adopt the concept of *domain adaptation*, which takes the traditional user-video interaction data as the source domain, the physiological data as the target domain, and adapts the model pre-trained on the source domain to the target domain via fine-tuning. Compared with training from scratch, only a small amount of data from the target domain is required. Considering the non-negligible gap between the two domains in our case, we propose a novel *cross-modality cross-context domain adaptation* (*CMCCDA*) scheme to fill the gap by introducing an extra "bridge domain". The adaptation is then performed in two incremental steps.

Finally, we address $\mathscr{C}_3$ by developing an energy-efficient adaptive encoding scheme. It adaptively encodes physiological signals in accordance with their entropy to significantly reduce the data size and thus the energy overhead for data transmission.

We highlight our contributions of the paper as follows:

- We introduce *Phyre*, a physiological-signal-enhanced video recommender system for VR. To our knowledge, this is the first video recommender system tailored for VR utilizing physiological signals.
- We integrate physiological signals into the mainstream recommendation framework and renovate the GCN learning paradigm to accommodate the new property of the user-video interaction graph. A novel domain adaptation approach is developed to address the data scarcity problem. Additionally, an energy-efficient adaptive encoding scheme is proposed to reduce the energy consumption of VR devices.
- We collect a physiological dataset for video recommendation in VR. It involves a total of 3,000 video sessions within 60 participants and 400 videos. This dataset will be open-sourced to the research community.

- We demonstrate through extensive evaluation that *Phyre* outperforms state-of-the-art schemes for video recommendation in VR by up to 68.0% in recommendation precision and up to 28.8% in the ranking quality.

Physiological data and recommender systems. The idea of involving physiological data in recommendation systems research has been explored in prior works [43, 70, 75, 165, 236, 292, 312]. Such data include facial expression, gaze, skin resistance, etc. These works mainly focus on quantifying correlations between physiological measures and users' preference [43, 70, 75], inferring video genres through facial expression analysis [165], acquiring viewers' attention time from eye tracking data [292], gaze prediction [312], and gaze clustering [236] in recommendation systems. However, how to utilize physiological data to generate recommendation results has rarely been investigated in a systematic way. This is partially due to a lack of public recommendation datasets available with the physiological data. More importantly, none of the above systems is designed for VR settings. In this work, we made an initial effort to bridge this gap.

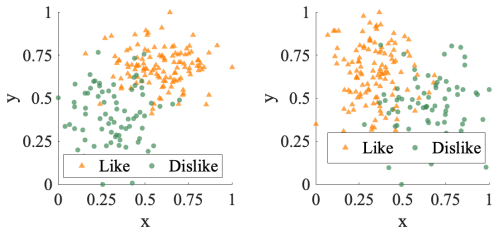
5.2 Related Work

Taxonomy of recommender systems. Over the past few decades, recommender systems have become a crucial technique in diverse domains, including e-commerce, social media, and content streaming [64, 201, 219, 232, 299, 309]. These systems are designed to predict and suggest items potentially liked by target users based on their historical preferences and behaviors. Based on how information is filtered for recommendations, these systems can be classified into three categories, namely *content-based filtering* [32, 179, 202], *collaborative filtering* [111, 141, 218, 231], and *hybrid filtering* [13, 40, 44, 249]. Among the above categories, collaborative filtering can capture complex and subtle patterns in user behavior and excels in its large-scale performance without requiring domain knowl-

edge *a priori* [122]. As a sub-category of this direction, graph-based collaboration filtering techniques first structure user-item interactions as *graphs*. Then, graph learning models are employed to extract collaborative information, which is further utilized to predict the preference of a target user on various items and make recommendations accordingly [85, 107, 281, 284, 299]. A representative state-of-the-art is PinSage [299], a recommendation framework for Pinterest utilizing GCN to learn from a user-item interaction graph for personalized recommendations. Prior works [85, 238, 281, 284] also fall into this category. As our work exploits users' physiological data for video recommendation, we adopt the graph-based collaborative filtering as our recommendation framework.

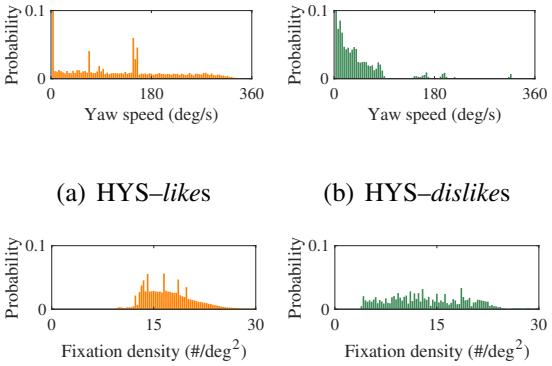
Video recommendation. Video recommendation is an important application of recommender systems. Related techniques have been widely employed by various video streaming platforms such as YouTube and TikTok [69, 100, 263]. Compared with the other tasks, video recommendation is unique due to its rich content and temporal dynamics [100, 107]. Extensive existing efforts have been devoted to tackling these characteristics [107, 117, 127, 172]. For example, Huang *et al.* [117] utilized video types and temporal factors to identify similar videos for recommendation. Jiang *et al.* [127] created fine-grained user interest groups based on users' interaction sequences and made recommendations based on the preferences of others from the same group. Recently, Han *et al.* [107] developed MTHGNN, a micro-video recommender system that considers the temporal and dynamic changes in users' preferences. Note that none of the above works utilizes physiological data to understand viewers' video preferences when making recommendations.

Physiological data and recommender systems. The idea of involving physiological data in recommendation systems research has been explored in prior works [43, 70, 75, 165, 236, 292, 312]. Such data include facial expression, gaze, skin resistance, etc. These works mainly focus on quantifying correlations between physiological measures and users' preference [43, 70, 75], inferring video genres through facial expression analysis [165], ac-



(a) Gaze feature distribution (b) Head movement feature distribution

Figure 5.1: 2D visualization of normalized feature distributions of gaze and head movement.



(a) HYS–*like*s (b) HYS–*dislike*s

(c) EFD–*like*s (d) EFD–*dislike*s

Figure 5.2: Exemplary features distributions.

quiring viewers' attention time from eye tracking data [292], gaze prediction [312], and gaze clustering [236] in recommendation systems. However, how to utilize physiological data to generate recommendation results has rarely been investigated in a systematic way. This is partially due to a lack of public recommendation datasets available with the physiological data. More importantly, none of the above systems is designed for VR settings. In this work, we made an initial effort to bridge this gap.

5.3 Measurement Study

Over decades, extensive research has been devoted to recognizing user's preferences on videos from their physiological responses such as gaze [58, 59, 110], head movement [129, 149], brainwave [59, 72, 178, 300], and heart rate [242].

Measurement setup. To validate that such a correlation also exists under the VR setting, we carry out an IRB-approved measurement study at a university lab. Ten subjects are recruited. Each is asked to watch the entire set of twenty videos wearing an HTC Vive Focus 3 VR device. After watching each video, subjects indicate whether they *like* or *dislike* the video. During the entire process, their physiological signals are recorded by the

onboard sensors and locally stored on the VR device. The analysis is performed over the collected dataset from 200 video-watching traces. We focus on two kinds of physiological data: gaze and head movement, which are captured by the onboard eye tracker and inertial measurement unit (IMU), respectively.

Results. Figure 5.1 presents the normalized distributions of two-dimensional features of gaze and head movement. These two-dimensional features are generated by applying an autoencoder to the raw readings and then casting the derived multi-dimensional embedding vectors onto two dimensions for visualization. The features marked with *like*s and those with *dislike*s are distributed in two distinctive clusters. Take gaze as an example: The mean values of its two-dimensional features are [0.63, 0.57] and [0.35, 0.34] (in [x, y]) for *like* and *dislike*, respectively. Their standard deviations are [0.18, 0.12] and [0.16, 0.22], respectively.

We further extract two calibrated features, namely head yaw speed (HYS) and eye fixation density (EFD). HYS, drawn from the IMU readings, denotes the average speed of a subject's head movement on the yaw axis. As shown in Figure 5.2(a) and 5.2(b), values of this feature are more evenly distributed when subjects like the video, whereas those are more concentrated on the lower end otherwise. EFD is the average number of gazes in a unit area for all fixations. We observe from Figure 5.2(c) and 5.2(d) that EFD approximately ranges between 12 and 24 gazes per unit area when the user likes the video; this value is more scattered otherwise.

The results are promising: Gaze and head movement exhibit patterns highly correlated with user interest in video content. They serve as evidence that such physiological data can be used as effective indicators of users' preference toward videos in VR, which will be exploited for video recommendation in the rest of this work.

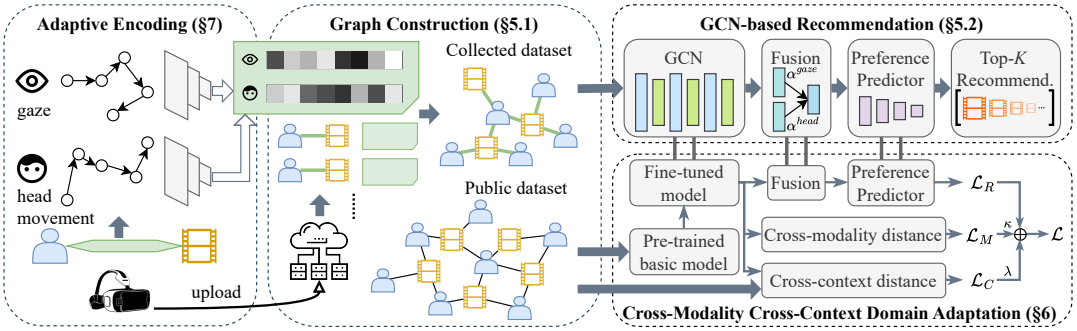


Figure 5.3: System architecture of *Phyre*.

5.4 System Overview

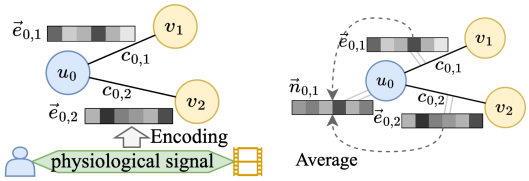
In this work, we propose *Phyre*, a novel video recommender system for VR by exploring viewers' physiological signals. *Phyre* harnesses the intrinsic correlation between viewers' preferences and physiological signals to enhance video recommendation. Figure 5.3 depicts the overall system architecture, which consists of four major components: adaptive encoding, graph construction, GCN-based recommendation, and CMCCDA. As a user watches a video, her physiological signals, i.e., gaze and head movement, are recorded and encoded by the VR device. The encoded embeddings are uploaded to the cloud server, where the user-video interactions are constructed into graphs (Section 5.5.1). Then, a GCN model is employed to learn representations from the graph, based on which the top-K videos are derived and recommended to the target user (Section 5.5.2). To train the model with the limited annotated physiological measures, we propose a novel domain adaptation strategy CMCCDA to deal with the non-negligible inter-domain distances (Section 5.6). An adaptive encoding algorithm is also developed to compress the raw physiological signals and reduce energy overhead for data communications at VR terminals (Section 5.7).

5.5 Graph-based Recommendation

Our task is to recommend a list of videos from a given video pool to a target user. The list consists of videos the user has not encountered and will likely align with her preference. Motivated by key observations from the measurement study, we propose to introduce physiological signals (i.e., gaze and head movement) as a new kind of user-video interaction metric to facilitate VR video recommendations. In the following, we first model user-video interactions into graphs. Then, we employ GCN as a graph learning tool, upon which a recommender system is built.

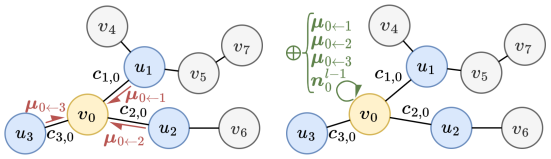
5.5.1 Graph Construction

We construct the entire dataset of all users, videos, and their interactions as a graph $\mathscr{G} = \{\mathscr{U}, \mathscr{V}, \mathscr{C}\}$, where the set of users are represented as graph nodes $\mathscr{U}$, the set of videos as nodes $\mathscr{V}$, and their connections as graph edges $\mathscr{C}$. An edge $c_{ij} \in \mathscr{C}$ connects a user node $u_i \in \mathscr{U}$ and a video node $v_j \in \mathscr{V}$; $c_{ij} = 1$ if u_i has watched v_j, and $c_{ij} = 0$ otherwise. We define the attribute of each edge as the *embedding*. To derive the embedding, an encoder is applied to the time-series physiological signals recorded during the video-watching session, as depicted in Figure 5.4(a); the encoder's design is detailed in Section 5.7. This embedding is a vector of features extracted to describe the user preference from the video watching session. Our definitions of edge attributes are intuitive: Physiological signals can reflect user-video interactions, as demonstrated in Section 5.3. With edge embeddings, we further define the node embedding $\boldsymbol{n}_i$ by taking the average of embeddings of all edges connected to that node: $\boldsymbol{n}_i = \frac{1}{|\mathscr{N}_i|} \sum_j \boldsymbol{e}_{ij}$, where $\mathscr{N}_i$ represents u_i's neighbor set and $\boldsymbol{e}_{ij}$ is the embedding of the edge between u_i and its neighbor v_j, as illustrated in Figure 5.4(b). The node's attribute is defined in such a way because physiological signals contain rich information regarding both the user and her watched video. Take a user node as an example: The signal



(a) Edge embedding (b) Node embedding

Figure 5.4: Defining edge and node embeddings.



(a) Message passing (b) Aggregation

Figure 5.5: Illustration of graph convolution operations.

may reveal the user's video preferences and watching habits, which can be used to profile the user.

With the constructed user-video interaction graph, we further derive a *node attribute array*, an *edge attribute array*, and an *adjacency matrix*, which will be used in the graph learning presented soon. A node attribute array $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ represents all node embeddings $\boldsymbol{X} = \{\boldsymbol{n}_i | \forall i \in [1,N]\}$. An edge attribute array $\boldsymbol{E} \in \mathbb{R}^{N \times N \times D}$ represents all edge embeddings $\boldsymbol{E} = \{\boldsymbol{e}_{ij} | \forall i,j \in [1,N]\}$. An adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ represents the connectivity between two arbitrary nodes $\boldsymbol{A} = \{c_{ij} \in \{0,1\} | \forall i,j \in [1,N]\}$. In the above definitions, N denotes the number of nodes in $\mathscr{G}$ and D is the cardinality of the embedding vector.

Discussions. Most existing video recommendation frameworks only consider traditional user-video interactions, such as video-watching duration or if a user *likes* that video. Due to their simple data format (i.e., binary or real values), these edge weights are conveniently formulated into the adjacency matrix to feed into the graph learning models. In contrast, (embeddings of) physiological measures here are high-dimension vectors, which cannot be represented as simple-value edge weights. Therefore, we incorporate them into the graph as edge embeddings $\boldsymbol{E}$. These edge embeddings provide richer information than simple-value edge weights and thus offer more valuable insights into users' preferences. The main job of the rest of this work is to develop suitable learning techniques to extract latent features from the new graph with a sophisticated structure.

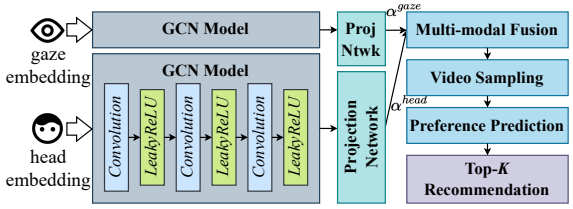


Figure 5.6: GCN-based video recommendation workflow.

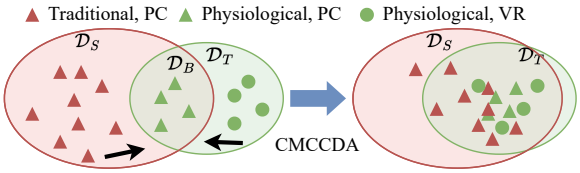


Figure 5.7: The illustration of the CMCCDA process; each point is a sample representation.

5.5.2 GCN-based Video Recommendation

5.5.2.1 Overview

Phyre is built on a classic graph-based recommendation framework, which consists of two stages: graph learning and top-K recommendation. *In the stage of graph learning*, to mine the complex relationships among users and videos from the constructed graph, we apply a GCN model, which takes as input the graph, i.e., $\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{E}$ as derived above, and produces the output as *node representations*. The representations in different modalities are then projected into the same space and fused. *In the stage of recommendation*, a subset of candidate videos is first sampled from the entire video pool. These are videos that the target user has not previously watched but may be interested in. Then, a preference predictor takes the target user's and each candidate video's *node representations* as an input pair and predicts the preference score. Finally, top-K videos with the highest preference scores are recommended to the user. Figure 5.6 illustrates this workflow.

To fit into our scenario, we make several renovations to the classic GCN-based recommendation framework, including modifying the message passing mechanism and some key calculations (i.e., *preference prediction*, *loss*, and *aggregation*). Next, we will introduce each step of our GCN-based recommender system in detail.

5.5.2.2 GCN Learning.

The first step of GCN applies convolution over the graph. It consists of the message passing and the aggregation steps – in these steps, the embeddings of each node's neighbors are propagated through connecting edges and integrated with its own embedding, as illustrated in Figure 5.5. As there are two different modalities, i.e., gaze and head movement data, we start by considering an arbitrary modality $m \in \mathscr{M}$.

Message passing. As the first step of the convolution operation, message passing propagates the information from each node to all neighbors through their connecting edges

$$\boldsymbol{\mu}_{j \leftarrow i}^{m} = \frac{1}{\sqrt{\left|\mathscr{N}_i\right| \left|\mathscr{N}_j\right|}} \left(\boldsymbol{W}_1^m \boldsymbol{n}_i^m + \boldsymbol{W}_2^m \left(\beta \boldsymbol{e}_{i,j}^m + \boldsymbol{n}_i^m \odot \boldsymbol{n}_j^m\right)\right) \tag{5.1}$$

where $\boldsymbol{n}_i^m$ represents node u_i's embeddings of modality m, $\odot$ stands for the element-wise multiplication, $\boldsymbol{W}_1^m$ and $\boldsymbol{W}_2^m$ are learnable matrices, and β is the weight. Conventional message passing allows each node to collect information from its immediate neighbors (i.e., the first term above) [156, 284], thereby integrating local neighborhood information into its representation.

Discussions. For effective feature extraction, we renovate the message-passing function to accommodate our unique graph structure, where physiological responses, serving as edge embeddings, contain information from both end nodes (user and video). To preserve such information in every convolutional layer, we integrate them into our message passing function as the second term. In addition, we introduce the third term above to encourage passing more information between similar neighbors. For example, if user u_i has a similar embedding as video v_j, it indicates that the characteristics of video v_j align well with u_i's preferences. This similarity can be leveraged to enhance v_j's feature representation by integrating more common information shared with u_i. Overall, our proposed design substantially changes the message-passing workflow in traditional GCN, a design unexplored in prior works.

Aggregation. Upon receiving all neighbor information, node v_j performs the following aggregation function on each layer l to derive the *collaborative information*

$$\boldsymbol{n}_j^{(l)m} = LeakyReLU\left(\boldsymbol{W}_3^m \boldsymbol{n}_j^{(l-1)m} + \sum_{i \in \mathcal{N}_j} \boldsymbol{\mu}_{j \leftarrow i}^{(l)m}\right) \tag{5.2}$$

where $\boldsymbol{n}_j^{(l-1)m}$ denotes the target node embedding in the previous convolutional layer, and $LeakyReLU(\cdot)$ is the activation function. On the lth convolutional layer, node v_j is updated by its information on the $(l-1)$th layer and the lth-layer neighbors' embeddings, which reflect the collaborative information.

Multi-modality attentional fusion. After several convolutional layers of message passing and aggregation, the output embeddings of multiple modalities are fused to derive the final embedding of each node. Considering the heterogeneous embedding space of each modality, we employ a projection network that maps the embedding of each modality into the common space, before passing it into the fusion layer, where the projection outputs are weighted by their modality-specific attention

$$\boldsymbol{n}_j = \sum_{m \in \mathcal{M}} \alpha^m H\left(\boldsymbol{n}_j^m\right) \quad \textbf{where} \quad \alpha^m = \frac{e^{-r^m}}{\sum_{i \in \mathcal{M}} e^{-r^i}} \tag{5.3}$$

where $\boldsymbol{n}_j$ stands for the final node representation after fusion, $H(\cdot)$ represents the projection network that maps all modalities to the same latent space, α^m denotes the attention for modality m, and r^m is the data compression ratio in m.

5.5.2.3 Recommendation.

After GCN learning, the final representations at all nodes possess sufficient information for video recommendations. A video sampler first samples a list of candidate videos from the video pool for the target user [107], from which a preference predictor estimates their preference scores, and recommends the candidate videos with the highest preference scores to the user.

The preference predictor consists of a few fully connected layers; given a target user u and a candidate video v, it takes the final representations of u and v as an input pair and produces the predicted preference score

$$\hat{y}_{uv} = F_\theta \left(\gamma \boldsymbol{n}_i \odot \boldsymbol{n}_j + \frac{1}{|\mathcal{N}_v|} \sum_{\boldsymbol{n}_i \in \mathcal{N}_v} \boldsymbol{n}_u \odot \boldsymbol{n}_i + \frac{1}{|\mathcal{N}_u|} \sum_{\boldsymbol{n}_j \in \mathcal{N}_u} \boldsymbol{n}_v \odot \boldsymbol{n}_j \right) \tag{5.4}$$

where $F_\theta(\cdot)$ is a multilayer perceptron (MLP) parametrized by θ, and γ is a weighting hyperparameter. In this way, we comprehensively formulate u's potential preference towards v by capturing 1) their direct similarity, 2) the similarity between u and users who have watched v, and 3) the similarity between v and videos watched by u, before feeding it to the MLP that maps the input vector to the final preference score.

Discussions. Compared with the preference prediction function in existing works, we propose to add the correlation between the users' and the videos' embeddings as the first term, leveraging the fact that they both fall into the same feature space. In contrast, they commonly reside in heterogeneous embedding spaces in previous works. The additional term directly encourages recommending videos to the target user that share similar embeddings with each other.

Finally, videos in the candidate list are ranked based on their predicted preference scores, and the top-K candidate videos are recommended to the target user.

5.5.2.4 Loss Function

The remaining piece is to decide the loss function for model training. To this end, we establish upon the classic Bayesian Personalized Ranking (BPR) loss [216], a commonly adopted loss function to train recommender systems, and propose our recommendation loss as follows

$$\mathscr{L}_R = -|y_{uv^p} - y_{uv^q}| \log \sigma \left(|\hat{y}_{uv^p} - \hat{y}_{uv^q}| \right) \tag{5.5}$$

where $\hat{y}_{uv^p}$ and $\hat{y}_{uv^q}$ are the predicted preference scores between the target user u and a random pair of two arbitrary videos v^p and v^q, respectively, with y_{uv^p} and y_{uv^q} representing their ground-truth preference scores.

5.6 Cross-Modality Cross-Context Domain Adaptation

To build the GCN-based recommendation model, it is crucial to gather a large labeled physiological dataset of the necessary diversity and volume so that the model can be properly trained. However, this is prohibitively infeasible as it would involve thousands of users/videos and up to a million interactions. As a reference, *MovieLens-1M*, a widely adopted public dataset for training video recommender systems, consists of 1 million interactions from over 6 thousand users and over 3 thousand videos [103].

To address this data scarcity issue, we propose to adopt the concept of *domain adaptation*. It allows a deep learning model trained in one *source domain* (i.e., traditional user-video interaction data of 2D videos, denoted by $\mathscr{D}_S$) to adapt to a different but related *target domain* (i.e., viewers' physiological data in watching VR videos, denoted by $\mathscr{D}_T$) via fine-tuning. Typically, domain adaptation needs a much smaller amount of data than training the whole model in the target domain from scratch. Nonetheless, the successful employment of domain adaptation requires the *distance* between the source and target domains to be within a certain threshold; otherwise, the performance will be degraded significantly. In our case, such distance is non-negligible as traditional user-video interaction (e.g., whether users finish watching videos, hit *like*s, etc.) and physiological data are two distinctive modalities. Additionally, they are across different contexts as $\mathscr{D}_S$ is for videos displayed on regular terminals, such as PCs and smartphones, whereas $\mathscr{D}_T$ is for VR videos. *The discrepancy between the two domains prohibits the direct adoption of traditional domain adaptation techniques.*

Bridge domain. To address this challenge, we propose a novel domain adaptation framework called *cross-modality cross-context domain adaptation* (*CMCCDA*). The idea is to introduce a *bridge domain* $\mathscr{D}_B$, which connects the source and target domains. Rather than directly adapting the original graph learning model (trained over $\mathscr{D}_S$) to the target domain (using $\mathscr{D}_T$), we propose to fine-tune the model by minimizing the representations' distance between $\mathscr{D}_S$ and $\mathscr{D}_B$ and subsequently their distance between $\mathscr{D}_B$ and $\mathscr{D}_T$. In doing so, the substantial distance between $\mathscr{D}_S$ and $\mathscr{D}_T$ is broken down into two manageable distances that can be bridged in two incremental steps.

To perform CMCCDA, we create our own *hybrid* dataset: Each subject watches some randomly selected videos in VR headsets. During each video session, they freely hit the *like* or the *share* button, pause, fast-forward, rewind, or skip the current video play as they like. In the meantime, the VR headset records the subject's physiological data throughout the session. We call it a hybrid dataset because it involves both the traditional user-video interactions and the physiological data. The former shares the same modality with $\mathscr{D}_S$, while the latter shares the same context with $\mathscr{D}_T$. Even better, these two kinds of data are aligned in each video. We denote this common information in the hybrid dataset, i.e., traditional interactions in the VR context, as $\mathscr{D}_B$, to facilitate bridging the two distinctive domains.

Cross-modality distance. We define the cross-modality distance as the distance between $\mathscr{D}_B$ and $\mathscr{D}_T$. It is computed as the average distance between the representation of each node in $\mathscr{D}_B$ and that of the corresponding node (i.e., same user/video) in $\mathscr{D}_T$. We call it "cross-modality" because nodes in $\mathscr{D}_B$ and nodes in $\mathscr{D}_T$ are associated with two modalities, the traditional interactions and physiological data, respectively. The distance between two corresponding nodes is denoted as $\Delta(\boldsymbol{z}_i^B, H'(\boldsymbol{z}_i^T))$, where $\boldsymbol{z}^B$ and $\boldsymbol{z}^T$ are the node representations, $\Delta(\cdot,\cdot)$ refers to any appropriate distance function, e.g., Euclidean distance, and $H'(\cdot)$ is a projection function to cast representations across modalities.

Cross-context distance. We define the cross-context distance as the distance between $\mathscr{D}_S$ and $\mathscr{D}_B$. Like the cross-modality distance, the cross-context distance involves the distance between representations of nodes from $\mathscr{D}_S$ and those from $\mathscr{D}_B$. Additionally, it also considers the difference in the graph structures of the two domains.

We first identify *common nodes* from $\mathscr{D}_S$ and $\mathscr{D}_B$. They are common videos in both the public dataset and the hybrid dataset. For each common node v_i, we identify its local graph covering its neighbor nodes in h hops and all edges involved, where h is an empirical value. Then, we calculate the distance between the common nodes' representations in two domains and weight it with the similarity between their local graphs $\Delta\left(\boldsymbol{z}_i^S, \boldsymbol{z}_i^B\right) \cdot \Gamma_i^{S,B}$, where $\Gamma_i^{S,B}$ is the similarity between the local graphs in $\mathscr{D}_S$ and $\mathscr{D}_B$, respectively.

It is challenging to efficiently derive the similarity between local graphs with existing methods. To address this, we propose to approximate this similarity by comparing their *graph representations*, which is defined as the weighted average of its node representations. To further enhance efficiency, given a common node v_o, we propose to approximate the learnable weight for each u_i (v_i) in its local graph using its *centrality* $\zeta_i = \frac{d_o d_i}{l_{o,i}^k} + \eta \sum_{j \in \mathscr{N}_i, j \neq o} \frac{d_i d_j}{l_{i,j}^k}$ where d represents the node degree and $\imath$ gives the length of the shortest path between two different nodes; η and k are hyperparameters. This centrality reflects how much the node contributes to the common node. Based on this, we derive the graph representations and compute the graph similarity.

$$\Gamma_i^{S,B} = C\left(\sum_{p \in \mathscr{G}_{i,h}^S} \zeta_p \boldsymbol{x}_p, \sum_{q \in \mathscr{G}_{i,h}^B} \zeta_q \boldsymbol{x}_q\right) \tag{5.6}$$

where $C(\cdot,\cdot)$ is an arbitrary similarity function, e.g., cosine similarity. $\mathscr{G}_{i,h}^S$ and $\mathscr{G}_{i,h}^B$ denote the local graphs of common node v_i in $\mathscr{D}_S$ and $\mathscr{D}_B$, respectively. $\boldsymbol{x}_p$ and $\boldsymbol{x}_q$ are the node embeddings before GCN.

Bridging two domains. Finally, we formulate the loss functions based on the cross-modality distance and the cross-context distance discussed above

$$\mathscr{L}_M = \frac{1}{|\mathscr{S}|} \sum_{p \in \mathscr{S}} \Delta\left(\boldsymbol{z}_p^B, H'\left(\boldsymbol{z}_p^T\right)\right) \tag{5.7}$$

$$\mathscr{L}_C = \frac{1}{|\mathscr{O}|} \sum_{i \in \mathscr{O}} \Delta\left(\boldsymbol{z}_i^S, \boldsymbol{z}_i^B\right) \cdot \Gamma_i^{S,B} \tag{5.8}$$

In the above equations, $\mathscr{S}$ is a set of randomly sampled nodes and $\mathscr{O}$ is the set of all common nodes. We derive two losses, $\mathscr{L}_M$ and $\mathscr{L}_C$, based on the cross-modality and cross-context distances, respectively. During CMCCDA, these two losses are computed incrementally in each iteration and back-propagated to update the model parameters through optimization. Minimizing the former encourages extracting representations with a smaller distance between $\mathscr{D}_B$ and $\mathscr{D}_T$, adapting the model from the bridge domain to the target domain. Similarly, minimizing the latter rewards learning similar representations between $\mathscr{D}_S$ and $\mathscr{D}_B$, with an emphasis on common nodes that have more similar local graphs. These losses jointly guide the model to decrease the distance of representations in all domains through fine-tuning, adapting it from $\mathscr{D}_S$ to $\mathscr{D}_T$.

Given such, the final loss for fine-tuning is derived as the weighted sum of three components

$$\mathscr{L} = \mathscr{L}_R + \kappa \mathscr{L}_M + \lambda \mathscr{L}_C \tag{5.9}$$

where κ and λ are tunable weights. Recall that $\mathscr{L}_R$ stands for our recommendation loss proposed in Section 5.5. $\mathscr{L}_M$ and $\mathscr{L}_C$ are defined above in Equation 5.7 and 5.8, respectively. The final loss helps to adapt the model from $\mathscr{D}_B$ to $\mathscr{D}_T$ and meanwhile enhances the recommendation accuracy.

We summarize the steps of CMCCDA as follows. First, data from all three domains are sampled and fed into three identical, weight-sharing GCN models, respectively. Then,

the outputs are utilized to compute the *cross-modality distance* between $\mathscr{D}_B$ and $\mathscr{D}_T$, and the *cross-domain distance* between $\mathscr{D}_S$ and $\mathscr{D}_B$. Finally, losses are derived based on these distances to fine-tune the GCN models.

5.7 Energy-efficient Adaptive Encoding

Recommender systems are typically deployed at cloud servers as their operations are resource-demanding. We thus adopt a similar strategy here. On the other hand, unlike traditional user-video interaction metrics (e.g., hitting *like* and watching duration), whose data size is very minimal, the size of the time-series multi-modal physiological signals is enormous for even minutes of video watching. As a result, consistently uploading the raw readings would consume significant energy overhead. It becomes a critical issue, especially for the battery-powered standalone VR headsets. Figure 5.8 shows the device's energy consumption breakdowns of one minute of video watching. Data transmission takes the most energy consumption at around 52%. Figure 5.9 further shows that the energy consumption grows linearly with the video length.

Motivated by our observation, we propose to diminish the uploaded data size so as to reduce the corresponding energy consumption at the VR headsets. Specifically, we develop an encoding scheme that compresses the raw signals into vector embeddings. They are then uploaded to the cloud and serve as inputs for our graph learning model. Given a physiological signal, the encoder divides it into segments depending on their entropy and encodes each segment into an embedding with an adaptive compression ratio; the signal embedding is the average of all segment embeddings. In this way, physiological segments with higher entropy are preserved with more information with a lower compression ratio; segments with lower entropy contain less information and are thus more aggressively compressed.

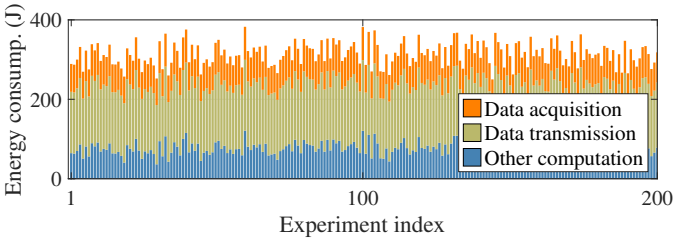


Figure 5.8: Empirical energy consumption breakdown on the VR headset.

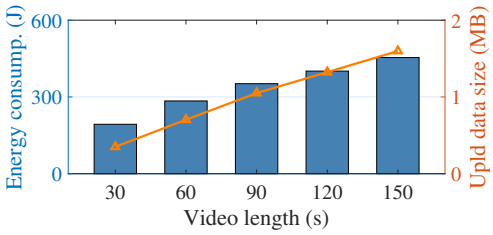


Figure 5.9: Energy consumption and data size vs video length.

Algorithm 2 outlines our encoder design. After a video session, the encoder takes as input the time-series physiological signal $\boldsymbol{\phi}$ and outputs a vector embedding $\boldsymbol{e}$. We first divide $\boldsymbol{\phi}$ into multiple segments $\boldsymbol{s}$: we apply a sliding window that continuously adds data points to a segment until its Shannon entropy reaches a threshold value. Then, each segment is passed into an LSTM model to generate a fixed-length embedding. Finally, all segment embeddings are averaged as the signal embedding $\boldsymbol{e}$. The entropy denotes the amount of information in one segment. The intuition is that the more informative segments should be preserved to benefit recommendation in later stages; in contrast, aggressive compression can be applied to segments with lower information. For example, given a random raw time-series physiological signal of 1 MB, its corresponding embedding only takes less than 1 KB according to our testing.

5.8 Evaluation

5.8.1 Evaluation Setup

We develop an Android VR app on a Focus 3 VR Headset running an Android-based OS and collect our hybrid dataset involving 60 participants, 400 videos, and 3000 interactions. Table 5.1 shows the demographic information of these participants. The VR app is used to play videos, enable controller interaction, and acquire physiological signals from users through the headset's embedded eye tracker and IMU sensor at a sampling

Algorithm 2: Energy-efficient Adaptive Encoding.

Input: Physiological signal $\boldsymbol{\phi}$; embedding length k; minimum segment length t; entropy threshold ε_θ; pre-trained model $LSTM$

Output: Signal embedding $\boldsymbol{e}$

$\boldsymbol{e} \leftarrow Zeros(k)$; $c \leftarrow 0$; $i \leftarrow 0$; $\boldsymbol{s} \leftarrow \boldsymbol{\phi}(0:t)$ // Initialize
while $i < |\boldsymbol{\phi}|$ **do**
  while $ShannonEntropy(\boldsymbol{s}) < \varepsilon_\theta$ **do**
    $\boldsymbol{s}.append(\boldsymbol{\phi}(i))$; $i \leftarrow i+1$; // Add next point
  $\boldsymbol{e} \leftarrow \boldsymbol{e} + LSTM(\boldsymbol{s}, k)$; // Add segment embedding
  $\boldsymbol{s} \leftarrow \boldsymbol{\phi}(i:i+t)$; $i \leftarrow i+t$; $c \leftarrow c+1$;
$\boldsymbol{e} \leftarrow \boldsymbol{e}/c$; // Average for the signal embedding

Table 5.1: Participants' demographic information.

Gender	#	%	Age range	#	%	Ethnicity	#	%	Education	#	%
Female	21	35	≤17	3	5	Asian	27	45	≤Bachelor	3	5
Male	38	63	18-25	25	42	Black/Afr.	9	15	Bachelor's	20	33
Other	1	2	26-35	19	32	Hisp./Lat.	8	13	Master's	27	45
			36-45	11	18	White	15	25	Doctorate	10	17
			≥45	2	3	Other	1	2			

rate of 120 Hz and 200 Hz, respectively. Videos played in the app are accessed via API from online resources such as YouTube and TikTok. They cover a wide range of topics and categories. Our cloud server is in charge of graph construction, model training, and making recommendations. The server is equipped with eight NVIDIA RTX A6000 GPUs and Intel Xeon Gold-5218R processors.

Before data collection, each participant is required to fill out the screening questionnaire and read and sign the consent form. Then, the participant is instructed by a researcher through the calibration phase and basic operations. During data collection, their physiolog-

Table 5.2: A list of public datasets adopted in *Phyre*.

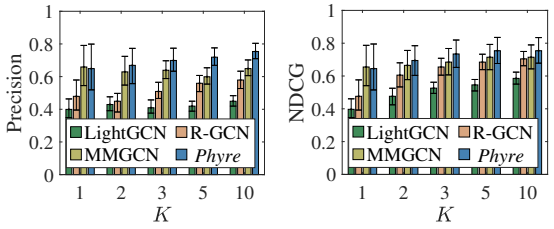
Public dataset	Number of users	Number of videos	Number of interactions
MovieLens-100K	1,000	1,700	100,000
MovieLens-1M	6,040	3,884	1,000,209
TikTok-1/50	1,434	29,662	95,426
TikTok-1/5	16,538	366,017	1,047,358

ical signals are captured in real-time. The participant watches 50 videos from a randomly sampled video set. Participants can freely interact with the video by hitting the *like* or the share button, fast-forwarding, rewinding the video play, or skipping the current video as they like. After a video play, the participant is asked to rate the preference score from 1 (lowest) to 5 (highest) based on how much they enjoy watching this video. After watching the entire video set, each participant is compensated with $10. All data collection phases are carried out in a university lab with normal lighting and environmental conditions. The entire experiment takes around 1 hour for each participant. The study meets all ethical requirements and holds active IRB approval at the researchers' university.

After acquiring the hybrid dataset, we randomly divide it into a training set and a testing set. The training set is used to fine-tune the base model. The evaluation below is based on the recommendation performance within the testing set. To pre-train the base model, we employ and study the public datasets as listed in Table 5.2 [102, 103, 262]; these datasets are commonly adopted in state-of-the-art video recommendation approaches.

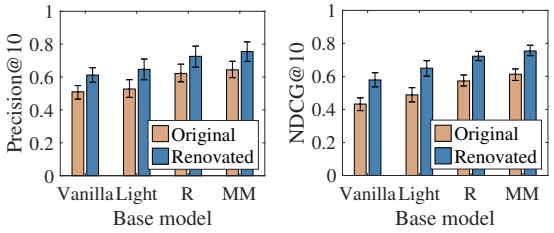
To demonstrate the superior performance of *Phyre*, we adopt the following state-of-the-art recommendation models as baselines for a comprehensive comparison.

- LightGCN [109], a GCN model with a simplified and concise design of GCN, especially tailored for recommendation tasks.
- R-GCN [233], a GCN framework focusing on relational modeling and graph construction, which has been effectively used for recommendation.



(a) Precision (b) NDCG

Figure 5.10: Overall comparison in recommendation performance.



(a) Precision@10 (b) NDCG@10

Figure 5.11: Ablation study of renovated GCN in recommendation performance.

- MMGCN [284], a recommendation framework that considers multiple modalities or types of information when generating embeddings of users and items.

5.8.2 Overall Performance

We showcase the overall performance of *Phyre* of all metrics and compare the result of each state-of-the-art baseline model in Figure 5.10. Generally, *Phyre* achieves higher recommendation precision and top-K ranking quality, indicated by the normalized discounted cumulative gain (NDCG), compared with all baseline models. Among all baseline models, MMGCN shows the best performance. For top-1 recommendation (K=1), *Phyre* maintains similar precision with MMGCN; as K grows to 10, *Phyre* outperforms all state-of-the-arts significantly by 0.11-0.31 (16.3-68.0%) in precision and 0.04-0.17 (5.6-28.8%) in NDCG. In summary, *Phyre* achieves superior performance compared with all state-of-the-art solutions for video recommendation in VR.

We also observe that the number of recommended items K in the top-K recommendation plays a crucial role in the recommendation performance. To investigate its impact, we change the value of K within $\{1, 2, 3, 5, 10\}$ and exhibit the corresponding performance$K$. Within 5 recommended items, increasing K would slightly increase the recommendation accuracy. This may be caused by the enlarged diversity and item space coverage, which

may better cater to the user's varied preferences and needs, reducing uncertainty with more recommended items. On the other hand, as K continuously increases, the recommendation precision and NDCG remain stable. This is potentially because recommending more items may introduce the lower-ranking items in the recommendation list, which may not be generated as accurately as the top-ranking ones, thus affecting the overall accuracy.

5.8.3 Impact of Base Models and Public Datasets

The model pre-trained on a public dataset plays an important role in *Phyre*'s performance as it determines the initial point of CMCCDA. We compare the recommendation precision of *Phyre* within four base models, namely vanilla GCN, LightGCN, R-GCN, and MMGCN. The graph learning models adopted in these works are all members within the GCN family and therefore share the common basic kernel structure with the vanilla GCN. This allows us to implement our interaction-preserving learning technique proposed in Section 5.5.2 within their message passing functions.

We pre-train these base models on each of the following public datasets. Table 5.3 demonstrates the result of their top-10 performance, i.e., precision@10. We observe that using MMGCN pre-trained on TikTok-1/50 renders the highest precision score (0.755), followed by R-GCN on TikTok-1/50 (0.751); the lowest precision (0.344) is obtained by adopting vanilla GCN on MovieLens-100K. A potential reason for the superior performance of MMGCN is the attentiveness to the user-video interaction, which better suits our scenario. TikTok datasets perform better than MovieLens as their videos have more similar genres and durations to ours. For optimal performance, we select MMGCN as the base model and pre-train it on TikTok-1/50 as the basis of *Phyre*.

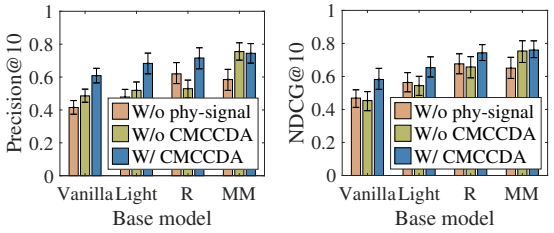
Table 5.3: Precision@10 with respect to different base models and datasets.

Base model	MovieLens-100K	MovieLens-1M	TikTok-1/50	TikTok-1/5
Vanilla GCN	0.344	0.393	0.570	0.620
LightGCN	0.393	0.437	0.627	0.699
R-GCN	0.408	0.455	0.699	0.751
MMGCN	0.436	0.489	0.742	**0.755**

5.8.4 Ablation Study

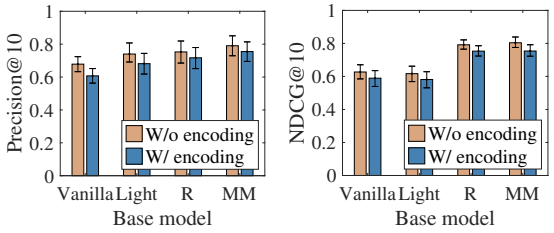
Renovated GCN. We investigate the effectiveness of the renovated GCN introduced in Section 5.5.2, one of the core techniques in *Phyre*, compared with the original GCN in the base models. We analyze the performance of *Phyre* and that of each base model using its original graph learning strategy. As illustrated in Figure 5.11, compared to each base model, *Phyre* significant improves the recommendation precision by 0.10-0.12. Similarly, a 0.14-0.16 improvement in NDCG from the base models also suggests the effectiveness of the proposed graph learning in *Phyre*. The major reason is its improved ability to excavate and preserve essential information from physiological signals, which suits our case.

CMCCDA. We study the efficacy of CMCCDA proposed in Section 5.6. Three strategies are compared: a) *Phyre* with CMCCDA, b) *Phyre* without CMCCDA, equivalent to using the pre-trained model without domain adaptation, and c) *Phyre* with only traditional interaction data. Figure 5.12 demonstrates the result. With a slight variance across datasets, we can clearly observe that our strategy, a) *Phyre* with CMCCDA, achieves the best result. Surprisingly, b) *Phyre* without CMCCDA renders even worse performance than c) *Phyre* without physiological signals. This may be due to the large domain distance, making the physiological signal data an overwhelming noise that does not improve but even deteriorates the pre-trained model's performance. This proves that CMCCDA is an indispensable technique in *Phyre* by "teaching" the model the knowledge of the VR context and physiological signals.



(a) Precision@10 (b) NDCG@10

Figure 5.12: Ablation study of CMCCDA in recommendation performance.



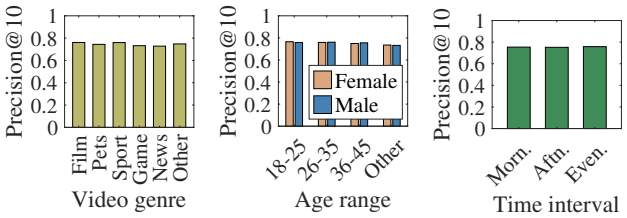
(a) Precision@10 (b) NDCG@10

Figure 5.13: Ablation study of adaptive encoding in recommendation performance.

Energy-efficient adaptive encoding. Lastly, we evaluate the energy-efficient adaptive encoding proposed in Section 5.7. We compare performance between *Phyre* with encoding vs. that without encoding, i.e., uploading raw signals and extracting embeddings later on the cloud. Here we focus on its performance drop, an inevitable side-effect of encoding due to certain data loss. We evaluate this drop from the optimal baseline, i.e., upload raw signals without energy limitations. As illustrated in Figure 5.13, *Phyre* with adaptive encoding achieves a comparable performance with the optimal baseline with only a marginal drop in precision (0.04-0.07) and NDCG (0.04-0.05). This result indicates that our adaptive encoding strategy preserves recommendation performance.

5.8.5 Robustness Against Impact Factors

It is important for *Phyre* to make recommendation videos with a robust, unbiased performance across different video categories, for various target user demographics, and at arbitrary time intervals within a day. To evaluate this, we divide videos into 6 genres (see Figure 5.14(a)), categorize participants based on gender and age range, and divide physiological signals based on their time collection into {morning, afternoon, evening}. Figure 5.14 demonstrates the results. First, *Phyre* exhibits similar performance across video genres with minimal stand deviation of 0.014. Notably, *film and animation* yields the highest



(a) Video genres (b) User age (c) Time of a day

Figure 5.14: Performance across impact factors.

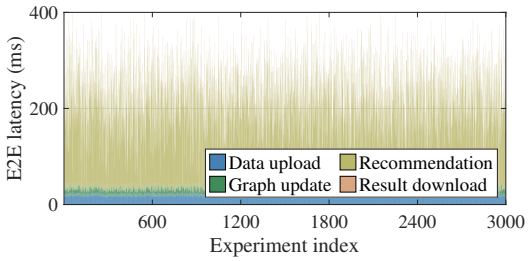


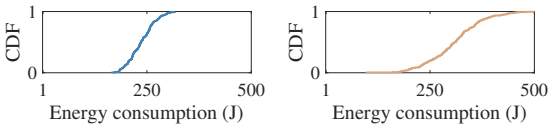
Figure 5.15: End-to-end latency breakdown.

precision at 0.761, while *news* sees the lowest at 0.728. Meanwhile, the recommendation precision of *Phyre* is consistently over 0.732 for all user demographics. Lastly, *Phyre* achieves robust performance at different time intervals within a day, with a marginal standard deviation of less than 0.01. These results indicate that *Phyre* can be used for all types of videos, users, and at any time of a day, without notable performance degradation.

5.8.6 System Overhead

End-to-end latency. The end-to-end latency of *Phyre* is defined as the time interval between a user finishes watching a video and when she receives a new recommendation from the server. It consists of four components: data uploading, graph updating, GCN-based recommendation, and recommendation result downloading. Note that data acquisition and encoding are not included in the end-to-end recommendation latency, as they are executed during the video-watching session. As shown in Figure 5.15, the end-to-end latency for *Phyre* ranges from 135 ms to 414 ms with an average of 225 ms, which is practically acceptable for real-world adoption.

Energy consumption. We evaluate the energy consumption (per minute of video play) of *Phyre* at VR terminals. The main operations include running the adaptive encoding algorithm and uploading the physiological embeddings. The CDF is plotted in Figure 5.16(a). The energy consumption ranges from 166.6 J to 316.9 J with an average of 237.4



(a) Encoding + uploading (b) Raw uploading

Figure 5.16: Energy consumption of encoding and uploading vs uploading raw signals.

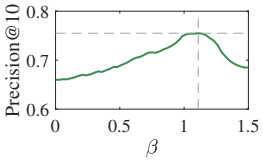


Figure 5.17: Impact of edge embedding weight β.

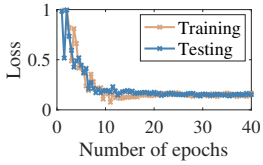


Figure 5.18: Loss with respect to the number of epochs.

J. We also test the energy consumption of uploading the raw signals without encoding, as shown in Figure 5.16(b). The energy consumption ranges from 137.9 J to 500.2 J with an average of 308.5 J. Therefore, a total of 71.1 J energy consumption is saved by applying our adaptive encoding scheme than directly uploading raw signals.

5.8.7 Micro Benchmarks

Edge embedding weight β. We study the impact of, β, an essential hyperparameter of *Phyre* on its recommendation precision. It controls the weight of the second term in the message passing function (Equation 5.1) of the convolutional layers. β determines how much influence the physiological signal casts on the graph learning. As shown in Figure 5.17, the optimal value is found at 1.11. We set β to this value.

Epochs. Lastly, we illustrate the training and testing losses with respect to the number of epochs in Figure 5.18. To save unnecessary training cost and avoid underfitting and overfitting, it is important to determine the ideal number of training epochs. As shown, both losses tend to stabilize after approximately 20 epochs. Therefore, we set it to 20.

5.9 Discussion and Future Work

Other physiological signals. This work aims to investigate the feasibility of incorporating physiological signals into VR video recommendation. We focus on gaze and head movement. Other data such as pupil size and facial expression can be acquired by com-

mercial VR devices and utilized in *Phyre* too. For example, the correlation between pupil size variations and user perception has been established [112, 315]. In our future work, we plan to extend *Phyre* by incorporating other modalities to further enhance its performance.

Privacy considerations. Uploading physiological signals to the cloud server may expose user privacy. Fortunately, in *Phyre*, only the signal embeddings are uploaded, rather than raw measures, which mitigates the privacy concern. Yet, prior work has pointed out that this still poses potential privacy threats [16]; for example, reconstruction attacks can be applied to reveal the original data. Existing privacy-preserving techniques such as homomorphic encryption, differential privacy, and federated learning can provide potential solutions to address this. We plan to integrate these approaches into our design in the future.

Other graph learning models. *Phyre* applies GCN as the basis for graph learning. There are several other graph learning models, such as Graph Autoencoder [139], GraphSAGE [106], and graph attention networks (GAT) [272], which have also demonstrated their advanced performance in a range of graph-based tasks. For example, GAT [272] introduces the attention mechanism into the graph neural networks, improving the adaptability of node importance during message passing. As part of our future work, we plan to modify *Phyre* by incorporating other graph learning models and compare their recommendation performances.

5.10 Conclusion

In this paper, we introduce *Phyre*, a physiological-signal-enhanced video recommender system for VR. To integrate physiological signals into the recommendation framework, we renovate the GCN learning paradigm to extract essential information from these signals. To address the data scarcity problem during model training, we propose a novel

domain adaptation strategy CMCCDA to bridge the discrepancy between the source and target domains. We further develop an energy-efficient adaptive encoding algorithm to improve energy efficiency on the VR device. We demonstrate through a comprehensive evaluation that *Phyre* outperforms state-of-the-art solutions by up to 68.0% in recommendation precision and up to 28.8% in the ranking quality.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This dissertation discusses the potential drawbacks of existing VR techniques, introduces several mechanisms to enhance the security and usability of VR systems and applications, and points out future research directions.

First, we develop two novel user authentication mechanisms designed for VR users, namely BlinKey and SoundLock. As a two-factor authentication scheme, Blinkey employs the user-designed blinking rhythm and unique biometrics exhibited in pupil size variations to fingerprint legitimate users, offering a secure, convenient, and deployable solution. Extending this, we propose SoundLock, an effortless, accurate, and revocable state-of-the-art biometric authentication solution based on a human's auditory-pupillary response mechanism. To recognize legitimate users, carefully designed features are extracted from pupil size reactions to auditory stimuli and used to verify the user's identity. Prototypes are developed to evaluate the security against multiple types of attacks and the usability in various real-world scenarios of proposed authentication schemes.

Second, we present EyeQoE, a QoE prediction model for 360-degree videos using subjects' ocular behaviors. To extract useful features from the behaviors, we propose a novel method that models them into graphs and then builds a GCN-based classifier to learn over graphs. EyeQoE is further equipped with advanced machine learning solutions, including a Siamese network to eliminate irrelevant factors through dedicated model training and a novel domain adaptation framework to enhance real-world performance. We implement EyeQoE and evaluate its performance via extensive in-field studies, which demonstrates its superior performance over state-of-the-art solutions.

Lastly, we introduce *Phyre*, a video recommender system tailored for VR users, leveraging viewers' physiological responses as they engage with VR videos to infer their preferences and thus make future recommendations. To this end, we integrate these new physiological user-video interaction measures into the mainstream recommendation framework and renovate the graph learning-based paradigm to accommodate the new changes. We further develop a novel domain adaptation approach named CMCCDA to address the data scarcity problem for model training and an energy-efficient adaptive encoding scheme to reduce energy consumption on the VR device. We collect a physiological dataset and demonstrate through extensive evaluation that *Phyre* significantly outperforms state-of-the-art schemes.

My future research envisions the following topics. First, extending my research regarding user authentication on VR, I plan to investigate other security threats in VR and propose practical countermeasures. To this end, I intend to conduct comprehensive studies to identify and mitigate these vulnerabilities against adversarial threats such as side-channel attacks, data leakage, and injection attacks. Second, I am dedicated to improving user experience in video streaming and other applications in the VR context with human-centered computing based on our insights of human biometrics and QoE. Lastly, I aim to leverage the VR technology to study and mitigate potential security threats in the real-world, such as autonomous vehicles and pedestrian safety issues, to enable and enhance future applications.

In summary, my research endeavors are dedicated to improving the security of VR devices against malicious attacks and enhancing the efficiency and usability of novel VR applications such as QoE assessment and video recommendation, aiming to bring the distant and "virtual" VR techniques into the tangible "reality" of everyone's daily lives, and applying them to address other real-world problems.

REFERENCES

[1] High-resolution vr headset for professionals - varjo vr-3. https://varjo.com/products/vr-3/, 2022.

[2] Neo 3 pro/pro eye. https://www.picoxr.com/us/neo3.html, 2022.

[3] Vive pro 2. https://www.vive.com/us/product/vive-pro2-full-kit/overview/, 2022.

[4] Vr eye tracking for business: Fove 0 eye tracking vr devkit. https://fove-inc.com/, 2022.

[5] Play station vr2. https://www.playstation.com/en-us/ps-vr2/, 2023.

[6] This is meta quest pro. https://www.meta.com/quest/quest-pro/, 2023.

[7] Vive focus 3. https://www.vive.com/us/product/vive-focus3/overview/, 2023.

[8] Vision pro. https://www.apple.com/apple-vision-pro/, 2024.

[9] Viar 360. Virtual reality in education – how are schools using vr?, 2017.

[10] Yomna Abdelrahman, Florian Mathis, Pascal Knierim, Axel Kettler, Florian Alt, and Mohamed Khamis. Cuevr: Studying the usability of cue-based authentication for virtual reality. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces*, pages 1–9, 2022.

[11] Michael Abehsera. 3 ways virtual reality will transform e-commerce, 2020.

[12] Andy Adler, Richard Youmaran, and Sergey Loyka. Towards a measure of biometric feature information. *Pattern Analysis and Applications*, 12(3):261–270, 2009.

[13] Gediminas Adomavicius and Jingjing Zhang. Impact of data characteristics on recommender systems performance. *ACM Transactions on Management Information Systems (TMIS)*, 3(1):1–17, 2012.

[14] Mayank Agarwal, Mahendra Mehra, Renuka Pawar, and Deven Shah. Secure authentication using dynamic virtual keyboard layout. In *Proceedings of the International Conference Workshop on Emerging Trends in Technology*, ICWET '11, page 288–291. Association for Computing Machinery, February 2011.

[15] Zahid Akhtar, Kamran Siddique, Ajita Rattani, Syaheerah Lebai Lutfi, and Tiago H. Falk. Why is multimedia quality of experience assessment a challenging problem? *IEEE Access*, 7:117897–117915, 2019.

[16] Mohammad Al-Rubaie and J Morris Chang. Privacy-preserving machine learning: Threats and solutions. *IEEE Security & Privacy*, 17(2):49–58, 2019.

[17] Fawaz Alsulaiman and Abdulmotale El Saddik. A novel 3d graphical password schema. In *Proceedings of the 2006 IEEE Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems*, pages 125–128, July 2006.

[18] Cemil Altın. Comparison of different time and frequency domain feature extraction methods on elbow gesture's emg. *European journal of interdisciplinary studies*, 2(3):35–44, 2016.

[19] Patricia Arias-Cabarcos, Thilo Habrich, Karen Becker, Christian Becker, and Thorsten Strufe. Inexpensive brainwave authentication: new techniques and insights on user acceptance. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 55–72, 2021.

[20] Bruno Arsioli and Pedro Dedin. Machine learning applied to multifrequency data in astrophysics: blazar classification. *Monthly Notices of the Royal Astronomical Society*, 498(2):1750–1764, 2020.

[21] Ilhan Aslan, Andreas Uhl, Alexander Meschtscherjakov, and Manfred Tscheligi. Mid-air authentication gestures: An exploration of authentication based on palm and finger motions. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, page 311–318. Association for Computing Machinery, November 2014.

[22] Nick Babich. How vr in education will change how we learn and teach, 2019.

[23] Lu Bai, Wu Lin, Abhishek Gupta, and Yew-Soon Ong. From multitask gradient descent to gradient-free evolutionary multitasking: a proof of faster convergence. *IEEE Transactions on Cybernetics*, 2021.

[24] Christos G Bampis, Zhi Li, and Alan C Bovik. Continuous prediction of streaming video qoe using dynamic networks. *IEEE Signal Processing Letters*, 24(7):1083–1087, 2017.

[25] John Barbur. Pupillary responses to grating stimuli. *Journal of Psychophysiology*, 1991.

[26] Claude Barral and Assia Tria. Fake fingers in fingerprint recognition: Glycerin supersedes gelatin. In *Formal to Practical Security*, pages 57–69. Springer, 2009.

[27] Jackson Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin*, 91(2):276, 1982.

[28] Roman Bednarik, Tomi Kinnunen, Andrei Mihaila, and Pasi Fränti. Eye-movements as a biometric. In *Scandinavian conference on image analysis*, pages 780–789. Springer, 2005.

[29] Mathias Benedek, Robert Stoiser, Sonja Annerer-Walcher, and Christof Körner. Eye behavior associated with internally versus externally directed cognition. *Frontiers in Psychology*, 8, 06 2017.

[30] Oliver Bergamin and Randy H Kardon. Latency of the pupil light reflex: sample rate, stimulus intensity, and variation in normal subjects. *Investigative Ophthalmology & Visual Science*, 44(4):1546–1554, 2003.

[31] Gordon Bill, Elisabeth Whyte, Jason Griffin, and Kathryn Scherf. Measuring sensitivity to eye gaze cues in naturalistic scenes: Presenting the eye gaze focus database. *International Journal of Methods in Psychiatric Research*, 29, 07 2020.

[32] Daniel Billsus and Michael J Pazzani. User modeling for adaptive news access. *User modeling and user-adapted interaction*, 10:147–180, 2000.

[33] Ariel Bogle. ebay launches a world-first virtual reality department store, 2020.

[34] Bhavana Borkar, Shiba Sheikh, and PD Kaware. 4d password mechanism. In *Imperial Journal of Interdisciplinary Research*, volume 2, pages 240–245, 2016.

[35] Fadi Boutros, Naser Damer, Kiran Raja, Raghavendra Ramachandra, Florian Kirchbuchner, and Arjan Kuijper. Iris and periocular biometrics for head mounted displays: Segmentation, recognition, and synthetic data generation. *Image and Vision Computing*, 104:104007, 2020.

[36] Steven Brand. How virtual reality is changing the manufacturing game, 2020.

[37] Davina Bristow, John-Dylan Haynes, Richard Sylvester, Christopher D. Frith, and Geraint Rees. Blinking suppresses the neural response to unchanging retinal stimulation. *Current Biology*, 15(14):1296 – 1300, June 2005.

[38] Encyclopaedia Britannica. Iris. 2020.

[39] John Lott Brown. Flash blindness. *American Journal of Ophthalmology*, 60(3):505–520, 1965.

[40] Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12:331–370, 2002.

[41] Hongyun Cai, Vincent Zheng, and Kevin Chang. A comprehensive survey of graph embedding: Problems, techniques and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30, 09 2017.

[42] Liang Cai and Hao Chen. Touchlogger: Inferring keystrokes on touch screen from smartphone motion. In *6th USENIX Workshop on Hot Topics in Security (HotSec 11)*, 2011.

[43] Christos Calcanis, Vic Callaghan, Michael Gardner, and Matthew Walker. Towards end-user physiological profiling for video recommendation engines. 2008.

[44] Erion Çano and Maurizio Morisio. Hybrid recommender systems: A systematic literature review. *Intelligent data analysis*, 21(6):1487–1524, 2017.

[45] Virginio Cantoni, Chiara Galdi, Michele Nappi, Marco Porta, and Daniel Riccio. Gant: Gaze analysis technique for human identification. *Pattern Recognition*, 48(4):1027–1038, 2015.

[46] Lucia Cascone, Carlo Medaglia, Michele Nappi, and Fabio Narducci. Pupil size as a soft biometrics for age and gender classification. *Pattern Recognition Letters*, 140:238–244, 2020.

[47] Raymundo Cassani, Marc-Antoine Moinnereau, and Tiago H. Falk. A neurophysiological sensor-equipped head-mounted display for instrumental qoe assessment of immersive multimedia. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2018.

[48] Pan Chan, Tzipora Halevi, and Nasir Memon. Glass otp: Secure and convenient user authentication on google glass. In *International Conference on Financial Cryptography and Data Security*, pages 298–308. Springer, 2015.

[49] Eunhee Chang, Hyun Taek Kim, and Byounghyun Yoo. Virtual reality sickness: a review of causes and measurements. *International Journal of Human–Computer Interaction*, 36(17):1658–1682, 2020.

[50] Supply Chain Game Changer. Virtual reality (VR) is enhancing e-commerce shopping!, 2020.

[51] Da Chen, Yuefeng Chen, Yuhong Li, Feng Mao, Yuan He, and Hui Xue. Self-supervised learning for few-shot image classification. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1745–1749, 2021.

[52] Sijia Chen, Yingxue Zhang, Yiming Li, Zhenzhong Chen, and Zhou Wang. Spherical structural similarity index for objective omnidirectional video quality assessment. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2018.

[53] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

[54] Yimin Chen, Jingchao Sun, Rui Zhang, and Yanchao Zhang. Your song your way: Rhythm-based two-factor authentication for multi-touch mobile devices. In *Proceedings of the 2015 IEEE Conference on Computer Communications*, pages 2686–2694, April 2015.

[55] Yuxin Chen, Zhuolin Yang, Ruben Abbou, Pedro Lopes, Ben Y Zhao, and Haitao Zheng. User authentication via electrical muscle stimulation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[56] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[57] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference*

on Computer Vision and Pattern Recognition (CVPR'05), volume 1, pages 539–546 vol. 1, 2005.

[58] Christoforos Christoforou, Spyros Christou-Champi, Fofi Constantinidou, and Maria Theodorou. From the eyes and the heart: a novel eye-gaze metric that predicts video preferences of a large audience. *Frontiers in psychology*, 6:118967, 2015.

[59] Christoforos Christoforou, Timothy C Papadopoulos, Fofi Constantinidou, and Maria Theodorou. Your brain on the movies: a computational approach for predicting box-office performance from viewer's brain responses to movie trailers. *Frontiers in neuroinformatics*, 11:72, 2017.

[60] Viviane Clay, Peter König, and Sabine Koenig. Eye tracking in virtual reality. *Journal of eye movement research*, 12(1), 2019.

[61] Jennifer Clopton. Virtual reality brings new vision to health care, 2020.

[62] John Cochary. Common noise levels. `https://noiseawareness.org/info-center/common-noise-levels/`, 2021.

[63] Thomas M Cover and Joy A Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

[64] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.

[65] Gabriela Csurka. *Domain Adaptation for Visual Applications: A Comprehensive Survey*. 09 2017.

[66] Ronaldo Martins da Costa and Adilson Gonzaga. Dynamic features for iris recognition. *IEEE transactions on systems, man, and cybernetics, part B (cybernetics)*, 42(4):1072–1082, 2012.

[67] Roberto Irajá Tavares da Costa Filho, Marcelo Caggiani Luizelli, Maria Torres Vega, Jeroen van der Hooft, Stefano Petrangeli, Tim Wauters, Filip De Turck, and Lu-

ciano Paschoal Gaspary. Predicting the performance of virtual reality video streaming in mobile networks. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 270–283, 2018.

[68] Sauvik Das, Gierad Laput, Chris Harrison, and Jason I Hong. Thumprint: Socially-inclusive local group authentication through shared secret knocks. In *Proceedings of the 2017 chi conference on human factors in computing systems*, pages 3764–3774, 2017.

[69] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296, 2010.

[70] Toon De Pessemier, Ine Coppens, and Luc Martens. Evaluating facial recognition services as interaction technique for recommender systems. *Multimedia Tools and Applications*, 79(31):23547–23570, 2020.

[71] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering, 2017.

[72] Yaling Deng, Ye Wang, Liming Xu, Xiangli Meng, and Lingxiao Wang. Do you like it or not? identifying preference using an electroencephalogram during the viewing of short videos. *PsyCh Journal*, 12(3):421–429, 2023.

[73] Statista Research Department. Global virtual reality device shipments by vendor, 2020.

[74] Chesner Désir, Simon Bernard, Caroline Petitjean, and Heutte Laurent. One class random forests. *Pattern Recognition*, 46(12):3490–3506, 2013.

[75] Yancarlos Diaz, Cecilia O Alm, Ifeoma Nwogu, and Reynold Bailey. Towards an affective video recommendation system. In *2018 IEEE International Conference on*

Pervasive Computing and Communications Workshops (PerCom Workshops), pages 137–142. IEEE, 2018.

[76] Nguyen Minh Duc and Bui Quang Minh. Your face is not your password face authentication bypassing lenovo–asus–toshiba. *Black Hat Briefings*, 4:158, 2009.

[77] Tho Nguyen Duc, Chanh Minh Tran, Tan Phan-Xuan, and Eiji Kamioka. Modeling of cumulative qoe in on-demand video services: Role of memory effect and degree of interest. *Future Internet*, 11:171, 2019.

[78] Andrew T. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer International Publishing, 2017.

[79] Simon Eberz, Giulio Lovisotto, Kasper B Rasmussen, Vincent Lenders, and Ivan Martinovic. 28 blinks later: Tackling practical challenges of eye movement biometrics. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1187–1199, 2019.

[80] Simon Eberz, Kasper Bonne Rasmussen, Vincent Lenders, and Ivan Martinovic. Preventing lunchtime attacks: Fighting insider threats with eye movement biometrics. In *The Network and Distributed System Security Symposium*, February 2015.

[81] Darragh Egan, Sean Brennan, John Barrett, Yuansong Qiao, Christian Timmerer, and Niall Murray. An evaluation of heart rate and electrodermal activity as an objective qoe evaluation method for immersive virtual reality environments. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2016.

[82] Malin Eiband, Mohamed Khamis, Emanuel Von Zezschwitz, Heinrich Hussmann, and Florian Alt. Understanding shoulder surfing in the wild: Stories from users and observers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 4254–4265, 2017.

[83] eMarkerter. Us virtual and augmented reality users 2020, 2020.

[84] Ulrich Engelke, Marcus Barkowsky, Patrick Le Callet, and Hans-Jürgen Zepernick. Modelling saliency awareness for objective video quality assessment. In *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 212–217, 2010.

[85] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. In *The world wide web conference*, pages 417–426, 2019.

[86] Wells Fargo. Biometric authentication, 2020.

[87] Michael Fauscette. Biometrics are coming so are security concerns, 2020.

[88] Zesong Fei, Fei Wang, Jing Wang, and Xiang Xie. Qoe evaluation methods for 360-degree vr video transmission. *IEEE Journal of Selected Topics in Signal Processing*, PP:1–1, 11 2019.

[89] Caleb Finch. Manufacturing with vr becoming a (virtual) reality, 2018.

[90] Lauren Fink, Jaana Simola, Alessandro Tavano, Elke B Lange, Sebastian Wallot, and Bruno Laeng. From pre-processing to advanced dynamic modeling of pupil data. 2021.

[91] International Organization for Standardization. *Image Safety - Reducing the Incidence of Undesirable Biomedical Effects Caused by Visual Image Sequences*. International Workshop Agreement. ISO, 2005.

[92] fotonVR. Is virtual reality headset harmful to the eyes? `https://fotonvr.com/is-virtual-reality-headset-harmful-to-the-eyes/`, 2020.

[93] Fove. Vr eye tracking for business: Fove 0 eye tracking vr devkit. `https://fove-inc.com/`, 2022.

[94] Markus Funk, Karola Marky, Iori Mizutani, Mareike Kritzler, Simon Mayer, and Florian Michahelles. Lookunlock: Using spatial-targets for user-authentication on

hmds. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

[95] Chiara Galdi and Michele Nappi. *Eye Movement Analysis in Biometrics*, pages 171–183. 01 2019.

[96] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, January 2016.

[97] Albizu Garcia. Is virtual reality the future of social networking?, 2019.

[98] Ceenu George, Mohamed Khamis, Daniel Buschek, and Heinrich Hussmann. Investigating the third dimension for authentication in immersive virtual reality and in the real world. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 277–285. IEEE, 2019.

[99] Ceenu George, Mohamed Khamis, Emanuel von Zezschwitz, Marinus Burger, Henri Schmidt, Florian Alt, and Heinrich Hussmann. Seamless and secure vr: Adapting and evaluating established authentication systems for virtual reality. NDSS, 2017.

[100] Cristos Goodrow. On youtube's recommendation system, 2021.

[101] Stephen Gossett. Virtual reality in education: An overview, 2020.

[102] GroupLens. Movielens 100k dataset. `https://grouplens.org/datasets/movielens/100k/`, 2023.

[103] GroupLens. Movielens 1m dataset. `https://grouplens.org/datasets/movielens/1m/`, 2023.

[104] Alan LV Guedes, Roberto G de A Azevedo, Pascal Frossard, Sérgio Colcher, and Simone Diniz Junqueira Barbosa. Subjective evaluation of 360-degree sensory experiences. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2019.

[105] Kelly S Hale and Kay M Stanney. *Handbook of virtual environments: Design, implementation, and applications*. CRC Press, 2014.

[106] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[107] Jinkun Han, Wei Li, Zhipeng Cai, and Yingshu Li. Multi-aggregator time-warping heterogeneous graph neural network for personalized micro-video recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 676–685, 2022.

[108] HandWiki. Damped sine wave. `https://handwiki.org/wiki/index.php?title=Damped_sine_wave&oldid=75474`, 2022.

[109] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648, 2020.

[110] Melanie Heck, Janick Edinger, Jonathan Bünemann, and Christian Becker. Exploring gaze-based prediction strategies for preference detection in dynamic interface elements. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 129–139, 2021.

[111] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[112] Eckhard H Hess. The role of pupil size in communication. *Scientific American*, 233(5):110–119, 1975.

[113] Bert Hoeks and Willem JM Levelt. Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research methods, instruments, & computers*, 25(1):16–26, 1993.

[114] Corey D Holland and Oleg V Komogortsev. Biometric identification via eye movement scanpaths in reading. In *2011 International Joint Conference on Biometrics*, pages 1–8, October 2011.

[115] Corey D Holland and Oleg V Komogortsev. Complex eye movement pattern biometrics: Analyzing fixations and saccades. pages 1–8, June 2013.

[116] HTC. Htc vive pro eye, 2020.

[117] Yanxiang Huang, Bin Cui, Jie Jiang, Kunqian Hong, Wenyu Zhang, and Yiran Xie. Real-time video recommendation exploration. In *Proceedings of the 2016 international conference on management of data*, pages 35–46, 2016.

[118] Ben Hutchins, Anudeep Reddy, Wenqiang Jin, Michael Zhou, Ming Li, and Lei Yang. Beat-pin: A user authentication mechanism for wearable devices through secret beats. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ASIACCS '18, page 101–115. Association for Computing Machinery, May 2018.

[119] Andrew Hutchinson. Facebook begins user testing of new 'horizon' vr social platform, 2020.

[120] Mordor Intelligence. Virtual reality (vr) market - growth, trends, and forecast (2020 - 2025), 2020.

[121] Mohsina Ishrat and Pawanesh Abrol. Eye movement analysis in the context of external stimuli effect. In *2017 International Conference on Informatics, Health Technology (ICIHT)*, pages 1–6, 2017.

[122] Folasade Olubusola Isinkaye, Yetunde O Folajimi, and Bolande Adefowoke Ojokoh. Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal*, 16(3):261–273, 2015.

[123] ITU-T. Definition of quality of experience (qoe), 2007.

[124] Anil K Jain, Karthik Nandakumar, and Arun Ross. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern recognition letters*, 79:80–105, 2016.

[125] Kapil Jain and Nirbhay Pherwani. Virtual reality based user authentication system. In *International Journal of Science Technology Engineering*, volume 4, pages 49–53, 2017.

[126] Samay Jain, Greg J Siegle, Chen Gu, Charity G Moore, Larry S Ivanco, Stephanie Studenski, J Timothy Greenamyre, and Stuart R Steinhauer. Pupillary unrest correlates with arousal symptoms and motor signs in parkinson disease. *Movement disorders*, 26(7):1344–1347, 2011.

[127] Hao Jiang, Wenjie Wang, Yinwei Wei, Zan Gao, Yinglong Wang, and Liqiang Nie. What aspect do you like: Multi-scale time-aware user interest modeling for micro-video recommendation. In *Proceedings of the 28th ACM International conference on Multimedia*, pages 3487–3495, 2020.

[128] Siddhartha Joshi, Yin Li, Rishi M Kalwani, and Joshua I Gold. Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron*, 89(1):221–234, 2016.

[129] Hanseul Jun, Mark Roman Miller, Fernanda Herrera, Byron Reeves, and Jeremy N Bailenson. Stimulus sampling with 360-videos: Examining head movements, arousal, presence, simulator sickness, and preference on a large sample of participants and videos. *IEEE Transactions on Affective Computing*, 13(3):1416–1425, 2020.

[130] Joakim Karlén. Eye-tracking is virtual reality's next frontier. `https://venturebeat.com/games/eye-tracking-is-virtual-realitys-next-frontier/`.

[131] Shunichi Kasahara, Shohei Nagai, and Jun Rekimoto. First person omnidirectional video: System design and implications for immersive experience. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, TVX '15, page 33–42, New York, NY, USA, 2015. Association for Computing Machinery.

[132] Conor Keighrey, Ronan Flynn, Siobhan Murray, and Niall Murray. A qoe evaluation of immersive augmented and virtual reality speech language assessment applications. 06 2017.

[133] Donghyun Kim, Sunghwan Choi, Sangil Park, and Kwanghoon Sohn. Stereoscopic visual fatigue measurement based on fusional response curve and eye-blinks. In *2011 17th International Conference on Digital Signal Processing (DSP)*, pages 1–6, 2011.

[134] Hak Gu Kim, Heoun-Taek Lim, Sangmin Lee, and Yong Man Ro. Vrsa net: Vr sickness assessment considering exceptional motion for 360 vr video. *IEEE transactions on image processing*, 28(4):1646–1660, 2018.

[135] Si Jung Kim, Teemu H Laine, and Hae Jung Suk. Presence effects in virtual reality based on user characteristics: Attention, enjoyment, and memory. *Electronics*, 10(9):1051, 2021.

[136] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[137] 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*, ETRA '10, page 187–190. Association for Computing Machinery, March 2010.

[138] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv e-prints*, page arXiv:1609.02907, September 2016.

[139] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

[140] Jeff Klingner, Rakshit Kumar, and Pat Hanrahan. Measuring the task-evoked pupillary response with a remote eye tracker. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, pages 69–72, 2008.

[141] Yehuda Koren, Steffen Rendle, and Robert Bell. Advances in collaborative filtering. *Recommender systems handbook*, pages 91–142, 2021.

[142] Nadia Kovics. Virtual reality in military, 2020.

[143] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[144] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[145] Alexander Kupin, Benjamin Moeller, Yijun Jiang, Natasha Kholgade Banerjee, and Sean Banerjee. Task-driven biometric authentication of users in virtual reality (vr) environments. In *International conference on multimedia modeling*, pages 55–67. Springer, 2019.

[146] Oscar D Lara and Miguel A Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys Tutorials*, 15(3):1192–1209, March 2013.

[147] Eric C. Larson, Cuong Vu, and Damon M. Chandler. Can visual fixation patterns improve image fidelity assessment? In *2008 15th IEEE International Conference on Image Processing*, pages 2572–2575, 2008.

[148] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19):2483–2498, 2007.

[149] Min-Seok Lee, Seok Ho Baek, Yoo-Jeong Shim, and Myeong-Jin Lee. Analysis of object-centric visual preference in 360-degree videos. *IEEE Access*, 9:98026–98038, 2021.

[150] Chen Li, Mai Xu, Xinzhe Du, and Zulin Wang. Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 932–940, New York, NY, USA, 2018. Association for Computing Machinery.

[151] Chen Li, Mai Xu, Lai Jiang, Shanyi Zhang, and Xiaoming Tao. Viewport proposal cnn for 360deg video quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[152] Jingjie Li, Kassem Fawaz, and Younghyun Kim. Velody: Nonlinear vibration challenge-response for resilient user authentication. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1201–1213, 2019.

[153] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks, 2017.

[154] Hsin-I Liao, Shunsuke Kidani, Makoto Yoneya, Makio Kashino, and Shigeto Furukawa. Correspondences among pupillary dilation response, subjective salience of sounds, and loudness. *Psychonomic bulletin & review*, 23(2):412–425, 2016.

[155] Feng Lin, Kun Woo Cho, Chen Song, Wenyao Xu, and Zhanpeng Jin. Brain password: A secure and truly cancelable brain biometrics for smart headwear. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, pages 296–309, 2018.

[156] Fan Liu, Zhiyong Cheng, Lei Zhu, Zan Gao, and Liqiang Nie. Interest-aware message-passing gcn for recommendation. In *Proceedings of the Web Conference 2021*, pages 1296–1305, 2021.

[157] Hantao Liu and Ingrid Heynderickx. Visual attention in objective image quality assessment: Based on eye-tracking data. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(7):971–982, 2011.

[158] Jian Liu, Chen Wang, Yingying Chen, and Nitesh Saxena. Vibwrite: Towards finger-input authentication on ubiquitous surfaces via physical vibration. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 73–87, 2017.

[159] Xinxiong Liu, Jing Zhang, Guoxiang Hou, and Zenan Wang. Virtual reality and its application in military. In *IOP Conference Series: Earth and Environmental Science*, volume 170, page 032155. IOP Publishing, 2018.

[160] Holger Lüdtke, Barbara Wilhelm, Martin Adler, Frank Schaeffel, and Helmut Wilhelm. Mathematical procedures in data recording and processing of pupillary fatigue waves. *Vision research*, 38(19):2889–2896, 1998.

[161] Alessandra Lumini and Loris Nanni. A clustering method for automatic biometric template selection. *Pattern Recognition*, 39(3):495–497, 2006.

[162] Shiqing Luo, Anh Nguyen, Chen Song, Feng Lin, Wenyao Xu, and Zhisheng Yan. Oculock: Exploring human visual system for authentication in virtual reality head-mounted display. In *2020 Network and Distributed System Security Symposium (NDSS)*, 2020.

[163] Jacqueline Lykstad, Vamsi Reddy, and Andrew Hanna. Neuroanatomy, pupillary dilation pathway. 2018.

[164] Dong Ma, Guohao Lan, Mahbub Hassan, Wen Hu, Mushfika B Upama, Ashraf Uddin, and Moustafa Youssef. Solargest: Ubiquitous and battery-free gesture recogni-

tion using solar cells. In *Proceedings of the 25th Annual International Conference on Mobile Computing and Networking*, MobiCom '19. Association for Computing Machinery, August 2019.

[165] Abhishek Mahata, Nandini Saini, Sneha Saharawat, and Ritu Tiwari. Intelligent movie recommender system using machine learning. In *Intelligent Human Computer Interaction: 8th International Conference, IHCI 2016, Pilani, India, December 12-13, 2016, Proceedings 8*, pages 94–110. Springer, 2017.

[166] Anthony J Mansfield and James L Wayman. Best practices in testing and reporting performance of biometric devices. 2002.

[167] Inés P Mariño and Joaquín Míguez. An approximate gradient-descent method for joint parameter estimation and synchronization of coupled chaotic systems. *Physics Letters A*, 351(4-5):262–267, 2006.

[168] Jesus Martínez-Navarro, Enrique Bigné, Jaime Guixeres, Mariano Alcañiz, and Carmen Torrecilla. The influence of virtual reality in e-commerce. *Journal of Business Research*, 100:475–482, 2019.

[169] Florian Mathis, Hassan Ismail Fawaz, and Mohamed Khamis. Knowledge-driven biometric authentication in virtual reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2020.

[170] Florian Mathis, John Williamson, Kami Vaniea, and Mohamed Khamis. Rubikauth: Fast and secure authentication in virtual reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2020.

[171] Sebastiaan Mathôt. Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, 1(1), 2018.

[172] Tao Mei, Bo Yang, Xian-Sheng Hua, and Shipeng Li. Contextual video recommendation by multimodal relevance and user feedback. *ACM Transactions on Information Systems (TOIS)*, 29(2):1–24, 2011.

[173] Gebremariam Mesfin, Nadia Hussain, Alexandra Covaci, and Gheorghita Ghinea. Using eye tracking and heart-rate activity to examine crossmodal correspondences qoe in mulsemedia. *AACM Transactions on Multimedia Computing, Communications, and Applications*, 15(2), June 2019.

[174] Abir Mhenni, Christophe Rosenberger, Estelle Cherrier, and Najoua Essoukri Ben Amara. Keystroke template update with adapted thresholds. In *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 483–488. IEEE, 2016.

[175] Microsoft. Microsoft hololens — mixed reality technology for business. `https://www.microsoft.com/en-us/hololens/`, 2022.

[176] Mixkit. Free assets for your next video project. `https://mixkit.co/`, 2022.

[177] Judi Moline et al. Virtual reality for health care: a survey. *Studies in health technology and informatics*, pages 3–34, 1997.

[178] Jinyoung Moon, Youngrae Kim, Hyungjik Lee, Changseok Bae, and Wan Chul Yoon. Extraction of user preference for video stimuli using eeg-based user responses. *ETRI Journal*, 35(6):1105–1114, 2013.

[179] Raymond J Mooney and Loriene Roy. Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 195–204, 2000.

[180] Y Morad, H Lemberg, N Yofe, and Y Dagan. Pupillography as an objective indicator of fatigue. *Current eye research*, 21(1):535—542, July 2000.

[181] Michael Morozov. Virtual reality in manufacturing, 2019.

[182] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. *CoRR*, abs/2001.09691, 2020.

[183] Kevin P Murphy et al. Naive bayes classifiers. *University of British Columbia*, 18(60):1–8, 2006.

[184] Muhammad-Sajid Mushtaq and Abdelhamid Mellouk. 5 - qoe and power-saving model for 5g network. In Muhammad-Sajid Mushtaq and Abdelhamid Mellouk, editors, *Quality of Experience Paradigm in Multimedia Services*, pages 127–160. Elsevier, 2017.

[185] Tahrima Mustafa, Richard Matovu, Abdul Serwadda, and Nicholas Muirhead. Unsure how to authenticate on your vr headset? come on, use your head! In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, pages 23–30, 2018.

[186] Satoshi Nakakoga, Kengo Shimizu, Junya Muramatsu, Takashi Kitagawa, Shigeki Nakauchi, and Tetsuto Minami. Pupillary response reflects attentional modulation to sound after emotional arousal. *Scientific reports*, 11(1):1–10, 2021.

[187] Yoshiko Nakamura. Measurement of pupillary unrest in eyestrain. *Japanese journal of ophthalmology*, 40(4):533–539, 1996.

[188] Toan Nguyen and Nasir Memon. Tap-based user authentication for smartwatches. *Computers and Security*, 78:174–186, September 2018.

[189] Nahumi Nugrahaningsih and Marco Porta. Pupil size as a biometric trait. In *International Workshop on Biometric Authentication*, pages 222–233. Springer, 2014.

[190] Oculus. Getting started with your oculus quest. `https://store.facebook.com/help/quest/articles/getting-started/getting-started-with-quest-2/`.

[191] Oculus. Facebook horizon, 2020.

[192] Jonny O'Dwyer, Niall Murray, and Ronan Flynn. Eye-based continuous affect prediction, 2020.

[193] Bank of America. Access your account securely with fingerprint sign-in, 2020.

[194] Internet of Business. Alibaba launches vr pay, gives virtual reality payments the nod, 2020.

[195] The Database of Useful Biological Numbers. Average duration of a single eye blink, 2001.

[196] Sean O'Kane. Tesla starts using in-car camera for autopilot driver monitoring - the verge, 2021.

[197] Marlis Ontivero-Ortega, Agustin Lage-Castellanos, Giancarlo Valente, Rainer Goebel, and Mitchell Valdes-Sosa. Fast gaussian naïve bayes for searchlight classification analysis. *Neuroimage*, 163:471–479, 2017.

[198] Randall C O'Reilly. Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural computation*, 8(5):895–938, 1996.

[199] Poojan Oza and Vishal Patel. Active authentication using an autoencoder regularized cnn-based one-class classifier. pages 1–8, 05 2019.

[200] Poojan Oza and Vishal M Patel. One-class convolutional neural network. *IEEE Signal Processing Letters*, 26(2):277–281, 2018.

[201] Deuk Hee Park, Hyea Kyeong Kim, Il Young Choi, and Jae Kyeong Kim. A literature review and classification of recommender systems research. *Expert systems with applications*, 39(11):10059–10072, 2012.

[202] Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. In *The adaptive web: methods and strategies of web personalization*, pages 325–341. Springer, 2007.

[203] Viswa Mani Kiran Peddinti and Prakriti Chintalapoodi. Domain adaptation in sentiment analysis of twitter. AAAIWS'11-05, page 44–49. AAAI Press, 2011.

[204] Ken Pfeuffer, Matthias J Geiger, Sarah Prange, Lukas Mecke, Daniel Buschek, and Florian Alt. Behavioural biometrics in vr: Identifying people from body motion and relations in virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[205] Paulo Henrique Pisani, Abir Mhenni, Romain Giot, Estelle Cherrier, Norman Poh, André Carlos Ponce de Leon Ferreira de Carvalho, Christophe Rosenberger, and Najoua Essoukri Ben Amara. Adaptive biometric systems: Review and perspectives. *ACM Computing Surveys (CSUR)*, 52(5):1–38, 2019.

[206] Norman Poh, Josef Kittler, Sebastien Marcel, Driss Matrouf, and Jean-Francois Bonastre. Model and score adaptation for biometric systems: Coping with device interoperability and changing acquisition conditions. In *2010 20th International Conference on Pattern Recognition*, pages 1229–1232. IEEE, 2010.

[207] Norman Poh, Ajita Rattani, and Fabio Roli. Critical analysis of adaptive biometric systems. *IET biometrics*, 1(4):179–187, 2012.

[208] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006.

[209] Simone Porcu, Alessandro Floris, Jan-Niklas Voigt-Antons, Luigi Atzori, and Sebastian Möller. Estimation of the quality of experience during video streaming from facial expression and gaze direction. *IEEE Transactions on Network and Service Management*, 17(4):2702–2716, 2020.

[210] Dale Purves, George J Augustine, David Fitzpatrick, William Hall, Anthony-Samuel LaMantia, and Leonard White. *Neurosciences*. De Boeck Supérieur, 2019.

[211] Sundaramurthi Rajarajan, K Kavitha Maheswari, R Hemapriya, and S Sriharilakshmi. Shoulder surfing resistant virtual keyboard for internet banking. *World Applied Sciences Journal*, 31(7):1297–1304, 2014.

[212] Christian Rathgeb and Andreas Uhl. A survey on biometric cryptosystems and cancelable biometrics. *EURASIP journal on information security*, 2011(1):1–25, 2011.

[213] Ajita Rattani, Biagio Freni, Gian Luca Marcialis, and Fabio Roli. Template update methods in adaptive biometric systems: A critical review. In *International Conference on Biometrics*, pages 847–856. Springer, 2009.

[214] Ievgen Redko, Amaury Habrard, Emilie Morvant, Marc Sebban, and Younès Bennani. 2 - domain adaptation problem. In Ievgen Redko, Amaury Habrard, Emilie Morvant, Marc Sebban, and Younès Bennani, editors, *Advances in Domain Adaptation Theory*, pages 21–36. Elsevier, 2019.

[215] Jacob Reimer, Matthew J McGinley, Yang Liu, Charles Rodenkirch, Qi Wang, David A McCormick, and Andreas S Tolias. Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature communications*, 7(1):1–7, 2016.

[216] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[217] ReportLinker. Virtual reality market size, share trends analysis report by technology, by device, by component, by application, by region and segment forecasts, 2022 - 2030, 2022.

[218] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186, 1994.

[219] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Techniques, applications, and challenges. *Recommender Systems Handbook*, pages 1–35, 2021.

[220] Ioannis Rigas, George Economou, and Spiros Fotopoulos. Biometric identification based on the eye movements and graph matching techniques. *Pattern Recognition Letters*, 33(6):786–792, 2012.

[221] Ioannis Rigas and Oleg Komogortsev. Biometric recognition via probabilistic spatial projection of eye movement trajectories in dynamic visual environments. *Information Forensics and Security, IEEE Transactions on*, 9:1743–1754, 2014.

[222] Giuseppe Riva and Brenda K Wiederhold. *The new dawn of virtual reality in health care: Medical simulation and experiential interface*, volume 219 of *SHTI '15*. IOS Press, 2015.

[223] Sol Rogers. Seven reasons why eye-tracking will fundamentally change vr, 2019.

[224] Lior Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33:1–39, 02 2010.

[225] Fabio Roli, Luca Didaci, and Gian Luca Marcialis. Adaptive biometric systems that can improve with use. *Advances in Biometrics*, pages 447–471, 2008.

[226] Marius Rubo and Matthias Gamer. Social content and emotional valence modulate gaze fixations in dynamic scenes. *Scientific Reports*, 8, 02 2018.

[227] Napa Sae-Bae and Nasir Memon. Distinguishability of keystroke dynamic template. *PloS one*, 17(1):e0261291, 2022.

[228] Kenneth S Saladin. *Anatomy and physiology*. McGraw-Hill, 2012.

[229] Débora Salgado, Felipe Martins, Thiago Braga Rodrigues, Conor Keighrey, Ronan Flynn, Eduardo Naves, and Niall Murray. A qoe assessment method based on eda, heart rate and eeg of a virtual reality assistive technology system. pages 517–520, 06 2018.

[230] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[231] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*, pages 291–324. Springer, 2007.

[232] J Ben Schafer, Joseph Konstan, and John Riedl. Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce*, pages 158–166, 1999.

[233] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer, 2018.

[234] Stefan Schneegass, Youssef Oualil, and Andreas Bulling. Skullconduct: Biometric user identification on eyewear computers using bone conduction through the skull. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1379–1384, 2016.

[235] Andy Schumann, Stephanie Kietzer, Juliane Ebel, and Karl Jürgen Bär. Sympathetic and parasympathetic modulation of pupillary unrest. *Frontiers in neuroscience*, 14:178, 2020.

[236] ABM Fahim Shahriar, Mahedee Zaman Moon, Hasan Mahmud, and Kamrul Hasan. Online product recommendation system by using eye gaze data. In *Proceedings of the International Conference on Computing Advancements*, pages 1–7, 2020.

[237] Yiran Shen, Hongkai Wen, Chengwen Luo, Weitao Xu, Tao Zhang, Wen Hu, and Daniela Rus. Gaitlock: Protect virtual and augmented reality headsets using gait. *IEEE Transactions on Dependable and Secure Computing*, 16(3):484–497, 2018.

[238] Chuan Shi, Binbin Hu, Wayne Xin Zhao, and S Yu Philip. Heterogeneous information network embedding for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 31(2):357–370, 2018.

[239] Ashutosh Singla, Stephan Fremerey, Werner Robitza, Pierre Lebreton, and Alexander Raake. Comparison of subjective quality evaluation for hevc encoded omnidi-

rectional videos at different bit-rates for uhd and fhd resolution. pages 511–519, 10 2017.

[240] Ivo Sluganovic, Marc Roeschlin, Kasper B Rasmussen, and Ivan Martinovic. Using reflexive eye movements for fast challenge-response authentication. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1056–1067, 2016.

[241] James D Smith, Gerald A Masek, Lester Y Ichinose, Takeshi Watanabe, and Lawrence Stark. Single neuron activity in the pupillary system. *Brain research*, 24(2):219–234, 1970.

[242] Tsz Yan So, Man Yi Erica Li, and Hakwan Lau. Between-subject correlation of heart rate variability predicts movie preferences. *PloS one*, 16(2):e0247625, 2021.

[243] Virtual Reality Society. Virtual reality in the military, 2017.

[244] Yunpeng Song, Zhongmin Cai, and Zhi-Li Zhang. Multi-touch authentication using hand geometry and behavioral information. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy*, pages 357–372, May 2017.

[245] Steam Community. Vive can cause snow blindness if you don't do this. `https://steamcommunity.com/app/358040/discussions/0/365163686083238173/`, 2016.

[246] Scott Stein. Watching me, watching you: How eye tracking is coming to vr and beyond. `https://www.cnet.com/tech/computing/watching-me-watching-you-how-eye-tracking-is-coming-to-vr-and-beyond/`.

[247] Scott Stein. Eye tracking is the next phase for vr, ready or not, 2020.

[248] Sophie Stephenson, Bijeeta Pal, Stephen Fan, Earlence Fernandes, Yuhang Zhao, and Rahul Chatterjee. Sok: Authentication in augmented and virtual reality. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1552–1552. IEEE, 2022.

[249] David H Stern, Ralf Herbrich, and Thore Graepel. Matchbox: large scale online bayesian recommendations. In *Proceedings of the 18th international conference on World wide web*, pages 111–120, 2009.

[250] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[251] Wei Sun, Weike Luo, Xiongkuo Min, Guangtao Zhai, Xiaokang Yang, Ke Gu, and Siwei Ma. Mc360iqa: The multi-channel cnn for blind 360-degree image quality assessment. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, 2019.

[252] Yule Sun, Ang Lu, and Lu Yu. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE Signal Processing Letters*, 24(9):1408–1412, 2017.

[253] Yagiz Sutcu, Elham Tabassi, Husrev T Sencar, and Nasir Memon. What is biometric information and how to measure it? In *2013 IEEE international conference on technologies for homeland security (HST)*, pages 67–72. IEEE, 2013.

[254] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[255] Elemer Szabadi. Functional organization of the sympathetic pathways controlling the pupil: light-inhibited and light-stimulated pathways. *Frontiers in Neurology*, 9:1069, 2018.

[256] Kenta Takahashi and Takao Murakami. A measure of information gained through biometric systems. *Image and Vision Computing*, 32(12):1194–1203, 2014.

[257] Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu. Adapting naive bayes to domain adaptation for sentiment analysis. volume 5478, pages 337–349, 04 2009.

[258] Qbit Technologies. How vr will revolutionize e-commerce, 2020.

[259] Kemal Tekin, Mehmet Ali Sekeroglu, Hasan Kiziltoprak, Sibel Doguizi, Merve Inanc, and Pelin Yilmazbas. Static and dynamic pupillometry data of healthy individuals. *Clinical and Experimental Optometry*, 101(5):659–665, 2018.

[260] Danny Thakkar. Biometric devices: Cost, types and comparative analysis. `https://www.bayometric.com/biometric-devices-cost/`, 2017.

[261] Sophie Thompson. Vr applications: 23 industries using virtual reality. `https://virtualspeech.com/blog/vr-applications`, 2022.

[262] TikTok. Tiktok dataset. `https://www.biendata.xyz/competition/icmechallenge2019`, 2019.

[263] TikTok. How tiktok recommends videos for you, 2020.

[264] Tobii. Tobii vr: Eye tracking technology in virtual reality. `https://vr.tobii.com/`, 2022.

[265] Huyen TT Tran, Nam Pham Ngoc, Tobias Hoßfeld, and Truong Cong Thang. A cumulative quality model for http adaptive streaming. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2018.

[266] Huyen TT Tran, Nam Pham Ngoc, Cuong T Pham, Yong Ju Jung, and Truong Cong Thang. A subjective study on qoe of 360 video for vr communication. In *2017 IEEE 19th international workshop on multimedia signal processing (MMSP)*, pages 1–6. IEEE, 2017.

[267] Kazuo Tsubota and Katsu Nakamori. Dry eyes and video display terminals. *New England Journal of Medicine*, 328(8):584–584, 1993. PMID: 8426634.

[268] Pascal WM Van Gerven, Fred Paas, Jeroen JG Van Merriënboer, and Henk G Schmidt. Memory load and the cognitive pupillary response in aging. *Psychophysiology*, 41(2):167–174, 2004.

[269] Jacolien van Rij, Petra Hendriks, Hedderik van Rijn, R Harald Baayen, and Simon N Wood. Analyzing the time course of pupillometric data. *Trends in hearing*, 23:2331216519832483, 2019.

[270] Varjo. High-resolution vr headset for professionals - varjo vr-3. `https://varjo.com/products/vr-3/`, 2022.

[271] Michael Velichko. Vr military training – the next step of combat evolution, 2019.

[272] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

[273] Visualise. Virtual reality in healthcare, 2020.

[274] Vive. Htc vive pro eye. `https://www.vive.com/eu/product/vive-pro-eye/overview/`, 2022.

[275] Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, and Georgios Piliouras. Efficiently avoiding saddle points with zero order methods: No gradients required. *Advances in Neural Information Processing Systems*, 32, 2019.

[276] Jane Wakefield. Bionic eyes: Obsolete tech leaves patients in the dark. `https://www.bbc.com/news/technology-60416058/`, 2022.

[277] Gordon L Walls. The evolutionary history of eye movements. *Vision Research*, 2(1-4):69–80, 1962.

[278] Chin-An Wang and Douglas P Munoz. Modulation of stimulus contrast on the human pupil orienting response. *European Journal of Neuroscience*, 40(5):2822–2832, 2014.

[279] Ding Wang, Haibo Cheng, Ping Wang, Xinyi Huang, and Gaopeng Jian. Zipf's law in passwords. *IEEE Transactions on Information Forensics and Security*, 12(11):2776–2791, 2017.

[280] Ding Wang, Qianchen Gu, Xinyi Huang, and Ping Wang. Understanding human-chosen pins: characteristics, distribution and security. In *Proceedings of the 2017*

ACM on Asia Conference on Computer and Communications Security, pages 372–385, 2017.

[281] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, pages 165–174, 2019.

[282] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[283] Tracy Watson. Vr social media: Is it the future of social interaction?, 2019.

[284] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1437–1445, 2019.

[285] Wikipedia. Blinking, 2020.

[286] Wikipedia. Spline interpolatoin, 2020.

[287] Jacob Otto Wobbrock. Tapsongs: Tapping rhythm-based passwords on a single binary sensor. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology*, UIST '09, page 93–96. Association for Computing Machinery, October 2009.

[288] Qunjian Wu, Ying Zeng, Chi Zhang, Li Tong, and Bin Yan. An eeg-based person authentication system with open-set capability combining eye blinking signals. *Sensors*, 18(2):335, 2018.

[289] Qinghan Xiao. Security issues in biometric authentication. In *Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop*, pages 8–13. IEEE, 2005.

[290] Xiaoyu Xiu, Yuwen He, Yan Ye, and Bharath Vishwanath. An evaluation framework for 360-degree video compression. *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2017.

[291] Mai Xu, Chen Li, Zhenzhong Chen, Zulin Wang, and Zhenyu Guan. Assessing visual quality of omnidirectional videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(12):3516–3530, 2019.

[292] Songhua Xu, Hao Jiang, and Francis CM Lau. Personalized online document, image and video recommendation via commodity eye-tracking. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 83–90, 2008.

[293] Jingteng Xue, Dong-Qing Zhang, Heather Yu, and Chang Wen Chen. Assessing quality of experience for adaptive http video streaming. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, 2014.

[294] Shilin Yan, Shan Chang, Jiacheng Wang, and Shanila Azhar. Using pupil light reflex for fast biometric authentication. In *Proceedings of the ACM Turing Celebration Conference-China*, pages 139–143, 2020.

[295] Vitor Yano, Alessandro Zimmer, and Lee Luan Ling. Extraction and application of dynamic pupillometry features for biometric authentication. *Measurement*, 63:41–48, 2015.

[296] Abid Yaqoob, Ting Bi, and Gabriel-Miro Muntean. A survey on adaptive 360° video streaming: Solutions, challenges and opportunities. *IEEE Communications Surveys Tutorials*, 22(4):2801–2838, 2020.

[297] Shanhe Yi, Zhengrui Qin, Ed Novak, Yafeng Yin, and Qun Li. Glassgesture: Exploring head gesture interface of smart glasses. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9. IEEE, 2016.

[298] Wenpeng Yin. Meta-learning for few-shot natural language processing: A survey. *arXiv preprint arXiv:2007.09604*, 2020.

[299] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 974–983, 2018.

[300] Shingchern D You and Chun-Wei Liu. Classification of user preference for music videos based on eeg recordings. In *2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech)*, pages 1–2. IEEE, 2020.

[301] Richard Youmaran and Andy Adler. Measuring biometric sample quality in terms of biometric feature information in iris images. *Journal of Electrical and Computer Engineering*, 2012, 2012.

[302] Rockefeller SL Young and Mathew Alpern. Pupil responses to foveal exchange of monochromatic lights. *JOSA*, 70(6):697–706, 1980.

[303] Matt Yu, Haricharan Lakshman, and Bernd Girod. A framework to evaluate omnidirectional video coding schemes. In *2015 IEEE International Symposium on Mixed and Augmented Reality*, pages 31–36, 2015.

[304] Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. Diverse few-shot text classification with multiple metrics. *CoRR*, abs/1805.07513, 2018.

[305] Zhen Yu, Hai-Ning Liang, Charles Fleming, and Ka Lok Man. An exploration of usable authentication mechanisms for virtual reality systems. In *2016 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, pages 458–460. IEEE, 2016.

[306] Vladyslav Zakharchenko, K. Choi, and J. Park. Quality metric for spherical panoramic video. In *Optical Engineering + Applications*, 2016.

[307] Ramtin Zargari Marandi, Pascal Madeleine, Øyvind Omland, Nicolas Vuillerme, and Afshin Samani. Eye movement characteristics reflected fatigue development in both young and elderly individuals. *Scientific Reports*, 8, 09 2018.

[308] Mei Zhang, Jinglan Wu, Huifeng Lin, Peng Yuan, and Yanan Song. The application of one-class classifier based on cnn in image defect detection. *Procedia Computer Science*, 114:341 – 348, 2017. Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS October 30 – November 1, 2017, Chicago, Illinois, USA.

[309] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019.

[310] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6:1–23, 2019.

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

[312] Qian Zhao, Shuo Chang, F Maxwell Harper, and Joseph A Konstan. Gaze prediction for recommender systems. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 131–138, 2016.

[313] Yufeng Zhou, Mei Yu, Hualin Ma, Hua Shao, and Gangyi Jiang. Weighted-to-spherically-uniform ssim objective quality evaluation for panoramic video. In *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pages 54–57, 2018.

[314] 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):1–29, 2020.

[315] Huadi Zhu, Tianhao Li, Chaowei Wang, Wenqiang Jin, Srinivasan Murali, Mingyan Xiao, Dongqing Ye, and Ming Li. Eyeqoe: a novel qoe assessment model for 360-degree videos using ocular behaviors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–26, 2022.

[316] Huadi Zhu, Mingyan Xiao, Demoria Sherman, and Ming Li. Soundlock: A novel user authentication scheme for vr devices using auditory-pupillary response. In *NDSS*, 2023.

[317] Wenjie Zou, Fuzheng Yang, Wei Zhang, Yi Li, and Haoping Yu. A framework for assessing spatial presence of omnidirectional video on virtual reality device. *IEEE Access*, 6:44676–44684, 2018.