UNIVERSITAT DE BARCELONA

Gaze Estimation with Spatiotemporal and Multimodal Deep Learning

Cristina Palmero Cantariño

Doctoral Thesis

Gaze Estimation with Spatiotemporal and Multimodal Deep Learning

Autor: Cristina Palmero Cantariño

Director: Dr. Sergio Escalera Guerrero

UNIVERSITAT DE BARCELONA

Gaze Estimation with Spatiotemporal and Multimodal Deep Learning

Programa de Doctorat en
Enginyeria i Ciències Aplicades

Autor: Cristina Palmero Cantariño

Director i Tutor: Dr. Sergio Escalera Guerrero

Departament de Matemàtiques i Informàtica

UNIVERSITAT DE BARCELONA

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisor, Sergio Escalera. He has been extremely supportive throughout all stages of this long journey, and has given me the opportunity to pursue different avenues of research and personal growth. His guidance, mentorship, and unwavering belief in my abilities have been instrumental in shaping the outcomes of this endeavor. Thx!

This long journey actually began before starting this PhD, with the iCARE project, halfway between the Netherlands and Belgium. I am grateful to Guido Lichtert, who taught me the importance of gaze during social interactions, in particular for supportive parent-child interactions. This realization was the spark to start pursuing the topic of this thesis. I also thank Astrid van Wieringen, Nico van der Aa, Andrew Spink, and Elsbeth van Dam for their support during that period, and all colleagues from Noldus and iCARE, all contributing to my appreciation for interdisciplinary research and to learning different ways of approaching it.

I am also deeply grateful to the Eye Tracking Machine Learning team at Meta Reality Labs Research in the USA for allowing me to challenge and hone my research skills as an intern, twice. I especially thank Sachin Talathi for his guidance and support, and for offering yet another perspective to machine learning research. I also thank Oleg Komogortsev for his humble demeanor and willingness to share his knowledge and insights generously. Also, thanks to all the colleagues and fellow interns who made the first stay in 2019 a truly memorable experience, and the second a more bearable one despite the events that unfolded in 2020.

In addition, I thank the EMPATHIC project for funding most of this period, and also the partners for letting us be part of such an interesting project. Special thanks to Manés Torres, the leader.

I would like to acknowledge all the past and current members of the Human Pose Recovery and Behavior Analysis research group (HuPBA) and *altell*, some of whom have accompanied me since the Master's early days. Special thanks to Albert, Dani, Marc, Chip, Carles, Javi (UDIVA's primary accomplice), Pau, Julio, Meysam, Sorina, Johnny, and German. I extend my appreciation to all the students and collaborators with whom I have had the privilege to work along the way, as they have played a pivotal role in my growth as a researcher.

Doing a PhD is certainly not easy, but it becomes hardly achievable without a strong support network. I thank all the friends who kept me going at different stages of this journey, especially my former master's colleagues, ñaas, flores, Ana, Hemel's Greek squad, and Gelida. Mención especial a Raquel, que me ha seguido acompañando en la distancia hasta este momento. A en Gerard i la Neus, la nostra segona familia a Londres. Y a Ricky, Jesús, Carlos y Albert, por acompañarme de fondo durante muchas tardes largas de trabajo.

Quiero agradecer a mi familia, en especial a mis padres, Luisa y Manuel, y a mi hermano Marc, por acompañarme también desde la distancia, y por darme vuestro apoyo incondicional a pesar de que este camino sea a veces difícil de entender. Gràcies també a la Magda i en Jesús, que han sigut partícips de tot aquest procés.

I finalment, gràcies Roger. Aquesta tesi també és teva.

UNIVERSITAT DE BARCELONA

Abstract

Departament de Matemàtiques i Informàtica

Doctor of Philosophy

Gaze Estimation with Spatiotemporal and Multimodal Deep Learning

by Cristina Palmero Cantariño

It is often said that *the eyes are the window to the soul*. The eyes and their behavior have sparked interest for centuries, and have been widely studied due to their link with multiple developmental, neurological, behavioral, cognitive, and clinical factors. Furthermore, the ability to accurately detect the line of sight has enabled many possibilities for consumer applications, such as human-computer interaction and gaze-contingent displays. Eye-tracking technology has evolved to the point where non-invasive, sufficiently accurate, and cost-effective camera-based approaches are becoming increasingly available, driven by the progressive miniaturization of electronics and breakthroughs in computer vision and deep learning. However, achieving universal applicability in eye tracking remains a challenge, primarily due to the influence of individual factors, varying environmental conditions, and the impact of sensor viewpoint or head pose shifts. Recent remote and portable eye-tracking devices often sacrifice robustness and accuracy when used in uncontrolled scenarios. In addition, they grapple with the need for rapid eye signal capture, a crucial requirement for specific applications. The promising potential of eye tracking motivates us to further enhance existing methods, striving for greater reliability, accuracy, and speed. In turn, as eye tracking becomes more ubiquitous, it encourages us to explore innovative applications that leverage its expanding capabilities.

This thesis approaches eye tracking from a computer vision and deep learning perspective, with the goal of: 1) increasing the accuracy and sampling rate of current gaze estimation approaches across different scenarios and devices; and 2) promoting the use of gaze input in emerging applications. For the first goal, we investigate the contribution of spatiotemporal and multimodal/multisensor cues for gaze estimation, both for remote cameras (e.g., desktop setting) and infrared, near-eye devices (e.g., head-mounted displays), across different sources of variability. To do so, we rely on the combination of convolutional-recurrent deep neural networks and feature-based and hybrid multimodal fusion. In particular, we address multimodality from two different angles. First, by combining appearance and shape cues (i.e., 3D facial landmarks) extracted from RGB face images to increase accuracy. And second, by combining the signal obtained by two different sensors (camera and photosensors) operating at the same or different sampling rates, to increase the accuracy and the effective sampling rate of the estimated gaze signal. We then move on to the second goal, for which we explore the use of gaze-related features along with other modalities, such as speech and facial expressions, for emotion expression recognition in a conversational human-machine interaction scenario. More concretely, we focus on the interaction between a simulated virtual coach and older adults, delving into the nuances of affective computing in this context.

Resum

Es diu que *els ulls són el reflex de l'ànima*. El comportament dels ulls ha despertat interès durant segles i ha estat àmpliament estudiat per la seva relació amb diversos factors del desenvolupament, neurològics, conductuals, cognitius i clínics. A més, la capacitat de detectar amb precisió la direcció de mirada ha obert nombroses possibilitats en diverses aplicacions de consum, com ara la interacció persona-ordinador i els dispositius de visualització contingents a la mirada. La tecnologia de seguiment d'ulls ha evolucionat fins al punt en què sistemes no invasius basats en càmeres de vídeo, prou precisos i rendibles, s'estan tornant cada cop més accessibles, impulsats per la miniaturització progressiva de l'electrònica i els avenços en visió per ordinador i aprenentatge profund. No obstant això, aconseguir una aplicabilitat universal en el seguiment d'ulls continua sent un repte, principalment a causa de la influència de factors individuals, condicions ambientals variables, i l'impacte de canvis en la posició del sensor respecte a l'ull o moviments del cap. Els dispositius remots i portàtils de seguiment d'ulls recents veuen sovint compromesa la seva robustesa i exactitud quan s'utilitzen en escenaris no controlats. Addicionalment, s'enfronten al repte de capturar el senyal ocular de manera ràpida, un requisit fonamental per a algunes aplicacions. El potencial prometedor del seguiment d'ulls ens motiva a millorar els mètodes existents, buscant més fiabilitat, exactitud i velocitat. Alhora, a mesura que el seguiment d'ulls es torna més ubic, ens impulsa a explorar aplicacions innovadores que aprofitin les seves capacitats en expansió.

Aquesta tesi aborda el seguiment d'ulls des d'una perspectiva de visió per ordinador i aprenentatge profund, amb l'objectiu de: 1) augmentar l'exactitud i la taxa de mostreig dels sistemes actuals d'estimació de mirada a diferents escenaris i dispositius; i 2) promoure l'ús del seguiment d'ulls en aplicacions emergents. Per al primer objectiu, investiguem la contribució d'informació espaciotemporal i de diferents modalitats i sensors per a l'estimació de la direcció de mirada, tant amb càmeres remotes (per exemple, en configuració de sobretaula) com amb càmeres infraroges a prop de l'ull (per exemple, en cascos de realitat virtual), tenint en compte diferents fonts de variabilitat. Per dur això a terme, ens basem en la combinació de xarxes neuronals profundes convolucionals-recurrents i fusió multimodal basada en característiques i híbrida. En particular, abordem la multimodalitat des de dos angles diferents. Primer, mitjançant la combinació d'informació d'aparença i geomètrica (punts facials de referència en 3D) extreta d'imatges facials RGB per millorar l'exactitud de l'estimació de mirada. I, en segon lloc, mitjançant la combinació del senyal obtingut per dos sensors diferents (càmera i fotosensors) que operen a la mateixa o diferent freqüència, per augmentar l'exactitud i la taxa de mostreig efectiva de la línia estimada de mirada. Després passem al segon objectiu, per al qual explorem l'ús de característiques relacionades amb el comportament ocular juntament amb altres modalitats, com ara la parla i les expressions facials, per al reconeixement d'expressions emocionals en un escenari d'interacció humà-màquina conversacional. Més concretament, ens centrem en la interacció entre un assistent virtual i gent gran, aprofundint en els matisos de la computació afectiva en aquest context.

Resumen

Se suele decir que *los ojos son el reflejo del alma*. El comportamiento de los ojos ha despertado interés durante siglos, siendo ampliamente estudiado debido a su relación con diversos factores de desarrollo, neurológicos, conductuales, cognitivos y clínicos. Asimismo, la capacidad de detectar con precisión la dirección de mirada ha abierto numerosas posibilidades en aplicaciones de consumo, como la interacción persona-ordenador y los dispositivos de visualización contingentes a la mirada. La tecnología de seguimiento ocular ha evolucionado hasta el punto en que sistemas no invasivos basados en cámaras de video, lo suficientemente precisos y rentables, se están volviendo cada vez más accesibles, impulsados por la miniaturización progresiva de la electrónica y los avances en visión por ordenador y aprendizaje profundo. Sin embargo, lograr una aplicabilidad universal en el seguimiento ocular sigue siendo un desafío, principalmente debido a la influencia de factores individuales, condiciones ambientales variables, y el impacto de cambios en la posición del sensor respecto al ojo o movimientos de cabeza. Los dispositivos remotos y portátiles de rastreo ocular recientes ven a menudo comprometida su robustez y exactitud cuando se utilizan en escenarios no controlados. Además, se enfrentan al reto de capturar la señal ocular de manera rápida, un requisito fundamental para algunas aplicaciones. El potencial prometedor del rastreo ocular nos motiva a mejorar los métodos existentes, buscando una mayor fiabilidad, exactitud y velocidad. A su vez, a medida que el seguimiento ocular se vuelve más ubicuo, nos impulsa a explorar nuevas aplicaciones que aprovechen sus capacidades en expansión.

Esta tesis aborda el seguimiento ocular desde una perspectiva de visión por ordenador y aprendizaje profundo, con el objetivo de: 1) aumentar la exactitud y la tasa de muestreo de los sistemas actuales de estimación de mirada en diferentes escenarios y dispositivos; y 2) promover el uso del seguimiento ocular en aplicaciones emergentes. Para el primer objetivo, investigamos la contribución de información espaciotemporal y de diferentes modalidades y sensores para la estimación de la dirección de mirada, tanto con cámaras remotas (por ejemplo, en configuración de sobremesa) como con cámaras infrarrojas cerca del ojo (por ejemplo, en cascos de realidad virtual), teniendo en cuenta diferentes fuentes de variabilidad. Para ello, nos basamos en la combinación de redes neuronales profundas convolucionales-recurrentes y en la fusión multimodal basada en características y híbrida. En particular, abordamos la multimodalidad desde dos ángulos diferentes. Primero, mediante la combinación de información de apariencia y geométrica (puntos faciales de referencia en 3D) extraída de imágenes faciales RGB para mejorar la exactitud de la estimación de dirección de mirada. Y, en segundo lugar, mediante la combinación de la señal obtenida por dos sensores diferentes (cámara y fotosensores) que operan a la misma o diferente frecuencia, para aumentar la exactitud y la tasa de muestreo efectiva de la línea estimada de mirada. Luego pasamos al segundo objetivo, para el cual exploramos el uso de características relacionadas con el comportamiento ocular junto con otras modalidades, como el habla y las expresiones faciales, para el reconocimiento de expresiones emocionales en un escenario de interacción humano-máquina conversacional. Más concretamente, nos centramos en la interacción entre un asistente virtual y personas mayores, profundizando en los matices de la computación afectiva en este contexto.

Contents

List of Figures

List of Tables

List of Abbreviations

3DMM	**3D M**orphable **M**odel
ALR	**A**daptive **L**inear **R**egression
AR	**A**ugmented **R**eality
AU	**A**ction **U**nit
CCS	**C**amera **C**oordinate **S**ystem
CNN	**C**onvolutional **N**eural **N**etwork
CV	**C**omputer **V**ision
DL	**D**eep **L**earning
DNN	**D**eep **N**eural **N**etworks
EOG	**E**lectro-**O**culo**g**raphy
FC	**F**ully **C**onnected
FER	**F**acial **E**xpression **R**ecognition
GDPR	**G**eneral **D**ata **P**rotection **R**egulation
GRU	**G**ated **R**ecurrent **U**nit
HCI	**H**uman-**C**omputer **I**nteraction
HCS	**H**ead **C**oordinate **S**ystem
HMD	**H**ead-**M**ounted **D**isplay
HMI	**H**uman-**M**achine **I**nteraction
IR	**I**nfra**r**ed
LED	**L**ight-**E**mitting **D**iode
LLD	**L**ow-**L**evel **D**escriptor
MAE	**M**ean **A**bsolute **E**rror
MFCC	**M**el-**F**requency **C**epstral **C**oefficients
ML	**M**achine **L**earning
MLP	**M**ulti**l**ayer **P**erceptron
LSTM	**L**ong **S**hort-**T**erm **M**emory
ODE	**O**rdinary **D**ifferential **E**quation
PCA	**P**rincipal **C**omponent **A**nalysis
PnP	**P**erspective-**n**-**P**oint
PoR	**P**oint **o**f **R**egard
PSOG	**P**hoto**s**ensor **O**culo**g**raphy
ReLU	**Re**ctified **L**inear **U**nit
RGB	**R**ed, **G**reen and **B**lue
RNN	**R**ecurrent **N**eural **N**etwork
SD	**S**tandard **D**eviation
SEM	**S**tandard **E**rror of the **M**ean
SER	**S**peech **E**motion **R**ecognition
VAD	**V**alence-**A**rousal-**D**ominance
VC	**V**irtual **C**oach
VOG	**V**ideo-**O**culo**g**raphy
VR	**V**irtual **R**eality
WoZ	**W**izard **o**f o**Z**

"The soul, fortunately, has an interpreter – often an unconscious but still a faithful interpreter – in the eye."

Charlotte Brontë, Jane Eyre

Chapter 1

Introduction

FIGURE 1.1: Frontal image of the human eye. Attribution: Vecteezy.com.

THE EYE is one of the most complex organs in many species, including humans. Eyes evolved from light-sensing cells to the camera-like form found in most vertebrates around 500 million years ago (Lamb, Collin, and Pugh Jr, 2007), which has been assumed to be an evolutionary advantage for fast locomotion, navigation, and detection of prey and predators (Parker, 2004). Our eyes can sense the world surrounding us, with a monocular field of view of about 120 degrees of visual angle (°), but highly acute information can only be gleaned from around 1-2°. Therefore, we must move our eyes to bring and maintain a particular area of the visible field in sharp focus. If the area of interest is large, we have to move our eyes quickly to extract all relevant information as fast as possible. And we do: eye movements are considered to be one of the fastest in the human body, achieving peak velocities of up to 700°/s (Leigh and Zee, 2015). Visual input is transferred to the brain, which processes and interprets the received information. This transfer can be carried out at an estimated speed of about 10^6 bps, similar to an Ethernet connection (Koch et al., 2006). Around one third of the cerebral cortex is primarily devoted to processing visual information (Van Essen, 2003), but many other cortical and subcortical areas are related to vision or eye movements (Leigh and Zee, 2015; Pouget, 2015).

The high contrast between the external parts of the eye, namely the *pupil* (the dark, round opening in the center of the eye, through which light enters the eye), the *iris* (the colored part of the eye that regulates the pupil aperture), and the *sclera* (the white outer layer of the eyeball), depicted in Figure 1.1, allows us to detect where another person is looking, quickly and from a distance, which has been linked to the evolution of social intelligence (Emery, 2000). Indeed, among humans, detecting

FIGURE 1.2: During gaze following, the child shifts their attention from the caregiver to the object that the caregiver is showing to them, leading to joint attention events. Accurate detection of such events is essential for parent-child interaction analysis. Reproduced from http://beforefirstwords.upf.edu/precursors-of-language/gaze-following/, used under CC BY-NC-ND 4.0.

someone else's direction of gaze is a crucial component of social interactions in many aspects, for instance, as a deictic (pointing) cue to guide behavior (Shepherd, 2010), turn-taking signaling during conversation (Ho, Foulsham, and Kingstone, 2015), or as a communication channel to convey our focus of attention, intentions, and even emotions (Itier and Batty, 2009). Nonetheless, its significance starts early in our lives: eye contact between babies and their caregivers is one of the first milestones, usually around seven weeks old (Haith, Bergman, and Moore, 1977). From at least four months of age, infants start following others' line of gaze, known as *gaze following* (see Figure 1.2). These behaviors are associated with critical aspects of infant development, such as language acquisition and developing a theory of mind (Brooks and Meltzoff, 2005; Itier and Batty, 2009; Clark and Casillas, 2015).

Oculomotor behavior, including gaze direction, eye movements, and pupillometry, has been extensively studied for more than two centuries (Wade, 2010). For starters, gaze direction is a measure of selective overt attention, providing valuable insights into what captures an individual's interest and focus in a given environment. Gaze behavior is task dependent (Yarbus, 1967), although it tends to be drawn toward visually salient and semantically meaningful stimuli (see Figure 1.3), such as faces, objects in motion, or items of personal relevance. By analyzing gaze patterns and changes in pupil size, scientists have been able to unravel cognitive processes related to perception, memory, decision-making, cognitive workload, and emotional responses, among others (Fogarty and Stern, 1989; Rayner, 1998; Liversedge and Findlay, 2000; Orquin and Loose, 2013; Mathôt, 2018). Furthermore, since eye movements are tightly coupled to the brain's processing of visual information, anomalies in oculomotor behavior can indicate underlying neurological and psychiatric disorders or cognitive impairments (Rommelse, Stigchel, and Sergeant, 2008; Klein and Ettinger, 2008; Leigh and Zee, 2015; Das et al., 2022). In fact, it is largely possible to determine which area of the brain is affected by observing specific changes in pupil size and eye movements, allowing for assessments that can be conducted in the doctor's office (see Figure 1.4). Consequently, oculometrics are considered potential biomarkers for the diagnosis, prognosis, monitoring, and treatment evaluation of many conditions, individually or in combination with other measures.

Eye tracking and gaze estimation

Given the importance of oculomotor behavior, it is crucial to provide an objective and reliable way of measuring it. The process of measuring the rotation or direction

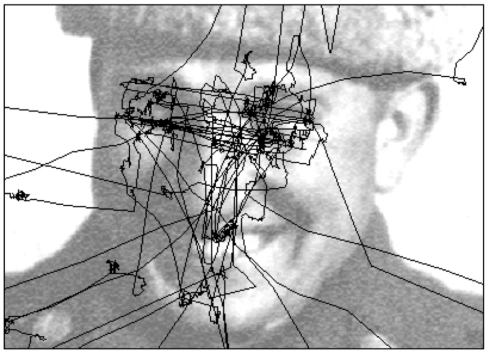


FIGURE 1.3: Example of the scan pattern (sequence of eye movements) of an observer during free viewing of a person's portrait. Reproduced from Tatler et al. (2010), used under CC BY-NC-ND 3.0.

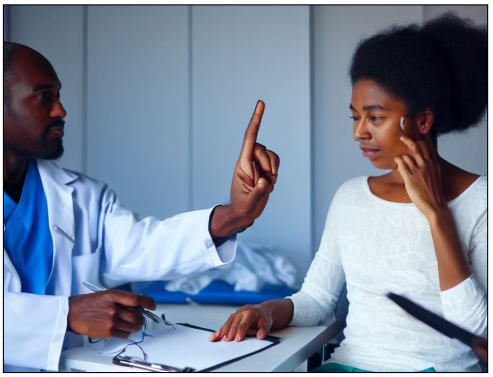


FIGURE 1.4: During a neurological exam, the healthcare provider elicits different eye movements using their finger as the gaze target. Generated with Microsoft Bing AI.

of the eye over time, or the point of gaze, is commonly called *eye tracking*. This word stems from tracking specific eye regions, like the iris or pupil, to do so. However, nowadays, the term generally refers to all devices that can perform such measures irrespective of the technique employed. More recent terms like *gaze tracking* or *gaze estimation* usually refer to more recent techniques employed by eye-tracking devices that estimate the direction of gaze without necessarily tracking specific regions.

Eye tracking has evolved in parallel with the study of oculomotor function and eye movements. First approaches in the 19th century consisted in attaching certain instruments to the eye, such as caps, suction cups, or plaster casts, which were extremely inconvenient for study participants (Płużyczka, 2018; Fletcher, Dunne, and Butler, 2022). In the 20th century, eye tracking progressed to more comfortable devices, like magnetic scleral search coils mounted as contact lenses (Robinson, 1963), or placing electrodes around the periocular region, known as electro-oculography (EOG) (Marg, 1951). Non-invasive eye tracking was possible thanks to advances in photography, and later cinematography and image processing, being the predecessor of the most used technique nowadays, *video-oculography* (VOG) (Płużyczka, 2018; Fletcher, Dunne, and Butler, 2022). VOG is a camera-based approach that records the visible part of the eye (e.g., as in Figure 1.1), and the recorded signal is then computer-processed to determine the horizontal, vertical, and sometimes torsional, movements of the eye. VOG has usually relied on carefully calibrated setups of one or multiple infrared (IR) high-resolution or near-eye cameras and dedicated light sources. These were required to model the geometry of the eye and enhance the aforementioned contrast between the pupil and the iris, so that traditional edge detectors could be applied to detect them (Hansen and Ji, 2010). Traditional VOG also depends on a user calibration stage prior to its usage, where the user has to look at specific gaze targets on a screen. This is done to estimate subject-specific eye parameters (for *model-based* methods) or map eye features to specific target locations (for *feature-based*). To date, most desktop eye trackers (Figure 1.5) and recent portable head-mounted devices (Figure 1.6) use this technology.

The last decade has witnessed a technological revolution with the advent of deep learning (DL), considerably improving performance in many areas and making possible others that had not been considered (LeCun, Bengio, and Hinton, 2015). Eye

FIGURE 1.5: Example of consumer remote/desktop eye tracker (Gazepoint). Used with permission of Springer, from Duchowski (2017); permission conveyed through Copyright Clearance Center, Inc.

FIGURE 1.6: Example of virtual reality headset equipped with eye tracking. Generated with Microsoft Bing AI.

tracking has been one of them. In particular, DL has fostered research in *appearance-based* gaze estimation, which timidly started during the 1990s with advances in computer vision (CV) and machine learning (ML). Appearance-based approaches directly map an image of the eye or face to a specific 2D location, such as a screen position, or to a 3D gaze direction vector. Hence, they do not require a high-resolution image of the eye or dedicated IR lighting setups. This has made eye tracking more accessible, enabling remote gaze tracking with regular cameras, such as webcams or smartphones. As feature extractors and mapping functions, convolutional neural networks (CNNs) have been proven to perform incredibly well for these tasks due to their locality biases (Zhang et al., 2015), powered by recently acquired large-scale datasets of pairs of eye/face images and associated gaze directions (Ghosh et al., 2021). Presently, both traditional and appearance-based VOG eye trackers usually incorporate one or more DL modules to tackle the main challenges in gaze estimation, which include:

- the large variations in eye/face appearances and anatomical differences across the human population;
- illumination, camera viewpoint, and head pose variability;
- the use of eyeglasses and/or makeup;
- sensor noise and image artifacts.

Whereas current state-of-the-art appearance-based methods exploit end-to-end deep networks to regress gaze from input eye or face images directly, model-based methods deploy per-pixel segmentation networks to extract the visible eye regions for further processing (Yiu et al., 2019). In addition, since deep networks are capable of learning complex relationships in the data and can generalize reasonably well across different appearances, DL has also fostered the creation of subject-independent gaze estimation models. This reduces the requirement of prior user calibration, making eye tracking even more user-friendly and straightforward to use. Figure 1.7 shows the standard pipeline for appearance-based 3D gaze estimation, from the input image to the final application.

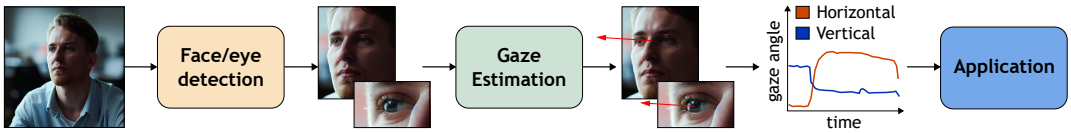


FIGURE 1.7: Standard pipeline for appearance-based 3D gaze estimation with full-face or eye-region input. Face generated with Microsoft Bing AI.

Eye-tracking applications

According to Duchowski (2002), eye-tracking applications can be broadly classified as *diagnostic* and *interactive*. Diagnostic applications use eye tracking to analyze eye movements while users perform specific tasks, often not requiring online processing, whereas interactive ones use the user's gaze direction as input in real time.

As diagnostic applications, we stress the importance of clinical research, practice, and rehabilitation, where oculometrics can be potential biomarkers (Larrazabal, Cena, and Martínez, 2019). Conditions associated with abnormal oculomotor behavior include: vestibular disorders (Leigh and Zee, 2015), learning disabilities (Pavlidis, 1985), stroke (Sand et al., 2013), mental disorders such as schizophrenia, depression, anxiety, and obsessive-compulsive disorder (Braff, 1993; O'Driscoll and Callahan, 2008; Armstrong and Olatunji, 2012; Carvalho et al., 2015), neurodevelopmental disorders such as attention-deficit hyperactivity disorder and autism spectrum disorder (Sweeney et al., 2004; Karatekin, 2007; Falck-Ytter, Bölte, and Gredebäck, 2013; Guillon et al., 2014; Chita-Tegmark, 2016), neurodegenerative diseases such as multiple sclerosis, Parkinson's, Huntington's, and types of dementia like Alzheimer's (Crutcher et al., 2009; Patel et al., 2012; Anderson and MacAskill, 2013; Das et al., 2022), and other signs of normal aging (Marandi and Gazerani, 2019).

Diagnostic applications are also present in numerous areas, including: user experience research, for example, to optimize digital interfaces and automotive designs (Jacob and Karn, 2003; Pan et al., 2004; Poole and Ball, 2006); advertising, marketing, and consumer retail research (Wedel and Pieters, 2017); in the educational field (Lai et al., 2013), for example, in medical education (Ashraf et al., 2018), to identify and improve search strategies among novice and experienced pathologists when reading and grading scans (Brunyé et al., 2019); psychology and neuroscience research (Hannula et al., 2010; Mele and Federici, 2012; Rahal and Fiedler, 2019); human performance, such as decision making in sports activities (Kredel et al., 2017), driving (Kapitaniak et al., 2015; Khan and Lee, 2019), or aviation (Peißl, Wickens, and Baruah, 2018); and others (Meißner and Oll, 2019; Hu, Wang, and Xu, 2022).

Regarding interactive applications, progress in eye tracking was a turning point for human-computer interaction (HCI) (Jacob, 1991; Majaranta and Bulling, 2014), gaze-contingent displays (Duchowski, Cournia, and Murphy, 2004), and assistive interfaces for people with disabilities or mobility limitations (Majaranta, 2011). Today, eye tracking is being increasingly used for gaming and for emerging augmented reality and virtual reality (AR/VR) headsets (Clay, König, and Koenig, 2019). For these devices, eye tracking can improve immersion and social presence (Oh, Bailenson, and Welch, 2018). Furthermore, gaze-contingent rendering drastically reduces power consumption by compressing the scene that falls on the visual periphery (Patney et al., 2016). Eye tracking is also increasingly used to detect drivers' fatigue and distractions (Ramzan et al., 2019). In human-machine interaction (HMI), eye tracking will allow embodied agents and robots to infer the user's *visual focus on attention* (the specific target the user is looking at) and the level of engagement (Palinko et al., 2016). And this is just a foretaste of what eye tracking can provide.

1.1 Motivation

One of the most used metrics to evaluate the quality of an eye tracker or a gaze estimation approach is *spatial accuracy*, also known as error, which indicates the offset between the real and estimated gaze signal. Accuracy is contingent upon the robustness of the eye-tracking approach to the different challenges that gaze estimation presents, and upon the quality of the result of the user calibration stage, if any. Another metric employed for eye trackers is *temporal resolution*, also known as frequency or sampling rate, which refers to the temporal granularity with which changes in eye rotation or gaze points can be detected and measured over time. This one is contingent upon the sensor used, the power requirements of the device, and the computational complexity of the technique employed. Accuracy and temporal resolution requirements depend on the application. For instance, a spatial accuracy of up to 0.1° and a temporal resolution of around 1 kHz are required to faithfully detect and track the fastest and smallest[1] eye movements, needed for some diagnostic oculometrics. An accuracy of 0.5-1° at 120 Hz is generally sufficient to measure basic eye movement patterns and gaze behavior for many interactive applications. These values can be further reduced for applications that require coarse estimates as input features for other tasks or to detect the visual focus of attention. For the latter, explicit gaze estimation is sometimes augmented or even replaced by head pose estimation, saliency detection, or context-aware approaches that combine these in addition to other modalities of the scene and environment, such as who is speaking in a group conversation (Massé, Ba, and Horaud, 2017; Siegfried and Odobez, 2021).

Historically, most diagnostic applications have been based on controlled, reductionist screen-based studies with expensive, dedicated desktop setups and head motion stabilizers to ensure high accuracy (Graham et al., 2022; Harston and Faisal, 2022). Thanks to recent advances in remote and portable eye tracking, research studies can be increasingly performed in real-world, more ecologically valid scenarios, such as daily living activities and VR environments, with free head and body movement (Callahan-Flintoft et al., 2021; Lamb et al., 2022). These advances have also encouraged the inclusion of eye tracking into consumer devices for interactive applications. However, such increase in accessibility still comes at the expense of lower robustness and accuracy due to the increase in variability and noise under uncontrolled conditions (headset movement or *slippage* during operation, illumination changes, etc., see Hessels et al., 2020). Furthermore, the temporal resolution of these new devices is generally lower than that of traditional ones, making them less suitable for certain applications. Although DL-based approaches hold promise for improving robustness to different populations and scenarios, achieving the vision of making eye tracking functional and valuable for everyone everywhere remains an open challenge. We hope to see this vision come to fruition across any type of device (i.e., desktop or head-mounted), camera location (i.e., remote or near-eye), and camera type (i.e., color or IR), encouraging the development of eye-tracking solutions tailored to the needs of existing and future applications.

For remote scenarios in particular, DL-powered appearance-based approaches enable the application of eye tracking using regular webcams or smartphones, reducing costs and making eye tracking accessible and scalable to a larger and diverse population, since no dedicated setups are required, and the requirement of personal

[1] For this case, high *precision* (how close measurements of the same real eye rotation are to each other) and *spatial resolution* (the smallest eye movement that can be resolved by the sensor) are also necessary. However, we emphasize the importance of accuracy and temporal resolution, as these are the metrics for which we will optimize throughout the thesis.

calibration can be substantially reduced or removed. This is not only beneficial for researchers who aim to expand their participant sample, but could be a turning point for healthcare: easier deployment on a larger scale would allow for more accessible early assessments for a number of conditions. For instance, eye tracking has been found to be useful for the early diagnosis of Alzheimer's, years before the actual clinical diagnosis, by being analyzed in naturalistic scenarios and daily activities (Beltrán et al., 2018). Recent comparisons between traditional IR remote (with or without restricted head movement) and new webcam- and smartphone-based eye trackers have confirmed that, despite lower accuracy, the new devices are reliable and their results are consistent with theoretical expectations for different tasks (Valliappan et al., 2020; Shehu et al., 2021; Wisiecka et al., 2022; Hutt and D'Mello, 2022). This can be extended to interactive applications, where off-the-shelf cameras could be seamlessly integrated with DL-based eye tracking for applications such as HCI/HMI and assistive gaze-contingent technologies outside controlled laboratory environments.

The potential applications and prospects for eye tracking are encouraging to continue improving existing DL-powered methods for greater robustness, accuracy, and speed. In turn, the democratization and progress of eye tracking prompt the exploration of novel applications that exploit its expanding capabilities and accessibility.

1.2 Thesis objectives

This thesis approaches gaze tracking from a CV/DL-, appearance-based perspective, with the goal of:

1. Increasing accuracy and sampling rate of current methods and devices to enhance their robustness and applicability;

2. Leveraging gaze input in emerging applications to promote its adoption.

We argue that *spatiotemporal* and *multimodal* DL can help us achieve such goals. In CV and ML/DL, exploiting temporal information and dynamics has been shown to be useful for a number of video-based tasks, increasing accuracy due to considering previous information and correlations in the data, and decreasing prediction jitter of individual frames (Hossain and Little, 2018; LaLonde, Zhang, and Shah, 2018; Wang et al., 2021b). Similarly, combining information from different image modalities or sensors tends to provide complementary information that enriches the representation in the feature space (Baltrušaitis, Ahuja, and Morency, 2018; Guo, Wang, and Wang, 2019). Temporal dynamics and multimodal information have previously been considered for model-based eye tracking approaches, but very sparingly. For instance, Haro, Flickner, and Essa (2000) and Hansen and Pece (2005) used state models and Kalman filters to track the pupil or iris along a sequence of frames. State-of-the-art appearance-based gaze estimation methods mainly rely on static features. However, intuitively, the temporal traces of eye gaze and head movements should contain useful information for estimating a given gaze point. Furthermore, to our knowledge, the combination of RGB and depth cameras was the only existing multisensor approach for remote-camera scenarios (Xiong et al., 2014; Funes-Mora and Odobez, 2016), and IR camera and photosensors for near-eye ones (Rigas, Raffle, and Komogortsev, 2017). Following this gap in the literature with respect to appearance-based approaches, this thesis aims to answer the following two research questions:

RQ_1. *Is temporal information beneficial for appearance-based gaze estimation?*

RQ_2. *Can the fusion of different modalities or sensors improve appearance-based gaze estimation performance, in terms of accuracy and/or sampling rate?*

Accessible eye tracking allows its integration with existing and new applications. Recent research demonstrates that gaze behavior can be leveraged as input for other CV/ML-powered applications, such as automatic speech, emotion, personality, or intention recognition (Cooke and Russell, 2008; Jang et al., 2014; Hoppe et al., 2018; Lim, Mountstephens, and Teo, 2020). Proper recognition of these aspects is crucial for emerging socially intelligent systems and embodied agents to provide personalized and empathic interactions. We find that, as is common in CV/ML research, most studies in these areas have been carried out in young adults. However, such systems are expected and are being conceived to reach other populations, namely infants and older adults. In this thesis, we focus on the emotion recognition task centered on older adults in a conversational HMI scenario, and aim to answer an additional research question:

RQ_3. *Is gaze-related information beneficial for emotion recognition in older adults, either alone or in combination with other modalities?*

We answer the three questions in four chapters devoted to gaze estimation applied to different tasks in remote (e.g., desktop) and near-eye (e.g., head-mounted) camera scenarios. We focus on subject-independent models, that is, generic gaze estimation models that can be used without any person calibration stage prior to their usage. Nevertheless, they could also be used as a prior model that is personalized ad hoc to increase gaze estimation accuracy. Three chapters are devoted to methodological analysis (RQ_1 and RQ_2), while one chapter focuses on the application of gaze tracking on an HMI task (RQ_3).

1.3 Thesis contributions

The contributions of each chapter are outlined below:

1. **Multimodal and spatiotemporal gaze estimation in remote-camera scenarios (RQ_1 and RQ_2).** We tackle the problem of subject- and head pose-independent 3D gaze estimation from remote RGB cameras by means of a spatiotemporal CNN-recurrent neural network (RNN). We propose to combine appearance cues from the face and eyes region, and shape cues from face landmarks, as individual streams in a CNN to estimate gaze in still images. Then, we exploit the dynamic nature of gaze by feeding the learned features of all the frames in a sequence to a many-to-one recurrent module that predicts the 3D gaze vector of the last frame. Our multimodal static and spatiotemporal solutions are evaluated on a wide range of head poses and gaze directions on the EYEDIAP dataset. Results show that adding facial shape cues regularizes the gaze estimates obtained. Furthermore, we demonstrate that spatiotemporal information is especially useful when head motion is present in non-screen-oriented scenarios. To our knowledge, this was the first approach leveraging shape cues and spatiotemporal information for appearance-based gaze estimation. This work has appeared in Palmero et al. (2018a) and Palmero et al. (2018b).

2. **Spatiotemporal gaze estimation in near-eye camera scenarios (RQ_1).** Despite the promising results obtained previously with off-the-shelf remote cameras

by leveraging sequential information, the magnitude of the contribution from eye movement traces specifically is yet unclear. These traces can be better captured with higher resolution/sampling rate imaging systems, in which more detailed information about the eye is obtained. We investigate whether temporal sequences of IR near-eye images, captured using a high-resolution, high-frame-rate head-mounted VR system, can contribute to enhancing the accuracy of an end-to-end appearance-based DL model for gaze estimation. In addition, we analyze how temporal information is beneficial for this task. Results demonstrate statistically significant benefits of temporal information, particularly for the vertical component of gaze. This work was carried out in collaboration with Meta Reality Labs Research and appears in Palmero, Komogortsev, and Talathi (2020).

3. **Single- and multirate sensor fusion for gaze estimation in near-eye camera scenarios (RQ_2).** The power requirements of camera-based gaze estimation can be prohibitive for high-speed operation with portable, battery-equipped devices. Recently, low-power sensor alternatives such as photosensors have been evaluated, being able to provide gaze estimates at high frequency with a trade-off in accuracy and robustness. Potentially, a hybrid approach that combines fast/low-fidelity and slow/high-fidelity sensors should be able to exploit their complementarity to track fast eye motion accurately and robustly. To validate the potential of this approach, and to foster research on this topic, we introduce OpenSFEDS, a multisensor near-eye gaze estimation dataset. The dataset contains more than 2M synthetic camera-photosensor image pairs in the form of synchronized videos sampled at 500 Hz with varied subject appearance and geometry, lighting, and camera position that mimic sensor shifts. We also formulate the task of sensor fusion for gaze estimation, proposing a framework based on appearance-based encoding and temporal eye state dynamics. We evaluate a set of sensor fusion baselines for single- and multirate operation on OpenSFEDS, achieving a statistically significant decrease in angular error when tracking fast eye movements with a multirate sensor fusion approach versus a gaze forecasting approach operating with a low-speed sensor alone. Furthermore, we analyze the robustness of the two sensors, individually and combined, against several sources of variability. To our knowledge, this was the first dataset providing synchronized image pairs at high frequency, and the first work proposing a multisensor, feature-based fusion framework for gaze estimation from an appearance-based perspective. This work was also carried out in collaboration with Meta Reality Labs Research and appears in Palmero et al. (2023b).

4. **Emotion expression recognition using facial expressions, speech, head pose, and gaze-related cues in a remote-camera scenario (RQ_3).** During the thesis period, we were part of the European EMPATHIC project, which aimed to design an emotionally expressive virtual coach (VC) capable of engaging healthy senior users to enhance well-being and promote independent aging. One of the core aspects of the system is its human sensing capabilities, allowing for the perception of emotional states to provide a personalized experience during the conversation. Within the context of the project, several partners participated in the research and development of the EMPATHIC-VC emotion expression recognition module, which receives information from the users' facial expressions, speech, head pose, and gaze dynamics, and combines it to estimate the current user's emotional state. Our team led the head pose, gaze

dynamics, and fusion submodules. To develop and evaluate the module, a corpus of older adults interacting with an initial version of the VC was collected and annotated, for which we participated in the design choices and deployment of the facial expression annotation process. In this thesis, we outline the development of the module, data collection, and annotation process, and provide a first methodological approach. With the latter, we provide an extensive study on discrete emotion expression recognition, wherein we investigate the role of the different modalities in this context, individually and combined. The collected corpus includes users from three countries, and was annotated separately for the audio and video channels with distinct emotional labels, allowing for a performance comparison across cultures and label types. The results confirm the informative power of the modalities studied for the emotional categories considered, with multimodal methods usually outperforming others. In particular, we find that gaze and head features provide redundant information for audio-based labels when used together, and their contribution is limited. By contrast, these features provide complementary information for video-based labels, with a significant contribution both individually and combined with speech and/or facial expressions. This work appears in Justo et al. (2020), Amorese et al. (2022), and Palmero et al. (2023a).

1.4 Publications

1.4.1 Main publications

The following publications are part of this thesis, either directly or indirectly. Among these works, some have been featured in prestigious CV and eye-tracking venues, including first-quartile journals. In particular, one of these publications has earned an honorable mention award (Palmero, Komogortsev, and Talathi, 2020), recognizing its contributions to the domain. Grayed-out entries correspond to work under review at the time of writing.

Journal papers

- *Cristina Palmero, Mikel deVelasco, Mohamed Amine Hmani, Aymen Mtibaa, Leila Ben Letaifa, Pau Buch-Cardona, Raquel Justo, Terry Amorese, Eduardo González-Fraile, Begoña Fernández-Ruanova, Jofre Tenorio-Laranga, Anna Torp Johansen, Micaela Rodrigues da Silva, Liva Jenny Martinussen, Maria Stylianou Korsnes, Gennaro Cordasco, Anna Esposito, Mounim A. El-Yacoubi, Dijana Petrovska-Delacrétaz, M. Inés Torres, and Sergio Escalera.* **Exploring Emotion Expression Recognition in Older Adults Interacting with a Virtual Coach**. Under review, 2023.

- *Cristina Palmero, Abhishek Sharma, Karsten Behrendt, Kapil Krishnakumar, Oleg V. Komogortsev, and Sachin S. Talathi.* **OpenEDS2020 Challenge on Gaze Tracking for VR: Dataset and Results.** Sensors 21, no. 14, pp. 4769, 2021.

- *Raquel Justo, Leila Ben Letaifa, Cristina Palmero, Eduardo Gonzalez-Fraile, Anna Torp Johansen, Alain Vázquez, Gennaro Cordasco, Stephan Schlögl, Begoña Fernández-Ruanova, Micaela Silva, Sergio Escalera, Mikel deVelasco, Joffre Tenorio-Laranga, Anna Esposito, Maria Korsnes, and M. Inés Torres.* **Analysis of the Interaction Between Elderly People and a Simulated Virtual Coach.** Journal of Ambient Intelligence and Humanized Computing 11, pp. 6125-6140, 2020.

International conferences and workshops

- *Cristina Palmero, Oleg V. Komogortsev, Sergio Escalera, and Sachin S. Talathi.* **Multi-Rate Sensor Fusion for Unconstrained Near-Eye Gaze Estimation.** In Proceedings of the 2023 Symposium on Eye Tracking Research and Applications, pp. 1-8. 2023.

- *Arya Farkhondeh, Cristina Palmero, Simone Scardapane, and Sergio Escalera.* **Towards Self-supervised Gaze Estimation.** In 33rd British Machine Vision Conference (BMVC), 2022.

- *Terry Amorese, Claudia Greco, Marialucia Cuciniello, Carmela Buono, Cristina Palmero, Pau Buch-Cardona, Sergio Escalera, Maria Inés Torres, Gennaro Cordasco, and Anna Esposito.* **Using Eye Tracking to Investigate Interaction Between Humans and Virtual Agents.** In 2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), pp. 125-132. IEEE, 2022.

- *Javier M. Olaso, Alain Vázquez, Leila Ben Letaifa, Mikel De Velasco, Aymen Mtibaa, Mohamed Amine Hmani, Dijana Petrovska-Delacrétaz, Gérard Chollet, César Montenegro, Asier López-Zorrilla, Raquel Justo, Roberto Santana, Jofre Tenorio-Laranga, Eduardo González-Fraile, Begoña Fernández-Ruanova, Gennaro Cordasco, Anna Esposito, Kristin Beck Gjellesvik, Anna Torp Johansen, Maria Stylianou Kornes, Colin Pickard, Cornelius Glackin, Gary Cahalane, Pau Buch, Cristina Palmero, Sergio Escalera, Olga Gordeeva, Olivier Deroo, Anaïs Fernández, Daria Kyslitska, Jose Antonio Lozano, M. Inés Torres, and Stephan Schlögl.* **The Empathic Virtual Coach: A demo.** In Proceedings of the 2021 International Conference on Multimodal Interaction, pp. 848-851, 2021.

- *Claudia Greco, Carmela Buono, Pau Buch-Cardona, Gennaro Cordasco, Sergio Escalera, Anna Esposito, Anais Fernandez, Daria Kyslitska, Maria Stylianou Kornes, Cristina Palmero, Jofre Tenorio Laranga, Anna Torp Johansen, and M. Inés Torres.* **Emotional Features of Interactions with Empathic Agents.** In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2168-2176, 2021.

- *Cristina Palmero, Oleg V. Komogortsev, and Sachin S. Talathi.* **Benefits of Temporal Information for Appearance-based Gaze Estimation.** In ACM Symposium on Eye Tracking Research and Applications, pp. 1-5, 2020.

- *Josep Famadas, Meysam Madadi, Cristina Palmero, and Sergio Escalera.* **Generative Video Face Reenactment by AUs and Gaze Regularization.** In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 444-451, 2020.

- *M. Inés Torres, Javier Mikel Olaso, César Montenegro, Roberto Santana, Alain Vázquez, Raquel Justo, José Antonio Lozano, Stephan Schlögl, Gérard Chollet, Nazim Dugan, M. Irvine, N. Glackin, C. Pickard, Anna Esposito, Gennaro Cordasco, Alda Troncone, Dijana Petrovska-Delacretaz, Aymen Mtibaa, Mohamed Amine Hmani, MS Korsnes, L. J. Martinussen, Sergio Escalera, Cristina Palmero, Olivier Deroo, Olga Gordeeva, Jofre Tenorio-Laranga, E. Gonzalez-Fraile, Begoña Fernández-Ruanova, and A. Gonzalez-Pinto.* **The EMPATHIC Project: Mid-Term Achievements.** In Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments, pp. 629-638, 2019.

– *Cristina Palmero, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera.* **Recurrent CNN for 3D Gaze Estimation using Appearance and Shape Cues.** In 29th British Machine Vision Conference (BMVC), 2018.

– *Cristina Palmero, Elsbeth A. van Dam, Sergio Escalera, Mike Kelia, Guido F. Lichtert, Lucas P. J. J. Noldus, Andrew J. Spink, and Astrid van Wieringen.* **Automatic Mutual Gaze Detection in Face-to-Face Dyadic Interaction Videos.** In Proceedings of Measuring Behavior, vol. 1, p. 2, 2018.

1.4.2 Other publications

The following publications have been carried out during the thesis period, but are not associated with the thesis.

Journal papers

– *Cristina Palmero, M. Inés Torres, Anna Esposito, and Sergio Escalera.* **Guest Editorial: Special Issue on Computer Vision and Machine Learning for Healthcare Applications.** Pattern Analysis and Applications 25, no. 3, pp. 489-492, 2022.

– *Ricardo Darío Pérez Principi, Cristina Palmero, Julio C. S. Jacques Junior, and Sergio Escalera.* **On the Effect of Observed Subject Biases in Apparent Personality Analysis from Audio-Visual Signals.** IEEE Transactions on Affective Computing 12, no. 3, pp. 607-621, 2019.

International conferences and workshops

– *Siyang Song, Micol Spitale, Cheng Luo, German Barquero, Cristina Palmero, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, Elisabeth André, Hatice Gunes.* **REACT2023: The First Multiple Appropriate Facial Reaction Generation Challenge.** In Proceedings of ACM Multimedia, 2023.

– *German Barquero, Sergio Escalera, and Cristina Palmero.* **Belfusion: Latent Diffusion for Behavior-driven Human Motion Prediction.** In Proceedings of the International Conference on Computer Vision, 2023.

– *Cristina Palmero, Julio C. S. Jacques Junior, Albert Clapés, Isabelle Guyon, Wei-Wei Tu, Thomas B. Moeslund, and Sergio Escalera.* **Understanding Social Behavior in Dyadic and Small Group Interactions: Preface.** In Understanding Social Behavior in Dyadic and Small Group Interactions, pp. 1-3. PMLR, 2022.

– *Cristina Palmero, German Barquero, Julio C. S. Jacques Junior, Albert Clapés, Johnny Núnez, David Curto, Sorina Smeureanu, Javier Selva, Zejian Zhang, David Saeteros, David Gallardo-Pujol, Georgina Guilera, David Leiva, Feng Han, Xiaoxue Feng, Jennifer He, Wei-Wei Tu, Thomas B. Moeslund, Isabelle Guyon, and Sergio Escalera.* **Chalearn LAP challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results.** In Understanding Social Behavior in Dyadic and Small Group Interactions, pp. 4-52. PMLR, 2022.

– *German Barquero, Johnny Núñez, Zhen Xu, Sergio Escalera, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero.* **Comparison of Spatio-Temporal Models for Human Motion and Pose Forecasting in Face-to-Face Interaction Scenarios.** In

Understanding Social Behavior in Dyadic and Small Group Interactions, pp. 107-138. PMLR, 2022.

- *German Barquero, Johnny Núnez, Sergio Escalera, Zhen Xu, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero.* **Didn't See That Coming: a Survey on Non-verbal Social Human Behavior Forecasting.** In Understanding Social Behavior in Dyadic and Small Group Interactions, pp. 139-178. PMLR, 2022.
- *David Curto, Albert Clapés, Javier Selva, Sorina Smeureanu, Julio C. S. Jacques Junior, David Gallardo-Pujol, Georgina Guilera, David Leiva, Thomas B Moeslund, and Sergio Escalera, and Cristina Palmero.* **Dyadformer: A Multi-Modal Transformer for Long-Range Modeling of Dyadic Interactions.** In Proceedings of the IEEE/CVF international conference on computer vision, pp. 2177-2188, 2021.
- *Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio C. S. Jacques Junior, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, David Leiva, and Sergio Escalera.* **Context-aware Personality Inference in Dyadic Scenarios: Introducing the UDIVA dataset.** In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1-12, 2021.
- *Julio C. S. Jacques Junior, Agata Lapedriza, Cristina Palmero, Xavier Baro, and Sergio Escalera.* **Person Perception Biases Exposed: Revisiting the First Impressions Dataset.** In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 13-21, 2021.

1.5 Further contributions

In addition to the main contributions of this thesis and the publications included above, we have contributed to the gaze estimation and eye-tracking communities in the form of code, novel datasets, and the co-organization of workshops and computational challenges.

Code and datasets

The code for Chapter 3 and the dataset of Chapter 5 are publicly available at `https://github.com/crisie`. The code for the gaze-related features of Chapter 6 will soon be available on the same website, while the data is available at `https://catalogue.elra.info/en-us/`.

Co-organization of workshops and challenges

We have participated in the organization of several workshops and computational challenges during the thesis period. We highlight two of them, led by Meta Reality Labs Research. First, the OpenEDS 2020 challenges held for the *OpenEyes: Eye Gaze in AR, VR, and in the Wild* workshop, in conjunction with the European Conference on Computer Vision in 2020, the objective of which was to foster advances in spatiotemporal gaze estimation and prediction (i.e., forecasting) and sparse temporal semantic segmentation using near-eye images from IR cameras, and which produced the associated publicly available OpenEDS2020 dataset (Palmero et al., 2020; Palmero et al., 2021b). And second, the OpenEDS 2021 challenges, held for the *OpenEDS 2021 Workshop on Eye Tracking for VR and AR: Sensors And Applications* in conjunction with the International Conference on Computer Vision in 2021.

1.6 Thesis outline

This thesis is divided into three parts, outlined below. All chapters are structured similarly, most of them including an introduction, specific related work, method, experimental results, limitations, and conclusions. Symbol definitions are not shared among chapters. While we dedicate a separate chapter to gaze estimation as the basis of this thesis, in the interest of completeness, we revisit certain term definitions not associated with gaze-specific topics across the different chapters.

- Prior to Part I, we include a background chapter, Chapter 2, reviewing the anatomy of the eye and eye movements that will appear throughout the thesis, a short history of eye tracking, a taxonomy of camera-based approaches, and an introduction to 3D gaze estimation.
- Part I consists of three chapters devoted to investigating the contribution of different sources of information for gaze estimation from a methodological perspective. Chapter 3 discusses the use of spatiotemporal and multimodal information for remote, off-the-shelf camera scenarios. Chapter 4 focuses on spatiotemporal information for IR-based near-eye camera scenarios. In a similar near-eye scenario, Chapter 5 studies single- and multirate sensor fusion to increase the accuracy and sampling rate of gaze tracking.
- Part II includes a single chapter, Chapter 6, exploring the use of gaze-related features for the task of emotion recognition in older adults when interacting with a virtual coach, individually and in combination with features from other modalities.
- Finally, Part III includes a single chapter, Chapter 7, with concluding remarks, future work, and discussion of ethical implications.

Chapter 2

Fantastic Eyes and How to Track Them

THE HUMAN EYE has been a subject of enduring interest for researchers due to its complex structure and dynamic functions. Over time, our understanding of the eye and its movements has evolved significantly, along with the evolution of technological approaches to measure them. In this background chapter, we introduce the eye structure and main eye movement types (Section 2.1). Then, we summarize the history of eye-tracking hardware (Section 2.2), and focus on the types of VOG approaches (Section 2.3). Finally, we detail the problem setting of 3D gaze estimation as a foundation for the next chapters (Section 2.4).

2.1 The eye and eye movements

The functioning of the human eye has long been compared to a camera (see Figure 2.1). The light reflected by an object passes through the pupil and the eye lens, which projects the light onto the *retina*, at the back of the eye. The retina contains *photoreceptors*, sensory cells that respond to light and can distinguish colors and luminance changes. When these photoreceptors sense light, they trigger action potentials that reach the optic disc and nerve, and are sent to the visual cortex and other parts of the brain, where visual information is processed. The intricate connection between the eyes and the brain constitutes the human visual system, which allows us to perceive our surroundings. Not all the retina is as color sensitive, though: there is a small pit, the *fovea*, which contains the highest amount of color photoreceptors (cones) and thus represents the area of highest visual acuity. This corresponds to around 1°, which is roughly the width of the thumbnail at arm's length (∼60 cm). The fovea is located approximately at $5.6° \pm 3°$ from the optic disc (Rohrschneider, 2004). The number of cones decreases as we move farther away from the fovea, but luminance photoreceptors (rods) start to appear, in what we call peripheral vision, which is more sensitive to motion and intensity changes. Consequently, we need to move our eyes to direct a specific area of the visible field of view toward the fovea, so that we can see it in high resolution and fine detail (Duchowski, 2017).

The dimensions of most of the eye structures are subject-dependent, such as the eyeball radii, the curvature of the cornea, the refraction index of the cornea, or the *kappa angle*. The kappa angle is the angle between the optical and visual axes. The *optical axis*, also known as *pupillary axis*, is the imaginary line that passes through the center of the pupil and the center of the eyeball, and thus can be estimated with model-based approaches without requiring personal calibration (see Section 2.3.1). However, it is the *visual axis*, also known as *line of sight*, the one that connects the object of interest with the fovea, and intersects with the optical axis at the center of

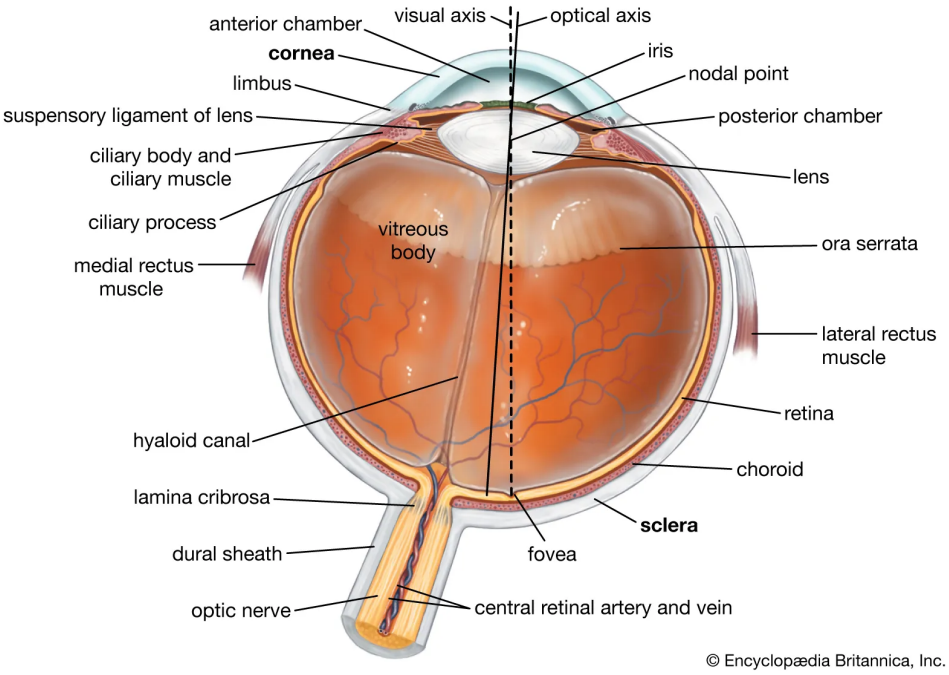


FIGURE 2.1: Cross section of the human eye. Reproduced from https://www.britannica.com/science/human-eye/Extraocular-muscles. © Encyclopædia Britannica, Inc.

corneal curvature. Gaze is directed along the visual axis. Thus, the kappa angle must be estimated for accurate gaze detection, which is usually done via user calibration. Calibration involves having the user look at specific (one or multiple) known points or targets in the 2D (or 3D) space while the eye tracker records their gaze data prior to starting the intended task. The term *line of gaze* differs in the literature by referring to either the optical axis (e.g., Hansen and Ji, 2010), or the visual axis (e.g., Model and Eizenman, 2010). We will adopt this term to denote the gaze direction in the 3D space to be estimated, regardless of the eye axis or coordinate system used (see Section 2.4).

There are six main types of eye movement: *fixational eye movements*, *smooth pursuit*, *saccade*, and *vergence* are used to maintain the visual target focused on the fovea, while *vestibulo-ocular* and *nystagmus* movements stabilize the eye when the head moves. Table 2.1 summarizes the main characteristics of each eye movement. These movements are carried out by three pairs of muscles that perform horizontal, vertical, and torsional motions. In this thesis, we will mostly refer to the three basic movements, depicted in Figure 2.2: fixations (stabilizing the fovea on a given stationary target), saccades (rapid movements between fixations), and smooth pursuit (slow movement that occurs when tracking a moving object).

Eye movements have particular dynamics and functioning (Robinson, 1968; Purves et al., 2001; Leigh and Zee, 2015). For instance, when fixating on a particular target, the eye is not still, but performs a series of miniature fixational eye movements. Otherwise, visual perception would fade completely because of neural adaptation. When a given stimulus elicits a saccade, the eye takes around 200 ms to initiate the movement toward the target location, which is usually referred to as latency. A post-saccadic oscillation is typically observed before the eye finally fixates on the target. Smooth pursuit eye movements are also characterized by an onset latency of 100-150 ms, and target tracking might not be consistent afterward:

TABLE 2.1: Main types of eye movements. Descriptions and typical characteristics are compiled from different sources (Collewijn and Kowler, 2008; Blignaut and Beelders, 2008; Holmqvist et al., 2011; Leigh and Zee, 2015; Duchowski, 2017; Graham et al., 2022), although values vary depending on the subject and experimental conditions.

Eye movement type	Description	Characteristics
Fixational	Eye movement that stabilizes the retina over a stationary target. Fixational eye movements include *microsaccades* (amplitude of 1-18 min arc, speed of 10°/s), *drift* (amplitude of 1.5-4 min arc, median speed of 4 min arc/s), and *tremor* (amplitude of 5-30 s arc, frequency of 90-200 Hz).	Duration: 50-600 ms Frequency: up to 3 Hz
Nystagmus	Involuntary, rhythmic oscillation of the eyes.	Amplitude: 2-3° Frequency: 2-3 Hz (depending on cause and type)
Saccade	Rapid, ballistic eye movements that reposition the fovea to a new location. They can be voluntary and reflexive. Visual information is not gathered during the saccadic movement (known as saccadic suppresion).	Duration: 10-100 ms Speed: 30-700°/s Amplitude: 1-30° Frequency: 4 Hz
Smooth Pursuit	Visually tracking a slowly, continuously moving target, where eyes can match the speed of the target.	Speed: up to 30°/s (for consistent tracking) Amplitude: depends on target
Vestibulo-ocular reflex	Reflex that stabilizes the eyes during head movement.	Gain*: 1.0 Speed: less than 10-ms lag after head movement
Vergence	Movement of both eyes in opposite directions to focus on distant or near targets.	Speed: 1-20°/s Amplitude: depends on target

* Ratio between eye and head motion.

this movement can combine catch-up saccades with predictive ones depending on the target velocity, the age of the user, and pharmacological conditions. The characteristics of saccades and smooth pursuit present a challenge to accurately map the eye rotation to the target position at a given time. Therefore, as one might expect, calibration is generally carried out during fixations, as the eyes are relatively stable and focus on a static target, which makes the mapping between target location and gaze direction easier. The reported accuracy of commercial eye trackers is typically measured during fixations as well.

2.2 Short history of eye tracking

The beginnings

The first known studies related to oculomotor anatomy and function date back to the late 17th century, when researchers provided the basis for future studies on the structure and function of the eye, comparing the latter with an optical device, or camera (Simon, 1975; Wade, 2010). Until the 19th century, eye behavior and eye movements were studied by pure observation or with *afterimages* of candle flames, an image that persists in the retina after extended exposure to the initial stimulus (Wade, 2015). Eye movements began to be related to sensorimotor functions due to the jerking eye movements observed during dizziness or vertigo. This movement is now known as one type of nystagmus (Wade, 2010). Researchers started to identify the need for accurate measurement of eye movements, and the only way this was possible was by attaching specific objects to the eye. One of the first known devices to

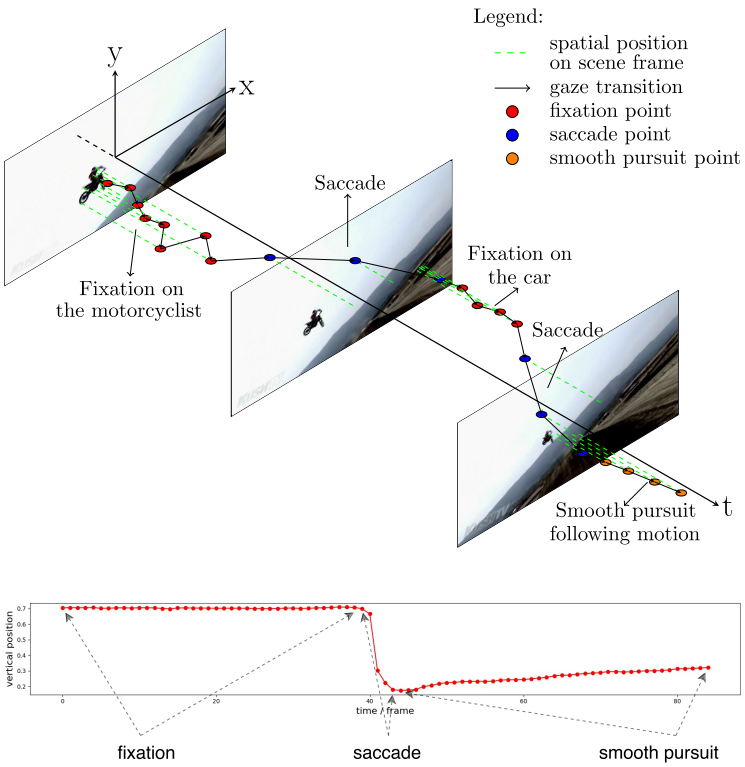


FIGURE 2.2: Representation of the dynamics of fixations, saccades, and smooth pursuit eye movements, while watching a video. Reproduced from Wang, Su, and Ji (2019), © 2019 IEEE.

objectively track eye movements was devised in the context of reading, for which a blunt point connected to a microphone was attached to the upper eyelid, and every time the bulge of the moving cornea bumped the microphone, an eye movement was counted (Wade and Tatler, 2008; Płużyczka, 2018). It was in this context that the term saccade was first coined. All subsequent approaches involved mechanical methods, such as attaching caps, suction cups, or plaster casts to the cornea (Huey, 1898; Delabarre, 1898). These would be connected to external devices where the eye movements would be recorded, such as kymographs (a cylindrical drum used to record measured changes in physiological phenomena). Apart from the extremely invasive nature of such approaches and the non-naturalistic experiments that they enabled, the temporal resolution of the recording devices (around 250 ms) was insufficient to measure fast eye movements (Huey, 1900), and head movement had to be restricted with bite bars to increase accuracy and spatial resolution. Still, mechanical invasive methods similar to these were used until the 1960s, as depicted in Figure 2.3c (Yarbus, 1967).

Toward non-invasive eye tracking

The first non-invasive optical tracker appeared in the early 20th century (Figure 2.3a), with which light reflected from the cornea surface was recorded onto a photosensitive photographic plate, allowing for the recording of horizontal eye movements, and later also vertical (Dodge and Cline, 1901; Judd, McAllister, and Steele, 1905; Dodge, 1926). This technique was first used to study eye movements in people with dementia, psychosis, and epilepsy, for which smooth pursuit eye movements were found to be different (Diefendorf and Dodge, 1908), and is the basis of most research

and commercial eye trackers. The technique was also used for reading and speaking research, among others (Płużyczka, 2018; Fletcher, Dunne, and Butler, 2022). With cinematographic advancements, photographic plates were later replaced by photographic tape and film cameras (Fitts, Jones, and Milton, 1950), first used to measure the eye movements of aircraft pilots during instrument-landing approaches.

All of these techniques still required a static head. The first mobile eye tracker was created in 1948, consisting of an optical device the participant held using a mouthpiece, enabling the recording of eye movements independently of head movements (Hartridge and Thomson, 1948). Later versions were tied to the forehead instead (Figure 2.3b). Around the mid-20th century, other eye-tracking devices appeared, which provided high accuracy and spatial resolution (Young and Sheena, 1975). EOG was one of them (Figure 2.3f), which allowed tracking of eye movements by measuring the corneo-retinal standing potential between the front and back of the eye by placing a series of electrodes on the skin around the eyes (Marg, 1951). Another was the magnetic scleral search coil (Figure 2.3g), mounted as a contact lens (Robinson, 1963).

In the 1970s, experimental psychology began studying the relationship between eye movements and cognitive (e.g., attention, memory) and linguistic processing. For instance, in the seminal work of Yarbus (1967) it was first determined that eye movements are task dependent, and that image saliency had an important role in attention (e.g., we first look at faces and eyes, followed by other areas where edges and contours predominate). This era was marked by improvements in eye-tracking technology, which provided more accurate measurements (Young and Sheena, 1975). The most prominent is the dual-Purkinje-image eye tracker (Cornsweet and Crane, 1973), an analog opto-electronic device that followed the principles of previous non-invasive eye trackers. More concretely, an IR beam was used to illuminate the eye, directed toward the pupil. This IR illumination, called *active*, is invisible to the human eye and, as such, does not distract participants or cause pupil contraction. The IR light is reflected by different structures of the inner and outer eye, causing at least four visible reflections called *Purkinje images*. The first (the corneal reflection, or *glint*) and the fourth images are tracked by dual-Purkinje-image eye trackers using a combination of lenses and servo-controlled mirrors. The first image is strong, while the fourth one is very weak. The distance between the two images is constant during eye translation, but changes during eye rotation, thus allowing for precise eye rotation measurement without interference from translational movements. These eye trackers provide extremely fast ($\sim$1 kHz) and accurate (up to 1 minute of arc) measurements, enabling them to determine the fastest and smallest eye movements. However, they require chin or forehead rests and bite bars to stabilize the head, which can be uncomfortable and difficult to achieve for many user groups such as infants, or clinical groups such as Parkinson's patients. Furthermore, they have a limited operational visual range (up to $\sim$15°) as the fourth image is occluded with extreme eye rotation. Dual-Purkinje eye trackers were conceived for desktop (also referred to as tower-mounted, as shown in Figure 2.3d) operation due to all the optics and motors they require, and are quite expensive as a consequence (more than 70,000$). They are still used today (sometimes with updated technology, see below) as high-end eye trackers for research or diagnostic applications that require very high accuracy and precise detection of saccades and microsaccades (Bowers and Poletti, 2017).

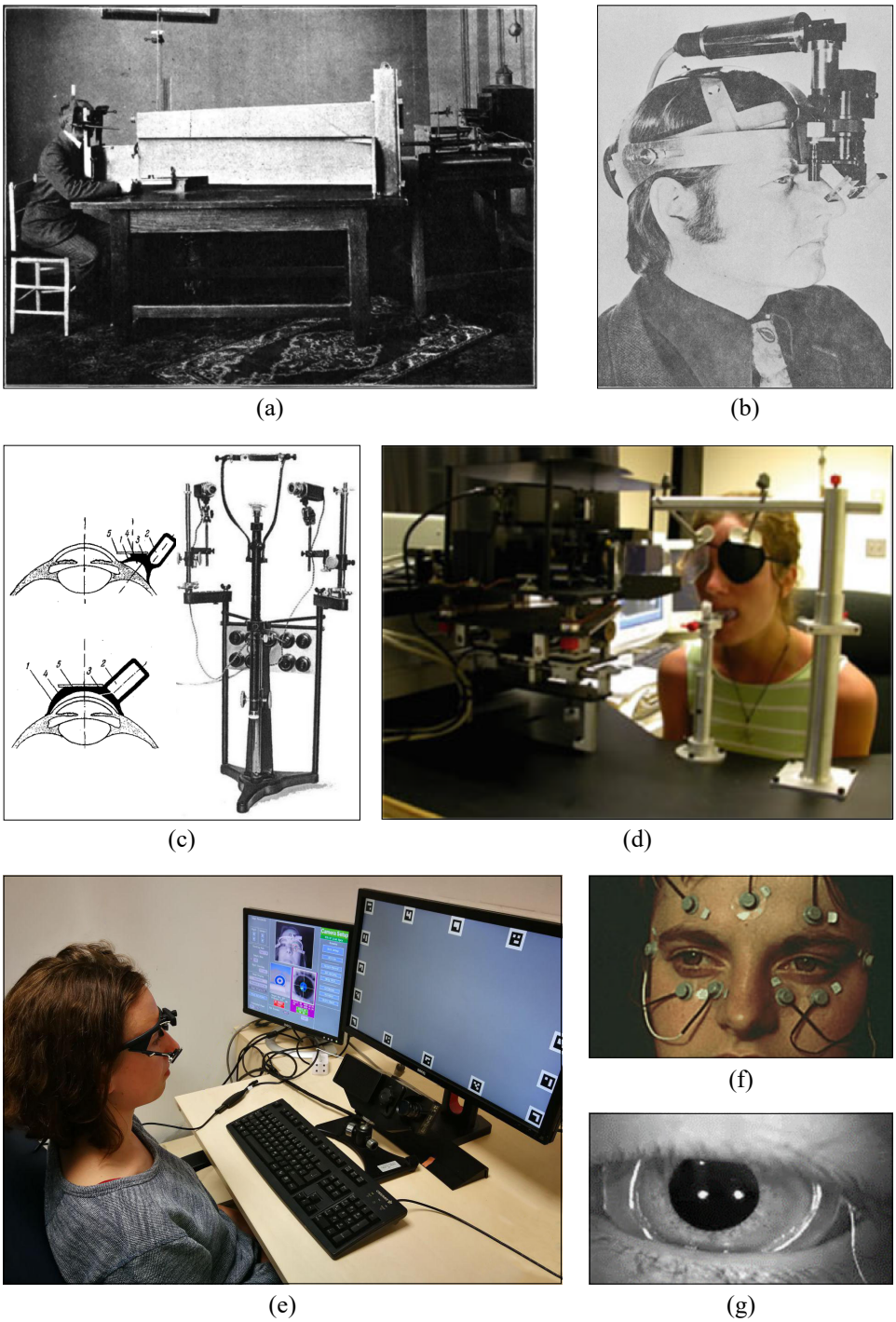


FIGURE 2.3: *The evolution of eye tracking*: (a) the photochronograph (Diefendorf and Dodge, 1908); (b) head-mounted eye tracker from the 1960s, combining corneal-reflex and front-facing cameras; (c) suction caps and recording device used by Yarbus (1967); (d) example of dual-Purkinje-image eye tracker; e) user employing a remote Eyelink 1000 eye tracker along with a mobile Pupil Labs tracker; (f) electrode placement in electro-oculography; e) search coil in use. Attributions: (a) is used with permission of Oxford University Press, from Diefendorf and Dodge (1908); (b) is reproduced from Young and Sheena (1975), (c) from Yarbus (1967), and (d), (f), and (g) from Hutton (2019), all with permission from Springer Nature; (e) is reproduced from Ehinger et al. (2019), under CC BY 4.0 (10.7717/peerj.7086/fig-1). Permissions of all images except (e) are conveyed through Copyright Clearance Center, Inc.

Video-oculography

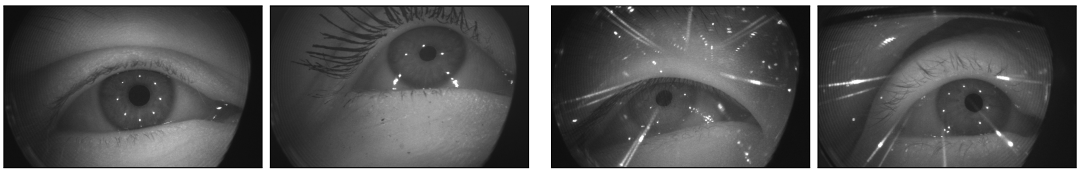
Technological advances in video capture and signal processing gave rise to camera- or video-based eye tracking in the 1980s, also known as VOG. VOG is currently the most widely used approach, being a much less obtrusive alternative while providing eye movement estimates with similar accuracy. Furthermore, it allows for faster data extraction and postprocessing compared to previous methods. VOG has usually relied on carefully calibrated setups of one or multiple IR cameras, and dedicated light sources, like LEDs. These would enhance the contrast between pupil and iris so that early CV edge detectors could be applied to detect and track them, and would also create glints. New versions of the dual-Purkinje-image eye tracker also incorporate cameras and digital signal processing techniques to track the images (Chamberlain, 1996). However, VOG approaches typically use glints only (the first Purkinje image) in combination with pupil or iris tracking. The best-dedicated desktop VOG trackers achieve accuracies of up to 0.5° at a lower cost than previous eye trackers (but are still on the order of thousands of dollars).

At that time, eye tracking began to be used in marketing and HCI research, with the development of gaze-contingent paradigms being a pivotal point (Fletcher, Dunne, and Butler, 2022). In addition, the high accuracy and increase of accessibility led to the design of eye movement metrics, or oculometrics, such as the amplitude and duration of a saccade, fixation counts, or heatmap and scanpath analysis (Mahanama et al., 2022). Still, a static head pose was required to maintain accuracy.

Where we stand

More recently, the miniaturization of electronics and continuous research in eye-tracking approaches have led to the development of consumer remote (Figure 1.5) or tethered head-mounted eye-tracking devices (Figure 1.6) that allow for small-to-large head movements with the former and almost free-head motion with the latter (Płużyczka, 2018). The accuracy, sampling rate, and robustness of these devices are close to static VOG setups (∼1-2° at 60-250 Hz) (Funke et al., 2016), but may degrade quickly with incorrect setup preparation for the former or headset slippage for the latter (Niehorster et al., 2018). During the last few years, battery-operated eye trackers have emerged, from more cumbersome head-mounted devices to lightweight, glasses-like form factors, which allow completely free movement (Kim et al., 2014). Beyond the impact of voluntary and involuntary camera shifts on performance, their most significant limitation is battery life: both cameras and image processing methods consume significant power, leading to a substantial reduction in the sampling rate (∼30-100 Hz) and/or simplification of approaches used, potentially resulting in compromised accuracy. Currently, prices range from 100$ to 10,000$ for low- and mid-end eye trackers, depending on the software and hardware they require, and the sampling rate, accuracy, spatial resolution, and robustness against head movements and other confounding factors they offer (Kasprowski, 2022). Figure 2.3e depicts an example of a recent high-end desktop eye tracker that allows head movement, along with a lightweight, portable tracker.

In general, relying on Purkinje images poses a problem for users wearing corrective lenses, as they can distort images or cause additional refraction glints or reflections as depicted in Figure 2.4, among other issues (note that corrective lenses will always make the eye look bigger or smaller regardless of the eye-tracking technique used, thus taking into account the glasses refractive index and other artifacts produced by them is required for optimal accuracy, e.g., Dahlberg, 2010). Furthermore,



(a) (b)

FIGURE 2.4: Examples of images (a) without glasses and (b) with glasses, captured with a VR headset equipped with an IR near-eye camera, representing different subject appearances and sensor shifts. Notice how the position of the glints changes with eye rotation, and how wearing glasses introduces additional reflections.

while IR lighting is ideal for indoor applications with controlled illumination, its performance decreases in uncontrolled scenarios and outdoor environments. For instance, environmental reflections may cause additional glints. Furthermore, IR light produced by natural sources such as sunlight can interfere with the eye-tracking IR illumination, leading to a drastic reduction in the quality of the eye image. Conventional approaches that leverage Purkinje images include model- and feature-based methods. In the last decade of the 20th century, researchers began investigating novel approaches to infer gaze without the need for dedicated eye-tracking devices and lighting, leveraging commercial cameras with natural illumination, or *passive*, and instituting what we now refer to as appearance-based methods (Hansen and Ji, 2010). In the next section, we dive into the different types of VOG approaches.

2.3 Taxonomy of camera-based approaches

VOG can be broadly classified into model-, feature-, and appearance-based methods. The choice of the method is usually tied to the device used, and primarily depends on the accuracy required, ease of use and access, and budget. We summarize the three types of approaches below, with an emphasis on appearance-based approaches as the main type considered in this thesis. We refer the reader to the comprehensive surveys of Hansen and Ji (2010), Kar and Corcoran (2017), Cazzato et al. (2020), Shehu et al. (2021), Ghosh et al. (2021), and Cheng et al. (2021) for a more detailed review of the literature.

2.3.1 Model-based methods

Model-based approaches, also called geometric-based, aim to fit a geometric 3D model of the eye to the pupil and/or iris, which can be detected with conventional edge detectors (Hansen and Ji, 2010) or recent deep segmentation approaches (Yiu et al., 2019; Kothari et al., 2021b). Such geometric model relies on the common physical structures of the eye, some of which are subject-specific. Nonetheless, the eye is usually approximated with a two-sphere model (for the eyeball and cornea, despite the fact that they are approximate ellipsoids) such as the LeGrand model (LeGrand and ElHage, 2013). A schematic of a simplified two-sphere model is depicted in Figure 2.5. The more simplified the model, the higher the error it may introduce to the system. Most subject-specific parameters of the eye model are usually fixed on the basis of anthropomorphic averages, such as the eyeball radii, while others, like the kappa angle, are generally estimated on a personal calibration stage. Some recent approaches aim to automate the calibration process (Model and Eizenman, 2010).

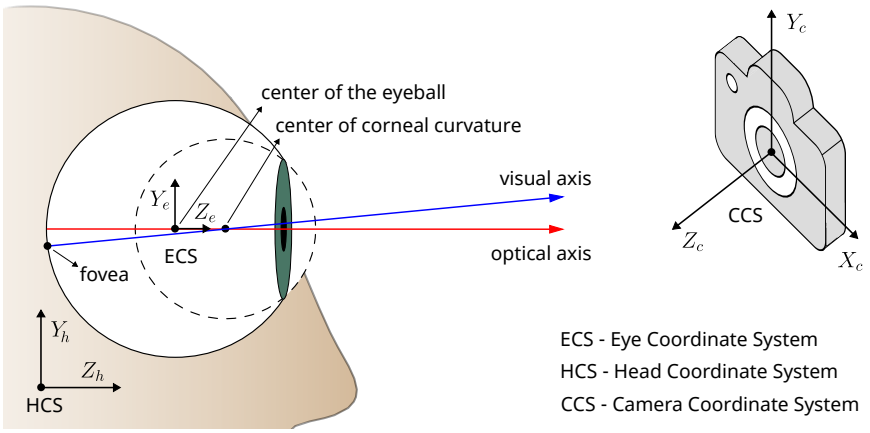


FIGURE 2.5: Schematic of a simplified two-sphere eye model (not to scale), commonly used in model-based gaze estimation. Here, the eye is portrayed to be looking at infinity, with its coordinate system sharing the same orientation as the HCS. The gaze direction can be represented in the HCS as the eye-in-head rotation, or in the CCS as a combination of the eye-in-head rotation and the head rotation with respect to the camera.

Model-based methods have been the standard for dedicated IR camera settings, leveraging Purkinje images to model the eye. The higher the number of cameras and light sources, the more parameters that can be faithfully detected without their explicit personal calibration (or with a lower number of target points), and the more robust against head motion and other sources of error they can be (Beymer and Flickner, 2003; Shih and Liu, 2004; Guestrin and Eizenman, 2006; Zhu and Ji, 2007; Villanueva and Cabeza, 2008). However, these systems require accurate geometric camera-lights and camera-screen calibration, and a high-resolution image of the eye, either by means of a near-eye camera placed on a headset, or a stationary desktop camera with a narrow field of view. More recent geometric approaches do not require dedicated systems and/or glints, and make use of 3D deformable eye-face models (Wang and Ji, 2017) with off-the-shelf RGB cameras and passive illumination instead. Eye movements can be detected from the estimated gaze traces, and the *point of regard* (PoR), or point of gaze, can simply be computed in the 3D space by intersecting the estimated line of gaze with the first object of interest in front of the user (e.g., specific point on screen). If we have the estimated lines of gaze of both eyes, we can also compute their intersection to help estimate the PoR in 3D spaces. The origin of the line of gaze is the estimated center of corneal curvature (or eyeball center if using the optical axis as a proxy for gaze instead).

2.3.2 Feature-based methods

Feature-based approaches usually rely on learning mapping functions between detected eye features and specific (usually 2D) target locations, to measure the PoR. One of the most widely used techniques is the *pupil center-cornea reflection*, also known as the *pupil-glint vector*, which uses the vector between the detected pupil center and the glint. Users are first required to look at specific points on the screen (which is also known as calibration, this time with a higher number of targets: usually 5 to 20), and the mapping between the vector and the target locations is usually learned with linear or polynomial regression. As such, these techniques are also called regression- or interpolation-based methods. Passive illumination approaches (Sesma, Villanueva, and Cabeza, 2012) and methods for recalibrating while using the system (Gomez and Gellersen, 2018) also exist.

As model-based approaches, they usually require either high-resolution images or near-eye cameras to have a high-fidelity view of the eye and thus ensure a high detection accuracy. But, by contrast, subject-specific parameters and gaze direction origin do not need to be explicitly inferred, and the geometric relationship between camera and light does not need to be known in advance. However, feature-based methods are highly affected by head movements, and thus require restraining head motion with chinrests or bite bars to ensure high accuracy in desktop settings. In head-mounted settings, however, headset slippage is more difficult to control. To overcome these issues, head motion and headset slippage compensation approaches have been proposed (Ji and Zhu, 2002; Zhu and Ji, 2005; Santini, Niehorster, and Kasneci, 2019). Another challenge found in both model- and feature-based approaches is the difficulty of properly detecting the pupil, which can be occluded by the eyelids and eyelashes, confused with mascara or eyeliner when using color-based segmentation approaches, and distorted with side camera views.

2.3.3 Appearance-based methods

Appearance-based methods directly map an image of the eye or face to a specific PoR location, such as a 2D screen position (known as 2D gaze estimation), or to a 3D gaze direction vector (known as 3D gaze estimation). The PoR can also be computed by performing 3D gaze estimation in addition to gaze origin estimation. Contrary to model- or feature-based approaches, appearance-based methods do not require high-resolution images or IR cameras, thus enabling remote gaze tracking with regular color cameras, like webcams, offering a trade-off between accuracy and accessibility. Different mapping functions have been explored, such as shallow neural networks (Baluja and Pomerleau, 1993), support vector machines (Zhu, Fujimura, and Ji, 2002), local interpolation (Tan, Kriegman, and Ahuja, 2002), gaussian processes (Williams, Blake, and Cipolla, 2006; Sugano, Matsushita, and Sato, 2013), adaptive linear regression (ALR) (Lu et al., 2011b) random forests (Sugano, Matsushita, and Sato, 2014; Huang, Veeraraghavan, and Sabharwal, 2017), or k-nearest neighbors (Wood et al., 2016b). Currently, CNNs variants are the state of the art for appearance-based gaze estimation (Zhang et al., 2015).

The first appearance-based models were subject-specific (Baluja and Pomerleau, 1993); that is, a single model was learned per user after a calibration stage, which resembles feature-based methods. However, with advances in ML and later in DL, it was shown that it was possible to learn useful features across subjects and appearances by learning from calibration data from multiple subjects (Mora and Odobez, 2013). This enabled the creation of subject-independent models, which could be applied to users not seen during training, removing the need for user calibration. As with any appearance-based ML/DL approach, these models need large variability during training to achieve generalization. Consequently, this finding boosted the creation of (large-scale) gaze estimation datasets, from laboratory conditions (Funes Mora, Monay, and Odobez, 2014a) to in-the-wild (Zhang et al., 2017c), first for remote-camera scenarios and later for near-eye ones (Kim et al., 2019).

The main challenges of appearance-based methods are head pose or camera viewpoint variability, illumination changes, subject variability with respect to eye geometry and appearance, and subject invariance without subject-specific calibration. While widely varied, large-scale datasets help in achieving some level of generalization for these challenges, it is still an open challenge to generalize to different scenarios and settings, and to capture the appearance and geometric variability of the entire population. To address these issues, very recent works have gone beyond

modeling gaze estimation as a fully supervised task and started exploring weakly-, self-, and unsupervised DL approaches to further increase generalization on different axes (Yu and Odobez, 2020; Kothari et al., 2021a). In addition, appearance-based approaches allow for the creation of generic subject-independent models that can be further improved by using a few calibration images of a target individual. This can be done for example via few-shot personalization (Park et al., 2019), adding a subject-specific model on top (Krafka et al., 2016), or via differential approaches (Liu et al., 2018). Currently, appearance-based approaches are capable of obtaining accuracies of 3-10° without calibration (depending on the dataset), further improving to up to around 2° upon a personal calibration stage (Zhang et al., 2020). Another line of work is the exploration of additional cues to increase accuracy and robustness, such as temporal and multimodal information as in the case of this thesis, or multiple camera views (Jindal and Manduchi, 2023). Furthermore, decreasing system complexity and increasing sampling rate for achieving real-time and/or high-speed eye tracking is also an important goal (Sewell and Komogortsev, 2010; Gudi, Li, and Gemert, 2020; Angelopoulos et al., 2021).

2.4 3D gaze estimation

This section provides an overview of essential components of the 3D gaze estimation task, setting the stage for the next chapters.

Problem setting

The goal of 3D gaze estimation is to estimate the line of gaze (i.e., gaze direction). This line of gaze can be represented as a 3D unit vector in Cartesian coordinates $\mathbf{g}_3 = (g_x, g_y, g_z)$, or as a 2D angle in spherical coordinates $\mathbf{g}_2 = (\theta, \varphi)$, where θ and φ correspond to the horizontal and vertical components of the gaze direction, respectively. The 2D gaze angle can be converted into a 3D unit vector as follows:

$$\begin{aligned} g_x &= \sin(\theta)\cos(\varphi), \\ g_y &= \sin(\varphi), \\ g_z &= \cos(\theta)\cos(\varphi), \\ \mathbf{g}_3 &= \mathbf{g}_3 / \|\mathbf{g}_3\|, \end{aligned} \tag{2.1}$$

and vice versa:

$$\begin{aligned} \theta &= \text{atan2}(g_x, g_z), \\ \varphi &= \text{asin}(g_y), \end{aligned} \tag{2.2}$$

such that (0, 0) corresponds to looking straight ahead. The signs of each component will depend on the orientation of the coordinate system of the line of gaze. This 2D angle representation is not to be confused with 2D gaze estimation, which aims to infer the 2D PoR (e.g., location on a screen), and is not covered explicitly in this thesis. Eye movements can be measured directly from the 2D angle representation.

Coordinate systems

The line of gaze can be estimated with respect to different coordinate systems. The two main coordinate systems that we use in this thesis are the head coordinate system (HCS) and the camera coordinate system (CCS), depicted in Figure 2.5.

In the HCS, the line of gaze refers to the eye-in-head rotation, being independent of head or camera pose. It may correspond to the eye's optical or visual axis when the input is a single eye image (i.e., monocular gaze estimation), or to a rough average of the visual/optical axes of both eyes when using full-face images as input. Of course, one can also infer the optical/visual axes of each eye for the latter type of input (i.e., binocular gaze estimation). By contrast, the line of gaze in the CCS is a combination of the head pose with respect to the camera and the eye-in-head rotation. Hence, the gaze direction will change with changes in camera viewpoint and head movements.

One can easily transform the eye-in-head rotation to the line of gaze in the 3D camera space. Let $\mathbf{g}_h \in \mathbb{R}^3$ represent the visual axis in the HCS, and $\mathbf{R} \in \mathbb{R}^{3x3}$ the head rotation matrix in the CCS. The 3D gaze direction in the CCS, represented by $\mathbf{g}_c \in \mathbb{R}^3$, can be computed as:

$$\mathbf{g}_c = \mathbf{R}\mathbf{g}_h. \tag{2.3}$$

Gaze origin

In 3D gaze estimation, the 2D or 3D PoR can be computed by the intersection of the line of gaze with the screen or a given object in the 3D space. To do so, it is also necessary to infer the 3D origin of the line of gaze $\mathbf{o} \in \mathbb{R}^3$.

When working with full-face images, which usually consider a single line of gaze, the gaze origin is usually represented as the center of the facial landmarks or as the midpoint between both eyes, depending on the dataset. To find this point, 2D facial alignment approaches are commonly employed (Baltrušaitis et al., 2018; Guo et al., 2020), which infer 2D facial landmarks. Then, given a calibrated camera (i.e., camera intrinsic parameters are known), a 3D face morphable model (3DMM, Blanz and Vetter, 1999) can be fit to the detected 2D landmarks and apply a perspective-n-point (PnP) algorithm (Fischler and Bolles, 1981) to estimate the 3D head position and orientation in the CCS, as well as the 3D eye position. The 3D head pose can also be obtained with higher precision by using motion capture or depth sensors, if available (Funes-Mora and Odobez, 2016; Stone et al., 2022). The origin estimation process via facial alignment can also be applied to eye-only images in a similar fashion, where the origin can be roughly estimated as the midpoint between the inner and outer eye corners. Nonetheless, for monocular gaze estimation in devices that require higher precision, eye model-based approaches are usually applied to find the center of the cornea (for the visual axis), or eyeball (for the optical axis). For either case, some recent approaches aim to estimate both gaze origin and line of gaze together (Kaur, Jindal, and Manduchi, 2022; Balim et al., 2023).

Accuracy metrics

Accuracy (or error) is usually computed as the angular error between the estimated $\hat{\mathbf{g}}_3$ and ground truth $\mathbf{g}_3$ gaze angles, in degrees. The angular error is defined as:

$$d(\mathbf{g}_3, \hat{\mathbf{g}}_3) = \arccos\left(\frac{\mathbf{g}_3 \cdot \hat{\mathbf{g}}_3}{\|\mathbf{g}_3\|\|\hat{\mathbf{g}}_3\|}\right). \tag{2.4}$$

Nonetheless, the 2D angle representation also allows for other accuracy metrics, such as the mean squared error or the mean absolute error (MAE). We use the latter in Chapter 4:

$$d(\mathbf{g}_2, \hat{\mathbf{g}}_2) = \frac{1}{d} \sum_{i=1}^{d} |g_i - \hat{g}_i|, \tag{2.5}$$

where $d = 2$.

Part I

Methods

Chapter 3

Multimodal and Spatiotemporal Cues for Remote Gaze Estimation

AS PREVIOUSLY introduced, gaze behavior is an important non-verbal cue for a myriad of applications, many of them requiring a fast deployment without the need for personal calibration and allowing free head and body movement. In this first methodological chapter, we tackle the problem of person- and head pose-independent 3D gaze estimation from remote, off-the-shelf cameras, using a multimodal convolutional-recurrent deep network. We propose to combine face, eyes region, and face landmarks as individual streams in a CNN to estimate gaze in still images. Then, we exploit the dynamic nature of gaze by feeding the learned features of all the frames in a sequence to a many-to-one recurrent module that predicts the 3D gaze vector of the last frame. Our multimodal static solution is evaluated on a wide range of head poses and gaze directions, achieving a significant improvement over the state of the art on the EYEDIAP dataset, further improved when the temporal modality is included.

3.1 Introduction

Many existing gaze tracking systems are operated under laboratory conditions with fixed head settings after a user-specific calibration stage (Hansen and Ji, 2010). Dedicated hardware, such as IR light sources or wearable devices, is usually employed, where the camera is near the subject and therefore, a high-resolution image is available. Despite the high accuracy of such systems, they are not suitable for assessing gaze behavior in naturalistic contexts or less constrained HCI/HMI tasks, where a non-obtrusive system is preferred. Examples include observational behavior studies between children and caregivers (Gardner, 2000), measurement of audience attention in public displays (Sugano, Zhang, and Bulling, 2016), or communication with VCs (Castellano et al., 2013). In such cases, remote camera-based systems offer a trade-off between usability and accuracy.

Recent gaze estimation research has focused on facilitating the use of eye tracking in general everyday applications under real-world conditions, using off-the-shelf remote RGB cameras and removing the need for personal calibration. In this setting, appearance-based methods, which learn a mapping from images to gaze directions, are the preferred choice (Shehu et al., 2021). These methods are commonly posed as supervised ML/DL problems. As such, they need large amounts of training data to be able to generalize well to in-the-wild situations, which are characterized by significant variability in head poses, face appearances, and lighting conditions. In recent years, CNNs have been reported to outperform classical ML methods (Zhang et al., 2015). However, most existing approaches are generally tested in restricted

HCI tasks, where users look at the screen or phone, featuring low head pose variability. It is not clear how these methods perform in a wider range of head poses.

On a different note, the majority of gaze estimation methods used to rely only on the static appearance of the eye region as input. Recent approaches have demonstrated that the use of the face along with a higher-resolution image of the eyes (Krafka et al., 2016), or even just the face itself (Zhang et al., 2017b), increases performance. Indeed, the whole-face image encodes more information than eyes alone, such as illumination and head pose. Nevertheless, gaze behavior is not static. Eye and head movements allow us to direct our gaze to target locations of interest. It has been demonstrated that humans can better predict gaze when shown image sequences of other people moving their eyes (Anderson, Risko, and Kingstone, 2016). However, it is still an open question whether this sequential information can increase the performance of automatic methods.

In this chapter, we investigate whether the combination of multiple cues extracted from the RGB camera signal benefits the gaze estimation task. In particular, we use face, eye region, and facial landmarks from still images. Facial landmarks model the global shape of the face and come at no cost, since face alignment is a common preprocessing step in many facial image analysis approaches (Jin and Tan, 2017), including gaze estimation. Furthermore, we present a subject-independent, head-pose-invariant recurrent 3D gaze regression network to leverage the temporal information of image sequences. The static streams of each frame are combined in a feature-based fashion using a multistream CNN. Then, all feature vectors are input to a many-to-one recurrent module that predicts the gaze vector of the last sequence frame. RNNs have a long history in the ML literature, being widely applied for sequential data modeling (Medsker and Jain, 2001; Salehinejad et al., 2017). Thus, we select them for our approach as they offer a natural path to addressing our hypothesis in an end-to-end fashion.

In summary, our contributions are two-fold. First, we present a CNN-Recurrent network architecture that combines appearance, shape, and temporal information for 3D gaze estimation. Second, we test static and temporal versions of our solution on the EYEDIAP dataset (Funes Mora, Monay, and Odobez, 2014a) (described in Section 3.4.1) in a wide range of head poses and gaze directions, showing consistent performance improvements compared to related appearance-based methods. To the best of our knowledge, this is the first third-person, remote camera-based approach that uses temporal information for this task. Table 3.1 outlines our main method characteristics compared to related work.

The remainder of this chapter is organized as follows. Section 3.2 reviews recent DL-powered, appearance-based gaze estimation approaches. Section 3.3 describes the proposed methodology and implementation details. Section 3.4 details the dataset used for the experiments, as well as the experimental evaluation for static and spatiotemporal approaches. Section 3.5 lists possible limitations of our approach and experimental evaluation. Finally, 3.6 concludes the chapter.

3.2 Related work

As described in Section 2.3, gaze estimation methods are generally classified into model-, feature-, or appearance-based (Hansen and Ji, 2010; Ferhat and Vilariño, 2016; Kar and Corcoran, 2017). Appearance-based methods learn a direct mapping from intensity images to gaze directions, thus being potentially applicable to relatively low-resolution images and mid-distance scenarios (up to 2 m approximately,

Method	3D gaze direction	Unrestricted gaze target	Full face	Eye region	Facial landmarks	Sequential information
Zhang et al. (2015)	✓	✗	✗	✓	✗	✗
Krafka et al. (2016)	✗	✗	✓	✓	✗	✗
Zhang et al. (2017b)	✓	✗	✓	✗	✗	✗
Deng and Zhu (2017)	✓	✓	✓	✓	✗	✗
Ours	✓	✓	✓	✓	✓	✓

TABLE 3.1: Characteristics of recent related work on person- and head pose-independent appearance-based gaze estimation using CNNs applied to remote-camera settings.

or as long as the eyes are visible). Main challenges of appearance-based methods for 3D gaze estimation are head pose, illumination, and subject invariance without user-specific calibration. To address head pose issues, some works proposed compensation methods (Lu et al., 2011a) and warping strategies that synthesize a canonical, frontal view of the face (Mora and Odobez, 2012; Funes-Mora and Odobez, 2016; Jeni and Cohn, 2016). Hybrid approaches based on analysis-by-synthesis have also been evaluated (Wood et al., 2016a).

Currently, data-driven methods are considered the state of the art for appearance-based gaze estimation with off-the-shelf cameras. Consequently, a number of RGB gaze estimation datasets have been introduced in recent years, either in controlled (Smith et al., 2013) or semi-controlled settings (Funes Mora, Monay, and Odobez, 2014a), in the wild (Zhang et al., 2015; Krafka et al., 2016), or consisting of synthetic data (Sugano, Matsushita, and Sato, 2014; Wood et al., 2015; Wood et al., 2016b). Zhang et al. (2015) showed that CNNs could outperform other mapping methods, using a multimodal CNN to learn the mapping from 3D head poses and eye images to 3D gaze directions. Krafka et al. (2016) proposed a multistream CNN for 2D gaze estimation, using individual eye and whole-face images, along with a face grid as input. As this method was limited to 2D screen mapping, Zhang et al. (2017b) later explored the potential of just using whole-face images as input to estimate 3D gaze directions. Using a spatial-weights CNN, their method was demonstrated to be more robust to facial appearance variation caused by head pose and illumination than eye-only methods. Although the method was evaluated in the wild, the subjects only interacted with a mobile device, thus restricting the head pose range. Deng and Zhu (2017) presented a two-stream CNN to disjointly model head pose from face images and eyeball movement from eye region images. Both were then aggregated into the 3D gaze direction using a gaze transform layer. The decomposition was aimed at avoiding the head-gaze correlation overfitting of previous data-driven approaches. They evaluated their approach in the wild with a wider range of head poses, obtaining better performance than previous eye-based methods. However, they did not test it on publicly annotated benchmark datasets.

Instead, we propose a multistream CNN-recurrent network for person- and head pose-independent 3D gaze estimation for a mid-distance scenario. We evaluate it on a publicly available dataset featuring a wider range of head poses and gaze directions than those restricted to screen interaction. Unlike previous methods, we also rely on explicit facial geometric cues, as well as temporal information inherent in sequential data.

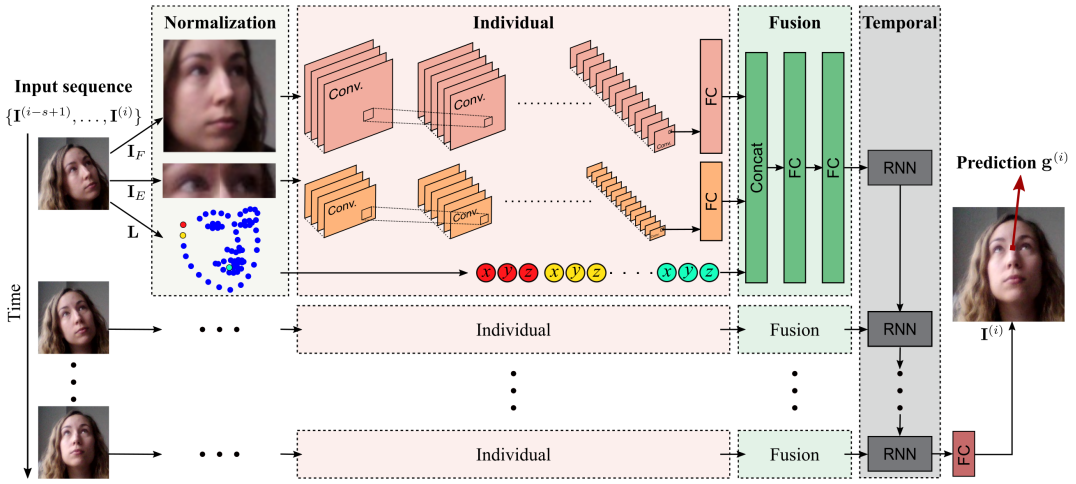


FIGURE 3.1: Overview of the proposed network. A multistream CNN jointly models full-face, eye region appearance, and face landmarks from still images. The combined extracted features from each frame are fed into a recurrent module to predict the gaze direction of the last frame of a video sequence.

3.3 Methodology

In this section, we present our approach for 3D gaze regression based on appearance and shape cues for still images and image sequences. First, we introduce the data modalities and formulate the problem. Then, we detail the normalization procedure prior to the regression stage, which has become a standard preprocessing step in remote, appearance-based gaze estimation. Finally, we explain the global network topology and implementation details. An overview of the system architecture is depicted in Figure 3.1.

3.3.1 Multimodal gaze regression

Let us represent the direction of gaze as a single 3D unit vector per face image $\mathbf{g} = (g_x, g_y, g_z) \in \mathbb{R}^3$ in the CCS, whose origin is the central point between the two eyeball centers. Assuming a calibrated camera, and known 3D head position and orientation, our goal is to estimate $\mathbf{g}$ from a sequence of images $\{\mathbf{I}^{(i)} \mid \mathbf{I} \in \mathbb{R}^{W \times H \times 3}\}$ as a regression problem, where i represents the image index, and W and H are the weight and height of an image, respectively.

Gazing to a specific target is achieved by a combination of eye and head movements, which are highly coordinated. Consequently, the apparent direction of gaze is influenced not only by the location of the pupil and iris within the eye socket, but also by the position and orientation of the face with respect to the camera. Known as the Wollaston effect (Wollaston, 1824), the same set of eyes may appear to be looking in different directions due to the surrounding facial cues (see Figure 3.2). Therefore, it is reasonable to state that eye images are not sufficient to estimate gaze direction. Instead, whole-face images can encode head pose or illumination-specific information across larger image areas than those available just in the eye region (Zhang et al., 2017b; Krafka et al., 2016).

One of the drawbacks of appearance-only methods is that global structure information is not explicitly considered. In that sense, facial landmarks can be used as global shape cues to encode spatial relationships and geometric constraints. Current

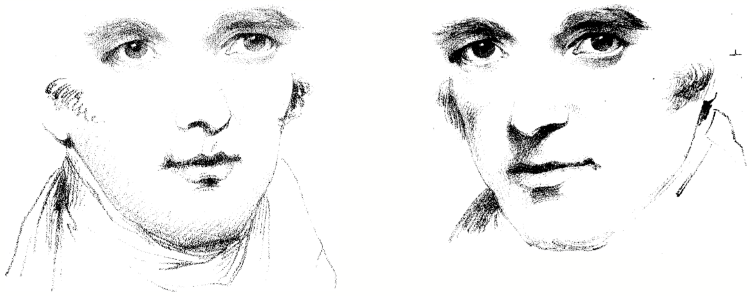


FIGURE 3.2: The exact same set of eyes is used in these two images; however, the perceived gaze direction is different, influenced by the head rotation and surrounding facial cues. This is known as the *Wollaston effect*. Used with permission of The Royal Society (U.K.), from Wollaston (1824); permission conveyed through Copyright Clearance Center, Inc.

state-of-the-art face alignment approaches are robust enough to handle large appearance variability, extreme head poses, and occlusions (Jin and Tan, 2017). Facial landmarks are mainly correlated with head orientation, eye position, eyelid openness, and eyebrow movement, which are valuable features for our task.

Therefore, in our approach, we jointly model appearance and shape cues (see Figure 3.1). The former is represented by a whole-face image $\mathbf{I}_F$, along with a higher resolution image of the eyes $\mathbf{I}_E$ to identify subtle changes. Due to dealing with wide head pose ranges, some eye images may not depict the whole eye, containing mostly background or other surrounding facial parts instead. For that reason, and contrary to previous approaches that only use one eye image (Sugano, Matsushita, and Sato, 2014; Zhang et al., 2015), we use a single image composed of two patches of centered left and right eyes. The shape modality is represented by 3D face landmarks following the Multi-PIE 68-landmark model scheme (Gross et al., 2010), denoted by $\mathbf{L} = \{(l_x, l_y, l_z)_c \mid \forall c \in [1, ..., 68]\}$. Finally, we also consider the dynamic component of gaze. We leverage the sequential information of eye and head movements such that, given the appearance and shape features of consecutive frames, it is possible to better predict the gaze direction of the current frame.

3.3.2 Data normalization

Before feeding the input to the network, a normalization step in the 3D space and the 2D image is usually carried out for appearance-based, remote-camera approaches. It is performed to reduce the appearance variability and, consequently, the degrees of freedom. In addition, it allows the gaze estimation model to be applied regardless of the original camera configuration. The normalization was first presented by Sugano, Matsushita, and Sato (2014) and later refined by Zhang, Sugano, and Bulling (2018) concurrently with our work. Figure 3.3 provides a schematic example of the normalization process in the 3D space.

Let $\mathbf{H} \in \mathbb{R}^{3x3}$ be the head rotation matrix, and $\mathbf{p} = [p_x, p_y, p_z]^T \in \mathbb{R}^3$ the reference face location (the red point in Figure 3.3) with respect to the original CCS. The goal is to find the conversion matrix $\mathbf{M}$:

$$\mathbf{M} = \mathbf{S}\mathbf{R}, \tag{3.1}$$

such that (a) the X-axes of the virtual camera and the head are on the same plane using the rotation matrix $\mathbf{R}$, and (b) the virtual camera looks at the reference location from a fixed distance d_n using the Z-direction scaling matrix $\mathbf{S}$:

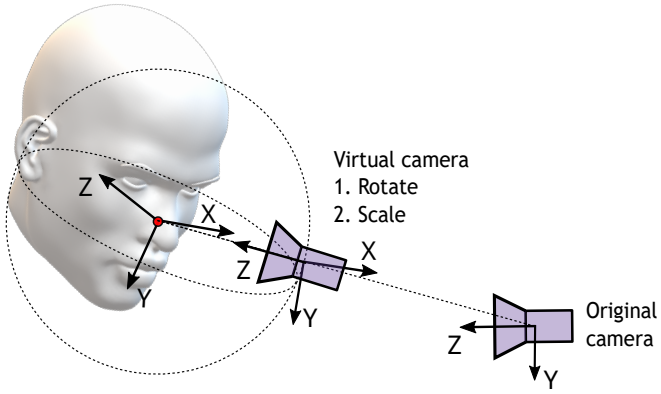


FIGURE 3.3: Schematic of the data normalization process in the 3D space. The original camera is transformed into a virtual camera with fixed intrinsic camera parameters by 1) rotating it such that the X-axes of the camera and head are on the same plane and the camera looks at a reference location (red landmark) of the face, and 2) scaling the camera such that it is always at a given distance to the reference location.

$$\mathbf{S} = diag(1, 1, d_n / \|\mathbf{p}\|). \tag{3.2}$$

$\mathbf{R}$ is computed as:

$$\mathbf{a} = \hat{\mathbf{p}} \times \mathbf{H}_x, \tag{3.3}$$

$$\mathbf{b} = \hat{\mathbf{a}} \times \hat{\mathbf{p}}, \tag{3.4}$$

$$\mathbf{R} = [\hat{\mathbf{b}}, \hat{\mathbf{a}}, \hat{\mathbf{p}}]^T, \tag{3.5}$$

where $\mathbf{H}_x$ denotes the x-axis of $\mathbf{H}$, and $\langle\, \hat{\cdot} \,\rangle$ is the unit vector.

This normalization translates into the image space as a cropped image patch of size $W_n \times H_n$ centered on $\mathbf{p}$ where the head roll rotation has been removed (see the *normalization* module of Figure 3.1). This is done by applying a perspective warping to the input image $\mathbf{I}$ using the transformation matrix $\mathbf{W}$:

$$\mathbf{W} = \mathbf{C}_n \mathbf{M} {\mathbf{C}_o}^{-1}, \tag{3.6}$$

where $\mathbf{C}_o$ and $\mathbf{C}_n$ are the original and virtual camera matrices, respectively.

The 3D gaze vector is also normalized as:

$$\mathbf{g}_n = \mathbf{R}\mathbf{g}. \tag{3.7}$$

After image normalization, the gaze direction can be represented in the 2D space. Therefore, $\mathbf{g}_n$ is further transformed to spherical coordinates (θ, φ) assuming unit length, following Equation 2.2.

3.3.3 Convolutional-Recurrent Neural Network

We propose a CNN-RNN regression network for 3D gaze estimation (see Figure 3.1). The network is divided into 3 modules: (1) *Individual*, (2) *Fusion*, and (3) *Temporal*.

First, the *Individual* module learns features from each appearance cue separately. It consists of a two-stream CNN, one devoted to the normalized face image stream,

FIGURE 3.4: Architecture of a full-face-only static gaze estimation network with a VGG-16 backbone (Parkhi, Vedaldi, and Zisserman, 2015). The numbers in the convolutional layers (in orange) correspond to the number of feature maps, while the numbers in the fully connected layers (green) correspond to the number of hidden units.

and the other to the joint normalized eyes image. Next, the *Fusion* module concatenates the extracted features of each appearance stream in a single vector along with the normalized landmark coordinates. Then, it learns a joint representation between modalities in a feature-based fusion fashion. Both *Individual* and *Fusion* modules, henceforth referred to as *Static* model, are applied to each frame of the sequence.

Finally, the resulting feature vectors of each frame are fed to the *Temporal* module, based on a many-to-one recurrent network. RNNs are characterized by their recurrent connections, enabling them to consider the context from previous inputs when processing the current input. A many-to-one network is one way of using recurrent networks, commonly applied to classification tasks (Donahue et al., 2015), wherein a sequence of feature vectors is used as input, one per time step, and the output of the final step is used as output of the network. It assumes that the same number of input steps will be used during training and testing. This *Temporal* module leverages the sequential information from the feature vectors to predict the normalized 2D gaze angles of the last frame of the sequence using a linear regression layer added on top of it. The combination of all modules is referred to as *Temporal* model.

3.3.4 Implementation details

Network details

Each stream of the *Individual* module is based on the VGG-16 deep network (Parkhi, Vedaldi, and Zisserman, 2015)[2], and consists of 13 convolutional layers, five max-pooling layers, and one fully connected (FC) layer with rectified linear unit (ReLU) activations. The full-face stream follows the same configuration as the base network, having an input of 224×224 pixels and a 4096D FC layer. By contrast, the input joint eye image is smaller, with a final size of 120×48 pixels, so the number of parameters is decreased proportionally. In this case, its last FC layer produces a 1536D vector. The output of each FC layer is concatenated along with the 204D landmark-coordinates vector, resulting in a 5836D feature vector. The *Fusion* module consists of two 5836D FC layers with ReLU activations and two dropout layers between FCs as regularization. Figure 3.4 depicts the architecture of a full-face-only stream network as an example. Finally, to model the temporal dependencies, we use a single gated recurrent unit (GRU) layer with 128 units. GRUs, as well as long short-term memory (LSTM) RNNs, are two types of RNN capable of considering longer-term dependencies than vanilla RNNs, reducing vanishing and exploding gradient issues (Graves, 2012; Chung et al., 2014). Both GRUs and LSTMs consist of gates that modulate the information flow, but GRUs have fewer parameters. Consequently, they often show better performance than LSTMs when the training data size is limited.

The network is trained in a stage-wise fashion. First, we train the *Static* model and the final regression layer end-to-end on each individual frame of the training

[2]VGG-16 implementation: https://keras.io/api/applications/vgg/.

data. The convolutional blocks and the first FC layer are pretrained with the VGG-Face dataset (Parkhi, Vedaldi, and Zisserman, 2015), whereas the remaining FCs are trained from scratch. The VGG-Face dataset contains over 2.6M full-face images from 2,622 identities in-the-wild, that is, with a wide variability of camera viewpoints, head poses, illumination conditions, image quality, etc. Second, the training data are rearranged by means of a sliding window with a stride of 1 frame to build input sequences. Each sequence is composed of $s = 4$ consecutive frames ($\sim$133 ms), with the gaze direction target being the gaze direction of the last frame of the sequence $\left(\{\mathbf{I}^{(i-s+1)}, \ldots, \mathbf{I}^{(i)}\}, \mathbf{g}^{(i)}\right)$. Using the rearranged training data, we extract features of each frame of the sequence from a frozen (i.e., network weights are not updated) *Individual* module, finetune the *Fusion* layers, and train both, the *Temporal* module and a new final regression layer from scratch. This way, the network can exploit the temporal information to further refine the fusion weights.

We trained the model using the ADAM optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.0001, dropout of 0.3, and batch size of 64 frames. The number of epochs was experimentally set to 21 for the first training stage and 10 for the second. We use the average Euclidean distance between the predicted and ground-truth 3D gaze vectors as loss function.

Input preprocessing

For this work, we used the head pose and eye locations in the 3D scene provided by the dataset (presented in Section 3.4.1). The 3D landmarks are extracted using the state-of-the-art method of Bulat and Tzimiropoulos (2017)[3], which is based on stacked hourglass networks (Newell, Yang, and Deng, 2016).

During training, the original image is preprocessed to get the two normalized input images. The normalized whole-face patch is centered 0.1 m ahead of the head center, and $\mathbf{C}_n$ is defined such that the image has size of 250×250 pixels. The difference between this size and the final input size allows us to perform random cropping and zooming to augment the data (explained in Section 3.4.1). Similarly, each normalized eye patch is centered at their respective eye center locations. In this case, the virtual camera matrix is defined so that the image is cropped to 70×58, while in practice, the final patches have size of 60×48. Landmarks are normalized using the same procedure and further preprocessed with mean subtraction and min-max normalization per axis. Finally, we divide them by a scaling factor w so that all coordinates are in the range $[0, w]$. This way, all concatenated feature values are in a similar range. After inference, the predicted normalized 2D angle is denormalized back to the original 3D space.

3.4 Experiments

In this section, we evaluate the cross-subject 3D gaze estimation task on a wide range of head poses and gaze directions. Furthermore, we validate the effectiveness of the proposed architecture comparing both static and temporal approaches. We report the error in terms of mean angular error (Equation 2.4) between predicted and ground-truth 3D gaze vectors. Note that due to the requirements of the temporal model, not all the frames obtain a prediction (e.g., the first frames of a video). Therefore, for a fair comparison, the reported results for static models disregard such frames when temporal models are included in the comparison.

[3]Face alignment model: https://github.com/1adrianb/face-alignment.

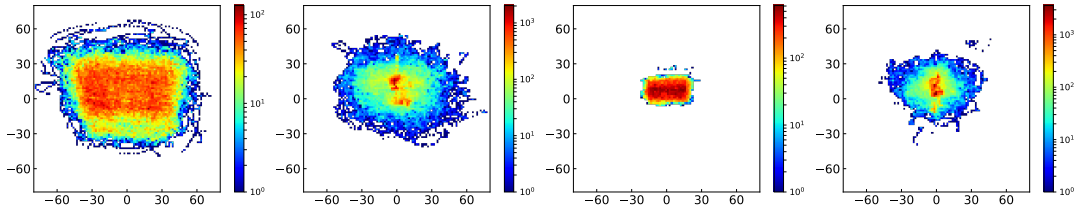
FIGURE 3.5: Sample images from the EYEDIAP dataset (Funes Mora, Monay, and Odobez, 2014a), corresponding to the screen-target *CS* (left) and floating-target *FT* (right) data subsets.

3.4.1 Dataset

Most publicly available 3D gaze estimation datasets focus on HCI (e.g., interacting with computers, tablets, or mobile phones) with a limited range of head pose and gaze directions. The EYEDIAP dataset (Funes Mora, Monay, and Odobez, 2014a)[4] is the only one containing video sequences with a wide range of head poses and showing the full face. Hence, we select this dataset for our evaluation. The dataset consists of 3-min videos of 16 subjects (12 male) without glasses seated in front of a computer screen looking at different types of targets. Videos were recorded with RGB-depth VGA, and RGB HD cameras. We select the VGA RGB videos at 30 fps and two types of targets: continuous *screen* targets on a fixed monitor (*CS*), and *floating* physical targets (*FT*). In the *CS* setting, the target is a circle shown on the screen that moves along a random trajectory for 2 s, while the *FT* setting entails following a 4-cm-diameter ball moving between the user and the screen. In the former setting, participants are seated at 80-90 cm from the camera, whereas in the latter setting, participants are seated at around 1.2 m to allow enough space to move the target (Funes Mora, Monay, and Odobez, 2014b). Figure 3.5 shows two sample image frames from *CS* and *FT*. For the *FT* setting, subjects 12-16 were recorded with two different lighting conditions. Videos are further divided into *static* (*S*) and *moving* (*M*) head pose for each of the subjects. Therefore, the *S* setting involves pure smooth pursuit eye movements, whereas the *M* setting involves a combination of head movements and smooth pursuit eye movements. Additionally, since the 3D target moves across the whole 3D space in the *FT* scenario, subtle vergence movements may also be present. The ground-truth gaze direction is computed as the vector from the mean of the provided eyeball positions to the target location in the 3D space.

For training and evaluation, we filtered out those frames that fulfilled at least one of the following conditions: (1) face or landmarks not detected by our landmark detector; (2) subject not looking at the target, which was provided by the dataset; (3) 3D head pose, eyes, or target location not properly recovered; and (4) eyeball rotations violating physical constraints ($|\theta| \leq 40°, |\phi| \leq 30°$) (MSC, 2000). Note that we purposely do not filter eye-blinking moments to capture more varied appearances with different levels of eyelid apertures, which may produce some outliers with higher estimation errors for the small number of frames where the eyes are fully closed.

[4]EYEDIAP dataset: https://www.idiap.ch/en/dataset/eyediap.



(a) $\mathbf{g}$ (FT) (b) $\mathbf{h}$ (FT) (c) $\mathbf{g}$ (CS) (d) $\mathbf{h}$ (CS)

FIGURE 3.6: Ground-truth eye gaze $\mathbf{g}$ and head orientation $\mathbf{h}$ distribution of the filtered EYEDIAP dataset for CS and FT settings, in terms of x- (horizontal) and y- (vertical) angles (degrees of visual angle).

Figure 3.6 shows the distribution of gaze directions and head poses for both filtered CS and FT cases.

We applied online data augmentation during training (i.e., a technique to artificially increase the number of training images and their variability by applying small modifications to the images) with the following random transformations: horizontal flip, shifts of up to 5 pixels, zoom of up to 2%, brightness changes by a factor in the range $[0.4, 1.75]$, and additive Gaussian noise with $\sigma^2 = 0.03$. When training the *Temporal* model, all frames of a given sequence were augmented in the same way.

3.4.2 Evaluation of static modalities

First, we evaluate the contribution of each static modality on the FT scenario. We divided the 16 participants into four groups, such that appearance variability was maximized while maintaining a similar number of training samples per group. Each static model was trained end-to-end, performing subject-independent (i.e., subjects in the training split are not included in the test split) 4-fold cross-validation using different combinations of input modalities. Since the number of fusion units depends on the number of input modalities, we also compare different fusion layer sizes. The effect of data normalization is also evaluated by training an unnormalized face model where the input image is the face bounding box with square size the maximum distance between 2D landmarks.

As shown in Figure 3.7, all models that use normalized full-face information as input achieve better performance than the eyes-only model. More specifically, the combination of face, eyes, and landmarks outperforms all other combinations by a small but significant margin (paired Wilcoxon test, Conover (1999), $p<.0001$). The standard deviation (SD) of the best-performing model is reduced compared to the face and eyes model, suggesting a regularizing effect due to the addition of landmarks. The unnormalized face-only model shows the largest error, proving the impact of normalization on reducing appearance variability. Furthermore, our results indicate that the increase in the number of fusion units is not correlated with a better performance.

To gain a better understanding of what the network is learning, we visualize the feature maps of the convolutional blocks of a full-face-only *Static* network (equivalent of NF-4096 from Figure 3.7) in Figure 3.9. As can be seen, the network not only detects eye features such as the pupil, but also surrounding facial cues, such as the

[5] As we are using one model per fold, results should have been calculated per fold (i.e., first computing the average error of all the samples, or subjects, per fold, and then averaging and computing the SD over the folds averages) and not per sample. However, the way the results were calculated does not affect the findings; instead, they help conceptualize the actual range of errors in this task.

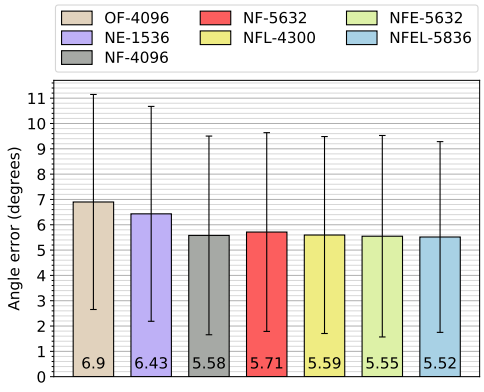


FIGURE 3.7: Performance evaluation (mean error + SD) of the *Static* network using different input modalities (*O - Not normalized, N - Normalized, F - Face, E - Eyes, L - 3D Landmarks*) and size of fusion layers on the *FT* scenario. Results are calculated with respect to all the samples of all the folds[5].

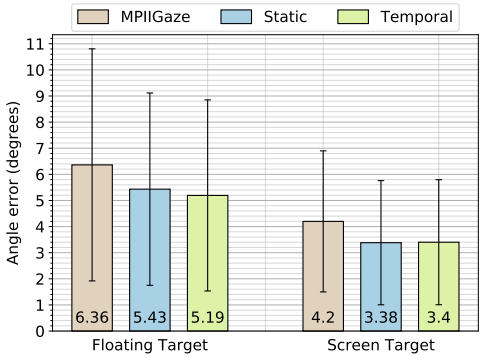


FIGURE 3.8: Performance comparison (mean error + SD) among state-of-the-art MPIIGaze method (Zhang et al., 2017b) and our *Static* and *Temporal* versions of the proposed network for *FT* and *CS* scenarios. Results are calculated with respect to all the samples of all the folds[5].

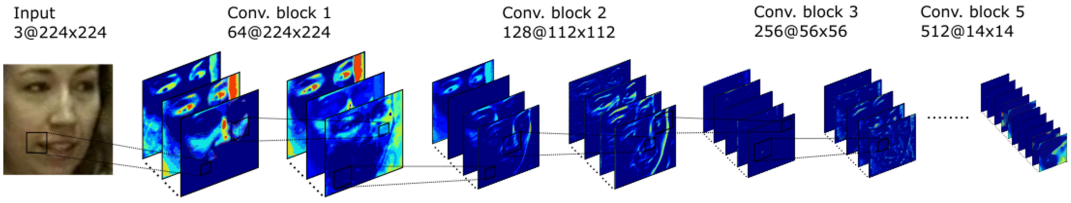


FIGURE 3.9: Visualization of the feature maps learned by a full-face-only *Static* network (equivalent of NF-4096 from Figure 3.7) for each convolutional block.

nose, eyebrows, cheeks, face contour, and even the hair (or background), which give information about the head pose.

3.4.3 Static gaze regression: comparison with existing methods

We compare our best-performing *Static* model with three baselines:

- **Head:** Treating the head pose directly as gaze direction.
- **PR-ALR:** Method that relies on RGB-depth data to rectify the viewpoint of the eyes image into a canonical head pose using a 3DMM. It then learns an RGB gaze appearance model using ALR (Mora and Odobez, 2012). Predicted 3D vectors for the *FT-S* scenario are provided by the EYEDIAP dataset.
- **MPIIGaze:** State-of-the-art full-face 3D gaze estimation method (Zhang et al., 2017b). It uses an AlexNet-based CNN model with spatial weights to enhance information in different facial regions. We finetuned the model[6] with the filtered EYEDIAP subsets using our training hyperparameters and normalization procedure.

[6]MPIIGaze full-face model: https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/gaze-based-human-computer-interaction/its-written-all-over-your-face-full-face-appearance-based-gaze-estimation.

	Participants																
Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**Avg.**
Head	23.5	22.1	20.3	23.6	23.2	23.2	23.6	21.2	26.7	23.6	23.1	24.4	23.3	24.0	24.5	22.8	23.3
PR-ALR	12.3	12.0	12.4	11.3	15.5	12.9	17.9	11.8	17.3	13.4	13.4	14.3	15.2	13.6	14.4	14.6	13.9
MPIIGaze	5.3	5.1	5.7	4.7	7.3	15.1	10.8	5.7	9.9	7.1	5.0	5.7	7.4	3.8	**4.8**	5.5	6.8
Static	**3.9**	**4.1**	**4.2**	**3.9**	**6.0**	**6.4**	7.2	**3.6**	**7.1**	**5.0**	5.7	6.7	**3.9**	4.7	5.1	**4.2**	**5.1**
Temporal	4.0	4.9	4.3	4.1	6.1	6.5	**6.6**	3.9	7.8	6.1	**4.7**	**5.6**	4.7	**3.5**	5.9	4.6	5.2
Head	19.3	14.2	16.4	19.9	16.8	21.9	16.1	24.2	20.3	19.9	18.8	22.3	18.1	14.9	16.2	19.3	18.7
MPIIGaze	7.6	6.2	5.7	8.7	10.1	12.0	12.2	6.1	8.3	5.9	6.1	6.2	7.4	4.7	4.4	6.0	7.3
Static	**5.8**	5.7	**4.4**	**7.5**	6.7	8.8	**11.6**	5.5	8.3	5.5	5.2	6.3	**5.3**	**3.9**	**4.3**	**5.6**	6.3
Temporal	6.1	**5.6**	4.5	**7.5**	**6.4**	**8.2**	12.0	**5.0**	**7.5**	**5.4**	**5.0**	**5.8**	6.6	4.0	4.5	5.8	**6.2**

TABLE 3.2: Gaze angular error comparison for *static* (top half) and *moving* (bottom half) head pose for each subject in the *FT* scenario, in degrees. Best results in bold.

In addition to the aforementioned *FT*-based evaluation setup, we also evaluate our method on the *CS* scenario. In this case, there are only 14 participants available, so we divided them into five groups and performed subject-independent 5-fold cross-validation. In Figure 3.8, we compare our method to MPIIGaze, achieving a statistically significant improvement of 14.6% and 19.5% on *FT* and *CS* scenarios, respectively (paired Wilcoxon test, $p<.0001$). We can observe that a restricted gaze target benefits the performance of all methods, compared to a more challenging unrestricted setting with a wider range of head poses and gaze directions.

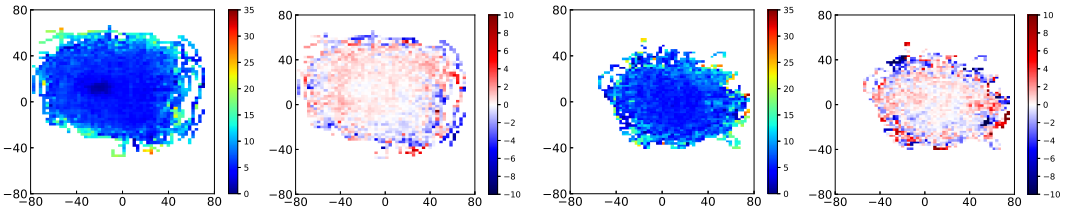
Note that the MPIIGaze authors report an error of 6.0° on the *screen target* sessions, which is higher than the error we obtain for their method. This performance improvement can be justified by the difference in training and evaluation data, the application of data augmentation on the training set, and the different normalization procedure (the MPIIGaze work uses the original normalization introduced by Sugano, Matsushita, and Sato, 2014, whereas in this work we use the refined version that appears in Zhang, Sugano, and Bulling, 2018).

Table 3.2 provides a per-subject comparison by performing leave-one-out cross-validation on the *FT* scenario for *S* and *M* separately. Results show that, as expected, facial appearance and head pose have a noticeable impact on gaze accuracy, with average error differences of up to 7.7° among participants.

3.4.4 Evaluation of the temporal network

In this section, we evaluate the contribution of adding the temporal module to the *Static* model. To do so, we trained a lower-dimensional version of the *Static* network with comparable performance to the original, reducing the number of units of the second fusion layer to 2918. First, we evaluated the effect of different recurrent architectures for the *Temporal* model on the *FT* scenario. In particular, we tested one (128 units) and two (256-128 units) LSTM and GRU layers, with one GRU layer obtaining slightly superior results (up to 0.12° of difference). We also assessed the effect of sequence length, setting s in the range $\{4, 7, 10\}$ (∼133, 233, and 333 ms, respectively), with $s = 7$ performing worse than the other two (up to 0.14°). We used the best configuration (one GRU layer and $s = 4$) for subsequent experiments, representing the configuration with the lowest computational footprint.

Results are reported in Figure 3.8 and Table 3.2. One can observe that using sequential information is helpful in the *FT* scenario, with our *Temporal* model outperforming our *Static* model by a statistically significant 4.4% (paired Wilcoxon test, $p<.0001$). This contribution is more noticeable in the *M* setting, proving that the



(a) Gaze space (b) Head orientation space

FIGURE 3.10: Angular error distribution (degrees) across gaze (a) and head orientation (b) spaces in the *FT* setting, in terms of x- (horizontal) and y- (vertical) angles (in degrees). For each space, we depict the *Static* model performance (left) and the contribution of the *Temporal* model versus *Static* (right). In the latter, positive difference (in red) means higher improvement of the *Temporal* model.

temporal model can benefit from, at least, head motion information. Indeed, the relative error increase between *S* and *M* is slightly higher for our *Static* model than for the *Temporal* one (23.5% vs 19.2% error increase, respectively), indicating that temporal information also makes the model more robust to head motion. By contrast, sequential information seems to be less meaningful in the *CS* scenario, where the obtained error is already very low for a subject-independent setting, and the amount and range of head movement declines. One difference between *CS* and *FT* scenarios, apart from the setting, is the range and velocity of the moving target, which can play a big role in the contribution of temporal information: if the range and/or speed are small/low, very small eye and/or head movements will occur during a sequence, which may be hardly perceptible; thus, adding temporal information will be less necessary. A post hoc analysis of the provided target data revealed that the mean velocity of the target in the *FT* scenario is 1.9-2.2 cm/frame, whereas in the *CS* scenario, it is 0.15-0.17 cm/frame. Hence, it is highly likely that the used *s* is too short to capture any meaningful motion changes in *CS*. The target movement range is wider in the *FT* scenario as well. These differences largely explain the fact that temporal information does not contribute to increasing accuracy for *CS*.

On a related note, we observe that our *Temporal* model also outperforms other state-of-the-art approaches on average in *CS* and *FT*, and also on average for the two head settings of *FT*. However, we note that the relative error increase between *S* and *M* head settings is lower for MPIIGaze (7.3% increase) than for our models, demonstrating the impact of the spatial weights mechanism of MPIIGaze on non-frontal faces.

3.4.5 Performance across gaze direction and head pose space

Figure 3.10 further explores the error distribution of the *Static* network and the impact of sequential information on *FT*. We can observe that the accuracy of the *Static* model is mostly constant across the head pose and gaze spaces (especially at the center) except for the extreme cases, which can be related to having fewer data in those areas. Compared to the *Static* model, the *Temporal* model particularly benefits gaze targets from mid-range upward. Its contribution is less clear for extreme targets, probably again due to data imbalance.

3.5 Limitations

The work presented in this chapter is not without limitations. The first stems from the inherent problem of data acquisition during eye movements: due to the usual smooth pursuit latency (see Section 2.1), and the size of the 3D target, the ground truth used might not always represent the real gaze direction, which can introduce noise in the learning process and evaluation. Another limitation entails the data normalization procedure. Since such normalization eliminates or modifies some head movements, it interferes with the proper modeling of head motion dynamics. This normalization may also affect specific eye-head motion dynamics.

With respect to the experimental evaluation, due to the computational resources and time required to perform all the cross-validation and leave-one-out experiments, all models were only trained once. Nonetheless, we ensured that the initialization of network parameters, data augmentation, and other stochastic elements were applied in the same way for all models (i.e., using the same seed). The time and computational restrictions also did not allow us to perform further evaluations on the *CS* scenario with different sequence lengths. Finally, the reported results depend on the evaluated hyperparameters, the types of data augmentation considered, and the backbone and facial alignment approach selected.

3.6 Conclusions

In this work, we studied the combination of full-face and eye images along with facial landmarks for person- and head pose-independent spatiotemporal 3D gaze estimation. To do so, we proposed a multistream convolutional-recurrent network that leverages the sequential information of eye and head movements, and geometric facial constraints. To our knowledge, this is the first attempt to exploit the temporal modality in the context of appearance-based gaze estimation from remote cameras.

Both static and temporal versions of our approach significantly outperform current state-of-the-art 3D gaze estimation methods on a wide range of head poses and gaze directions. We showed that adding geometry features to appearance-based methods has a regularizing effect on accuracy. Adding sequential information further benefits the final performance compared to static-only input, especially from mid-range upward, and in those cases where head motion is present and where interaction is not restricted to a screen-target scenario. The effect in very extreme head poses is not clear due to data imbalance, suggesting the importance of learning from a continuous, balanced dataset including all head poses and gaze directions of interest, or including data imbalance regularization techniques during training. In addition, accuracy is highly subject-specific, which calls for more robust methods capable of maintaining accuracy levels throughout the entire population (in terms of appearance and geometry variability caused by age, gender, race, skin and iris color, makeup, etc.).

Chapter 4

Benefits of Temporal Information for Near-eye Gaze Estimation

In Chapter 3, we learned that leveraging sequential information shows promising results when applied to remote or low-resolution image scenarios with off-the-shelf cameras. However, the increase in gaze estimation accuracy was mostly observed when the head could move freely, which may lead us to conclude that useful sequential information comes mainly from head movements. The magnitude of the contribution from eye movement traces specifically is yet unclear. These traces can be better captured with higher resolution/sampling rate imaging systems, in which more detailed information about the eye is obtained.

In this chapter, we investigate whether temporal sequences of eye images, captured using high-resolution, high-frame-rate IR cameras, can be leveraged to enhance the accuracy of an end-to-end appearance-based DL model for gaze estimation. Whereas Chapter 3 focused on smooth pursuit eye movements, this chapter considers fixations and saccades. We follow a similar methodology to the one used in Chapter 3. Results demonstrate statistically significant benefits of temporal information for this setting.

4.1 Introduction

State-of-the-art appearance-based gaze estimation methods mainly rely on static features. However, gaze is a dynamic process; depending on the task, we perform different eye and head movements. In addition, the gaze direction at a certain point in time is strongly correlated with the gaze direction of the previous time steps. Thus, it is safe to say that the temporal trace of eye gaze contains useful information for estimating a given gaze point. Following this line of reasoning, few appearance-based gaze estimation works have started to leverage temporal information and eye movement dynamics to increase gaze estimation accuracy with respect to static-based methods. This possibility was first explored by us in Chapter 3 (Palmero et al., 2018b), proposing to feed the learned static features of all the frames of a sequence to a many-to-one recurrent module to predict the gaze direction of the last sequence frame, improving the state of the art on head-pose independent gaze estimation. Later, Wang, Su, and Ji (2019) relied on a semi-Markov approach to model the gaze dynamics of fixations, saccades, and smooth pursuit movements; per-frame gaze estimates were first computed using a CNN and then refined using the learned dynamic information. Bidirectional recurrent methods have also been introduced (Kellnhofer et al., 2019; Zhou et al., 2019), although their applicability is limited to offline methods, as they rely on past and future information. Despite these initial explorations confirming the benefits of temporal information, these

works are based on RGB low-to-mid-resolution images and low framerate capture systems (∼30 fps), which do not allow to accurately capture some of the eye movement dynamics, especially from saccades, which are characterized by a very high velocity. Therefore, it is yet unclear how and why temporal information improves gaze estimation accuracy for different eye movements.

In this chapter, we investigate whether temporal sequences of eye images, captured at a higher frame rate (100 Hz) with a VR head-mounted display (HMD) mounted with two synchronized eye-facing IR cameras, can be leveraged for gaze estimation. Furthermore, we evaluate which eye movements benefit more from the additional temporal information. We focus specifically on fixations and saccades, two of the most prominent eye movements (Komogortsev et al., 2010). As in Chapter 3, we compare the results obtained with a spatiotemporal model based on a many-to-one CNN-recurrent approach, in contrast to those obtained with a static-only CNN model. We evaluate our hypothesis on a newly constructed dataset collected using the above-mentioned VR-HMD, in which 84 subjects of varied appearance were recorded performing a stimulus-elicited saccade task in a VR scenario. Results show that leveraging temporal information of eye image sequences for gaze estimation significantly improves accuracy, in particular for the vertical component of gaze. To the best of our knowledge, this work presents the first study systematically demonstrating the benefits of temporal information for appearance-based gaze estimation using eye image captures with a high-resolution, high-frame-rate camera system, evaluated on different eye movements.

The remainder of this chapter is organized as follows. Section 4.2 defines the methodological approach followed in the study. Section 4.3 describes the dataset, experimental protocol, and results of adding temporal information to a static baseline. In addition, we measure the contribution of temporal information with respect to different eye movement types. Section 4.4 reviews the limitations of the study. Finally, Section 4.5 concludes the chapter.

4.2 Methodology

In this section, we describe the proposed methodology to evaluate the benefits of sequential information for appearance-based gaze estimation models applied to near-images from IR cameras.

4.2.1 Spatiotemporal gaze estimation

In this work, the spatiotemporal gaze estimation task is posed as a regression problem based on monocular eye images. The methodological approach follows that of Chapter 3, adapted for the present use case. First, spatial features are extracted for each frame $\mathbf{I}_i \in \mathbb{R}^{WxH}$ of a sequence using a static CNN backbone g, where i is the frame index, and W and H denote the width and height of the frame, respectively. The sequence of per-frame features is then fed to a many-to-one recurrent module r to learn sequential dependencies. The recurrent module produces a vector of spatiotemporal features, which is used to finally regress the gaze direction of the last frame of the sequence, $\mathbf{y}_t \in \mathbb{R}^2$, such that:

$$\mathbf{y}_t = f(r(g(\mathbf{I}_{t-s+1}), \ldots, g(\mathbf{I}_t))), \tag{4.1}$$

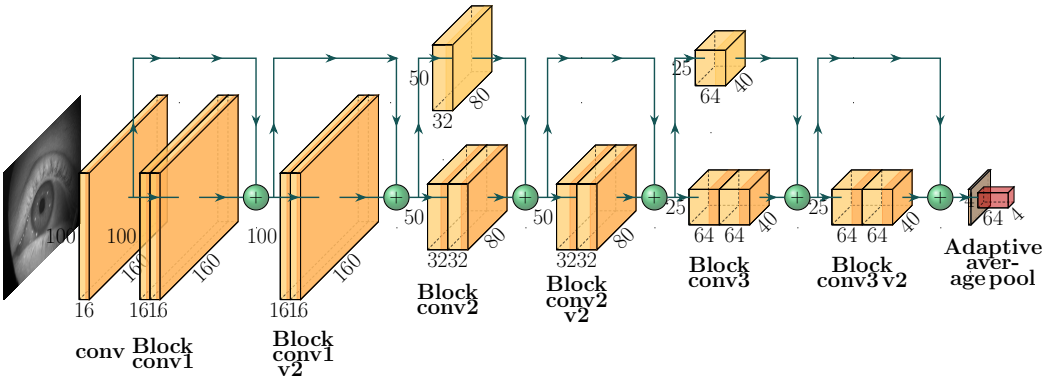


FIGURE 4.1: Architecture of the backbone network used for static gaze estimation, based on a modified ResNet.

where f denotes the regression function, t corresponds to the last frame of the sequence, and s to the number of frames in a sequence. In this work, the 3D gaze direction corresponds to the visual axis of the eye, the origin of which is the center of corneal curvature. For simplicity, the gaze direction is expressed by 2D spherical coordinates, representing yaw (horizontal) and pitch (vertical) angles (see Section 2.4).

4.2.2 Network architecture

As backbone (depicted in Figure 4.1), we use a modified ResNet architecture (He et al., 2016) with 13 convolutional layers and a final adaptive average pooling layer at the end to decrease the final feature vector size to 64x4x4. The sequence of feature vectors is flattened to serve as input for the recurrent module.

The recurrent module consists of a single-layer plain LSTM (Greff et al., 2016) with 32 units. The LSTM is unrolled into s time steps, depending on the input sequence length. We also considered a GRU cell in preliminary experiments, but LSTM performed better in this scenario. The output of the recurrent module is further fed to an FC layer with ReLU activation function, which produces a 32D vector. Finally, an FC layer with linear activation function (i.e., regression) produces the estimated 2D gaze angles.

4.2.3 Training strategy

The network is trained in a stage-wise fashion. First, the static backbone, coupled to a 32-hidden-unit FC layer and a 2-hidden-unit FC regression layer, is trained end-to-end from scratch on each individual frame of the training data to learn static features. This network is referred to as *Static1* (or *S1*) in Section 4.3. Second, the FC layers are discarded, and the recurrent module, coupled to new 32-hidden-unit and 2-hidden-unit FC layers, is added to the pretrained static backbone. The new architecture is further trained end-to-end, finetuning the convolutional backbone while training recurrent layer and new FC layers from scratch. By further finetuning the backbone weights, the convolutional module is able to learn useful features derived from the sequential information captured by the recurrent module. For this second stage, however, the training data are rearranged using a sliding window with stride 1, to build input eye image sequences compatible with the many-to-one architecture. Each input sequence is composed of s consecutive frames. This second network is referred to as *S1+LSTM* in Section 4.3.

The network was trained using ADAM optimizer (Kingma and Ba, 2014), empirically setting the learning rate to 0.0005, batch size to 32, and weight decay to 0.00001. The learning rate parameter was found to have a large influence on the final accuracy, with higher values not allowing a proper learning of the LSTM. CNN weights were initialized from a uniform distribution. For the LSTM module, input weights were initialized using Xavier uniform, while an orthogonal initialization was used for the hidden weights. Biases were set to 0. Early stopping on the validation set was used to select the best model for each training stage, with a maximum number of epochs of 150 for the first stage and 30 for the second. Finally, we used the L1 loss for both training stages, as preliminary experiments showed it to yield slightly lower error than the L2 loss.

4.3 Experiments

In this section, we describe the experimental setup and evaluate the effectiveness of the spatiotemporal model in comparison to a static-only version for different window lengths and eye movements.

4.3.1 Dataset

The study is based on a newly constructed dataset of 100 × 160-pixel eye-image sequences captured using a VR-HMD, with two synchronized eye-facing IR cameras at a frame rate of 100Hz under constant illumination[7]. The dataset consists of 84 subjects with wide appearance variability in terms of ethnicity, gender (70% male), and age (20-70 years old), with some of them wearing glasses (26%), contact lenses, and/or make-up (13%). Subjects were recorded while gazing at a moving target on a blank screen. Each recording consisted of a set of patterns with 1-s-long randomly located target fixations at different depths (50-600 cm) and instantaneous (0.1 s) target transitions to elicit saccades.

Ground-truth eye-gaze vectors were obtained using a classical user-calibrated glint-based model (Guestrin and Eizenman, 2006). While this approach poses some limitations on the ground truth quality (see Section 4.4), it still allows us to soundly evaluate our hypotheses. The ground-truth vectors correspond to the visual axis of each eye, in the headset coordinate system, the origin of coordinates of which is located at the midpoint between the eye boxes of left and right eye.

Frames with no valid gaze, or for which the subject was distracted, were discarded, causing most of the recordings to be divided into smaller non-contiguous sequences. To keep consistency, the remaining data was further processed by randomly selecting 10 non-overlapping sequences of 100 contiguous frames (1 s) each, thus having 1K frames per recording and a total of 168K eye-region images. Therefore, each sequence can contain fixations only, saccades, or a combination thereof. Figure 4.2 shows the gaze angle distribution and sample eye images from the dataset.

4.3.2 Experimental protocol

To perform the experimental evaluation, the processed dataset was partitioned into subject-independent train, validation, and test sets following a 5:1:2.4 ratio. The

[7] A second version of this dataset was later employed for the OpenEDS 2020 challenges held in the *ECCV 2020 OpenEyes: Eye Gaze in AR, VR, and in the Wild* workshop, and associated publicly available OpenEDS2020 dataset (Palmero et al., 2020; Palmero et al., 2021b).

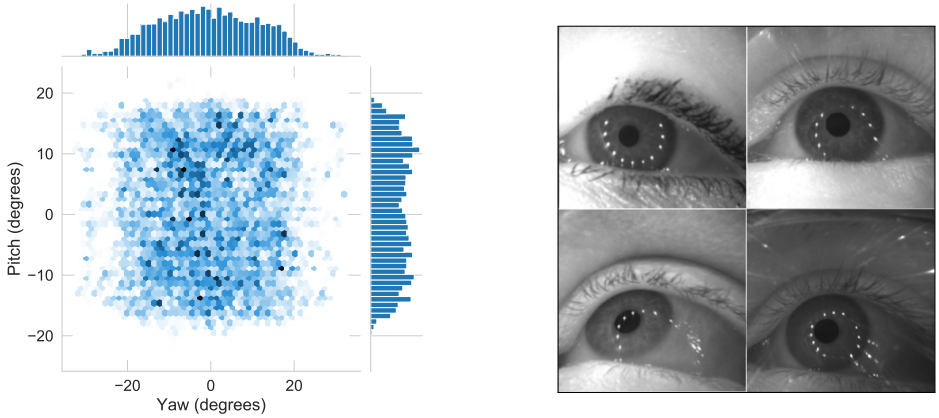


FIGURE 4.2: Gaze distribution (left) and sample eye images (right) from the dataset.

evaluated models were trained on the training split, using the validation split to optimize model hyperparameters. Right-eye images and their corresponding ground truth gaze vectors were horizontally flipped to mimic left-eye data. This way, the same model can be used for both eyes, while augmenting the number of data samples. Contrary to remote camera-based appearance-based gaze estimation, no data normalization is applied for near-eye images, as head pose has no influence on the viewpoint of the images, and near-eye devices are usually tight to a single camera configuration. In practice, however, the eyes may not be perfectly centered in the image due to headset slippage or anatomy differences across subjects.

Experimental results are reported on the test split, using the MAE between estimated and ground truth 2D gaze angles as main metric (Equation 2.5). Due to the sliding window approach followed by the spatiotemporal model, the first frames of a sequence do not obtain a gaze estimation. Therefore, results are reported on the subset of the test split that does obtain gaze estimates for all evaluated models[8].

4.3.3 Addition of temporal information to the baseline static model

First, we use the initial static model (*Static1*) as baseline and compare it to the proposed spatiotemporal model. In particular, we evaluate four versions of the spatiotemporal model, each of them trained with different sliding window lengths s in the range $\{5, 10, 15, 20\}$ (equating to 50, 100, 150, and 200 ms, respectively), to assess the effect of the amount of frames used on the final accuracy.

Table 4.1 shows the performance of the evaluated models with respect to each axis individually and simultaneously. We can observe that all spatiotemporal models significantly outperform the static baseline, with up to 19.78% mean error improvement (paired Wilcoxon test, $p<.0001$). While the error for the horizontal gaze component is higher than for the vertical for all models, the addition of temporal information decreases the error by up to 16.91% for the former and 23% for the latter, evidencing that such information is more beneficial for the vertical axis. This is an important contribution with respect to classical pupil-based methods, as they usually have less accuracy on this axis due to occlusions of the limbus caused by the eyelids and eyelashes. The higher error and SD obtained for the horizontal gaze component with all models might have been caused by the wider and less represented

[8]The longest window length evaluated in this study is 20 frames; therefore, 81% of the test split is used to report experiment results. Results from models based on smaller window lengths, evaluated on a consequently larger test subset, did not deviate significantly from the results reported herein.

TABLE 4.1: Mean absolute error (degrees) between ground truth and estimated gaze angles for the different evaluated models, reported on the test set. Standard deviation in brackets. Best results in bold.

Method	Window	Yaw	Pitch	Mean
Static1 (*S1*)	1	4.02 (4.22)	3.26 (2.67)	3.64 (2.59)
Static2	1	4.26 (4.93)	3.36 (2.72)	3.81 (2.93)
S1+LSTM	5	3.46 (4.03)	2.57 (2.14)	3.01 (2.41)
S1+LSTM	10	3.41 (3.95)	2.55 (2.14)	2.98 (2.39)
S1+LSTM	15	**3.34 (3.98)**	**2.51 (2.07)**	**2.92 (2.37)**
S1+LSTM	20	3.39 (3.99)	2.51 (2.08)	2.95 (2.39)

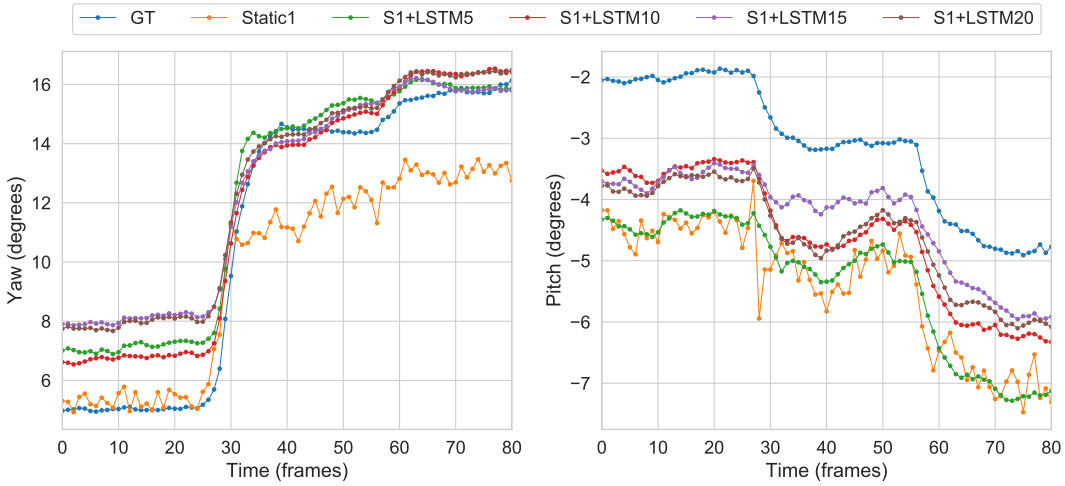


FIGURE 4.3: Example of ground truth (GT) and estimated gaze traces for one sequence of the dataset.

range of gaze directions on the horizontal axis, as can be observed in Figure 4.2 (left). Regarding window length, we can observe that the increase in performance peaks around $s = 15$ frames (i.e., 150 ms) and then decreases, indicating that longer-term dependencies are not required to obtain further accuracy gains.

It could be argued that the decrease in error is due to the larger complexity of the spatiotemporal models, as the addition of the LSTM layer highly increases the number of parameters. To validate this possibility, we trained a second static model (*Static2* in Table 4.1), adding a 128-hidden-unit FC layer between the two FC layers from *Static1* model to compensate for the difference in the number of parameters between baseline and spatiotemporal models. Results show that, in spite of the smaller number of parameters, even *Static1* outperforms *Static2*, suggesting that *Static2* may be overfitting to the training data. This indicates that the increase in complexity is not correlated with a lower error.

Figure 4.3 further illustrates the effects of leveraging temporal information, with an example of ground truth and estimated gaze angles during a fixation-saccade transition, for horizontal and vertical axis. We can clearly see the smoothing effect caused by adding temporal information as opposed to the noisy static estimation. Furthermore, spatiotemporal estimates are able to more accurately follow the saccade-to-fixation transition (frames 30 to 40). Indeed, using consecutive frames allows the network to better discard noisy features and learn a more robust representation for eye gaze estimation.

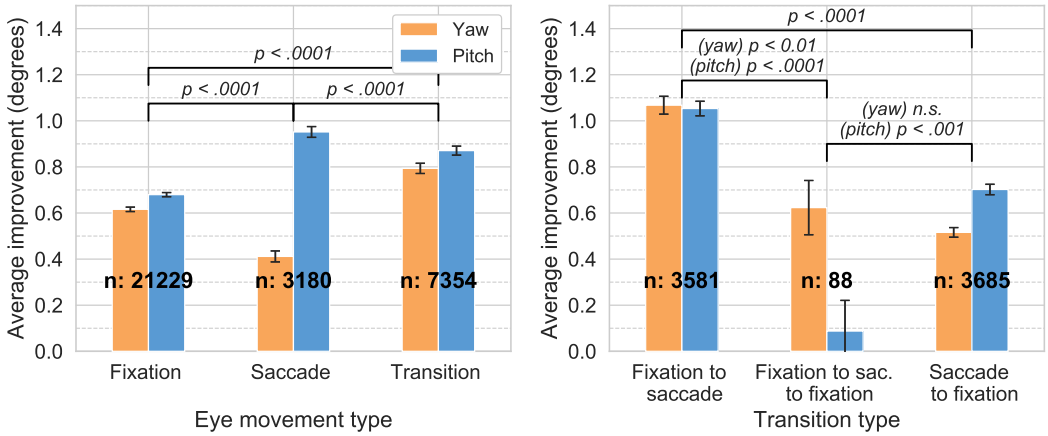


FIGURE 4.4: Average improvement of temporal (*S1+LSTM20*) over static (*Static1*) models per axis (horizontal and vertical), for different eye movement (left) and transition (right) types. Error bars indicate standard error of the mean. Significance computed using the two-sided Kolmogorov-Smirnov test (Massey Jr, 1951). *n:* indicates the number of sequences for each eye movement or transition type.

4.3.4 Contribution of temporal information wrt. eye movement type

Here, we evaluate the contribution of temporal over static models with respect to different eye movements types. To do so, we use the 20-frame-window spatiotemporal model (*S1+LSTM20*) as the reference temporal model for this experiment.

Since our dataset has a mixture of eye movements, to perform this evaluation, we manually annotated the test split based on visual inspection of ground truth gaze angles with the following labels per each 20-frame input sequence: *fixation*, when the eye was virtually static; *saccade*, if it only included a saccade movement; *transition*, if it included a combination of fixation and saccades; and *other*, when the eye status could not be clearly classified. *Transition* sequences were further divided into *fixation to saccade*, *saccade to fixation*, or *fixation to saccade to fixation*, according to their order in time.

Figure 4.4 depicts the contribution of temporal information with respect to the static baseline for each label and gaze component. As shown previously, the performance of the vertical component substantially improves when temporal information is added for fixations and saccades, compared to that of the horizontal component. In particular, we observe a substantially higher improvement for saccades, with an average difference of 0.44 degrees between components. With respect to transitions, the horizontal component shows a similar improvement to the vertical component for *fixation-to-saccade* transitions. This suggests that the spatiotemporal model is indeed taking into account the information coming from the first frames of the input sequence, being more beneficial when the sequence starts with a more stable eye gaze and more correlated eye images to better discern between noisy and important features for the gaze estimation task. As a matter of fact, this improvement is even higher than the one obtained on *saccade*-only sequences, demonstrating that the model is able to learn more representative features of the eye when being presented with a fixation bout first, particularly for the horizontal axis.

4.3.5 Effect of appearance

As a final experiment, we evaluated the differences in obtained error across subjects. While the error indeed varied, we did not find significant differences with respect to the considered subjects characteristics (i.e., age, gender, glasses, and makeup). The fact that using glasses does not cause a significant decrease in accuracy is also an important contribution with respect to classical glint-based methods, since glints can be distorted in such cases. Other sources of variability might have had a greater impact on performance, such as subject-dependent gaze dynamics, but were not identified in this study.

4.4 Limitations

This study offers an initial insight as to how and why temporal information benefits gaze estimation when different eye movement types are considered, specifically fixations, saccades, and transitions among them. These movements were elicited using a pattern-based task, which poses a limitation on the directions, velocities, and motion trajectories available. Eye dynamics are task-dependent, thus, a more complete study should contain subjects performing other tasks, including natural behaviors.

We note that the obtained results are linked to the selected methodology, static backbone, and loss used for training. Other backbones would pose different static priors which would affect the final obtained accuracy. As in Chapter 3, all models were trained a single time; therefore, model consistency has not been measured. Finally, even using state-of-the-art glint-based methods to obtain ground-truth eye gaze vectors, this process poses a lower bound on the obtained gaze estimation error, as we are trying to approximate a model to values that can be inherently noisy. Again, this is a limitation present in most of the gaze estimation literature, which evidences the need for better ways to gather accurate ground-truth gaze data.

4.5 Conclusions

In this chapter, we have analyzed the effect of leveraging sequential information for appearance-based gaze estimation applied to IR near-eye cameras in a VR scenario, using previous contiguous image frames along with the current image frame to be estimated. We leverage DL techniques, building a spatiotemporal model consisting of a static CNN network followed by a recurrent module to learn sequential dependencies in eye movements. The dataset consists of high-resolution eye-image sequences, consisting of 84 subjects performing a stimulus-elicited fixation and saccade task, captured at 100 Hz. Results have shown a significant improvement of the spatiotemporal model in comparison to a static-only approach, producing a less noisy estimation. The model is able to learn robust features, with increased accuracy when transitioning from fixation to saccade. In addition, temporal information has been demonstrated to be particularly beneficial in improving the accuracy of vertical axis estimates.

Therefore, we can conclude that temporal information benefits appearance-based gaze estimation. In addition, the large differences in performance with respect to gaze axes obtained in this study give rise to considering approaches based on independent models for each gaze component, as opposed to usual jointly trained methods, to improve final gaze estimation accuracy.

Chapter 5

Single- and Multirate Sensor Fusion for Near-Eye Gaze Estimation

IN PREVIOUS CHAPTERS, we demonstrated the utility of temporal information for gaze estimation in both remote- (Chapter 3) and near-eye- (Chapter 4) camera scenarios. In this chapter, we continue with a near-eye setting that can be found on AR/VR HMDs and smart glasses. Specifically, in addition to providing high accuracy and robustness to appearance variability, such portable eye-tracking devices are also expected to provide a high-speed gaze signal in a power-efficient manner while being robust to sensor slippage. However, the power requirements of VOG for high-speed operation can be prohibitive. Recently, alternatives with low-power sensors have been evaluated, providing gaze estimates at high frequency with a trade-off in accuracy and robustness. We posit that a hybrid approach that combines fast/low-fidelity and slow/high-fidelity sensors should be able to exploit their complementarity to track fast eye motion accurately and robustly.

In this chapter, we investigate the potential of *single-* (both sensors operating at the same sampling rate) and *multirate* (each sensor operating at a different sampling rate) sensor fusion (as a variant of multimodal fusion) for increasing the accuracy and/or sampling rate of portable eye-tracking devices. Considering the difficulty of collecting real varied data associated with accurate ground truth during eye movements, we synthesize a multisensor dataset, which can provide exact ground truth gaze vectors with varied appearance and camera position. This time, the dataset includes fixations, saccades, and smooth pursuit movements. We follow the spatiotemporal appearance-based pipeline of Chapters 3 and 4, as well as the feature-based fusion approach of Chapter 3 for our methodological approach. In addition to the naive concatenation fusion considered in Chapter 3, we evaluate three additional fusion approaches with increasing complexity. Finally, unlike the many-to-one recurrent network used in previous chapters, here we use a many-to-many network, which better matches real-world online operation. Obtained results show significant accuracy improvements when tracking fast eye movements with a multirate sensor fusion approach compared to a gaze prediction approach that operates with a low-speed sensor alone.

5.1 Introduction

VOG is currently employed by most commercial eye trackers, with camera sensors operating at sampling rates ranging from 30 Hz for portable devices to 2 kHz for desktop alternatives. For many diagnostic applications (see Chapter 1), obtaining

a high-frequency, high-quality eye-tracking signal is essential for detecting and determining the start-end times of different eye movements, especially fast and/or small movements such as saccades and microsaccades (Holmqvist et al., 2011). Furthermore, for gaze-contingent applications, such as foveated rendering (Patney et al., 2016), increasing the sampling rate is equally important to sustain optimal user experience. While tethered eye trackers can support high-frequency, high-quality eye tracking, power and processing speed constraints limit the sampling rate and the accuracy for eye tracking with current battery-operated portable eye trackers or emerging AR/VR headsets. Another requirement specific to portable devices is robustness of eye tracking to headset slippage, i.e., sensor shifts with respect to the user's head. In other words, gaze accuracy should not be affected by headset shifts, whether caused by changes in facial expressions or talking, manual adjustment, or ordinary wear and tear impacting the fit profile of head-worn devices. Addressing the slippage problem is a current research direction (Santini, Niehorster, and Kasneci, 2019; Niehorster et al., 2020).

In the last decade, there has been an effort to investigate alternative low-power sensors such as photosensors (Li, Liu, and Zhou, 2017; Rigas, Raffle, and Komogortsev, 2018; Li et al., 2020), which measure the amount of reflected light when the eye rotates. These are capable of providing sampling rates higher than 1 kHz; however, the accuracy of eye tracking with these sensors is still not on par with those obtained using VOG. Two recent works have studied the utility of fusing fast/low-fidelity and slow/high-fidelity sensors, combining near-eye cameras with photosensors (Rigas, Raffle, and Komogortsev, 2017) or event-based cameras (Angelopoulos et al., 2021) for fast eye tracking. Such hybrid approaches have proven to be capable of achieving high sampling rates while providing eye-tracking gaze accuracy comparable to VOG, within the constraints of the considered sensor shift ranges and appearance variability in a lab setting. This motivates further research into sensor fusion strategies for improving the accuracy and robustness of fast eye tracking in unconstrained scenarios. Multirate DL-based sensor fusion has proven to be successful in other areas like autonomous driving, characterized by high appearance and sensor location variability (Cho et al., 2014; Chen et al., 2019; Fayyad et al., 2020). We believe that this success can be leveraged for gaze estimation, wherein one can envision a hybrid approach that exploits the spatiotemporal fusion of signals captured by a combination of slow/high-fidelity and fast/low-fidelity sensors to track eye motion accurately at the sampling rate of the high-speed sensor (see Figure 5.1). However, the difficulty of obtaining precise eye gaze ground truth and time-synchronized input data are some of the main limitations to advancing research in this domain.

In this chapter, we explore the potential of sensor fusion for increasing accuracy, robustness, and temporal resolution of near-eye gaze estimation in unconstrained scenarios. While the main focus is on multirate gaze estimation, we also investigate single-rate operation as a lower error bound. First, to spawn research on the topic and validate our hypotheses, we present the *Open Sensor Fusion Eye Dataset* (OpenSFEDS), a synthetic dataset consisting of camera-photosensor eye-image pairs following a diverse appearance and sensor position variability for head-mounted eye-tracking devices. It has been created using a simulation framework based on UnityEyes (Wood et al., 2015; Wood et al., 2016b; Wood et al., 2016a), allowing for the synthesis of wide appearance and camera position variability that mimic sensor shifts. The dataset includes a *Static* subset of 450K image pairs, with a uniform gaze distribution covering $\pm 25^{\circ}$ of visual angle, and a *Temporal* subset, consisting of 1.8K, 2 s sequences of camera-photosensor image pairs at 500 Hz with gaze extracted from

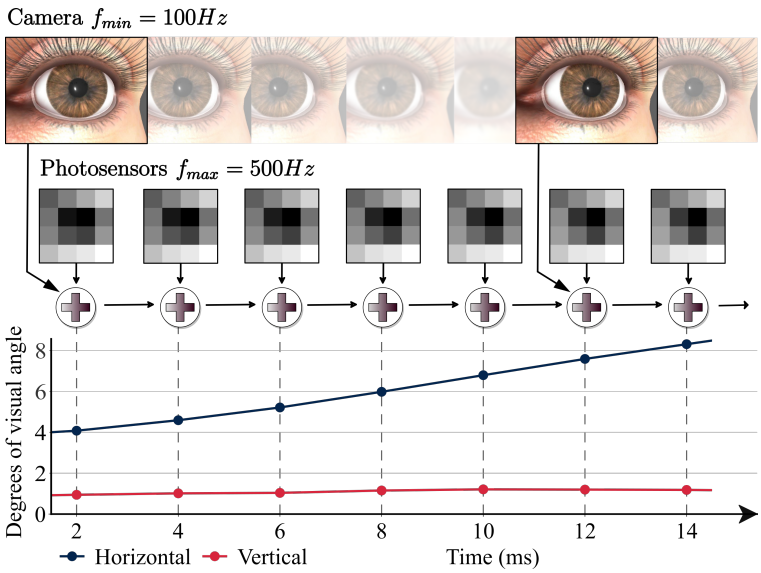


FIGURE 5.1: *Illustration.* Leveraging the multisensor data featured in the proposed *OpenSFEDS* dataset, our temporal fusion-based gaze estimation framework combines a set of fast/low-fidelity photosensors with a slow/high-fidelity camera to track fast eye movements accurately at the sampling rate of the high-frequency sensor.

real eye movement traces of the GazeBase dataset (Griffith et al., 2021). For single-sensor ET, synthetic datasets have long enabled the development and evaluation of new approaches and devices (Wood et al., 2016b; Kim et al., 2019; Nair et al., 2020), since they allow for obtaining highly accurate gaze ground truth for highly varied populations and different eye movement types, virtually impossible to obtain otherwise. In fact, UnityEyes and other simulators have been proven to include the essential eye geometry required to (pre)train gaze estimation models (Wang, Zhao, and Ji, 2018) and inform hardware design (Li et al., 2020), with findings on synthetic data comparable to those on real data (Garde et al., 2020). Furthermore, it is only via simulation that one can capture perfectly synchronized data with no (or controlled) system latency or noise, thus isolating the effect of fusing sensor signals versus single-sensor operation on the final gaze error, and allowing for the computation of lower error bounds. To the best of our knowledge, OpenSFEDS is the first dataset containing paired multisensor eye-tracking data at high frequency, and also the first featuring photosensor data.

Second, we formulate the task of sensor fusion for gaze estimation, and propose a framework that disentangles data encoding, fusion, and eye-state dynamics. Similarly to previous chapters, we rely on deep convolutional encoders to map the data from different sensors to a lower-dimensional feature embedding, perform temporal feature fusion of the available embeddings at any given time, and then estimate eye state based on the fused signal. We also evaluate a set of feature-based fusion baseline strategies capable of dealing with missing data. With this framework, we provide a first methodological evaluation on OpenSFEDS, offering an initial insight as to how leveraging multiple sensors for eye tracking, operating at the same or different sampling rates, can increase the accuracy of the estimated gaze signal and/or effective sampling rate with respect to a single sensor.

The remainder of this chapter is organized as follows. Section 5.2 reviews related work about video and *photosensor oculography* (PSOG), hybrid eye tracking, DL-based sensor fusion, and near-eye gaze estimation datasets. Section 5.3 presents the

design considerations and data subsets of the proposed OpenSFEDS dataset. Section 5.4 formulates the problem of sensor fusion for gaze estimation for single- and multirate scenarios, and describes the proposed gaze estimation framework along with the baseline fusion modules. Section 5.5 describes implementation details and the experimental protocol followed to evaluate static and temporal approaches for single- and multirate sensor fusion on the OpenSFEDS dataset, and discusses the results obtained. Finally, Section 5.6 concludes the chapter. Limitations of the dataset and experimental evaluation are outlined in their respective sections.

5.2 Related work

This section reviews related work along five axes: video and photosensor oculography, hybrid eye tracking, DL-based sensor fusion, and near-eye gaze estimation datasets.

5.2.1 Video-oculography

Model- and feature-based methods are commonly used for near-eye gaze estimation (Hansen and Ji, 2010). Instead, appearance-based approaches have been mostly applied to RGB, remote-camera eye tracking, boosted by remarkable gains in gaze accuracy using DL (Zhang, Sugano, and Bulling, 2019; Cazzato et al., 2020). Specifically, CNNs are the go-to approach for unsupervised, weakly supervised, or self-supervised gaze regression (Krafka et al., 2016; Zhang et al., 2017b; Fischer, Chang, and Demiris, 2018; Yu and Odobez, 2020; Kothari et al., 2021a; Farkhondeh et al., 2022), while LSTM-RNNs have been explored for spatiotemporal gaze estimation and prediction (Palmero et al., 2018b; Kellnhofer et al., 2019; Park et al., 2020; Palmero, Komogortsev, and Talathi, 2020; Palmero et al., 2021b). Due to their performance, deep appearance-based methods have gained more attention lately for near-eye settings to increase robustness against headset slippage, pupil occlusions caused by eyelids or eyelashes, or light reflections on eyeglasses (Tonsen et al., 2017; Zhang, Sugano, and Bulling, 2019; Yiu et al., 2019; Palmero, Komogortsev, and Talathi, 2020).

5.2.2 Photosensor oculography

Classic PSOG uses just few photosensors (e.g., four) and simple additive math for gaze estimation (Rigas, Raffle, and Komogortsev, 2018). PSOG-based eye tracking can provide high gaze accuracy when recording stationary subjects on a chinrest; however, the eye-tracking accuracy drops substantially when the sensors shift relative to the head. ML-based PSOG approaches have been proven to alleviate this issue by means of multilayer perceptrons (MLPs) to learn shift-invariant mappings between photosensor intensity values and gaze output (Li, Liu, and Zhou, 2017; Zemblys and Komogortsev, 2018; Li et al., 2020), and more recently, low-complexity CNNs to leverage the existing spatial structure of the photosensor array (Griffith, Katrychuk, and Komogortsev, 2019; Katrychuk, Griffith, and Komogortsev, 2020). Most of these works have used proprietary synthetic data to evaluate the effects of horizontal and vertical sensor shifts of up to $\pm$ 3 mm. Z-axis shifts or other variability factors have not been evaluated yet under an ML framework.

5.2.3 Hybrid eye tracking

The combination of VOG and PSOG has previously been proposed in Rigas, Raffle, and Komogortsev (2017), successfully reducing power consumption compared to VOG alone. In that work, PSOG was used as a backbone for providing high-speed (1 kHz) gaze estimation by least-squares regression, whereas the low-speed signal from VOG (5 Hz) was employed to detect sensor shifts to correct the final gaze estimates. More recently, Angelopoulos et al. (2021) presented a hybrid event-based eye-tracking system, combining low-speed VOG (25 Hz) with high-frequency events (>10 kHz) from an event camera. Fusion consists in an online 2D pupil fitting method that updates a parametric model every one or few input events. Gaze is directly regressed from the parametric pupil model, thus being susceptible to gaze accuracy degradation due to sensor slippage. While both approaches are based on model- or feature-based VOG, our work explores deep appearance-based PSOG-VOG and feature-level fusion to leverage the complementarity in the two signals.

5.2.4 Deep learning-based sensor fusion

Multimodal fusion has been extensively studied in many areas (Atrey et al., 2010; Baltrušaitis, Ahuja, and Morency, 2018; Feng et al., 2020; Fayyad et al., 2020). Fusion is usually performed at a feature level, i.e., fusing the feature representations from different modalities, or at a decision level, i.e., combining the outputs of sensor-specific networks (Ramachandram and Taylor, 2017). Decision-based approaches may be preferred for sensors operating at different sampling rates due to their modularity, but they cannot leverage the complementarity or redundancy of the modalities. For example, in the field of autonomous driving, it is common to have sensors operating at different sampling rates (Caesar et al., 2020). The framework proposed by Chen et al. (2019) for odometry is the closest to our approach. It consists in an end-to-end architecture for slow/high-fidelity (camera) and fast/low-fidelity (inertial measurement unit, or IMU) feature-based sensor fusion. However, their method aggregates all IMU features captured between two contiguous camera frames and fuses them with the visual features, such that the regressed output is provided at the camera sampling rate. Instead, our eye-tracking framework operates at the high-speed sensor frequency, thus dealing with missing data when performing fusion.

5.2.5 Near-eye gaze estimation datasets

The success of deep appearance-based gaze estimation is in part due to the proliferation of large-scale datasets with image-gaze ground truth pairs (McMurrough et al., 2012; Winkler and Subramanian, 2013; Funes Mora, Monay, and Odobez, 2014a; Zhang et al., 2017c; Porta et al., 2019; Zhang et al., 2020). A number of synthetic and real-world datasets with near-eye image sequences have emerged over the last four years (e.g., NVGaze by Kim et al., 2019; Gaze-in-wild by Kothari et al., 2020; OpenEDS2020 by Palmero et al., 2021b; MEMD by Fuhl and Kasneci, 2021; sGiW by Chaudhary et al., 2022), denoting the importance of modeling spatiotemporal eye movement dynamics. They consist of images sampled at or up to 120 Hz, sufficient to track most eye movements but not comparable to the operational rate of commercial high-frequency eye trackers. To our knowledge, the event-based eye-tracking dataset (Angelopoulos et al., 2021) is the only dataset that provides data from two sensors, including eye images at 25 Hz and events at 10 kHz. However, the fact that both data types are provided with different sampling rates does not allow for the evaluation of signal complementarity or eye-tracking performance as a

function of sampling rate. Our proposed dataset is the first to include near-eye image pairs from camera and photosensor sampled at 500 Hz, thus allowing for such types of evaluations.

5.3 The OpenSFEDS dataset

In this section, we introduce OpenSFEDS[9], a new dataset consisting of synthetic pairs of near-eye images and photosensor intensity values from 60 identities with high variability in terms of sensors placement, periocular appearance and geometry, and illumination. The dataset has been created to facilitate research on multirate, multisensor gaze estimation and prediction. Nonetheless, this dataset is the first to include high-frequency eye-image sequences and photosensor information associated with ground-truth gaze vectors, following a diverse variability distribution for head-mounted eye-tracking devices. Therefore, we anticipate it will also be beneficial for research in unconstrained, spatiotemporal gaze estimation.

5.3.1 Design

The dataset has been created using UnityEyes (Wood et al., 2016b; Wood et al., 2015; Wood et al., 2016a), an open-source rendering-based simulation framework that synthesizes images of the left periocular region. This tool has previously been used to create other synthetic gaze estimation datasets (Porta et al., 2019), and has been proven successful for training appearance-based DL approaches (Gou et al., 2017; Park et al., 2018). For this work, we have modified the Unity scene of UnityEyes[10], provided by the authors, to control the following parameters: eye rotation, periocular morphology, pupil size, camera location and orientation, illumination, as well as iris and skin texture.

Sensors configuration

The sensors are located on a sensor grid, simulated to be placed *on-axis* (i.e., orthogonal to the optical axis when the eye is looking at infinity) on a glasses-like form factor and co-located with the glasses lens. The sensor grid coordinate system is z-axis parallel to the optical axis of the eye when it is looking straight ahead, and y-axis perpendicular to it. Its origin is vertically aligned with the optical axis and aligned in Z with the corneal plane at 11mm physical *eye relief* (i.e., the distance between the cornea apex and the grid when looking straight ahead) as default distance. The sensor grid coordinate system is left-handed, i.e., positive x-axis toward the nose, positive y-axis upward, and positive z-axis toward the forward direction.

Inspired by recent studies on optimal photosensor positions to maximize eye-tracking signal quality (Zemblys and Komogortsev, 2018), the grid consists of a 2D planar array of 4x4 *photodiode cameras* equally spaced at 0.5 cm, and an HD camera in the center of the grid. The HD camera captures an RGB image of the eye of 480 px x 640 px. The photodiode cameras also capture an RGB image of small portions of the eye region, with a resolution of 512 px x 512 px. Photodiode images are converted to grayscale and transformed into *intensity* values using a transfer function with a Gaussian-like kernel as in Rigas, Raffle, and Komogortsev (2017), which captures the radiance as a function of the angle and multiplies it with the image. More

[9] More information about the dataset can be found at https://github.com/crisie/opensfeds.
[10] UnityEyes: https://www.cl.cam.ac.uk/research/rainbow/projects/unityeyes/.

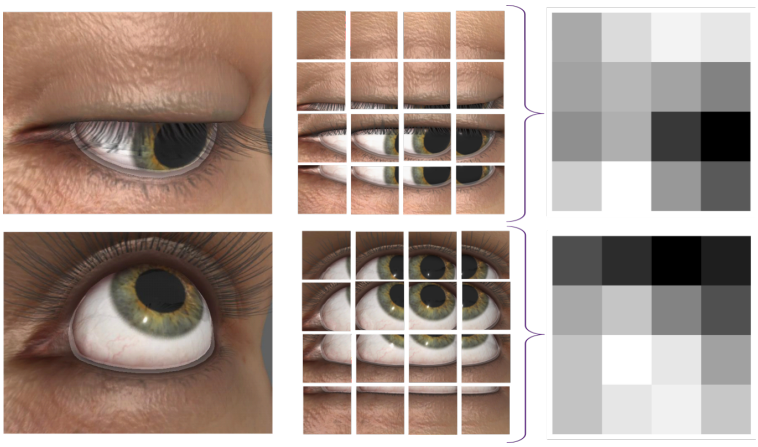


FIGURE 5.2: Example of identity from OpenSFEDS, with different illumination, gaze angle, and sensor shift. Left: image from main camera. Middle: images captured from photosensor cameras. Right: intensity values computed from photosensor cameras.

specifically, the transfer function was provided by a photodiode vendor[11]. The field of view of the HD camera is 70°, while the field of view of the photodiode cameras is 40.475°. Both camera and photosensor array cover the same field of view, so the information gathered from each is equivalent. Figure 5.2 shows an example of the captured images and transformed photosensor intensities. As can be seen, the photosensor intensities form a 2D image that is visually similar to a single low-resolution camera image, like the ones used by Tonsen et al. (2017). However, as commented, photosensors are significantly faster than digital cameras, as well as more sensitive to small changes in light intensity caused by small pupil movements. Representing the photosensor intensities as 2D images allows us to leverage the existing spatial structure with CNNs.

Definition of subject identities

We defined the 60 identities such that their appearance variability was maximized while keeping the number of unique identities as small as possible to reduce the dataset size. UnityEyes includes 20 skin textures and five iris textures, and controls the periocular region morphology by means of a set of 20 principal component analysis (PCA) coefficients. Consequently, we chose to have three identities per skin texture and 12 identities per iris texture, which sums up to 60 subject identities.

To select the morphology features of our identities, we first generated a set of 1,000 random identities using UnityEyes' random generator, each containing unique PCA coefficients and other identity features. We applied k-means (Kanungo et al., 2002)[12] to the 1,000 sets of PCA coefficients to select 60 representative samples of periocular morphology. To these selected samples, we then applied an approximation of the farthest neighbors algorithm (Agarwal, Matoušek, and Suri, 1992) with Euclidean distance as metric, to find the most differing morphologies in the shape-coefficients space and assigned the same skin texture to them, resulting in three periocular morphologies per skin texture. Prior to repeating the same procedure for the iris texture, we decreased the distance of the periocular morphologies with the same

[11] Technical specifications of simulated photodiode: https://www.vishay.com/docs/81962/vemd2000.pdf.

[12] K-means implementation: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

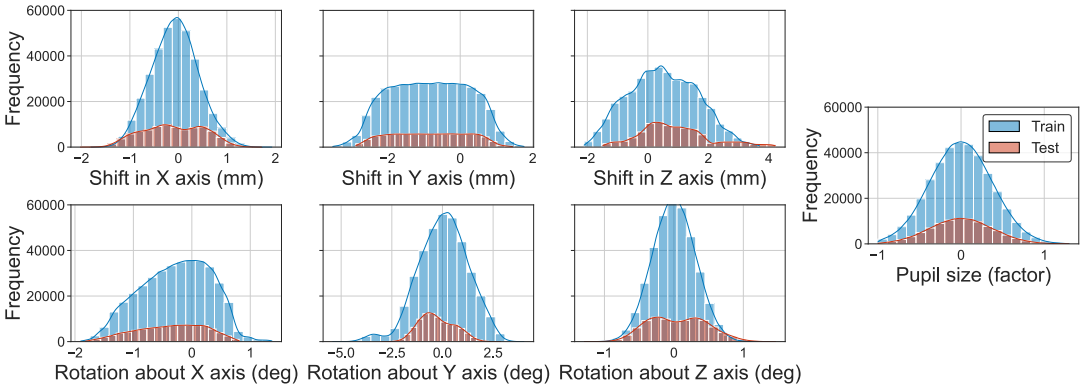


FIGURE 5.3: Histograms of the combined (*fixed* + *variable*) variability featured in the *OpenSFEDS-Static* subset (*OpenSFEDS-Temporal* follows the same distribution but with higher frequency). Shifts and rotations are given with respect to the default sensor grid position. Types of variability not included in this figure follow a uniform distribution.

assigned skin texture by a factor of two, to minimize the number of morphologies with the same iris and skin texture. This process gave us the 60 identities with their associated periocular morphology, skin, and iris texture.

The iris size was taken from the original, randomly generated identity features associated with the 60 selected samples. Bottom and top eyelashes length, and the eye rotation around its optical axis, were sampled from a uniform distribution. The latter was performed to increase variability with respect to the scleral veins appearance, as these are part of the iris texture and have the same location for all textures.

Generating further variability

We define two types of variability: 1) *fixed* per subject identity, related to per-person head anatomy differences (e.g., appearance, eye relief, nose bridge); and 2) *variable*, simulating sensor slippage and other variable scene and subject factors (e.g., illumination, pupil size). For the former, starting from the previously defined subject identities, we apply a per-identity translation and rotation transformation (6DOF) to the sensor grid to simulate differences in head geometry. For the latter, the sensor location variable variability is applied to the sensor grid after the fixed one, simulating sensor shifts by applying a second transformation to the grid with 6DOF. The distributions from which the different transformations were sampled (see Figure 5.3) were obtained empirically. Scene illumination is modeled as a directional light and follows a uniform distribution. Pupil size follows a Gaussian distribution.

5.3.2 Data subsets

The dataset contains two data subsets: *OpenSFEDS-Static*, consisting of static camera-photosensor image pairs, and *OpenSFEDS-Temporal*, consisting of image-pair sequences of eye movements. The subsets are divided into subject-independent train and test splits with ratio 4:1, stratified with respect to skin and iris textures, resulting in 48 subjects for training and 12 for test. The subject identities are the same for both subsets. Each image pair is associated with a ground-truth 2D gaze angle (i.e., horizontal and vertical eye rotation) given in the eye-in-head spherical coordinate system (i.e., HCS), with the eyeball center as the origin of gaze direction.

OpenSFEDS-Static

The *Static* subset can be used to investigate sensor shift-invariant approaches under different types of variability for static VOG, PSOG, or hybrid solutions. It consists of 450K non-contiguous image pairs with subjects looking at a 2D grid of range $\pm 25^\circ$, with gaze targets uniformly distributed at 0.5° on horizontal and vertical axes. We create this subset by generating samples for three full gaze grids per identity, and then randomly subsampling the data to 25% of the data size. The variable variability is assigned to each sample independently. Figure 5.3 depicts the per-sample combined (i.e., fixed + variable) variability of the subset. Figure 5.4 shows examples of the extremes and mean of the variability range for each variability source with a continuous distribution, featuring varied subject identity appearance (i.e., skin and iris textures, iris size, and eyelashes length) and gaze directions.

OpenSFEDS-Temporal

The *Temporal* subset facilitates research in single- and multirate sensor fusion leveraging spatiotemporal information. It consists of 30, 2-s image-pair sequences per identity sampled at 500 Hz, which sums up to 1800 sequences and 1.8M images. Although the image pairs are time-synchronized, assuming no system latency, they could also be leveraged to develop approaches for non-synchronized sensor fusion use cases. Gaze trajectories used in this subset are extracted from the GazeBase dataset (Griffith et al., 2021)[13], which contains real-world, monocular eye movement recordings at 1000 Hz, captured by 322 participants performing a battery of gaze-elicited tasks using a commercial eye tracker (EyeLink 1000). In particular, we selected the Balura game task (Komogortsev, Ryu, and Koh, 2012), for which stimuli varied across participants and recording sessions, and which included instances of fixations, smooth pursuit, and saccades. During this game, blue and red balls moving at a slow fixed speed were presented on a screen, and participants were asked to remove all red balls from the display area as quickly as possible by fixating on them. If after fixating on a specific ball such ball was not removed, participants were instructed to move their gaze away from it and fixate on it again.

To build the dataset sequences, we first gathered the 100 best participants throughout the whole dataset in terms of the average tracking accuracy provided by the eye tracker, which was kindly provided by the dataset authors. Then, we shortlisted up to 60 2-s non-overlapping valid subsequences (i.e., valid in terms of not including blinks or missing data) among the best Balura-based recordings per participant, selecting the subsequences randomly. Finally, we selected the best 60 participants that had at least 30 2-s valid subsequences. This way, each participant corresponds to one subject identity of our dataset. The final selected subsequences were downsampled from 1000 Hz to 500 Hz, dropping 1 every 2 samples in order to decrease the size of the final dataset while allowing for accurate tracking of saccadic eye movements (Mack, Belfanti, and Schwarz, 2017; Raynowska et al., 2018). Gaze trajectories are densely distributed within $\pm 25^\circ$ of visual angle (see Figure 5.5). The generated images for each specific sequence have fixed variable variability, that is, no sensor slippage or changes in illumination or pupil size take place during a sequence.

[13] GazeBase dataset: https://figshare.com/articles/dataset/GazeBase_Data_Repository/12912257.

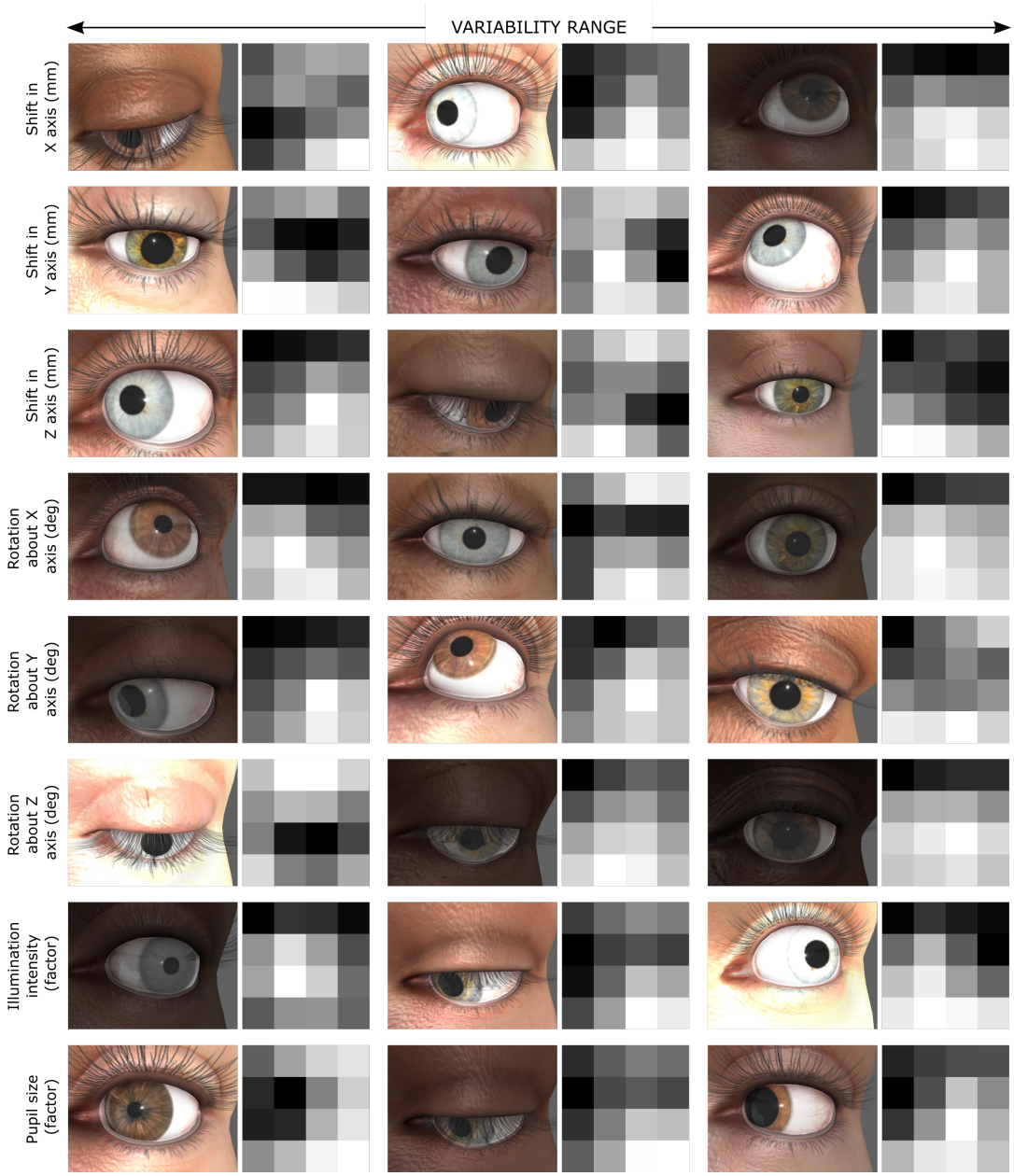


FIGURE 5.4: Example of variability featured in the OpenSFEDS dataset through samples included in the *Static* data subset. Each row depicts the dataset samples with most extreme values (left and right) and with average value (center) for each source of variability with continuous distribution, i.e., sensor shifts and rotations, illumination intensity, and pupil size. Each sample is also characterized by other variability factors, such as those fixed per identity, i.e., iris size, eyelashes length, and skin and iris textures, and variable ones, i.e., gaze direction.

5.3.3 Further considerations

The dataset is not tied to any particular existing device configuration, thus based on a generic setup. Nonetheless, the dataset has been rigorously generated following the requirements and conditions of existing real systems. In addition, to increase generalization to different environments, our dataset mimics an eye tracker operating under visible light (i.e., passive illumination), that is, not requiring IR illumination. Using visible light also makes OpenSFEDS comparable to remote-camera

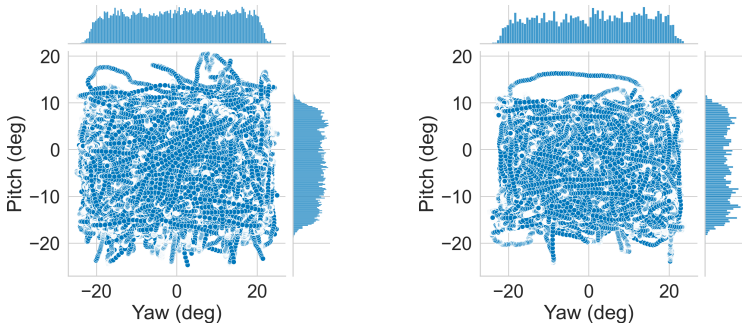


FIGURE 5.5: Gaze distribution of train (left) and test (right) splits of *OpenSFEDS-Temporal*. Yaw: horizontal eye movement. Pitch: vertical eye movement.

datasets, which are usually captured with RGB cameras.

The eye model included in UnityEyes is a simplified version of a real eye that approximates most of the characteristics required by VOG theory (Guestrin and Eizenman, 2006); however, it assumes that the visual and optical axes are equivalent, that is, the angle kappa is zero, among other simplifications. Other works based on UnityEyes have modified such eye model to increase anatomical accuracy (Kim et al., 2019; Porta et al., 2019). The GazeBase dataset does not provide the geometric eye parameters of the participants. Therefore, for simplicity, we opted to use the original UnityEyes eye model, as it does not impact the applicability of our proposed dataset for the objectives considered.

5.4 Sensor fusion for gaze estimation

In this section, we formulate the problem of sensor fusion for gaze estimation for single- and multirate scenarios. We specifically focus on the case of two sensors, but the formulation could be derived for any arbitrary number. We also propose a general feature-based fusion framework (see Figure 5.6) and present four baseline approaches valid for both scenarios.

5.4.1 Problem statement

Let us consider two sensors such that $s \in S = \{C, P\}$. C is a high-fidelity sensor that captures near-eye images $\mathbf{I}_t^c \in \mathbb{Z}^{C_W \times C_H}$ at a low sampling rate f_c. P is a low-fidelity sensor that captures low-resolution eye data $\mathbf{I}_t^p \in \mathbb{R}^{P_W \times P_H}$ at a sampling rate f_p ($> 2f_{Nyquist}$[14]), such that $f_c \leq f_p$. Both sensors measure complementary signals for the underlying eye state $\mathbf{x}_t \in \mathbb{R}^{x_D}$, thus they can be used to infer the line of gaze $\mathbf{y}_t \in \mathbb{R}^2$. C is more informative about the 3D geometry of the eye, appearance, or illumination changes, and thus captures the eye state more accurately. By contrast, P captures small/fast eye movements with high temporal and spatial resolution, but with a less accurate approximation of the global eye state.

We claim that, by leveraging both sensors for eye tracking: 1) for single-rate scenarios ($f_c = f_p$), we can exploit their spatiotemporal complementarity to increase the accuracy of the output gaze signal with respect to a single sensor; and 2) for multirate operation ($f_c < f_p$), we can use P, informed by C, to track fast eye motion accurately with a higher sampling rate. The goal is to find a function $g_F(\cdot)$ that

[14]The Nyquist frequency refers to the highest frequency present in the input signal, i.e., eye motion (Shannon, 1949).

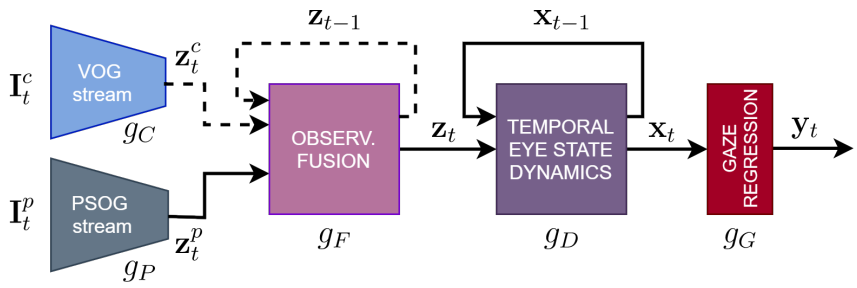


FIGURE 5.6: Overview of the proposed sensor fusion framework for gaze estimation. Dashed lines denote optional inputs.

combines the two modalities with the possibility of the sensors operating at different sampling rates. Ideally, the added model complexity resulting from the fusion of both modalities should be minimal for the final system to be eligible for portable devices. Despite that, in this work, we focus on feasibility of fusion rather than computational complexity. We also assume that: a) $f_p = nf_c$ where $n \in \mathbb{N}$; b) both sensors operate at a constant but possibly different frame rate; and c) the latency of the capture and processing systems is 0. However, our formulation could be adapted for other cases.

5.4.2 Gaze estimation framework

Let $\mathbf{y}_t$ be the 2D (horizontal and vertical) gaze angle at some discrete-time t. We pose the gaze estimation task as a regression problem, in which a mapping from a high-dimensional signal ($\mathbf{I}_t^c$ or $\mathbf{I}_t^p$) to $\mathbf{y}_t$ is learned. We consider two independent non-linear encoders that map the signal captured by each sensor to a low-dimensional feature vector of dimension Z_s, $\mathbf{z}_t^s \in \mathbb{R}^{Z_s}$, such that:

$$\mathbf{z}_t^s = g_s(\mathbf{I}_t^s). \tag{5.1}$$

The independent modeling allows for autonomous sensor operability.

We consider $\mathbf{z}_t^s$ to be *observations* of the underlying eye state $\mathbf{x}_t$. Such eye state may consist not only of its rotation, velocity, and acceleration, but also of other parameters deemed to be important to infer $\mathbf{y}_t$, such as the current type of eye movement being performed. Since eye movement is inherently a dynamic process, and following the previous evidence about the importance of sequential information for gaze estimation, we model the eye state as a non-linear dynamical system, such that:

$$\mathbf{x}_t = g_D(\mathbf{z}_t, \mathbf{x}_{t-1}), \tag{5.2}$$

where $\mathbf{z}_t$ is the fused signal from g_F (see Section 5.4.3) that combines the available sensor observations at time t. This framing resembles a state estimation problem, in which the goal is to estimate $\mathbf{x}_t$ based on a set of observations. Finally, gaze is computed from the eye state, such that:

$$\mathbf{y}_t = g_G(\mathbf{x}_t). \tag{5.3}$$

5.4.3 Baseline fusion modules

Here, we describe four feature-based baseline approaches to model the sensor fusion function $g_F(.)$ for the case $f_c \leq f_p$.

– **Naive fusion.** As a starting point, we consider the naive feature fusion function used in Chapter 3, such that:

$$\mathbf{z}_t = h_n([\mathbf{z}_t^C, \mathbf{z}_t^P]), \tag{5.4}$$

where $[\cdot, \cdot]$ denotes feature concatenation. Irrespective of the sensors sampling rate, when there is no signal from C at a given t such that $\nexists \mathbf{z}_t^C$, we set $\mathbf{z}_t^C = \mathbf{0}$, so that the function learns to perform feature selection under missing data.

– **XOR fusion.** We could hypothesize that when $\exists \mathbf{z}_t^C$, $\mathbf{z}_t^P$ may not provide enough new information, thus P could be deactivated to further decrease power consumption at t. XOR fusion is an alternative to Naive fusion, in which only one of the modalities is used as input to h_n. That is, when $\exists \mathbf{z}_t^C$, the input to h_n is $[\mathbf{z}_t^C, \mathbf{0}]$, and $[\mathbf{0}, \mathbf{z}_t^P]$ otherwise.

– **MR-MAG fusion.** Naive and XOR fusion approaches do not include any additional mechanism to inform $\mathbf{z}_t^P$ when $\nexists \mathbf{z}_t^C$. We could consider an *observation memory* that stores and propagates the previous fused signal $\mathbf{z}_{t-1}$, which encompasses information from past C and P observations, to enrich the current time step's signal and also help regularize it in case of noisy observations. Therefore, it could be used not only to inform $\mathbf{z}_t^P$, but also $\mathbf{z}_t^C$. To implement this mechanism, we take inspiration from the Multimodal Adaptation Gate (MAG) unit (Rahman et al., 2020), originally proposed for multimodal language models to enrich a lexical input vector with visual and acoustic information. We repurpose this idea and consider $\mathbf{z}_{t-1}$ as a location in the observation space, to be shifted based on the new information provided by $\mathbf{z}_t^C$ and $\mathbf{z}_t^P$. When $\nexists \mathbf{z}_t^C$, only $\mathbf{z}_t^P$ is used to shift the signal. We refer to this version as MR-MAG (Multirate MAG).

Following the original MAG formulation, we first obtain a gating vector per sensor ($\boldsymbol{\alpha}_t^s$) to highlight the relevant information of each observation conditioned on the previous fused signal, such that:

$$\boldsymbol{\alpha}_t^s = g_{\boldsymbol{\alpha}}^s([\mathbf{z}_{t-1}, \mathbf{z}_t^s]). \tag{5.5}$$

Then, we aggregate the observation signals available at t to synthesize a displacement vector:

$$\mathbf{z}'_t = \sum_{s \in G_t} \boldsymbol{\alpha}_t^s \circ h_m^s(\mathbf{z}_t^s), \tag{5.6}$$

where $G_t \subset S$ is the subset of sensors that provide signal at t and $\circ$ denotes the Hadamard product. Finally, a weighted summation is performed between $\mathbf{z}_{t-1}$ and the displacement vector based on the proportion of their magnitudes, such that:

$$\mathbf{z}_t = \mathbf{z}_{t-1} + \lambda \mathbf{z}'_t, \tag{5.7}$$

where $\lambda = \min(\beta(\|\mathbf{z}_{t-1}\|^2 / \|\mathbf{z}'_t\|^2), 1)$ and β is a hyperparameter. The scaling factor λ is used such that the effect of the shift $\mathbf{z}'_t$ is within a desirable scale.

– **MR-MRAG fusion.** The original MAG does not fuse the accompaniment modalities based on their relative contribution with respect to each other.

To do so, we slightly modify Equation 5.6 to incorporate *re-adaptation* gating vectors $\boldsymbol{\alpha}_t^{Fs}$, such that:

$$[\boldsymbol{\alpha}_t^{Fc}, \boldsymbol{\alpha}_t^{Fp}] = g_{\boldsymbol{\alpha}}^F([\mathbf{z}_t^C, \mathbf{z}_t^P]), \tag{5.8}$$

where $g_{\boldsymbol{\alpha}}^F$ is modeled as $g_{\boldsymbol{\alpha}}^s$. The displacement vector is then computed as:

$$\mathbf{z}'_t = \sum_{s \in G_t} \boldsymbol{\alpha}_t^{Fs} \circ \boldsymbol{\alpha}_t^s \circ h_m^s(\mathbf{z}_t^s). \tag{5.9}$$

We refer to this variant as MR-MRAG (Multirate Multimodal Re-Adaptation Gate). Note that when $\nexists \mathbf{z}_t^C$, Equation 5.8 is not applied, $\boldsymbol{\alpha}_t^{Fc} = \mathbf{0}$, and $\boldsymbol{\alpha}_t^{Fp} = \mathbf{1}$.

5.5 Experimental evaluation

In this section, we evaluate the performance of DL-based VOG and PSOG approaches under the different sources of variability considered in *OpenSFEDS-Static* and *OpenSFEDS-Temporal*, and the feasibility and utility of single- and multirate sensor fusion on *OpenSFEDS-Temporal*. With these evaluations, we also validate the adequacy of the proposed two data subsets. We describe the experimental procedure used and discuss the results obtained for each approach (i.e., unimodal or multimodal, for single-rate or multirate versions). The approaches evaluated are based on the framework presented in Section 5.4, being C the camera sensor (*VOG*), and P the photosensor array (*PSOG*).

5.5.1 Evaluation protocol

For hyperparameter tuning, and architecture and model selection, we create a validation set from the provided training split with 12 subjects randomly selected by stratifying on skin texture, resulting in a final training set of 36 subjects. By architecture, we refer to a specific combination of number of layers, hidden units, and hyperparameters for different choices within pre-fusion and fusion modules that will be described next (Section 5.5.2). By contrast, by model, we refer to the set of trained weights that a given architecture contains after a given training epoch for a given approach. The models are trained on the new training splits of each data subset (*Static* or *Temporal*).

We perform grid search to find the best architecture per approach, which is selected based on the best average angular error (i.e., gaze accuracy, Equation 2.4) on the validation set averaged over three independent runs to account for the stochasticity of the learning process. Performance results of the selected models for each approach are compared on the test split. Performance is reported as the angular error between the estimated and ground-truth gaze direction. To do so, the 2D gaze angle in spherical coordinates is converted into a 3D unit vector following Equation 2.1.

5.5.2 Implementation details

Input preprocessing

We first convert all camera images $\mathbf{I}^c$ to grayscale, normalize values into [0,1] range, and downsample to 80 px x 60 px, which gave similar accuracy in preliminary experiments as 160 px x 120 px while expediting training time. Higher resolutions were

not evaluated. The intensity values of the photosensors data $\mathbf{I}^p$ are also normalized into [0,1] range by means of min-max scaling, setting $\min_P = 0$ as theoretical minimum value and $\max_P = 40$ based on the overall maximum value of the training set.

Architecture of pre-fusion models

To model the backbone encoder, g_C, we employ a 14-layer ResNet-based network, previously used for both static and spatiotemporal near-eye gaze estimation in Chapter 4 (Palmero, Komogortsev, and Talathi, 2020; Palmero et al., 2021b), followed by a 64-hidden-unit FC layer with ReLU non-linearity. For g_P, following previous research on shift-invariant PSOG mappings (Griffith, Katrychuk, and Komogortsev, 2019), we propose a 3-layer CNN with kernel of 3 px x 3 px and $8l_P$ channels per layer, where l_P is the layer index, followed by an adaptive average pooling and a 32-hidden-unit FC layer with ReLU non-linearity. This architecture was the top performer in a preliminary evaluation with {2, 3, 4} convolutional layers, {8, 16, 32} number of channels for the first layer, and {16, 32, 64} FC sizes. For all dynamic models, right after each encoder g_s, which produce feature vectors of size $n_C = 64$ and $n_P = 32$, we add an intermediate module g_{I_s} to further evolve such output features and make the different unimodal and multimodal approaches comparable in terms of parameter complexity. We follow two possible strategies to model g_{I_s}: 1) include $L_s \in \{0, 1, 2\}$ additional FC layers with $n_s/2^{l_s}$ hidden units each, where n_s is the size of the output feature vector from g_s and l_s the 1-based FC layer index, i.e., if $L_s = 0$, no FC layer is added; and 2) include $L_s \in \{1, 2\}$ FC layers with final output vector of $n_f = \{16, 32, 64\}$ units (and intermediate FC size of $(n_s + n_f)/2$ for the first layer if $L_s = 2$), thus having same output size for each modality. With the former, we obtain observations of different sizes to be fed to the fusion module. By contrast, with the latter, we ensure that the observations from both sensors are of the same size. All FCs are followed by Tanh activation. For simplicity, we henceforth assume that the output of g_{I_s} is $\mathbf{z}_t^s$ with dimension Z_s.

Naive and XOR fusion

To model h_n, we follow two possible strategies: 1) an identity function, such that the fusion is based on simple feature concatenation; and 2) a learnable MLP of $M = \{1, 2\}$ layers with final fused signal of size $Z = \{8, 16, 32, 64\}$. If $M = 2$, the intermediate FC has size of $(Z_C + Z_P + Z)/2$ and is followed by ReLU activation. The last layer is followed by Tanh activation. We henceforth refer to l_n as the 1-based layer index of h_n.

MR-MAG and MR-MRAG fusion

Following the original MAG implementation[15], $g_{\boldsymbol{\alpha}}^s$ is modeled as an FC layer with ReLU non-linearity, h_m^s is an FC layer of size Z = {16, 32, 64} with linear activation function, and LayerNorm is applied to the output $\mathbf{z}_t$. The β parameter is optimized within the range {0.01, 0.1, 0.25, 0.5, 0.75, 1, 10, 100}.

[15]Original open-source implementation of Multimodal Adaptation Gate (MAG): https://github.com/WasifurRahman/BERT_multimodal_transformer.

Architecture of post-fusion modules

We model g_D as a vanilla 1-layer LSTM with 32 hidden units and $\mathbf{x}_0 = \mathbf{0}$, and g_G as a one FC layer with linear activation function. As opposed to previous chapters where the temporal module was a many-to-one network, for this work, we use a many-to-many network. In other words, we have an output gaze estimate for each image input at each time t.

Weights initialization

Convolutional layers are initialized following He et al. (2015). Batch normalization layers (Ioffe and Szegedy, 2015) are initialized with default configuration (gamma set to 1 and beta set to 0). FC layers are initialized with Xavier's uniform distribution (Glorot and Bengio, 2010) and bias set to 0. In regards to LSTM layers, hidden-hidden weights are initialized with orthogonal matrices (Saxe, McClelland, and Ganguli, 2014; Vorontsov et al., 2017) and input-hidden weights with Xavier's uniform initialization, while forget gate biases are set to 1 and remaining biases to 0 (Jozefowicz, Zaremba, and Sutskever, 2015).

Training strategy

We perform stage-wise training. First, we train from scratch static VOG (VOG_S) and PSOG ($PSOG_S$) models on the *Static* subset, using g_C and g_P encoders, respectively, followed by g_G. This pretraining allows for learning useful gaze and overall eye features with different variability despite the gaze range being sparsely distributed. Next, for dynamic fusion models, we discard the trained g_G, add selected g_F and g_D modules, and a new g_G on top, and train the whole sensor fusion network (SF_D, Section 5.4.2) on the *Temporal* subset. More specifically, in this second stage, the new modules are trained from scratch, while g_P is completely fine-tuned, and for g_C only the last convolutional block and FC layer are finetuned. We also train unimodal dynamic models for comparison (VOG_D and $PSOG_D$), using the same modules as SF_D except g_F.

All networks were trained using ADAM (Kingma and Ba, 2014) optimizer, with weight decay set to 10^{-6}, and L1 loss. We empirically set the initial learning rate to 10^{-5}, reducing it by a factor of 0.5 when learning stagnates. We used early stopping based on the validation error to select the best-performing weights for each architecture, with a maximum number of 150 epochs. The static networks were trained with a batch size of 32. The dynamic models were trained using backpropagation through time with a temporal window of 100 time steps (i.e., 200 ms), batch size of 8, and gradient clipping of norm 2. For multirate fusion, camera frames available are shifted with stride of 1 every epoch, such that all possible camera-photosensor pairs are seen during training.

5.5.3 Performance of static models on *OpenSFEDS-Static*

The performance of dynamic models depends on the static independent models VOG_S and $PSOG_S$. Here, we report the accuracy of such static models on the *Static* subset. VOG_S obtains an average angular error of $1.849^\circ \pm 1.017$ SD, on par with other works that use synthetic data for subject-independent, near-eye gaze estimation (Kim et al., 2019). By contrast, $PSOG_S$ obtains an error of $4.225^\circ \pm 3.046$, higher than other appearance-based PSOG methods like the one proposed by Griffith, Katrychuk, and Komogortsev (2019), which obtained around 1.3°. This difference can

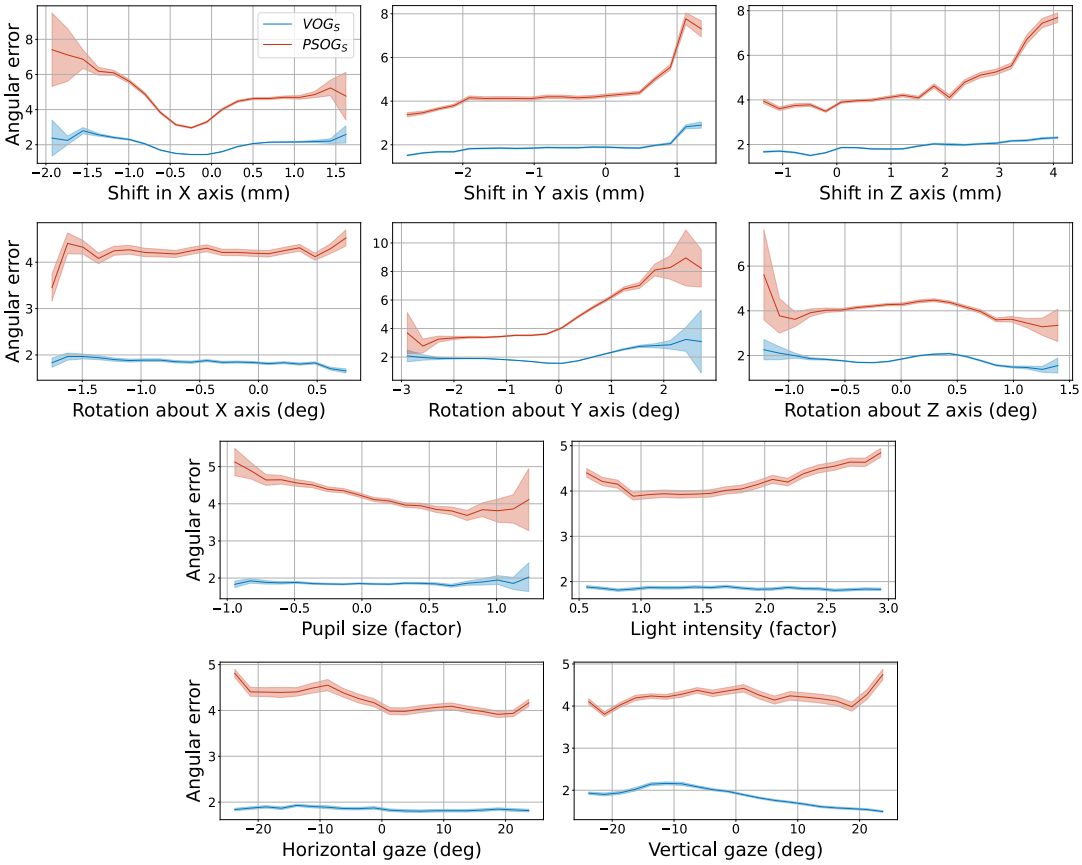


FIGURE 5.7: Effect of selected sources of variability on the performance of static models VOG_S and $PSOG_S$ on *OpenSFEDS-Static*. Reported as average angular error (degrees) $\pm$ 95% confidence interval.

be explained by the increase in variability in our dataset compared to the data used for previous PSOG studies (see Section 5.2). Figure 5.7 shows the effect of some of the sources of variability with continuous distribution on the performance of the static approaches. As can be seen, VOG_S is relatively robust to all sources, while $PSOG_S$ is severely affected by sensor shifts and rotation about the Y axis. For discrete distribution variability, we find that identity appearance plays an important role on $PSOG_S$ performance, with average errors ranging from 2 to 8°. Particularly, darker irises obtain a lower error, likely due to the higher contrast with the sclera.

5.5.4 Performance of temporal fusion approaches on *OpenSFEDS-Temporal*

Assuming that the eye state is always updated at a given frequency f_{max}, we hypothesize that multirate sensor fusion can be proven useful if the error obtained by a multirate fusion approach is within two expected error bounds: a lower bound with $C + P$ (SF_D) operating at a given f_{max}, and an upper bound with C (VOG_D) operating at f_{min}, i.e., performing eye state forecasting when no C signal is available ($\mathbf{z}_t^C = \mathbf{0}$). To evaluate this hypothesis, we train two versions of each dynamic model: one with C operating at f_{max} (single-rate scenario), and another with C operating at f_{min} (multirate scenario). For $PSOG_D$, we only train the f_{max} version, assuming P always operates at the eye-state frequency. For all experiments, we define $f_{max} = f_{\mathbf{x}} = f_P = 500$Hz, being the sampling rate of the dataset, and $f_{min} = 100$Hz,

TABLE 5.1: Results of evaluated sensor fusion approaches (SF_D) for single- and multirate scenarios compared to unimodal baselines (VOG_D and $PSOG_D$), reported as angular error (degrees) on the *OpenSFEDS-Temporal* test split, averaged among three independent runs. p50: 50th percentile. p95: 95th percentile. Total: error over the whole test split. Eye movement categories: *F/SP* - Fixation/smooth pursuit; *Sacc.* - Saccade; *Other* - uncategorized.

Approach	Single-rate ($f_c = f_p = f_{max}$)						Multirate ($f_c = f_{min}, f_p = f_{max}$)					
	Total Avg.	**Total p50**	**Total p95**	**F/SP Avg.**	**Sacc. Avg.**	**Other Avg.**	**Total Avg.**	**Total p50**	**Total p95**	**F/SP Avg.**	**Sacc. Avg.**	**Other Avg.**
CAM_D	1.752	1.580	3.630	1.718	1.884	1.993	1.891	1.696	3.901	1.842	2.133	2.075
$PSOG_D$	4.075	3.603	8.718	4.052	4.197	4.134	4.075	3.603	8.718	4.052	4.197	4.134
SF_D **Naive**	**1.730**	1.584	3.493	**1.701**	1.838	1.956	1.830	1.664	3.850	1.801	**1.948**	2.005
SF_D **XOR**	-	-	-	-	-	-	**1.795**	**1.635**	3.745	**1.758**	1.963	**1.993**
SF_D **MR-MAG**	1.752	1.593	3.472	1.723	1.869	1.943	1.836	1.706	**3.697**	1.798	1.997	2.070
SF_D **MR-MRAG**	1.733	**1.577**	**3.381**	1.710	**1.823**	**1.908**	1.829	1.654	3.779	1.795	1.970	2.046

as a usual sampling rate for portable eye trackers. Nevertheless, other sampling rates could be considered and studied in the future for lower-power eye tracking.

An ideal approach should be able to precisely follow the eye trajectory during fast gaze changes without increasing noise during steady gaze. To evaluate the performance of our approaches with respect to different eye movement types, we classify each frame into belonging to a *saccade*, *fixation/smooth pursuit* ("*F/SP*"), or *other* (unclassified) events by means of a velocity-based I-VT algorithm (Komogortsev et al., 2010)[16], with a saccade threshold empirically set to 70°/s. The *F/SP* category includes both fixation and pursuit events since their velocity ranges overlap, and thus their distinction is still an open research problem (Komogortsev and Karpov, 2013).

In addition to the average error, which is the standard measure used by the CV/ML appearance-based gaze estimation community, we also measure the 50th (i.e., median) and 95th percentiles for the whole test split. As observed in previous chapters, the SD of gaze estimation errors is usually large, indicating that the distribution of errors is skewed or exhibits a significant spread. Therefore, the average error leads to an incomplete portrayal of the estimation performance. By contrast, p50 offers a more robust central tendency, and p95 offers a measure of robustness for the most complicated samples while accounting for potential outliers. The latter is particularly relevant when considering real-world applications where certain critical tasks demand high-confidence estimates for the entire population.

Table 5.1 summarizes the results obtained for both single- and multirate scenarios. We apply a one-sided Wilcoxon signed-rank test (Conover, 1999) to all the models to further assess performance differences. Results are discussed next.

Single-rate fusion

Naive and MR-MRAG fusion strategies outperform VOG_D for most reported metrics ($p<.0001$). While average and p50 results are very similar, the reduction in error is better observed for p95, with an error decrease of up to 6.9%, indicating a regularizing effect when combining both modalities. In particular, we find that MR-MRAG surpasses the original MR-MAG ($p<.0001$) and obtains the best p95 value overall. We stress the importance of improving performance for p95, since it better represents which method is more stable and provides higher accuracy especially on challenging cases. For the rest of metrics, all approaches show comparable performance. Figure 5.8a depicts an example of ground truth and estimated gaze traces

[16] I-VT software: https://userweb.cs.txstate.edu/~ok11/emd_offline.html

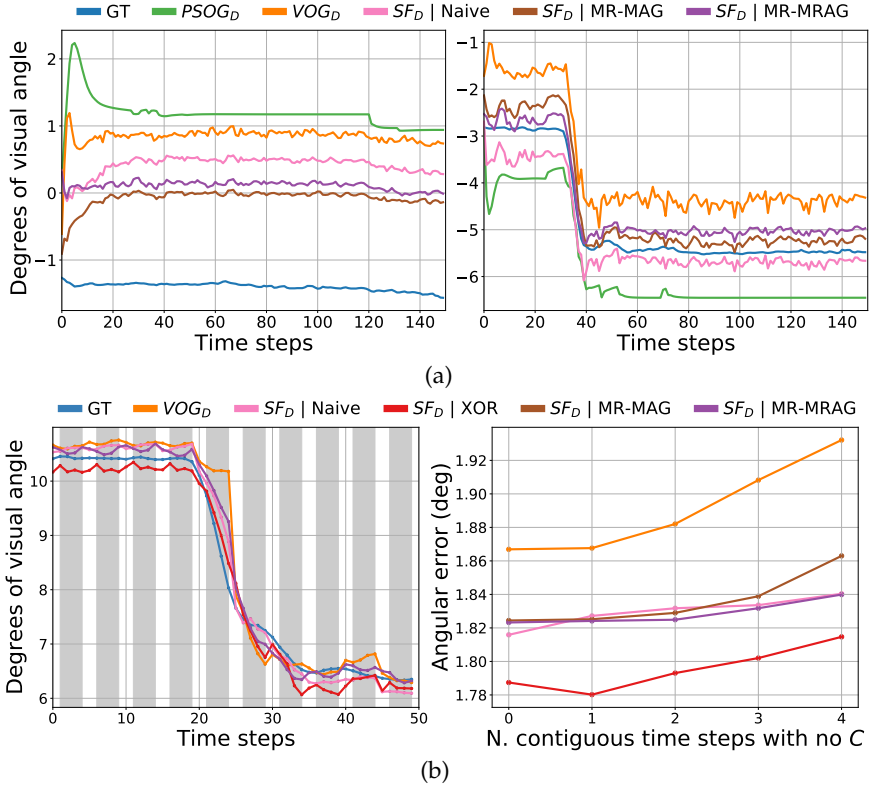


FIGURE 5.8: (a) *Single-rate fusion.* Estimated horizontal gaze traces of unimodal (VOG_D and $PSOG_D$) vs. fusion (SF_D) approaches compared to the ground truth (GT). Plots depict the initial segments of two independent sequences. (b) *Multirate fusion.* Left: Example of GT and estimated traces of unimodal gaze forecasting vs. fusion-based gaze estimation. Gray areas denote time steps with no C (camera) signal. Right: Average error with respect to the number of contiguous time steps with no C (C available at $t = 0$).

from the evaluated approaches. We notice that fusion approaches generally follow the ground-truth trace better overall, with a more stable eye state at the beginning of the sequence. Additionally, Figure 5.9 depicts the effect of different sources of variability on the performance of VOG_D, $PSOG_D$ and $SF_D|MR\text{-}MRAG$. We can observe a modest reduction in error between SF_D and VOG_D, denoting the regularizing effect caused by informing C with P.

Multirate fusion

As shown in Table 5.1, most of the evaluated fusion approaches outperform the forecasting baseline VOG_D ($p<.0001$). Interestingly, the simple XOR is the best-performing fusion approach for most metrics, with an average error decrease of 5.1% ($p<.0001$). However, as we can see in Figure 5.8b (right), the other fusion approaches seem to better exploit the information coming from C when the C signal is available. The best p95 error is obtained by MR-MAG (5.5% decrease). In regards to the different eye movements, the highest accuracy increase is observed for saccades (8.7% obtained by Naive fusion). The lower performance increase for F/SP can be better understood by analyzing the estimated traces. As shown in Figure 5.8b (left), these are slightly noisier during fixations (up to $t = 20$). For saccades (from $t = 20$ onward), VOG_D is less accurate following the trajectory when there is no C signal, producing a step-like trace. By contrast, the fusion approaches better model such

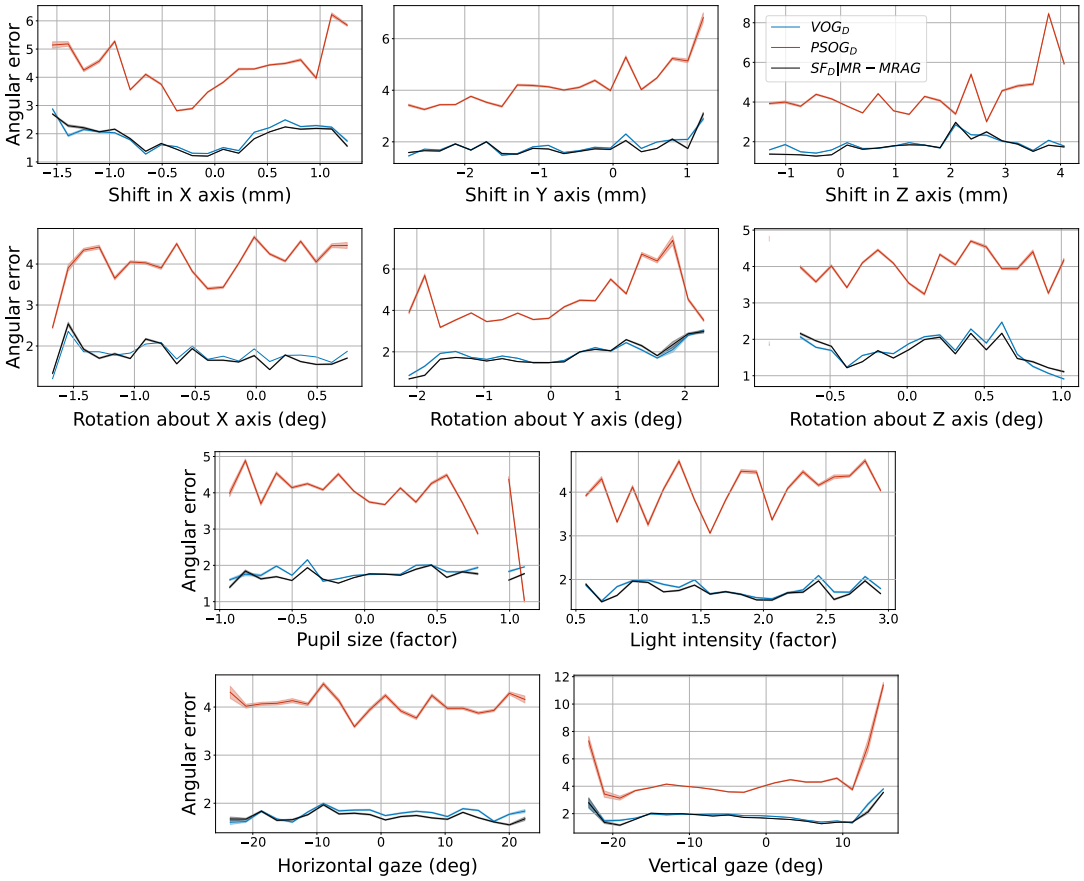


FIGURE 5.9: Effect of selected sources of variability on the performance of single-rate versions (i.e., C and P operating at 500 Hz) of the dynamic models VOG_D, $PSOG_D$, and SF_D | MR-MRAG on *OpenSFEDS-Temporal*. Reported as average angular error (degrees) ± 95% confidence interval.

fast eye motion, clearly benefiting from the added information from P. Note that for slow eye movements we should not expect large improvements regardless of the method used. However, fast and transient movements like saccades need fast-sensor captures to be tracked faithfully.

Architecture details

Table 5.2 illustrates the details of the best-performing architectures of all dynamic models. One might argue that the difference in accuracy between VOG_D operating at 100 Hz and the multirate approaches is caused by the slight difference in the number of parameters. However, note that other architectures with a higher number of parameters were also evaluated by grid search for VOG_D (100 Hz) and resulted in lower performance than the one reported.

Discussion

After analyzing the obtained results, we can determine that SF_D MR-MRAG (500 Hz) is our lower error bound, while VOG_D (100 Hz) is our upper error bound. The performance of most evaluated multirate sensor fusion approaches is within such bounds; therefore, we validate our hypothesis that the addition of P for multirate

TABLE 5.2: Architecture and parameter details of the evaluated dynamic models. These were the best-performing architectures and/or hyperparameters per approach, selected in a grid-search fashion based on the best average angular error on the validation set averaged over three independent runs. For s-specific parameters, C is reported first and P second.

	Single-rate ($f_c = f_p = f_{max}$)						**Multirate ($f_c = f_{min}, f_p = f_{max}$)**					
Approach	Num. **Params**	g_{I_s} (l_s - n. units)	Z_s	h_n (l_n - n. units)	Z	β	Num. **Params**	g_{I_s} (l_s - n. units)	Z_s	h_n (l_n - n. units)	Z	β
CAM_D	261K	1 - 64; 2 - 64	64	-	64	-	247K	1 - 16	16	-	16	-
$PSOG_D$	17K	1 - 16	16	-	16	-	17K	1 - 16	16	-	16	-
SF_D **Naive**	266K	0 0	64 32	0	96	-	284K	1 - 64; 2 - 64 1 - 48; 2 - 64	64 64	0	128	-
SF_D **XOR**	-	-	-	-	-	-	272K	0 1 - 16; 2 - 8	64 8	1 - 68; 2 - 64	64	-
SF_D **MR-MAG**	260K	0 1 - 16	64 16	-	16	0.75	279K	1 - 32 1 - 16	32 16	-	64	100
SF_D **MR-MRAG**	296K	0 1 - 16	64 16	-	64	1	272K	1 - 32 1 - 32	32 32	-	32	1

fusion is effective, improving gaze estimation performance especially for fast movements. Furthermore, we can also confirm that the fusion of both sensors in a single-rate setting is useful for increasing accuracy. Nonetheless, in Table 5.1, we observe that the best performance obtained for multirate operation is slightly lower than the performance of VOG_D at 500 Hz, with 3.48% performance loss for p50 and 1.85% for p95, motivating future research to better exploit the complementarity or redundancy of the signals to perform fusion with missing data. One of our accuracy bottlenecks is the static encoder; we anticipate improvements on this part may translate to better representations for feature-based fusion.

5.5.5 Limitations

This experimental evaluation offers an initial insight as to how leveraging multiple sensors for eye tracking, operating at different sampling rates, can increase the accuracy and effective frequency of the estimated gaze signal with respect to a single sensor when evaluated in a simulated but unconstrained scenario. Performance is dependent upon the architectures used, and the sensors configuration (e.g., their relative position to each other). Additionally, the results reported herein refer to an ideal case of no sensor noise, no system latency, and perfectly synchronized sensor signals, which currently can only be accomplished via simulation. Therefore, the reported accuracy can be regarded as a lower bound for gaze estimation error, conditioned on the baseline fusion modules, gaze estimation framework used, and photorealism of the synthetic data. Nonetheless, this work opens the door to future analyses on computational complexity, real-time performance, signal perturbation, and other real-world issues, and to future developments to transfer the findings to real-world setups.

5.6 Conclusions

In this chapter, we explored the feasibility of combining fast/low-fidelity (photosensors) and slow/high-fidelity (camera) sensors to increase the accuracy and sampling rate of portable eye-tracking systems. We presented OpenSFEDS, a dataset of 2.25M synthetic eye-image pairs captured with a camera and a set of photosensors, featuring variability in appearance and sensor locations. We also evaluated different

baseline fusion strategies for single- and multirate operation under a spatiotemporal appearance-based gaze estimation framework. Results confirm the usefulness of informing photosensors with camera signal to track fast eye movements for multirate operation, and the regularizing effect of informing camera with photosensors signal for the single-rate case, both under the considered synthetic scenario.

We hope this work serves as a stepping stone for future innovations in sensor fusion for eye tracking. We anticipate OpenSFEDS to be instrumental in enabling further research on the topic, and be potentially useful as a testbed for other multirate fusion developments. In turn, we expect to motivate the development of multisensor devices to transfer the findings and methodology advancements from the synthetic domain to real-world setups.

Part II

Applications

Chapter 6

Emotion Expression Recognition in Older Adults Interacting with a Virtual Coach

METHODOLOGICAL ADVANCES in eye tracking are motivated by their current and emerging applications, which were introduced in Chapter 1. As previously discussed, one significant breakthrough in the democratization of gaze tracking involves the potential utilization of off-the-shelf cameras, such as webcams, powered by CV/DL appearance-based gaze estimation approaches as the one presented in Chapter 3, which eliminate the need for personal calibration and make eye tracking more accessible. Despite the relatively lower accuracy and sampling rate of such approaches compared to dedicated eye trackers, their gaze estimates can still be useful for applications that do not require the exact detection or start-end determination of specific eye movements. Instead, researchers are actively investigating webcam-based gaze features computed from raw gaze estimates for diagnostic and interactive applications, such as emotion expression recognition (O'Dwyer, Murray, and Flynn, 2018) or mind wandering detection (Hutt et al., 2023).

Part of this thesis has been carried out under the umbrella of the H2020 EMPATHIC project (*Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly*)[17]. This project aimed to design an emotionally expressive simulated VC capable of engaging healthy senior users to enhance well-being and promote independent aging. One of the core aspects of the system is its human sensing capabilities, allowing for the perception of emotional states from audio and video channels to provide a personalized experience. Such a VC system calls for a non-intrusive approach that could be deployed in any house with just a computer, tablet, or smartphone with a built-in or external camera. Therefore, the project provides a unique setting for adding and evaluating remote camera-based gaze tracking to measure attention and help to recognize emotional states, in combination with speech from audio and facial expressions from video (i.e., multimodal emotion recognition), or individually. Although gaze-based emotion recognition has been previously considered in the literature (see O'Dwyer, Murray, and Flynn, 2018 and Lim, Mountstephens, and Teo, 2020 for extensive reviews), this is the first time gaze-based emotion recognition is evaluated on an HMI task with older adults as the target age group. Since gaze can be represented in different ways, such as with respect to the CCS and HCS, but also as a measure of attention, we study the contribution of these three representations for the emotion recognition task. Furthermore, we add head-related features due to the relationship between head pose and gaze

[17] EMPATHIC website: http://www.empathic-project.eu/.

direction. As a first study, we assess the contribution of the four representations together as a single modality, and each of them separately when in combination with audio and/or video.

In summary, this chapter outlines the development of the EMPATHIC-VC emotion expression recognition module, encompassing data collection, annotation design choices, and a first methodological approach, all tailored to the project requirements. With the latter, we provide an extensive study on discrete emotion expression recognition, wherein we investigate the role of various modalities in this context, individually and combined: speech from audio, and facial expressions, gaze, and head dynamics from video. The collected corpus includes users from three countries, and was annotated separately for the audio and video channels with distinct emotional labels, allowing for a performance comparison across cultures and label types. Results confirm the informative power of the modalities studied for the emotional categories considered, with multimodal methods generally outperforming others. In particular, we find gaze to be particularly useful for video-based labels, and to a lesser extent for audio-based labels.

6.1 Introduction

Emotion recognition plays a pivotal role in conversational HMI (Jaimes and Sebe, 2007; McKeown et al., 2011), enabling systems to perceive and respond to users' emotional states (Justo et al., 2020; Vázquez et al., 2023). Research in affective computing has long proved the possibility of detecting emotion expressions with data-driven approaches that use different input modalities, mainly linguistic, acoustic, and facial expressions (Poria et al., 2017; Rouast, Adam, and Chiong, 2019). Other modalities such as gaze have also been considered, but more sparsely (Soleymani et al., 2011; O'Dwyer, Murray, and Flynn, 2018). Multimodal approaches have also shown promising results in enhancing recognition accuracy and robustness, where each modality can provide complementary information (D'mello and Kory, 2015).

The literature has progressively shifted from initial attempts at recognizing acted, contextless prototypical expressions to more spontaneous reactions and in-the-wild data. However, the latter presents numerous challenges that are the subject of ongoing research. In addition to the increased appearance and behavioral variability, an important consideration is that spontaneous emotions are more subtle and difficult to disambiguate, significantly differing from acted emotions in surface representation (Schuller et al., 2019) and in the lack of exaggeration. As a consequence, the emotional space is smaller than in acted emotions and the emotions are closer to each other (Chakraborty, Pandharipande, and Kopparapu, 2017; De Velasco, Justo, and Torres, 2022). Natural contexts also suffer from a high imbalance in emotional categories, negatively affecting the learning process of data-driven approaches. In conversational HMI settings, users tend to exhibit less intense emotional responses than when interacting with other humans due to the limited emotional capacity of artificial agents, resulting in more neutral expressions (De Velasco, Justo, and Torres, 2022). Such challenges are not exclusive to unimodal models, as improvements achieved by multimodal models are also considerably reduced when dealing with natural, spontaneous data (D'mello and Kory, 2015). Visual-based emotion recognition is further affected by the *speaking effect*, for which facial deformations caused by speaking can be confused with an emotional expression (Mariooryad and Busso, 2015). Another challenge lies in establishing a reliable gold standard. This is usually achieved through perceptual

annotation procedures that sum up the inherently subjective judgment of each rater, which tend to involve low agreement and could result in discrepancies between emotions expressed and perceived (Zeng et al., 2009).

Most emotion recognition research has focused mainly on young adults. However, the global aging of the population is generating new socioeconomic challenges, and older adults, in particular, have been positioned as a target group and clear beneficiary of technological development in HMI (Justo et al., 2020; Demaeght et al., 2022; Alos et al., 2022; Olaso et al., 2023; McTear et al., 2023). In that sense, understanding the emotional states of older adults is of special importance, as it can lead to the development of customized virtual coaching applications, virtual companions, and healthcare technologies that foster active aging and independent living. This poses additional computational challenges, since aging changes facial features, voices, and speaking styles, among other non-verbal behaviors (Magai et al., 2006; Fölster, Hess, and Werheid, 2014). For example, a higher intensity in facial expressions and speech is associated with a higher emotion recognition accuracy; however, older subjects display less intense vocal and facial expressions compared to younger subjects (Levenson et al., 1991; Ma et al., 2019). Therefore, models trained in other age groups do not perform optimally for recognition in this specific age group (Ma et al., 2019), and models trained on data from this age group tend to show lower performance than other groups (Wang et al., 2015; Lopes et al., 2018).

The lack of public databases including older adults hinders progress on this front. Existing facial expression recognition (FER) datasets that include this age group are mainly based on images of posed prototypical expressions with actors (Ebner, Riediger, and Lindenberger, 2010; Yang et al., 2020). The same is true for speech emotion recognition (SER) datasets and audiovisual counterparts (Wang, Zhang, and Liao, 2014; Wang et al., 2016). To the best of our knowledge, only two datasets that focus on older adults provide non-acted data: the INTERSPEECH 2020 Computational Paralinguistics Challenge (ComParE) Elderly Emotion Sub-Challenge dataset for SER (Schuller et al., 2020), which includes personal narratives; and ElderReact (Ma et al., 2019), containing monologue Youtube videos in which older adults react to specific items spontaneously, but possibly exaggerating their responses due to the nature of the videos. Very few non-acted, interaction-oriented datasets include a subset of older adults (Kossaifi et al., 2019), and we are not aware of any HMI dataset including such population.

This chapter presents a comprehensive study on computational, non-verbal discrete emotion expression recognition in interactions between older adults and a simulated VC, as a specific case of HMI scenario. The work was developed as part of the European EMPATHIC project (Torres et al., 2019; Olaso et al., 2021), which aimed to explore and validate new interaction paradigms for empathic, expressive, and advanced VCs to improve independent, healthy-life-years of this age group. As part of the project, 157 participants over 65 years old from Spain, France, and Norway were recorded interacting with an initial version of the EMPATHIC-VC in a Wizard of Oz (WoZ) paradigm, as a first stage for gathering data to develop the system and studying the interaction between older adults and VCs. Under this framework, we first describe the annotation procedure and methodological choices tailored to the project requirements. Then, by means of a DL-based approach, we investigate the contribution of different modalities for emotion expression recognition, including speech, facial expressions, eye gaze, and head dynamics, both individually and combined, in various evaluation scenarios.

This framework provides two unique features that we exploit in our work. First,

it allows us to perform a comparative analysis across cultures (where culture influences not only emotional expression but also the annotation procedure) and languages as well as multicountry versus country-specific training, two aspects that have received limited attention (Kossaifi et al., 2019). Second, the project involved a channel-specific annotation, providing two distinct sets of emotional labels, audio- and video-based (the rationale for which is explained in Section 6.3). Thus, we consider speech from audio and facial expressions from video as main modalities, while eye gaze and head movements act as additional modalities that can be extracted from video. We assess the effectiveness of the main modalities in recognizing their associated labels, and the possible performance improvement when being combined with the remaining (*auxiliary*) modalities. Additionally, we conduct a cross-channel evaluation, wherein the main modality of one label type and the auxiliary modalities are individually employed to recognize the labels derived from the other channel. This offers insight into the transferability and adaptability of modalities across label types, with which we can discover the presence of shared emotional cues. Lastly, we analyze performance differences between training and evaluating on spoken and silent instances, to understand how the presence or absence of speech affects performance for this age group.

The chapter is organized as follows. Section 6.2 reviews current trends in affective computing for emotion recognition with the modalities considered in this work, with an emphasis on gaze features and multimodal approaches. Section 6.3 describes the corpus and the annotation protocol. Section 6.4 details our computational approach. Section 6.5 describes the evaluation protocol and results of all evaluation scenarios, which are discussed in Section 6.6. Finally, Section 6.7 concludes the chapter.

6.2 Related work

In this section, we first summarize the two main models of emotion used in affective computing. Second, we review related computational approaches for the automatic recognition of emotional states from speech, facial expressions, gaze, and head cues. Finally, we discuss multimodal approaches using such cues, with an emphasis on works featuring older adults.

6.2.1 Models of emotion

Expressions of emotion are generally represented by two main different models: a categorical or discrete model, and a dimensional or continuous model. The categorical model identifies a set of discrete emotional categories, ranging from the basic Ekman emotions (Ekman, 1999) (*happy*, *surprised*, *contempt*, *sad*, *fearful*, *disgusted*, and *angry*), to a larger set with more specific and realistic affective states. Indeed, ordinary communication involves a variety of complex feelings that cannot be characterized by a reduced, fixed set of categories (Gunes and Pantic, 2010a). Therefore, such categories are usually selected considering the task at hand. For instance, categories such as *bored*, *frustrated*, *delighted*, *calm*, *satisfied*, or *excited* are more applicable to HMI scenarios than most basic states (Calvo and D'Mello, 2010).

Given the complexity of the emotional semantic space, a number of researchers (Gunes and Pantic, 2010a; Schuller et al., 2011) are more in favour of adopting a dimensional model such as the circumplex model of affect (Russell, 1980). In the dimensional model, each affective state is represented by a point in a 2D space,

where the *valence* dimension represents the polarity of the emotion, i.e., a positive or negative value along a continuum, and the *arousal* dimension represents the degree of emotional activation. i.e., values vary from low to high along a continuum. Other versions include a third dimension, *dominance*, which represents a sense of control over the situation while experiencing the emotion. The valence, arousal, and dominance (VAD) model has been widely exploited for audio/video-based emotion recognition (Valstar et al., 2014; Gunes and Pantic, 2010a; De Velasco, Justo, and Torres, 2022), allowing for the encoding of slight emotional changes over time (Valstar et al., 2014).

Both models have their own advantages and drawbacks. For instance, emotional categories may not consider intensity and exhibit fuzzy boundaries. Conversely, dimensional models introduce more subjectivity in emotion scaling across raters. Ultimately, the choice depends on the task objectives. In our case, we are interested in detecting prespecified events of interest that are expected to occur during the interaction, for which the EMPATHIC-VC system can react and adapt to, in a practical and interpretable way. Thus, the categorical model better fits the system needs.

6.2.2 Emotions from speech

The speech signal captures the speaker's communicative intention, encompassing not only the words spoken but also the intonation, prosody, pauses, and other paralinguistic elements that contribute to the message. In the same way, speech provides a lot of information about the speaker, their accent, profile, speaking style, current emotional state, and even reveals states of depression or anxiety (Huang et al., 2019; De Velasco Vázquez et al., 2023).

The most commonly used features for SER are based on low-level descriptors (LLDs), such as zero-crossing rates, pitch, formants, energy, jitter, shimmer, spectral centroids, Mel-frequency cepstral coefficients (MFCC), flux, etc., as well as on their descriptive statistics or functionals (e.g., mean, SD, quartiles) (Huang et al., 2019; Panda, Malheiro, and Paiva, 2020; De Velasco, Justo, and Torres, 2022). Some works have proposed the standardization of the feature sets. However, only GeMAPS, which contains a combination of the previously mentioned LLDs and functionals, and the feature sets proposed in the ComPaRE challenge series, which are variations of GeMAPS, have become a reference (Schuller et al., 2013; Eyben et al., 2015). Alternatively, the spectrogram has also been used as a sequence of features represented as an image, which has been demonstrated to be specifically useful for feeding CNNs (De Velasco Vázquez et al., 2023). More recently, the first framework for self-learning rich representations of speech was published, called Wav2Vec (Baevski et al., 2020), which was initially used for SER in English by Luna-Jiménez et al. (2022) and in Spanish by De Velasco, Justo, and Torres (2022). Shortly afterward, new frameworks were proposed, including Hubert (Hsu et al., 2021) used by Pastor et al. (2022), UniSpeech (Wang et al., 2021a), and WavLM (Chen et al., 2022). A comparison of self-supervised representations for SER can be found in De Velasco (2023).

Similarly to other domains, the rise of DL also caused a gradual transition from traditional classifiers to deep neural networks (DNNs) for SER (Singh and Goel, 2022; De Lope and Graña, 2023). MLPs, similar to traditional classifiers, can benefit from fixed-length inputs and are useful for classification or regression tasks in small datasets (De Velasco, 2023). Current approaches also feature CNNs (De Velasco, Justo, and Torres, 2022), RNNs (Wang et al., 2020; De Velasco, 2023), Transformers (Morais et al., 2022; Wagner et al., 2023) and, more recently, combinations

of different DNNs (Atmaja, Sasou, and Akagi, 2022; De Velasco, Justo, and Torres, 2022).

6.2.3 Emotions from facial expressions

Facial expressions are considered one of the most significant means for humans to express their emotions and intentions in their daily communication (Ekman, 1999).

FER systems can be divided into two main categories according to the type of input they rely on: static-image and dynamic-sequence (Li and Deng, 2020). In static-based methods, the feature representation is based only on the spatial information associated with a single image, whereas dynamic-based ones consider the temporal relation among contiguous frames as well as the facial deformation dynamics.

In turn, and similarly to SER, FER approaches can also be divided into conventional and DL-based approaches. The former is usually composed of three major steps: face and landmarks detection, feature extraction, and emotion classification. These conventional algorithms usually extract face-based handcrafted features such as pixel intensities (Mohammadi, Fatemizadeh, and Mahoor, 2014), local binary patterns (Shan, Gong, and McOwan, 2009), Gabor filters (Liu and Wechsler, 2002), and histograms of oriented gradients (Mavadati et al., 2013). These handcrafted features, however, often lack enough generalizability in in-the-wild settings, characterized by a considerable variation with respect to image resolution, camera view, scene lighting, background, and subject head pose, in addition to the wide variability of subject appearances and individual factors (e.g., age, ethnicity). In contrast to conventional approaches, DL-based approaches are used as a conjoint feature extraction tool and facial expression classifier, reducing the dependency on preprocessing techniques and human-expertise-based feature extraction. The same neural network approaches discussed for SER have been applied to FER with similar results; hence, we avoid repeating them here. We refer the reader to the surveys of Corneanu et al. (2016), Ko (2018), and Li and Deng (2020) for a comprehensive review of the state of the art. As an example of approach related to the one used in this work, we highlight the work of Mollahosseini, Chan, and Mahoor (2016), one of the first to demonstrate the capability of CNNs to recognize Ekman emotions by outperforming traditional methods on popular posed and spontaneous expression datasets.

6.2.4 Emotions from eye gaze and head pose

Extensive behavioral and neuroscience literature has confirmed a relationship between eye state, gaze direction, head pose, and facial expressions, on the perception of emotions and mental states (Graham and LaBar, 2012). This link is subject to the type of stimulus and situation, or personal attributes such as culture, gender, age, or personality.

For the eye region, features that have been studied or used the most in affective computing are: pupil size, blinks, gaze direction, direct/averted gaze, extracted patterns of eye movement events, and eye aperture/closure. More specifically, pupil size is related to emotional processes (e.g., arousal, excitement) (Kreibig, 2010). However, it can also be sensitive to other confounding factors if they are not taken into account, such as cognitive processes (e.g., attention, workload), illumination, and pathological or pharmacological conditions (Spector, 1990; Bradley et al., 2008). Blink rate has also been associated with affective responses and attention (Lang, Bradley, and Cuthbert, 1990; Bentivoglio et al., 1997). Gaze direction and direct/averted gaze have been shown to modulate emotion processing in humans,

particularly when facial expressions are more ambiguous (Graham and LaBar, 2007), affecting emotion category and intensity recognition (Milders et al., 2011). We refer the reader to O'Dwyer, Murray, and Flynn (2018) and Lim, Mountstephens, and Teo (2020) for a detailed review of affective computing approaches leveraging such features.

Dedicated head-mounted or desktop eye-tracking systems with high-resolution, high-frequency cameras are generally required to extract these features with high accuracy and precision, particularly pupil size, blinks, and eye movement events. However, this is impractical for many everyday scenarios or HMI settings such as the EMPATHIC-VC, where a non-obtrusive or lower-cost approach is preferred. For such scenarios, regular cameras can now be used to estimate eye gaze and approximate the location of pupil and eye landmarks by means of appearance- or model-based methods. As commented in previous chapters, appearance-based gaze estimation has improved significantly during the past decade, boosted by DL advances (Ghosh et al., 2021). Still, since these estimated gaze trajectories are noisier and sampled with a lower frequency than with special equipment, it is not possible to accurately determine specific gaze events, but landmarks can be approximated rather reliably if the eye region has enough resolution. Therefore, a number of works compute functionals from the raw or smoothed estimated gaze trajectories over a time window, or compute features (e.g., eye closure, pupil size) based on specific eye landmarks instead (Ramirez, Baltrušaitis, and Morency, 2011; Alghowinem et al., 2016; O'Dwyer, Flynn, and Murray, 2017; O'Dwyer, Murray, and Flynn, 2018; Van Huynh et al., 2019; Abdou et al., 2022). Blinks can usually be detected via dedicated appearance-based methods (Cortacero, Fischer, and Demiris, 2019), or by detecting the action unit (AU) #45 (Baltrušaitis et al., 2018).

On a related note, head rotation plays an important role in stabilizing the line of gaze to fixate on objects of interest. This eye-head coordination has been widely studied for decades (Zangemeister and Stark, 1981; Guitton and Volle, 1987). The literature also indicates a relationship between head pose dynamics and expression and perception of different emotional and mental states (El Kaliouby and Robinson, 2005; Lhommet and Marsella, 2014; Mignault and Chaudhuri, 2003; Hess, Adams, and Kleck, 2007; Busso et al., 2007), being particularly related to emotional intensity (Karg et al., 2013). Some works have relied on head pose categorizations such as head tilts, nods, and shakes, for affect recognition (El Kaliouby and Robinson, 2005; Kapoor, Burleson, and Picard, 2007; Gunes and Pantic, 2010b; Eyben et al., 2011), which usually require specific action detectors. By contrast, more recent approaches directly use temporal 3D rotational angles (yaw, pitch, and roll) to describe head motion trajectories, as well as angular displacement, velocity, acceleration, and window-based functionals computed from such trajectories (Hammal and Cohn, 2014; Hammal, Cohn, and Messinger, 2015; Adams et al., 2015; Alghowinem et al., 2016; Samanta and Guha, 2017; Li et al., 2017), dynamic features based on the discrete Fourier transform (Ding, Shi, and Deng, 2018), or clustered sequences of kinemes (Samanta and Guha, 2020). With systems based on regular cameras, head orientation can be extracted with approaches ranging from appearance-based methods to model-based 3D head registration (Khan et al., 2021).

Due to their relationship, a handful of works have combined head and gaze features together for emotion recognition (Xue et al., 2021). Although gaze and/or head features have been proven to be sufficient for specific affective categories and dimensions in some scenarios (O'Dwyer, Murray, and Flynn, 2019; Samanta and Guha, 2020), they are generally added to facial or speech modalities to provide complementary rather than redundant information.

6.2.5 Multimodal emotion recognition

With significant advancements in multimodal ML (Liang, Zadeh, and Morency, 2022), multimodal emotion recognition has gained considerable momentum lately (see Zeng et al., 2009; D'mello and Kory, 2015; Poria et al., 2017 and Rouast, Adam, and Chiong, 2019 for exhaustive surveys on the topic). By leveraging the complementary information of multiple modalities, multimodal systems can achieve higher accuracy and reliability compared to unimodal systems.

As introduced in Section 5.2.4, multimodal fusion methods are broadly classified into feature-based, decision-based, and hybrid approaches. Feature-based fusion consists in combining the features extracted from different modalities, with methods that range from naive feature concatenation to, more recently, attention-based approaches. It allows learning from crossmodal correlations; however, an alignment among modalities is required since they may have different sampling rates (if coming from different sensors) or representations (e.g., video frames versus audio segments), and the complementary information may not be time-synchronized. Furthermore, not all modalities may be available at all times. Recent emotion recognition works have tackled the temporal alignment problem with attention-based approaches (Tsai et al., 2019), or the missing data at inference problem by means of modality dropout during training (Chumachenko, Iosifidis, and Gabbouj, 2022). Instead, decision-based fusion combines the scores or predictions of unimodal models for a final multimodal prediction, thus alleviating the alignment and incomplete data problems but disregarding crossmodal correlations. Finally, hybrid approaches combine feature- and decision-based fusion. According to the literature, the best fusion type for emotion recognition is task- and dataset-dependent.

Most multimodal emotion recognition works combine at least paralinguistic and facial expression features (Wu, Lin, and Wei, 2014). The fusion of acoustic and linguistic information has also been demonstrated to improve recognition performance (Schuller, 2018; Atmaja, Sasou, and Akagi, 2022). Gaze and head cues are usually combined with other features like facial information (Cohn et al., 2004; Kapoor, Burleson, and Picard, 2007; Wu et al., 2019; Javadi and Lim, 2021) and/or speech cues (Ramirez, Baltrušaitis, and Morency, 2011; Alghowinem et al., 2016; Alhargan, Cooke, and Binjammaz, 2017; O'Dwyer, Flynn, and Murray, 2017; O'Dwyer, 2019; Abdou et al., 2022). Nonetheless, their use is less explored compared to audiovisual fusion (O'Dwyer, Murray, and Flynn, 2018). One of the few works that combine speech, facial expressions, and gaze features is that of Abdou et al. (2022), which uses GeMAPS features extracted from audio, and a subset of the gaze functionals proposed by O'Dwyer, Murray, and Flynn (2019) and facial features extracted from a pretrained CNN from video.

The ComParE challenge recently drew attention to emotion recognition for older adults, in which participants could leverage acoustic and linguistic features (Sogancioglu et al., 2020; Boateng and Kowatsch, 2020). However, the work of Ma et al. (2019) is one of the few addressing discrete emotion recognition using the modalities considered in this work for such age group. More specifically, they extract gaze and head features, facial AUs (including blink), and facial landmarks from the visual channel, and voice quality, MFCCs, and prosody features from the audio channel of the ElderReact dataset. Features are extracted per frame, and the mean and SD of each feature are computed for the entire video clip, with a single prediction per clip. In their work, the bimodal (audio-video) model is the top performer for all emotions except for *fear*, for which audio-only models proved better than video and bimodal, and for *sadness*, for which the visual model performs better. Using the same

dataset, Sreevidya, Veni, and Murthy (2022) propose to process the audio features by means of a 1D-CNN, and create a spectrogram from the raw audio signal using a pretrained CNN. They also further process the video features by means of another CNN, and use the raw, sampled video data with a third CNN. In their work, audio features worked better than the spectrogram for all emotions except for *sad*, while the raw video worked better for all except for *fear* and *happiness*. Finally, Jannat and Canavan (2021) outperformed the results of Ma et al. (2019) by a large margin using only the visual modality with a Siamese network and a contrastive loss.

6.3 EMPATHIC WoZ Corpus

In this section, we describe the subset of the EMPATHIC WoZ Corpus considered for this work, and the protocol followed for the annotation of speech from audio and facial expressions from video.

6.3.1 Data collection

The target population of the EMPATHIC project was defined as healthy older adults based on the following inclusion criteria: 1) above the age of 65 or turning 65 within the year 2019; 2) hearing and sight are good (with or without glasses/hearing aid); 3) living independently at home; and 4) read, write, and speak the testing language fluently. Recruitment[18] involved participants from Spain, France, and Norway. A total of 157 participants (105 female) were recruited and participated in the first recording sessions of the project, of which 153 are included in the corpus. Participants are distributed as follows: 78 Spanish (54 female, mean age 69.5), 44 French (28 female, mean age 73.5), and 31 Norwegian (21 female, mean age 74.8). The overall mean age was 71.8 years (SD=$\pm$6.8). All participants were properly informed and signed an informed consent prior to enrolling in the study. Hereinafter, we refer to the Spanish subset as SP, the French as FR, the Norwegian as NO, and the complete data as WHOLE (or WH).

We used the WoZ paradigm for data acquisition, which is commonly used when building technology based on natural language and other artificial intelligence-driven applications (Schlögl, Doherty, and Luz, 2015). The key principle of the WoZ method is that study participants believe they are interacting with an autonomous system, while actually the actions of the system are controlled by a human (i.e., the *wizard*). This wizard is usually situated in a different room and connected to the study setting through a remote network connection. Consequently, WoZ sessions require a minimum of two researchers, i.e., the wizard controlling the technology, and an additional supervisor dealing with all the participant-related tasks (e.g., welcoming, informed consent, questionnaires, debriefing). Both researchers received relevant training before the recordings took place.

The interaction sessions combined different questionnaires and interaction with the EMPATHIC-VC, detailed in Justo et al. (2020). The setup consisted of a computer equipped with a webcam, a microphone, and an Internet connection (see Figure 6.1). At the beginning of the session, participants chose one of five available visual representations of agents for their EMPATHIC-VC session. During the interaction, participants were alone with the VC in order to avoid bias or undesired interactions with

[18]This trial protocol was approved by appropriate Institutional Review Boards of Spain (Ethical and Scientific Research Committee of University of the Basque Country -UPV/EHU. Cod: PI2018152), France (Commission Nationale de l'Informatique et des Libertés-CNIL- Cod: 2182146) and Norway (Ethical and Scientific Research Committee of the Oslo University Hospital).

FIGURE 6.1: Setup with a participant during an interaction session.

the supervisor. Two dialogues of 5–10 min each were completed. The first served as an introduction to the system and thus did not focus on any specific issues. The second focused on the user's nutrition habits and potential goals.

6.3.2 Definition of labels

The emotional labels used for the EMPATHIC project correspond to the users' perceived expressions of emotion. The procedure for selecting such emotion categories followed a three-step data-driven approach.

First, we considered the 27 categories defined in Cowen and Keltner (2017), which are based on the self-reported emotional states elicited by around 200 short videos over a population of nearly 1,000 people. The list defines a rich semantic space of emotions, which includes categories such as *amusement* that were found to capture well the subjective emotional experience.

As a second step, we removed the categories that were highly unlikely to be encountered during the interaction between the user and the EMPATHIC-VC, and added some labels that might potentially be perceived in these interactions. We worked with the target languages simultaneously, i.e., Spanish, French, and Norwegian, to which we added Italian and German, in order to provide accurate terms to express the same feelings in different languages, considering that cultural context can be accounted for by the translation that native speakers of each language can provide relative to the *Lingua Franca* (which in our case was English). The selected 18 labels were: *relieved*, *bored*, *excited*, *calm*, *sad*, *amused*, *puzzled*, *pleased*, *interested*, *tense*, *surprised*, *concerned*, *enthusiastic*, *skeptical*, *embarrassed*, *tired*, *delighted*, and *annoyed*.

Finally, we ran a set of pilot experiments on SP. Our goal was two-fold: 1) shorten the previous list by only considering the subset of emotions perceived during the interaction; 2) assess to what extent we could match totally or partially this list to the list of basic emotions defined by Ekman, which are typically featured in visual-based discrete emotion recognition datasets. This pilot, as well as the posterior results of the annotation procedures, defined the final labels to be considered for audio and video channels, which are presented next.

6.3.3 Annotation protocol

Few works are found in the literature aimed at establishing the amount of emotional information provided through the different audio and video channels. In particular, the study of such channels separately and their combination concludes that the latter does not always yield the best perception results, as might be otherwise expected (e.g., Kossaifi et al., 2019). Previous studies have established that the emotional information provided by each channel or combination strongly depends on the specific emotion, context, and language (Esposito, 2009). For instance, there is vast evidence in the literature that, with respect to dimensional models, arousal can be better detected from the audio channel, while valence is much better estimated from the video channel (Russell, Bachorowski, and Fernández-Dols, 2003). It has also been suggested that humans, when posed with the task of decoding emotional states, selectively attend to emotional cues that align closely with their personal and cultural experiences, thus minimizing the cognitive effort required for emotional processing during the task. Consequently, when annotating perceived emotional expressions from audio and video simultaneously, raters tend to explore the most familiar channel: if rater and rated person are culturally and/or language akin, the rater tends to exploit the auditory signal, whereas when they are culturally distant, they tend to rely more on visual cues (Riviello and Esposito, 2012).

One of the salient attributes considered in the EMPATHIC project is culture and cultural differences, so it was important that the annotation be carried out separately per country by native speakers to be able to capture subtle culture-specific emotional cues. Therefore, in order to avoid the annotators' reliance toward a single channel, we decided to separate channels at the annotation level, having different annotators for each channel. This, in turn, results in a richer variety of emotional information from different perceptual channels, which can be later leveraged by the EMPATHIC-VC system. We employed instructed annotators to be able to control the whole procedure and update it if necessary. Preliminary trials showed that annotators preferred having access to the entire video or speech file instead of annotating isolated snippets due to the presence of context, which helped them make more accurate estimations of the users' emotional state.

The annotation process consisted in determining the start and end times of all events associated to given emotions categories throughout a WoZ interaction. To ensure a high inter-rater agreement, we employed a sequential annotation process. Initially, each annotator received a set of files to annotate independently. Subsequently, the within-country inter-rater agreement was calculated with an ad hoc measure based on event overlap. If the agreement score fell below a predefined threshold, annotators engaged in discussions and reannotated the files. Conversely, if the threshold was met, the annotators received the remaining set of files and continued the process of monitoring, discussing, and reannotating until the desired level of agreement was attained.

Audio annotations

These were carried out by listening to the audio signal in Transcriber[19] with nine native annotators, three per country. For the specific case of audio, the perceived emotions were labeled in terms of categorical and VAD models. The categorical labels were: *calm/tired/bored*, *pleased/amused*, *puzzled*, *sad*, and *tense*. The first two labels

[19] https://transcriber.en.softonic.com/

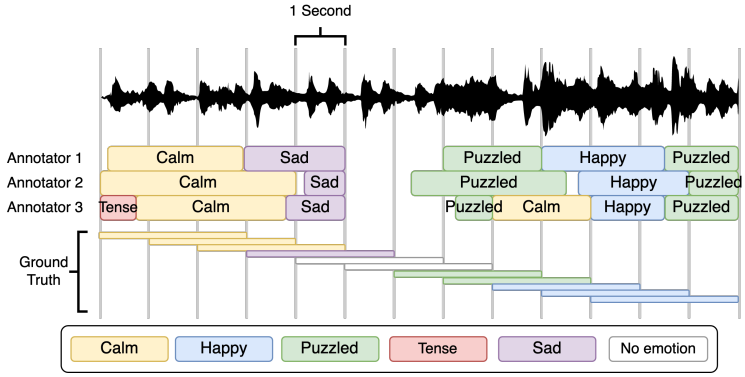


FIGURE 6.2: Representation of the segmentation of annotated emotion expression categories to create the gold standard for the audio modality. *Happy* corresponds to the *pleased/amused* category.

consist of a combination of similar categories, which was decided after the first annotation rounds, as they were highly confused among annotators. For simplicity, we henceforth refer to them as *calm* and *pleased*. Parts of the audio signal with no annotated label are not categorized. The labels assigned to the dimensional VAD model were also discretized for simplicity, and defined as: 1) *positive*, *neutral*, and *negative*, for valence; 2) *excited*, *slightly excited*, and *neutral*, for arousal; and 3) *dominant*, *neither dominant nor dominated*, and *defensive*, for dominance.

The inter-annotator agreement for the categorical annotation was computed with Cohen's Kappa for each pair of annotators at the millisecond level. SP and NO scored an average coefficient of 0.792 and 0.692, respectively, which indicates substantial agreement (McHugh, 2012), while FR scored 0.554, indicating moderate agreement. Once the entire corpus was labeled by all the annotators, we combined their annotations using 3-s segments with a 1-s stride, as depicted in Figure 6.2. To assign an emotion to each segment (i.e., the gold standard, or *ground truth*), the majority emotion was assigned if that emotion spanned a specific percentage of the whole audio segment. Otherwise, the segment was left without annotation, referred to as *discarded*. As a result, each segment has four different annotations: one categorical and three for the VAD model.

Video annotations

The annotation was carried out with an in-house software by six native annotators, two from each country. Annotators were instructed to watch muted videos, taking into account only facial expressions and head movements, thus disregarding out-of-face information such as body or hand movements. In addition, they were instructed to watch a short snippet of each user's video (up to 1 min) to familiarize themselves with the user's baseline facial expression. For video annotations specifically, a cross-country calibration was performed after the first set of files was annotated for a small, random subset of videos. This was done to ensure a common understanding of the instructions, and of the minimum intensity an expression should have to be categorized as such, which can be more objectively determined across countries than for audio-based annotations.

The categorical labels considered were: *sad*, *annoyed/angry* (henceforth referred to as *angry*), *surprised*, *happy/amused* (henceforth referred to as *happy*), *pensive*, and *other*. The first four are part of the Ekman's basic emotions (Ekman, 1999). *Pensive* is a mental state rather than an emotional expression; however, it was included in

TABLE 6.1: Number of audio segments extracted from speech emotional annotations of the EMPATHIC WoZ Corpus, per label and per country.

	Calm	Pleased	Puzzled	Sad	Tense	Discarded	Silence
Spain	38359	833	1022	151	81	4607	37910
France	19875	445	453	1	11	2819	15978
Norway	13960	474	44	0	0	1775	15764

our model as it was found to be a frequent facial expression during the conversation when users prepared their responses, as in previous HMI-oriented works (Steininger et al., 2002; El Kaliouby and Robinson, 2005). This facial expression is characterized by the eyes looking sideways and darting from side to side. Similarly to audio-based annotations, some categories were combined into a single label due to being often confused by annotators. Annotators were instructed to annotate as one of the first five categories those segments in which it was clear for them that the expression was present. The label *other* was used to denote either those segments in which an expression was occurring but which was not included in our expression list, or when more than one expression from the list was present simultaneously. Finally, all non-labeled instances were considered to be a *neutral* expression, denoting the baseline face as well as calmed, quiet, or very subtle emotional expressions which do not exceed the consensual expression thresholds.

Post-hoc inter-rater reliability was computed at frame level by means of Cohen's kappa coefficient, achieving a value of 0.7 for SP and FR, and 0.68 for NO, which indicates substantial agreement. We used the intersection between the two annotators to create the final gold standard. Frames with no intersection were discarded for automatic processing, representing around 8% of the total number of frames.

6.3.4 Analysis of labels

A thorough analysis of corpus annotations is reported in Greco et al. (2021). Here, we summarize the findings, with an emphasis on the categorical labels that will be used in our evaluation.

The number of final audio segments per emotional category is detailed in Table 6.1. As can be seen, *calm* is the most frequent emotion with around 95% of the samples, with respect to instances where the user is speaking and disregarding *discarded*, whereas *sad* and *tense* are quasi absent. Specifically for NO, users rarely showed a *puzzled* expression. With regards to the VAD model, we highlight the following differences: 1) around 30% of FR segments and only 3-4% of SP and NO segments are marked with *slightly excited* for the arousal dimension, while the rest is *neutral*; 2) SP segments are rather divided between *positive* and *neutral* valence; 3) about 25% of FR segments have *positive* valence, while for NO they are mainly *neutral*; and 4) participants in the three datasets are often *neither dominant nor dominated*.

Table 6.2 provides the distribution of emotion categories from video corresponding to spoken (top) and silence (bottom) instances separately. The reported quantities do not include the 0.3% of frames that are not matched to any audio segment, which mainly happened at the end of the video due to audio-video length mismatch. Similarly to audio annotations, video annotations lead to highly imbalanced results. *Pensive* was the most frequent manually labeled expression, appearing 11% of the time, followed by *happy*, present in 2% of the total images. Despite these findings,

TABLE 6.2: Number of frames extracted from the video emotional annotations of the EMPATHIC WoZ Corpus, per label and per country, corresponding to spoken (top) and silence instances (bottom).

	Neutral	Happy	Pensive	Surprise	Angry	Sad	Other
Spain	864112	8163	186735	115	0	0	0
	824162	3028	13427	56	0	0	0
France	484061	28118	98646	162	103	0	693
	421980	14385	12915	107	28	0	278
Norway	345876	11945	67859	72	0	0	239
	317166	6253	17303	68	0	0	415

the *neutral* category clearly dominates over all categories, appearing around 87% of the time.

As observed, the main challenges encountered in the EMPATHIC WoZ corpus are: 1) the imbalance between the different emotion classes; and 2) the imbalanced number of subjects across countries and limited data samples, particularly for audio. The former indicates that the interaction with the VC did not lead users to experience strong emotions like *sad*, *angry*, and *surprise*, and is in line with what is usually observed in real, spontaneous HMI interactions. In addition, many of the users may have an *a priory* positive attitude since they are volunteers to participate in the experiment. The reduced number of audio samples is partly caused by the amount of time that users had to wait for the WoZ to respond. The high class imbalance can be a problem for data-driven models to properly learn any discriminative information for the minority classes. Hence, for this study, we reduce the number of categories to the three most represented for each label type. That is, for audio, we maintain *calm*, *pleased*, and *puzzled*, whereas for video, we keep *neutral*, *happy*, and *pensive*.

Table 6.3 depicts the relationship among audio-video labels for the three countries, using the audio segments as reference and computing the most repeated video category for the valid frames within the start-end times of an audio segment. We find that audio-based *calm* and video-based *neutral* coincide 66-70% of the time. However, there is no evident one-to-one correspondence for the remaining cases. Given that each channel contributes distinct information, we choose to retain the two label types as independent entities, allowing the system to estimate both of them at each time step.

6.4 Methodology

In this section, we describe our methodology and training strategy for data-driven recognition of emotional states using different modalities. The methodological choices depend on the requirements of the EMPATHIC-VC system, which follow those of common multiagent systems (Jaimes and Sebe, 2007). The *human sensing* module, which includes emotion recognition, is one of the multiple system modules that must communicate timely with a *dialogue manager*. The manager controls the conversation flow by integrating the information from *human sensing* and other modules to transfer the appropriate VC reactions to the *natural language generation* and avatar animation modules (Olaso et al., 2021; Vázquez et al., 2023). In the final system, some modules would be located in remote servers, and thus data transfers would be done via network. Therefore, efficiency in the whole process is

TABLE 6.3: Contingency table for audio-video labels. Percentage computed over the total of rows and columns per country, using the audio segments as the unit of measure.

	Neutral	Happy	Pensive	Surprise	Angry	Sad	Other
			Spain				
Calm	69.75	0.3	15.26	0.01	0	0	0
Pleased	1.2	0.32	0.16	0	0	0	0
Puzzled	1.85	0.004	0.51	0.003	0	0	0
Sad	0.16	0	0.03	0	0	0	0
Tense	0.18	0	0.03	0	0	0	0
Discarded	8.43	0.14	1.64	0	0	0	0
			France				
Calm	66.83	2.85	14.2	0.02	0.02	0	0.07
Pleased	1.06	0.76	0.05	0	0	0	0
Puzzled	1.45	0.02	0.51	0	0	0	0
Sad	0.005	0	0	0	0	0	0
Tense	0.054	0	0	0	0	0	0
Discarded	9.73	0.97	1.36	0.001	0	0	0.04
			Norway				
Calm	70.33	1.14	14.88	0.02	0	0	0.048
Pleased	2.01	0.75	0.11	0	0	0	0.006
Puzzled	0.19	0	0.07	0	0	0	0
Sad	0	0	0	0	0	0	0
Tense	0	0	0	0	0	0	0
Discarded	8.66	0.91	0.87	0	0	0	0.001

crucial to ensure a seamless and natural interaction. Consequently, we prioritize independent, lightweight computational submodules for each channel, which can operate asynchronously and produce estimates at the lowest granularity level for further processing.

Figure 6.3 shows an overview of our methodological pipeline. In summary, we first extract features from the different modalities. More specifically, for the main modalities (i.e., speech from audio and facial expressions from video), we train individual models for their respective labels on all the available data to learn rich emotional features. In parallel, we extract additional features from video, namely *looking-at-VC*, head, 3D gaze, and eye movement information. Since the features of each modality are extracted at different time resolutions (i.e., audio features every 3 s, facial features at every frame, and additional features every 1.5 s), we apply fixed modality synchronization tailored to each label type, which allows us to perform the cross-channel and multimodal evaluation. Finally, the previously extracted features are combined and further evolved with an MLP to recognize the user's emotional state for the audio- and video-based labels separately.

6.4.1 Speech features from audio

In this work, we only consider those audio segments with an associated emotion while the user is speaking and for which the avatar speaks less than one-third of the segment duration.

First, we use the WavLM speech model (Chen et al., 2022)[20] to extract the acoustic

[20]WavLM model: https://github.com/microsoft/unilm/tree/master/wavlm.

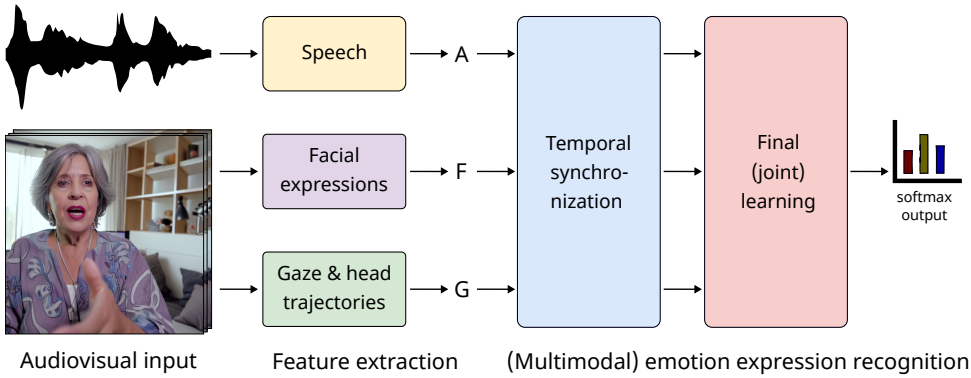


FIGURE 6.3: Overview of the methodological pipeline for emotion expression recognition. The person depicted in the image is not an actual participant of the corpus (generated with Microsoft Bing AI).

information of each segment using the raw signal waveform. WavLM was trained on 94K hours of English-spoken audios extracted from three large-scale speech datasets, and can obtain high performance in SER, among other tasks. We extract the features of its last hidden states, which output a 1024D vector every 1/50 s. This results in 150 feature vector instances per segment. We compute the average for the time dimension to reduce the final feature length to 1024.

Then, we feed such features to four two-layer MLPs: one for categorical emotional state recognition and three for the dimensional model. Although the main goal is to perform categorical recognition, we decided to include the dimensional model to leverage the available annotations and enrich the feature representation, given the relationship between the two models (Russell, 1980). We also chose to train them separately since each task may converge at a different rate; thus, a multitask learning approach would not be optimal for all the outputs. The first layer reduces the 1024 extracted features to 64 with ReLU as the activation function, while the second one is in charge of extracting the logits for the prediction of the emotional states via softmax. Cross-entropy is used as a loss function, and the Adam optimizer is used to train the four networks with a learning rate of 0.001 over 5K iterations. To deal with the imbalance of the data, the sampling probability for the samples of the minority classes is four times higher than that for *calm*.

Finally, the prediction scores of the four models are concatenated to the computed WavLM features in a hybrid fusion fashion, resulting in a 1034D feature vector. This way, we preserve the generic speech representation and augment it with a reduced set of domain-specific information. We refer to this feature set, and consequently to this modality, as A.

6.4.2 Facial expression features from video

For this work, we adopt a static-based approach. For each video frame, we first detect faces using FaceBoxes (Zhang et al., 2017a)[21] and estimate 68 facial landmarks in the image space by means of 3DDFA_v2 (Guo et al., 2020)[22]. Less than 1% of the data is lost in these steps. Using these landmarks, the face is rotated, scaled, and cropped to obtain a normalized RGB image of 224x224 pixels. Then, we use the

[21] FaceBoxes model: https://github.com/sfzhang15/FaceBoxes.
[22] 3DDFA_v2 model: https://github.com/cleardusk/3DDFA_V2.

Xception CNN model (Chollet, 2017) pretrained on ImageNet (Deng et al., 2009)[23] to extract discriminative features from the face images, and add four FC layers to the top of the network, each followed by ReLU and dropout (with a rate of 0.5), in addition to a final softmax layer for FER. During optimization, we found that the best strategy was to freeze the first 70 layers of Xception and finetuning the last 10. Consequently, we finetune such layers and train the added ones from scratch on both spoken and silent instances. According to this transfer learning scheme, we get a total of 23.6M parameters, where 16.5M are trainable, and the remaining 7M are fixed (non-trainable). Training is based on the Adam optimizer, with a learning rate of 0.001. To tackle the class imbalance issue, we use a weighted cross-entropy loss function where the weight of each emotion class is associated with the inverse frequency in the training set.

Finally, we extract the output features from the last hidden layer, resulting in a 256D vector. We refer to this feature set and modality as F.

6.4.3 Additional features from video (gaze and head pose)

We also use the video stream to compute a series of additional features based on per-frame estimated gaze vectors and head poses. In this work, gaze estimation consists in identifying the line of gaze in the 3D space with respect to the CCS, where the line of gaze is a single 3D gaze direction vector, the origin of which is the center of the head (due to the dataset used, see below). Similarly, head pose estimation consists in identifying the 3D pose of the head (i.e., yaw, pitch, and roll) with respect to the CCS. Due to the lack of camera calibration, we used default camera parameters with the same focal length and zero distortion for all setups.

First, we leverage the preprocessing for facial expressions to detect faces and landmarks. We then fit a face 3DMM (Huber et al., 2016)[24] to the detected 2D landmarks and apply PnP (Lepetit, Moreno-Noguer, and Fua, 2009)[25] to estimate the 3D head position and orientation. The 3D head pose is used to normalize the face image (see Section 3.3.2), which is then used as input for the gaze estimation model. Although none of the existing gaze estimation datasets features older adults as the target age group, it is important that our gaze estimation model is trained on a dataset with wide subject variability to maximize generalization to our target population. Currently, the largest-scale dataset is ETH-XGaze (Zhang et al., 2020), featuring 110 participants (63 male, aged between 19 and 41 years old) with different ethnicities, age, gender, and accessories such as eyeglasses, in addition to including a wide range of head poses and gaze directions. However, this dataset does not provide video sequences; thus, it does not enable the use of spatiotemporal gaze estimation models like the one presented in Chapter 3. For this reason, we opt to use the static ETH-XGaze baseline method instead, which is based on ResNet-50 (He et al., 2016)[26]. A visual examination of the estimated gaze direction using this model showed that performance was primarily impacted for users wearing colored lenses or eyeglasses with substantial reflection caused by the computer screen, which hinders proper perception of the eye regardless of the method and dataset used. We did not detect blinks or pupil size due to their low reliability in our scenario. Nonetheless, blinks are implicitly included in the estimated gaze trajectories, as when there

[23] Xception model: https://keras.io/api/applications/xception/.
[24] 3DMM model: https://github.com/patrikhuber/eos.
[25] PnP implementation: https://docs.opencv.org/4.x/d5/d1f/calib3d_solvePnP.html.
[26] ETH-XGaze dataset and model: https://ait.ethz.ch/xgaze.

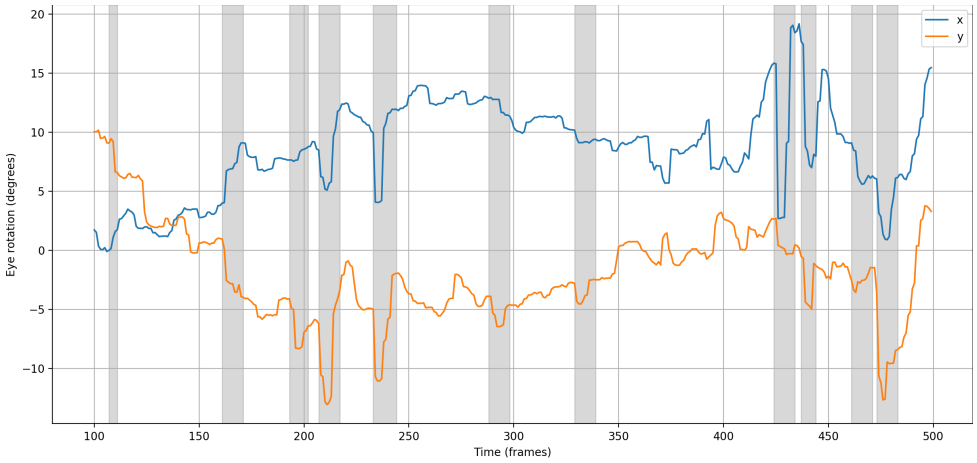


FIGURE 6.4: Estimated eye-in-head rotation traces for a snippet of an SP recording, postprocessed with a median filter of window size 5. 'x' and 'y' refer to the horizontal and vertical movement, respectively, in degrees of visual angle. Shadowed areas depict blinking events.

is a blinking event, the estimated gaze direction follows a unique fast, downward-upward trajectory (see Figure 6.4).

To reduce noise, the estimated head pose and eye gaze trajectories are post-processed with a median filter of window size 5 frames. Combining head pose $\mathbf{h} = (y, p, r)$ and eye gaze $\mathbf{g} = (x_g, y_g)$ vectors (converted from 3D direction vector to 2D angles following Equation 2.2), we further convert the line of gaze into eye-in-head gaze angles, that is, mimicking eye rotation in the HCS: $\mathbf{e} = (x_e, y_e)$. Figure 6.4 depicts an example of extracted eye-in-head gaze angle trajectories. The three data sources are filtered to discard invalid data, including frames with incorrectly detected faces, or for which head or eye movements are not anatomically plausible, that is: eye rotation larger than 40° (Shin et al., 2016), or faster than 860°/s between consecutive frames, which is the highest peak angular speed that saccades have been reported to reach (Bahill, Clark, and Stark, 1975); and head movements faster than 700°/s between consecutive frames, based on existing research on maximum rotation speed for voluntary motion (Grossman et al., 1988). Following other works on eye and head pose data processing (Bulling et al., 2010; Samanta and Guha, 2020), gaze, eye, and head trajectories are processed using a sliding window of 1.5 s and stride of 1, and centered at the half of every second throughout the video. Frames with invalid data are not taken into account when performing the aggregation. Windows smaller than 0.5 s or for which more than 50% of the frames are invalid are discarded, which represents around 2% of the windows.

For each window, a vector of 227 features is extracted, containing information from the three sources of information represented as functionals of the trajectories, selected based on the literature (Holland and Komogortsev, 2011; Baranes, Oudeyer, and Gottlieb, 2015; Hoppe et al., 2018), and a complementary attention measure as fourth source. In addition, due to the effect of the glasses on the resulting eye trajectories, we add a manually annotated ternary flag as fifth source to denote whether the participant is wearing glasses, and, if so, whether the eyes are clearly visible. The resulting 228D feature set is referred to as G. We evaluate all sources together and separately to assess their effect on performance, except the glasses one, which is always used in combination with other sources. The individual feature subsets are detailed below.

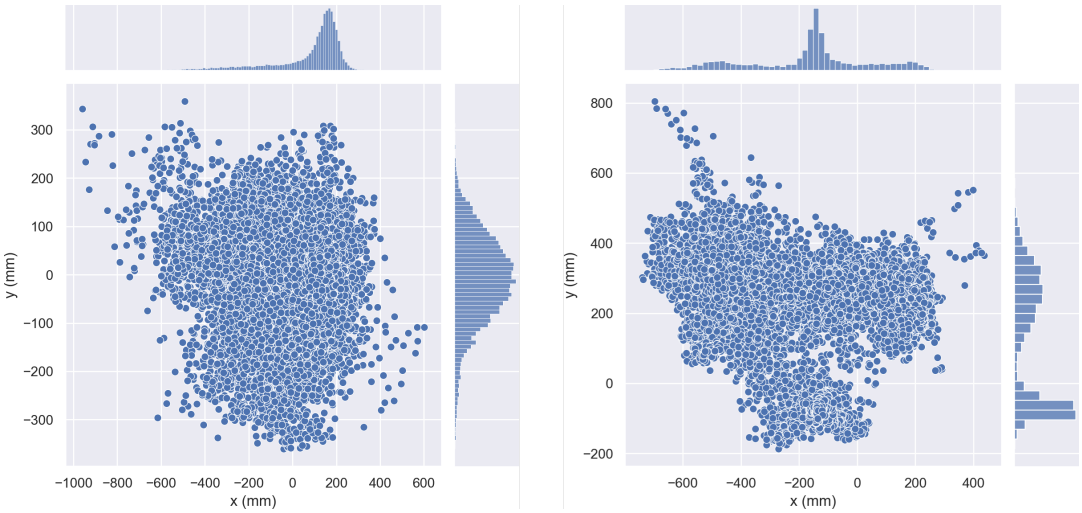


FIGURE 6.5: 2D distribution of gaze points, computed as the intersection of the per-frame 3D gaze direction vectors in the CCS with the 3D plane of the camera, as a proxy of the screen plane. We assume that the EMPATHIC-VC is located at the highest density zone. As can be seen, such zone varies among 2 example videos (from the SP set): for the graph on the left, the highest density zone is located around (200, 0), while for the graph on the right, it is around (-150, -100). Plot ranges differ.

Looking at Virtual Coach

The EMPATHIC-VC system can use this feature as an overt measure of attention (i.e., if the user's gaze is oriented toward the VC) to estimate whether the user is engaged with the VC. As the relative camera-screen position is unknown, some assumptions are made to estimate this feature. The first assumption is to consider that the camera was near the screen and roughly centered on the screen's horizontal axis. As the setup changed within and across countries, it is not possible to assume a single configuration for all the videos. Therefore, we compute the 3D plane of the camera as a proxy of the 3D plane of the screen, and intersect the 3D gaze vectors in the CCS onto that plane per video. This produces a single distribution of 2D gaze points per video (see Figure 6.5), which does not necessarily follow a Gaussian, and for which it is assumed that the zone with the highest density of points represents the VC location. That is, the second assumption considers that participants were looking at the VC most of the time. This assumption holds for most of the videos. However, for some videos for which the participant tends to avert their gaze, the highest density of points is located in other parts of the scene. As defining a per-video zone would require manual inspection of all videos, we decided to continue with an automatic approach that would work in most cases. For future developments, if prior setup calibration is not feasible, it would suffice with a first verification step (i.e., similar to a user calibration stage) in which the participant would be asked to look at the VC for a small amount of time (e.g., 1-2 seconds) at the beginning of an interaction session. That would be used as a future reference to infer where the VC is located in the 3D space.

To find the zone with the highest density, first filter points that fall outside $\pm$300 cm of the center of the plane. Then, we find gaze point clusters near the center of the plane using the Mean Shift clustering algorithm (Comaniciu and Meer, 2002)[27] with

[27] Mean Shift implementation: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html.

TABLE 6.4: Functionals computed for each element of the additional modalities (3D gaze vector, eye rotation, and head pose). $x.$ and $y.$ correspond to horizontal and vertical 3D gaze/eye rotation components, respectively. y, p, and r correspond to head yaw, pitch, and roll, respectively.

Element	Functionals
$x.$, $y.$, abs $\Delta x.$, abs $\Delta y.$, abs $\Delta x./t$, abs $\Delta y./t$, abs $\Delta\mathbf{g}/t$, abs $\Delta\mathbf{e}/t$, y, p, r, abs Δy, abs Δp, abs Δr	min, max, mean, SD, range (except for abs $\Delta x.$ and abs $\Delta y.$), 25th perc, 50th perc, 75th perc, IQR
$\Delta x.$, $\Delta y.$, $\Delta\mathbf{g}$, $\Delta\mathbf{e}$, Δy, Δp, Δr, abs $\Delta y/t$, abs $\Delta p/t$, abs $\Delta r/t$	mean, SD

SD: standard deviation; perc: percentile; IQR: interquantile range.

a bandwidth value estimated per video, and select the cluster with the highest number of points. To account for possible noisy estimates of the line of gaze, head pose, and VC's position, we assign weights to each gaze point based on its Mahalanobis distance to the cluster's distribution weighted by the cluster's inverse covariance. The weights are assigned to each gaze point $\mathbf{p}_i$ such that:

$$w(\mathbf{p}_i) = \begin{cases} 1, & \text{if } d(\mathbf{p}_i, \mathbf{c}) \leq thr_1 \\ (1 - d(\mathbf{p}_i, \mathbf{c}))/thr_2, & \text{if } thr_1 < d(\mathbf{p}_i, \mathbf{c}) \leq thr_2 \\ 0, & \text{otherwise,} \end{cases} \tag{6.1}$$

where $d(\mathbf{p}_i, \mathbf{c})$ is the Mahalanobis distance between the gaze point $\mathbf{p}_i$ and the cluster $\mathbf{c}$, and the thresholds are set to $thr_1 = 1$ and $thr_2 = 4$ SDs. For the second case of the piecewise function, points that belong to the cluster are transformed to be in the range $[0.7, 1)$, whereas points that do not belong to it are transformed to be in the range $(0, 0.7)$. These values were found empirically.

Per-point weights are further binned and converted into a 6D one-hot encoding vector denoting the likelihood of looking at VC from lower to higher. Per-valid-frame vectors are averaged over a time window, producing a 6D feature vector per window. Combined with the glasses flag, this feature set is referred to as G_{vc}.

3D gaze direction

For the line of gaze $\mathbf{g}$ (gaze direction in the CCS), we compute functionals of the following elements, which are described in Table 6.4: per-component (i.e., x_g, y_g) gaze angles, per-component angle differences (e.g., Δx_g) and their magnitude (e.g., abs Δx_g) between any two consecutive frames, direction ($\Delta\mathbf{g}$) and speed ($\Delta\mathbf{g}/t$) of the gaze vector between any two consecutive frames, and per-component speed (e.g., $\Delta x_g/t$) between any two consecutive frames. This results in a 67D feature vector. Combined with the glasses flag, this feature set is referred to as G_{3dg}.

Eye rotation

We compute the same statistics for the eye-in-head rotation $\mathbf{e}$ (gaze direction in the HCS) as for the 3D gaze vector, resulting in a 67D feature vector. Combined with the

TABLE 6.5: Number of audio segments from the EMPATHIC WoZ Corpus used for the evaluation of the final models, per label and per country.

	Calm	Pleased	Puzzled
Spain	33089	669	898
France	16534	316	369
Norway	12392	353	29

glasses flag, this feature set is referred to as G_{eye}.

Head rotation

Finally, we compute functionals (see Table 6.4) for the following: per-component (i.e., y, p, r) head pose angle, per-component angle differences (e.g., Δy) and their magnitude (e.g., abs Δy) between any two consecutive frames, and per-component speed (e.g., $\Delta y/t$) with respect to any two consecutive frames. This results in an 87D feature vector. Combined with the glasses flag, this feature set is referred to as G_h.

6.4.4 Temporal synchronization of modalities

In order to effectively integrate and analyze the multimodal data captured from different sources, we employ a fixed modality synchronization approach per label type.

For the audio-based evaluation, for which the system would output an estimate every 3 s, we compute the average and SD of the available per-frame F features within an audio segment, resulting in a 512D vector. This provides a robust facial expression descriptor that is less susceptible to accidental fluctuations despite disregarding facial temporal dynamics. Preliminary experiments evaluated a second version, consisting in concatenating the features of the most central frame of each second of the audio segment, hence maintaining such dynamics. However, the former version outperformed the latter for the majority of settings. Regarding G, we use the window aligned to the center of the audio segment, thus discarding those windows at the extremes of the segment.

Conversely, for video-based evaluation, the temporal resolution is increased to frame level. Thus, each G window and A segment are used multiple times and matched to different frames. In particular, we associate each frame with a specific G window and A segment based on its closest proximity to the central timestamp of the respective window and segment.

With this matching, all F frames, G windows, and A segments have an audio- and video-based label assigned. For each evaluation case, feature sets that do not have correspondence due to missing data of any of the modalities are omitted. The final amount of data is around 86% of the original data for audio, and 98% for video. Table 6.5 includes the number of data samples per class and country used for the final evaluated models with audio-based labels. Similarly, Table 6.6 includes the number of data samples per class and country used for the final evaluated models with video-based labels when training with speech instances (left), and silence instances (right). Compared to Tables 6.1 and 6.2, the class ratios with respect to the original sample size are generally maintained.

TABLE 6.6: Number of video frames from the EMPATHIC WoZ Corpus used for the evaluation of the final models, per label and per country, corresponding to spoken (left) and silent (right) instances.

	Neutral		**Happy**		**Pensive**	
Spain	858976	815634	8163	3028	184693	13233
France	471343	410344	27709	14213	93378	12225
Norway	321622	299338	10930	5031	67313	17291

6.4.5 Final models

The extracted features from a given modality are normalized according to the range of the training set and fed to a 2-layer MLP with ReLU activation and dropout of 0.5, followed by a dense layer with softmax for classification of a given label type. We evaluate three low-complexity MLP configurations, from lowest to highest: 1) 100 hidden units for the first MLP layer, and 20 for the second, referred to as *low* (L); 2) 200 and 40, referred to as *mid* (M); and 3) 500 and 100, referred to as *high* (H). For multimodal evaluation, the feature sets of the different modalities are concatenated before being fed to the MLP. We evaluated other attention-based fusion approaches in preliminary experiments, such as self- and crossmodal attention (Rajan, Brutti, and Cavallaro, 2022). However, their performance was equivalent to concatenation, so we proceed with the latter for the experimental evaluation.

We tackle data imbalance by randomly sampling instances of each class with the same probability. Additionally, due to the small sample size of the audio-based evaluation, we employ an oversampling strategy such that each sample of the minority class (*pleased* for SP and FR, and *puzzled* for NO and WH) is utilized around three times per epoch. To maintain an approximate balance between classes, the other classes are sampled a similar number of times. The training samples per epoch are thus set to 5418 samples for SP, 2556 for FR, 234 for NO, and 10494 for WH. Conversely, since the sample size for video-based evaluation is considerably larger but also contains higher redundancy, we set the training sample size to 7500 for all countries. Samples are randomly selected; thus, at the end of the training stage, all samples from the minority classes are seen multiple times, while for *neutral*, only a fraction is seen.

All evaluated models are trained with cross-entropy loss, Adam optimizer, learning rate of 0.0001, and batch size of 64. We empirically set the number of training epochs to 100 for all countries and evaluations except for NO with audio-based labels, for which we train for 200 epochs.

6.5 Experimental evaluation

In this section, we present a comprehensive experimental evaluation to assess the impact of different modalities on the recognition performance of emotional states for audio and video labels.

6.5.1 Research questions

The characteristics of the EMPATHIC-VC scenario allow us to evaluate the contribution of the different modalities for the considered emotions in various contexts. First, we separately consider the evaluation scenario with audio-based labels and that with video-based labels. We have a main modality for each label type: A for

audio-based and F for video-based labels. We refer to the remaining modalities (e.g., F and G for audio-based evaluation) as auxiliaries for that evaluation scenario. Main and auxiliary modalities can be combined to improve performance. Each evaluation is performed in each country individually (SP, FR, and NO) and on WH. The latter allows us to evaluate trends of the complete set of data and quantify the effect of training with country-specific data in comparison to a larger multicountry set. The audio-based scenario only includes data where the user is speaking. By contrast, for the video-based scenario, we can compare the performance of evaluating spoken content with that of silent content. Furthermore, as for the country-oriented evaluation, we can assess the effect of training the final video-based model with speaking-status-specific data in comparison to with all data.

On this basis, our aim is to answer the following research questions in the context of our scenario:

Q_1. Can the main modality for a given label type obtain the same discriminative power for all the classes considered?;

Q_2. Can the auxiliary modalities achieve similar performance to the main modality?;

Q_3. Is multimodality beneficial?;

Q_4. Which subset of G features contributes more toward the task?;

Q_5. Are there noteworthy differences in performance among countries?;

Q_6. Does training with data from multiple countries prove beneficial with respect to country-specific training?;

Q_7. For video-based evaluation, does training with spoken and silent instances prove beneficial with respect to spoken/silent-specific training?;

Q_8. Are there any performance differences between spoken and silence instances?;

Q_9. Are there any performance differences between audio- and video-based evaluation?

6.5.2 Evaluation protocol

We build 10-fold subject-independent training and test splits for each country subset (SP, FR, NO) and a fourth with the data of all countries (WH) following approximately a 9:1 ratio. The folds for WH contain the same subjects as the per-country folds. Architecture selection (i.e., the number of MLP hidden units) and hyperparameter tuning are carried out independently per experiment based on random validation subpartitions of the training splits. Each experiment is characterized by the specific modalities, speaking status, and country sets used for training, as well as the label type. For each experiment, the best configuration over all folds is selected, and then the final models are trained using the complete per-fold training split. Tables 6.7, 6.8 and 6.9 report the best configuration for each model of audio, video-under-speech, and video-under-silence evaluations, respectively. We perform 10-fold cross-validation three times following the same splits for all models to account for the stochasticity of the data sampling and whole learning process.

Performance is measured per fold by means of the *unweighted average accuracy*, also known as unweighted average recall, which gives the same weight to the accuracy of each class regardless of the number of samples for each class. Per-class

TABLE 6.7: Complexity of the best MLP configuration for each evaluated audio-based model. L: 100 and 20 hidden units per layer. M: 200 and 40. H: 500 and 100.

Modality	Spain	France	Norway	Whole
A	L	L	H	H
F	M	L	H	L
G	H	H	H	L
A+F	M	L	H	M
A+G	M	L	H	L
A+F+G	H	L	H	M
A+G_{vc}	H	M	H	H
A+G_{3dg}	M	L	H	M
A+G_{3eye}	H	M	H	M
A+G_{h}	H	L	H	H
A+F+G_{vc}	L	L	H	H
A+F+G_{3dg}	L	L	M	L
A+F+G_{3eye}	L	L	H	M
A+F+G_{h}	L	L	M	L

accuracy is thus equivalent to per-class recall (i.e., the number of samples predicted correctly out of the total number of samples for a given class). Note that the test splits of some folds do not contain all classes, especially *puzzled* for NO. In such cases, the average accuracy is computed only for the classes that have at least one sample in the test split. We also perform multiple pairwise comparisons with the corrected repeated k-fold cross-validation t-test (Bouckaert and Frank, 2004) to test for statistically significant differences ($p<.05$) among average accuracy results. We control for the false discovery rate using the BKY correction (Benjamini, Krieger, and Yekutieli, 2006)[28], grouped by country subset.

Country sets used for training and testing are denoted as *training country→testing country* (e.g., WH→SP to denote training with WH and testing with SP). The WH models selected for the country-specific comparison are the ones that worked better for the WH validation sets, so the reported performance would be different if the best models were selected for each country independently. Likewise, models trained on WH with silence and speech data are also those that worked better for the WH validation sets.

6.5.3 Audio-based emotion expression recognition results

Tables 6.10, 6.11, 6.12, and 6.13 show the results for the different experiments with audio-based labels on the WH, SP, FR, and NO data subsets, respectively. We report below the results with respect to each research question.

Main modality

For WH, A alone obtains a higher performance for *calm*, followed by *pleased* and then *puzzled*, correlated with the number of samples per class. Furthermore, *puzzled* gets more confused with *calm* than *pleased*. Country-wise, we see similar trends, except that, for SP, *puzzled* obtains higher accuracy than *pleased*. Additionally, for FR, the

[28]BKY correction implementation: https://www.statsmodels.org/dev/generated/statsmodels.stats.multitest.multipletests.html.

TABLE 6.8: Complexity of the best MLP configuration for each model evaluated on the video-under-speech scenario. L: 100 and 20 hidden units per layer. M: 200 and 40. H: 500 and 100.

Modality	Spain	France	Norway	Whole
Training on speech data:				
F	H	L	L	H
A	H	H	M	H
G	H	H	M	H
F+A	L	L	L	L
F+G	H	L	M	H
F+A+G	M	L	L	L
F+G_{vc}	H	L	M	L
F+G_{3dg}	H	L	L	H
F+G_{eye}	H	L	L	H
F+G_{h}	H	L	H	H
F+A+G_{vc}	M	L	M	M
F+A+G_{3dg}	H	L	M	H
F+A+G_{eye}	L	L	M	M
F+A+G_{h}	L	L	M	H
Training on all data (speech+silence):				
F	H	L	L	H
G	H	M	M	H
F+G	H	L	M	H
F+G_{vc}	H	L	M	M
F+G_{3dg}	H	L	M	H
F+G_{eye}	H	L	M	H
F+G_{h}	H	L	L	H

accuracy for *pleased* is less stable than for *puzzled*. For NO, there is a substantial difference in performance across classes due to data imbalance. More concretely, only 0.23% of the NO dataset belongs to *puzzled* instances, and some folds do not contain any test instance of this class, making this subset harder to evaluate. For this country, *pleased* is more often confused with *calm* than for the other countries.

In general, the results are more stable across runs than across folds. The mean SD across folds for WH is 3.3%, while the mean SD across runs is around 0.8%. Thus, as a general note, changes in the standard error of the mean (SEM) mainly denote higher variability across folds.

Auxiliary modalities

For WH, despite the high accuracy obtained by F for *pleased* and *puzzled*, confusion patterns reveal that *calm* is mostly confused with *puzzled*, which does not occur for *pleased*. This implies that F provides information that is particularly discriminative and potentially correlated with the audio modality specifically for the latter class. By further analyzing the F predictions and their correlation to facial expression categories, we observe that 97% of the *pleased* predictions correspond to audio segments where the majority facial expression is *happy*, despite only 30% of the audio segments matching to a *happy* expression having the *pleased* annotation. Thus, the same

TABLE 6.9: Complexity of the best MLP configuration for each model evaluated on the video-under-silence scenario. L: 100 and 20 hidden units per layer. M: 200 and 40. H: 500 and 100.

Modality	Spain	France	Norway	Whole
Training on silence data:				
F	H	M	L	H
G	M	L	L	H
F+G	L	L	L	L
F+G_{vc}	M	L	L	M
F+G_{3dg}	H	H	L	H
F+G_{eye}	H	M	L	H
F+G_{h}	L	M	L	M
Training on all data (speech+silence):				
F	H	L	M	L
G	L	L	L	H
F+G	H	L	L	H
F+G_{vc}	M	L	M	L
F+G_{3dg}	H	H	L	H
F+G_{eye}	H	H	H	H
F+G_{h}	L	L	H	H

TABLE 6.10: Audio-based results on WHOLE, reported as unweighted average accuracy ± SEM over 10 folds and three runs per fold. Bold: best accuracy per group. Underlined: best accuracy overall.

Modality	Calm Accuracy	Pleased Accuracy	Puzzled Accuracy	Average Accuracy
A	76.42 ± 0.8	63.54 ± 1.3	59.89 ± 2.3	66.61 ± 0.6
F	15.9 ± 1.3	**69.91 ± 3.2**	**63.41 ± 3.0**	**49.74 ± 0.9**
G	**25.26 ± 1.6**	49.74 ± 2.3	44.66 ± 2.0	39.89 ± 0.9
A+F	75.68 ± 0.5	**67.59 ± 1.6**	**61.65 ± 2.2**	**68.31 ± 0.5**
A+G	76.62 ± 0.7	62.38 ± 1.4	58.52 ± 2.3	65.84 ± 0.6
A+F+G	**76.93 ± 0.7**	67.11 ± 1.7	60.41 ± 2.7	68.15 ± 0.6
A+G_{vc}	77.07 ± 0.6	61.84 ± 1.1	59.39 ± 2.2	66.1 ± 0.6
A+G_{3dg}	76.18 ± 0.7	61.83 ± 1.2	**60.48 ± 2.3**	66.16 ± 0.5
A+G_{eye}	76.34 ± 0.6	61.98 ± 1.4	60.19 ± 2.1	66.17 ± 0.6
A+G_{h}	**77.74 ± 0.8**	**63.02 ± 1.3**	58.61 ± 2.4	**66.46 ± 0.7**
A+F+G_{vc}	76.77 ± 0.6	**68.7 ± 1.5**	59.13 ± 2.4	**68.2 ± 0.6**
A+F+G_{3dg}	76.38 ± 0.7	66.62 ± 1.8	60.22 ± 2.4	67.74 ± 0.6
A+F+G_{eye}	75.5 ± 0.5	67.17 ± 1.6	**61.66 ± 2.2**	68.11 ± 0.5
A+F+G_{h}	**77.04 ± 0.6**	67.26 ± 1.6	59.93 ± 2.4	68.08 ± 0.5

features that correspond to the *happy* facial expression are related to the facial features corresponding to a *pleased* speech. On the contrary, G's results are slightly over random performance, indicating that G alone is not informative enough to recognize the classes considered.

Trends are overall maintained country-wise, except that, for SP, *calm* benefits

TABLE 6.11: Emotion recognition results for audio-based labels trained on the SPAIN (left) and WHOLE (right) training subsets and evaluated on the SPAIN test subset (train→test), reported as unweighted average accuracy ± SEM over 10 folds and three runs per fold. Bold: best accuracy per group. Underlined: best accuracy overall.

	SPAIN → SPAIN				WHOLE → SPAIN			
Modality	**Calm Accuracy**	**Pleased Accuracy**	**Puzzled Accuracy**	**Average Accuracy**	**Calm Accuracy**	**Pleased Accuracy**	**Puzzled Accuracy**	**Average Accuracy**
A	73.57 ± 1.8	58.17 ± 1.9	60.74 ± 2.4	64.16 ± 0.6	74.06 ± 1.3	60.48 ± 2.4	63.91 ± 3.0	66.15 ± 0.9
F	**52.61 ± 3.3**	**59.61 ± 4.7**	23.88 ± 3.6	**45.37 ± 1.7**	19.42 ± 2.0	**57.42 ± 5.2**	**66.76 ± 3.4**	**47.87 ± 1.3**
G	43.45 ± 2.5	29.95 ± 3.6	**35.67 ± 2.7**	36.36 ± 1.0	**27.02 ± 1.9**	34.36 ± 3.8	48.15 ± 2.5	36.51 ± 1.2
A+F	72.12 ± 2.1	**61.27 ± 2.7**	**64.56 ± 2.4**	65.98 ± 0.7	73.25 ± 1.1	**63.16 ± 2.6**	**65.35 ± 2.6**	67.25 ± 0.8
A+G	**74.98 ± 1.7**	53.4 ± 2.0	61.03 ± 2.1	63.14 ± 0.6	74.63 ± 1.2	57.43 ± 2.2	62.18 ± 3.0	64.75 ± 0.9
A+F+G	73.49 ± 1.3	60.57 ± 3.0	64.33 ± 2.2	**66.13 ± 0.7**	**75.44 ± 1.2**	62.45 ± 2.8	64.04 ± 3.2	**67.31 ± 1.0**
A+G_{vc}	**73.86 ± 1.5**	55.54 ± 2.0	62.87 ± 2.3	64.09 ± 0.7	75.18 ± 1.1	59.2 ± 2.3	63.68 ± 3.1	**66.02 ± 1.0**
A+G_{3dg}	73.6 ± 1.5	55.3 ± 2.0	61.93 ± 2.3	63.61 ± 0.6	73.66 ± 1.3	57.66 ± 2.2	**64.75 ± 3.0**	65.36 ± 0.8
A+G_{eye}	72.91 ± 1.5	55.07 ± 2.0	62.01 ± 2.8	63.33 ± 0.6	73.94 ± 1.1	57.99 ± 2.3	64.41 ± 2.9	65.45 ± 0.9
A+G_h	73.37 ± 1.6	**57.7 ± 2.0**	**63.44 ± 2.5**	**64.84 ± 0.6**	**75.89 ± 1.4**	**59.23 ± 2.4**	61.94 ± 3.1	65.69 ± 0.9
A+F+G_{vc}	**72.37 ± 2.1**	64.47 ± 2.7	61.14 ± 2.6	65.99 ± 0.7	**75.52 ± 1.0**	**64.18 ± 2.8**	62.72 ± 3.2	**67.47 ± 1.0**
A+F+G_{3dg}	72.23 ± 1.8	63.21 ± 2.9	**61.75 ± 2.5**	65.73 ± 0.7	74.25 ± 1.1	62.14 ± 2.7	64.08 ± 3.0	66.82 ± 0.8
A+F+G_{eye}	72.35 ± 1.9	**65.53 ± 3.0**	59.26 ± 2.6	65.71 ± 0.7	73.67 ± 1.0	62.77 ± 2.7	**65.36 ± 3.0**	67.27 ± 0.8
A+F+G_h	72.23 ± 2.0	64.82 ± 2.9	61.67 ± 2.0	**66.24 ± 0.7**	74.88 ± 1.1	62.41 ± 2.5	63.78 ± 3.0	67.02 ± 0.8

TABLE 6.12: Emotion recognition results for audio-based labels trained on the FRANCE (left) and WHOLE (right) training subsets and evaluated on the FRANCE test subset (train→test), reported as unweighted average accuracy ± SEM over 10 folds and three runs per fold. Bold: best accuracy per group. Underlined: best accuracy overall.

	FRANCE → FRANCE				WHOLE → FRANCE			
Modality	**Calm Accuracy**	**Pleased Accuracy**	**Puzzled Accuracy**	**Average Accuracy**	**Calm Accuracy**	**Pleased Accuracy**	**Puzzled Accuracy**	**Average Accuracy**
A	72.34 ± 1.8	64.04 ± 3.9	62.45 ± 2.1	66.28 ± 1.2	76.62 ± 1.7	53.76 ± 4.7	60.61 ± 2.4	63.66 ± 1.5
F	29.13 ± 2.1	**44.7 ± 4.1**	**63.49 ± 2.8**	**45.77 ± 1.0**	12.06 ± 0.9	**64.96 ± 4.5**	**58.84 ± 3.1**	**45.29 ± 1.3**
G	**51.85 ± 2.7**	31.56 ± 4.9	43.64 ± 3.4	42.35 ± 1.3	**22.87 ± 2.0**	51.97 ± 4.8	44.16 ± 2.4	39.67 ± 1.5
A+F	73.5 ± 2.3	59.92 ± 4.6	**61.97 ± 2.5**	65.13 ± 1.5	76.37 ± 1.6	50.62 ± 5.0	**63.11 ± 2.4**	**63.37 ± 1.5**
A+G	73.38 ± 2.4	**62.77 ± 4.1**	59.3 ± 2.1	**65.15 ± 1.2**	76.37 ± 1.8	**53.58 ± 5.0**	59.64 ± 2.4	63.2 ± 1.7
A+F+G	**74.08 ± 2.4**	55.34 ± 4.5	61.06 ± 2.5	63.49 ± 1.5	**76.4 ± 1.6**	51.0 ± 5.1	62.5 ± 2.5	63.3 ± 1.5
A+G_{vc}	**74.01 ± 2.2**	60.18 ± 4.5	60.81 ± 2.1	65.0 ± 1.3	76.76 ± 1.9	52.69 ± 4.4	60.53 ± 2.2	63.33 ± 1.4
A+G_{3dg}	71.76 ± 2.3	63.57 ± 4.6	60.13 ± 2.1	65.15 ± 1.4	76.56 ± 1.6	53.51 ± 4.7	**62.14 ± 2.1**	64.07 ± 1.4
A+G_{eye}	73.66 ± 2.7	61.71 ± 4.4	60.27 ± 2.2	65.21 ± 1.3	76.25 ± 1.8	52.78 ± 4.6	61.43 ± 2.4	63.49 ± 1.4
A+G_h	72.21 ± 2.2	**64.02 ± 4.0**	**61.1 ± 2.1**	**65.78 ± 1.1**	**77.24 ± 1.8**	**54.16 ± 4.7**	61.41 ± 2.3	**64.27 ± 1.4**
A+F+G_{vc}	**73.89 ± 2.3**	59.06 ± 4.0	60.46 ± 2.3	64.47 ± 1.4	75.88 ± 1.7	**53.58 ± 5.1**	63.17 ± 2.8	**64.21 ± 1.6**
A+F+G_{3dg}	73.41 ± 2.2	59.46 ± 4.2	61.12 ± 2.4	64.66 ± 1.3	76.7 ± 1.7	51.37 ± 4.8	61.91 ± 2.4	63.33 ± 1.5
A+F+G_{eye}	73.12 ± 2.5	56.26 ± 4.5	60.74 ± 2.4	63.37 ± 1.5	74.87 ± 1.7	52.66 ± 4.8	**64.44 ± 2.4**	63.99 ± 1.5
A+F+G_h	73.19 ± 2.4	**59.75 ± 4.3**	**63.23 ± 2.8**	**65.39 ± 1.5**	**77.24 ± 1.6**	51.7 ± 5.1	61.47 ± 2.3	63.47 ± 1.5

more from F and *puzzled* from G. Nevertheless, G and F results are generally less stable than those of A. For SP particularly, accuracy remarkably increases for *puzzled* when training on WHOLE while decreasing to a similar degree for *calm*. As a matter of fact, *puzzled* obtains higher accuracy with F than A for WH→SP, while *pleased* decreases with respect to SP→SP. For FR, with F, *puzzled* obtains the highest accuracy among all models with FR→FR, contrary to SP and NO. *Pleased* accuracy notably increases when training on WHOLE. The confusion patterns of G are similar to those for SP; however, for F, *calm* is greatly confused with *puzzled*. Finally, for NO→NO, *pleased* gets greatly confused with *neutral* with G. By contrast, with F, *pleased* obtains the highest accuracy among all models both when training with NO and with WHOLE. *Pleased* accuracy also increases notably with G when adding more training data.

TABLE 6.13: Emotion recognition results for audio-based labels trained on the NORWAY (left) and WHOLE (right) training subsets and evaluated on the NORWAY test subset (train→test), reported as unweighted average accuracy ± SEM over 10 folds and three runs per fold. Bold: best accuracy per group. Underlined: best accuracy overall.

	NORWAY → NORWAY				WHOLE → NORWAY			
Modality	**Calm Accuracy**	**Pleased Accuracy**	**Puzzled Accuracy**	**Average Accuracy**	**Calm Accuracy**	**Pleased Accuracy**	**Puzzled Accuracy**	**Average Accuracy**
A	87.79 ± 0.8	53.69 ± 4.1	20.83 ± 11.0	63.12 ± 1.9	79.85 ± 1.1	66.52 ± 4.1	30.09 ± 11.2	66.59 ± 2.2
F	43.79 ± 3.7	**66.34 ± 3.7**	**29.17 ± 12.6**	**49.89 ± 1.8**	12.5 ± 1.5	**81.6 ± 3.9**	**23.15 ± 12.1**	**43.33 ± 2.6**
G	**63.09 ± 2.2**	29.79 ± 4.0	1.39 ± 0.7	40.02 ± 1.8	**21.69 ± 1.8**	60.5 ± 4.3	12.5 ± 6.5	36.83 ± 2.4
A+F	88.12 ± 0.9	60.96 ± 3.9	27.78 ± 12.7	67.3 ± 2.0	79.43 ± 1.3	75.16 ± 4.3	**30.56 ± 11.1**	**70.13 ± 2.2**
A+G	88.04 ± 0.6	53.17 ± 4.0	26.39 ± 12.9	63.88 ± 2.0	**79.77 ± 1.2**	68.06 ± 3.9	16.67 ± 8.9	65.59 ± 2.3
A+F+G	**88.24 ± 0.8**	**61.02 ± 3.6**	**29.17 ± 12.5**	**67.7 ± 2.1**	79.61 ± 1.2	**76.7 ± 3.9**	20.83 ± 9.6	69.75 ± 2.2
A+G_{vc}	**87.74 ± 0.8**	54.1 ± 3.8	26.39 ± 12.9	64.1 ± 1.8	80.06 ± 0.9	65.6 ± 4.0	10.42 ± 5.7	63.72 ± 2.1
A+G_{3dg}	87.69 ± 0.8	**54.81 ± 3.9**	**28.24 ± 12.6**	**64.62 ± 1.9**	79.98 ± 0.8	66.7 ± 3.7	31.25 ± 12.0	66.94 ± 2.3
A+G_{eye}	87.57 ± 0.8	54.18 ± 3.8	25.46 ± 13.0	63.91 ± 2.0	80.06 ± 1.2	66.6 ± 3.8	**31.71 ± 12.3**	**66.96 ± 2.2**
A+G_{h}	87.48 ± 0.8	53.87 ± 4.0	24.77 ± 11.9	63.67 ± 1.9	**80.97 ± 0.8**	**67.37 ± 3.3**	15.51 ± 7.7	65.57 ± 2.0
A+F+G_{vc}	87.75 ± 1.0	61.87 ± 3.9	27.78 ± 12.7	67.49 ± 2.0	79.07 ± 1.2	**76.88 ± 4.1**	**31.71 ± 11.9**	**71.03 ± 2.3**
A+F+G_{3dg}	85.99 ± 0.9	**62.9 ± 3.9**	30.09 ± 12.3	67.52 ± 2.0	79.62 ± 1.2	75.35 ± 4.0	27.55 ± 11.2	70.01 ± 2.1
A+F+G_{eye}	**88.29 ± 0.8**	62.05 ± 3.9	29.63 ± 12.4	**68.19 ± 2.1**	79.11 ± 1.3	75.0 ± 4.2	31.25 ± 12.0	70.14 ± 2.2
A+F+G_{h}	86.11 ± 0.8	61.23 ± 3.9	**30.56 ± 12.2**	66.97 ± 2.1	**80.52 ± 1.4**	75.34 ± 4.0	27.55 ± 11.6	70.43 ± 2.2

Statistical tests confirm significant differences for the following pairwise comparisons per country set. For WH: all unimodal (A, F, G) and unimodal vs bimodal (F vs A+F, G vs A+G) pairwise comparisons are significantly different ($p<.0001$). For SP→SP: A vs F, A vs G, F vs A+F, and G vs A+G ($p<.0001$). For WH→SP: all unimodal and unimodal vs bimodal comparisons ($p<.0001$). For FR→FR: A vs F, A vs G, F vs A+F, and G vs A+G ($p<.0001$). For WH→FR: all unimodal (F vs G obtaining $p=0.037$) and unimodal vs bimodal comparisons ($p<.0001$). For NO→NO: A vs G, G vs A+G ($p<.0001$). Finally, for WH→NO: A vs F, A vs G, F vs A+F, and G vs A+G ($p<.0001$).

Multimodality

Overall, incorporating F to A improves performance over A alone. By contrast, incorporating G seems detrimental on average. The best multimodal approach for WH, A+F, achieves a 2.5% relative performance increase over A. Class-wise, we observe that adding G increases performance and stability for *calm*, while adding F is beneficial for *pleased* and *puzzled*, despite the fact that the stability of the latter decreases.

SP and NO follow similar trends to WH class-wise, although for them, G is also beneficial for *puzzled* but to a lesser extent than A. More specifically, for SP, we observe that adding F to A moderately improves performance overall, and adding G to F+A increases it further, with A+F+G being the top performer. Adding G is beneficial for *calm*, with F+G obtaining the highest accuracy overall, while adding F is beneficial for *pleased*, and both modalities are beneficial for *puzzled*. For NO, A+F+G obtains the highest performance as well, which is the highest increase in performance (7.3%) caused by multimodal fusion across all countries. Contrary to SP and NO, adding more modalities is detrimental for FR. Class-wise, for FR→FR, *calm* obtains the highest accuracy with A+F+G, while for WH→FR, *puzzled* obtains the highest accuracy with A+F. However, these class-wise accuracy increases seem to come from pure accuracy redistribution, not increased discriminative power. We discuss these results in the *Comparison across countries* segment below.

Statistical tests confirm that, for WH, A+F vs A+G (p=.038) and A+G vs A+F+G (p=.024) differ significantly, while for SP, only A+F vs A+G are significantly different (p=.008). No statistical differences are found for FR and NO.

Eye and head feature subsets

For all countries, incorporating a single subset of G features (we henceforth denote any of the G subsets as G_s) produces slightly better results than the entire G set. However, the performance differences across the different subsets are minimal, and we find no statistically significant differences. This suggests that the different subsets contain redundant information for the audio-based emotion recognition task, with the combination of all features being detrimental. What is more, models containing G_s features are the top performers overall for all countries and training regimes (i.e., training per country and with WH), except for FR→FR and WH.

More specifically, for SP and FR, the top-performing feature subset on average is usually G_h, with A+F+G_h being the top performer overall for SP→SP and A+G_h for WH→FR. A+F+G_{vc} is the top performer overall for WH→SP. For NO, A+F+G_{eye} and A+F+G_{vc} are the top performers overall for NO→NO and WH→NO, respectively. Class-wise, for WH, G_h also seems more informative for *calm*, with A+G_h obtaining the best *calm* performance overall. However, this pattern does not replicate in the country subsets, for which the top-performing subset alternates between G_h and G_{vc}. There are no consistent patterns for the other classes across countries.

Comparison across countries

Figure 6.6 depicts per-country average accuracy results. With respect to multimodality, for SP and NO, we observe a similar trend to that of WH, with A+F and A+F+G outperforming A alone. For FR, however, accuracy decreases as the number of features increases, obtaining the highest average accuracy overall with A (66.28%), which evidences variations among countries. First, FR obtained the lowest inter-agreement score (Section 6.3.3). Analyzing the FR dataset distribution, we find that the average SD of the FR audio features is slightly larger than that of the other countries (0.68 for FR, 0.65 for SP, and 0.63 for NO), indicating that the data are more dispersed in the feature space. Furthermore, the best models for FR are those with the lowest number of parameters, as opposed to NO, which requires the largest number of parameters (Table 6.7), despite having a similar sample size (Table 6.5). Additionally, when reducing the number of features using only G_s instead of the complete G feature set, accuracy reaches the levels obtained with the rest of multimodal alternatives. What is more, A+G_h for WH→FR obtains the highest accuracy for such scenario (64.27%), higher than A alone. These results indicate that adding more features to FR increases the risk of overfitting, thus decreasing performance.

Class-wise, in contrast to WH results, both G and F aid in *puzzled* recognition for SP, while both aid in *calm* recognition for FR. Additionally, FR obtains the highest accuracy for *pleased* with A alone, while for the rest of the classes and countries, multimodal models outperform A. With respect to the auxiliary modalities, NO and FR appear to leverage F and G better than SP, respectively. Per-class accuracies are redistributed with respect to WH.

Expanding training data including other countries

Figure 6.6, along with Tables 6.11, 6.12, and 6.13, also depict the effect of training with the WH dataset instead of each country separately. As can be seen, adding

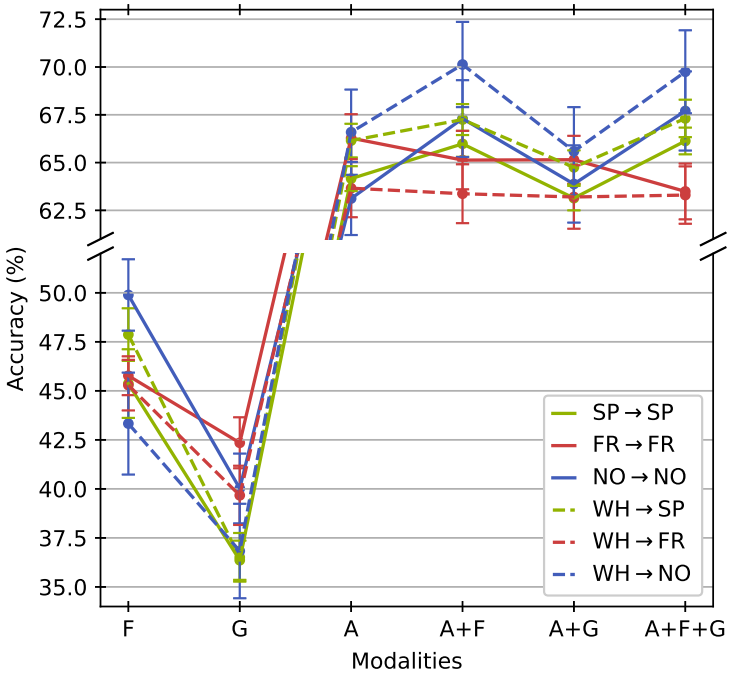


FIGURE 6.6: Per-country audio-based results, training on either SP, FR, NO, or WH training sets, and evaluating on SP, FR and NO test sets. Reported as unweighted average accuracy ± SEM over 10 folds and three runs per fold.

more data from different countries consistently improves accuracy on average for SP and NO with A and multimodal models, although stability decreases, and no statistically significant differences are found. NO trained on WH obtains the highest accuracy overall (71.03% with A+F+G_{vc}). For FR, however, we find the opposite effect again, with a decrease in accuracy of up to 4.1% with A, where *pleased* is the most affected class (although A+G_h obtains the highest accuracy for FR with a 64.27%). By contrast, the behavior of the auxiliary modalities is the opposite, with F obtaining the highest accuracy for *pleased*, which additionally increases with respect to FR→FR. Considering previous findings, we hypothesize that the FR audio feature distribution of the *pleased* class is significantly different from that of SP and NO; thus, adding more data is detrimental. Another difference comes from the arousal distribution of FR, being the country with the highest number of annotated *excited* instances (61.75% compared to 12-18% for the other countries).

Continuing with class-wise results, SP obtains performance increases for all classes in a similar proportion, benefiting from the increased data variability and sample size. By contrast, *calm* performance decreases for NO, which may be caused by the significant increase in the number of training instances for the minority classes (around 279% and 4,369% increase for *pleased* and *puzzled*, respectively). With respect to auxiliary modalities, the general trend shows that adding more data hurts performance. By further analyzing confusion patterns, we observe that these are mostly reversed when training with WH, especially affecting discrimination between *calm* and the other classes for G and *calm* and *puzzled* for F. For the latter modality, though, *pleased* is still recognized accurately.

TABLE 6.14: Video-based results under speech on WHOLE, reported as unweighted average accuracy ± SEM over 10 folds and three runs per fold. Bold: best accuracy per group. Underlined: best accuracy per training type. Italics: best accuracy overall.

Modality	Neutral Accuracy	Happy Accuracy	Pensive Accuracy	Average Accuracy
Training on speech data:				
F	73.98 ± 1.0	72.2 ± 2.4	57.87 ± 2.1	68.02 ± 1.1
A	39.07 ± 1.1	**63.57 ± 2.1**	66.67 ± 1.5	56.44 ± 0.6
G	**51.13** ± 1.8	46.33 ± 2.4	**74.47 ± 2.2**	**57.31 ± 0.5**
F+A	74.5 ± 0.9	**73.39 ± 2.4**	60.95 ± 1.9	69.61 ± 1.1
F+G	***76.45 ± 1.1***	71.58 ± 2.4	66.86 ± 2.2	71.63 ± 1.1
F+A+G	76.36 ± 1.0	73.27 ± 2.4	**68.52 ± 1.9**	***72.72 ± 1.0***
F+G_{vc}	**75.41 ± 1.0**	71.65 ± 2.2	**66.34 ± 2.0**	**71.13 ± 1.0**
F+G_{3dg}	74.41 ± 1.1	72.03 ± 2.2	62.42 ± 2.4	69.62 ± 1.1
F+G_{eye}	73.83 ± 1.0	71.84 ± 2.3	62.46 ± 2.2	69.38 ± 1.1
F+G_h	73.96 ± 1.0	**72.42 ± 2.3**	58.79 ± 2.2	68.39 ± 1.1
F+A+G_{vc}	**75.94 ± 1.0**	73.08 ± 2.4	**67.73 ± 1.7**	**72.25 ± 1.0**
F+A+G_{3dg}	75.63 ± 1.0	73.21 ± 2.4	64.05 ± 2.2	70.96 ± 1.0
F+A+G_{eye}	75.06 ± 0.9	73.13 ± 2.5	64.13 ± 2.1	70.77 ± 1.1
F+A+G_h	75.19 ± 0.9	**73.44 ± 2.4**	60.83 ± 2.0	69.82 ± 1.1
Training on all data (speech + silence):				
F	70.44 ± 1.1	73.44 ± 2.2	61.58 ± 2.1	68.49 ± 1.1
G	42.45 ± 2.0	55.56 ± 2.1	*76.96 ± 2.0*	58.32 ± 0.5
F+G	73.55 ± 1.1	73.11 ± 2.2	70.56 ± 2.1	72.41 ± 1.0
F+G_{vc}	**71.99 ± 1.0**	73.01 ± 2.1	**70.39 ± 1.9**	**71.8 ± 1.0**
F+G_{3dg}	71.37 ± 1.2	73.25 ± 2.1	66.1 ± 2.4	70.24 ± 1.0
F+G_{eye}	70.71 ± 1.1	73.36 ± 2.2	65.85 ± 2.3	69.97 ± 1.0
F+G_h	70.1 ± 1.1	***73.69 ± 2.2***	63.42 ± 2.3	69.07 ± 1.1

6.5.4 Video-based emotion expression recognition under speech

Tables 6.14, 6.15, 6.16, and 6.17 show results for video-based labels under speech trained and evaluated on WH, SP, FR, and NO, respectively, using two different training regimes: 1) training on samples where the user is speaking (*speech data*); and 2) training on all samples irrespective of speaking status (*speech+silence*). We report the results below.

Main modality

For WH, F alone obtains similar accuracy for *neutral* and *happy*, while comparatively struggles with *pensive*, which is highly confused with *neutral*. This behavior is not proportional to the number of instances since *pensive* has more than *happy*. Nonetheless, *happy* performance is slightly less stable than that of *pensive*. Country-wise, however, the performance gap is found between *neutral* and the minority classes instead. More specifically, for SP, the accuracy for *happy* and *pensive* is reduced and similar. This time, the accuracy distribution does follow the per-class sample size. The two minority classes are confused with *neutral* at a similar rate. For FR, the accuracy gap between the minority classes and *neutral* is higher for SP due to an increase

TABLE 6.15: Emotion recognition results for video-based labels trained on the SPAIN (left) and WHOLE (right) training subsets under speech only or speech and silence instances, and evaluated on the SPAIN test subset under speech (train→test), reported as unweighted average accuracy ± SEM over 10 folds and three runs per fold. Bold: best accuracy per group. Underlined: best accuracy per training type. Italics: best accuracy overall.

	SPAIN → SPAIN				WHOLE → SPAIN			
Modality	**Neutral Accuracy**	**Happy Accuracy**	**Pensive Accuracy**	**Average Accuracy**	**Neutral Accuracy**	**Happy Accuracy**	**Pensive Accuracy**	**Average Accuracy**
Training on speech data:								
F	72.56 ± 2.6	59.06 ± 3.8	60.45 ± 3.2	64.09 ± 1.3	76.75 ± 1.8	57.39 ± 4.9	60.36 ± 3.4	65.14 ± 1.5
A	52.91 ± 1.5	**56.33 ± 5.2**	67.11 ± 1.7	**58.95 ± 1.5**	51.27 ± 1.6	**50.99 ± 5.1**	71.2 ± 2.0	58.35 ± 1.5
G	**58.95 ± 1.6**	26.54 ± 3.6	**76.51 ± 2.6**	55.08 ± 1.3	**58.37 ± 2.5**	37.71 ± 3.5	**79.12 ± 3.1**	**59.3 ± 1.4**
F+A	75.25 ± 2.0	**60.06 ± 5.3**	63.45 ± 2.8	66.41 ± 1.5	76.17 ± 1.9	**61.92 ± 5.5**	65.26 ± 3.3	68.04 ± 1.7
F+G	76.27 ± 2.0	57.96 ± 3.5	70.43 ± 2.6	68.55 ± 1.1	***79.22 ± 1.8***	56.68 ± 5.0	70.11 ± 3.0	69.14 ± 1.5
F+A+G	***77.4 ± 2.1***	58.28 ± 5.1	**71.34 ± 2.8**	***69.34 ± 1.4***	79.12 ± 1.7	61.6 ± 5.5	**70.7 ± 3.0**	***70.84 ± 1.7***
F+G_{vc}	**76.15 ± 1.9**	60.07 ± 3.5	66.34 ± 3.1	**67.73 ± 1.3**	**79.11 ± 1.6**	56.07 ± 5.0	**67.04 ± 3.2**	**67.86 ± 1.6**
F+G_{3dg}	74.61 ± 2.1	58.23 ± 3.4	66.3 ± 2.8	66.65 ± 1.1	77.63 ± 1.8	55.91 ± 4.9	66.44 ± 3.2	67.13 ± 1.5
F+G_{eye}	73.75 ± 2.2	***60.75 ± 3.6***	**67.02 ± 2.8**	67.34 ± 1.2	76.55 ± 1.9	56.34 ± 5.0	66.95 ± 3.1	67.06 ± 1.5
F+G_h	72.78 ± 2.3	60.55 ± 3.6	60.31 ± 3.0	64.57 ± 1.3	76.67 ± 2.0	**57.09 ± 5.0**	61.25 ± 3.4	65.31 ± 1.5
F+A+G_{vc}	**76.68 ± 2.0**	59.74 ± 4.9	**68.78 ± 2.9**	**68.62 ± 1.5**	**77.99 ± 1.6**	61.68 ± 5.5	**70.05 ± 2.8**	**70.22 ± 1.7**
F+A+G_{3dg}	76.41 ± 2.1	57.66 ± 5.1	68.18 ± 2.8	67.74 ± 1.5	77.99 ± 1.7	61.84 ± 5.5	68.11 ± 3.1	69.68 ± 1.7
F+A+G_{eye}	75.37 ± 2.0	**60.49 ± 5.4**	67.4 ± 2.8	67.96 ± 1.6	77.49 ± 1.8	61.87 ± 5.5	67.96 ± 3.2	69.45 ± 1.7
F+A+G_h	75.5 ± 2.0	60.19 ± 5.3	63.03 ± 2.9	66.41 ± 1.5	76.59 ± 1.9	***62.17 ± 5.5***	65.57 ± 3.1	68.35 ± 1.7
Training on all data (speech + silence):								
F	69.33 ± 2.7	59.12 ± 3.6	64.06 ± 3.2	64.19 ± 1.4	73.14 ± 2.0	58.42 ± 4.9	64.06 ± 3.2	65.44 ± 1.5
G	47.92 ± 1.8	40.8 ± 4.5	*81.83 ± 2.4*	57.46 ± 1.6	50.65 ± 2.8	45.1 ± 3.6	*82.04 ± 2.8*	59.91 ± 1.2
F+G	73.79 ± 2.2	57.89 ± 3.5	75.25 ± 2.4	69.33 ± 1.0	77.09 ± 1.8	57.59 ± 5.0	73.21 ± 2.8	69.73 ± 1.5
F+G_{vc}	**73.09 ± 2.0**	59.47 ± 3.4	70.77 ± 3.0	**68.01 ± 1.3**	**75.82 ± 1.6**	57.21 ± 4.9	**71.27 ± 2.9**	**68.51 ± 1.5**
F+G_{3dg}	72.02 ± 2.3	57.26 ± 3.8	70.84 ± 2.8	67.02 ± 1.2	75.27 ± 1.8	57.06 ± 5.0	69.83 ± 3.0	67.86 ± 1.5
F+G_{eye}	70.92 ± 2.4	59.16 ± 4.2	**71.44 ± 2.7**	67.41 ± 1.3	74.22 ± 1.9	57.49 ± 5.0	70.12 ± 2.9	67.7 ± 1.5
F+G_h	68.56 ± 2.5	**59.75 ± 3.8**	65.99 ± 2.8	64.78 ± 1.3	73.16 ± 2.1	**58.14 ± 5.1**	65.88 ± 3.2	65.97 ± 1.5

in confusion. Nonetheless, it is worth mentioning that, when training with WHOLE, *happy* performance almost equals that of *neutral*. Finally, for NO, trends are similar to FR, although the gap between *neutral* and the minority classes is even larger, highly likely due to the decrease in sample size.

The SD across folds is 6.2% for WH, higher for this scenario than for the audio-based but more consistent, and it is even higher per country, reaching values of 15% for NO. By contrast, the SD across runs is around 0.05%. This unveils the large variability across subjects.

Auxiliary modalities

For WH, A and G obtain accuracy results closer to the main modality than for the audio-based scenario, with G slightly outperforming F on average. Remarkably, G achieves the highest accuracy overall for *pensive*. A appears to be informative for *pensive* as well but to a lesser extent, also outperforming F, and is more informative than G for *happy*. Nonetheless, G is less stable than A class-wise, although on average, they are more stable than F.

Trends are maintained across countries class-wise. However, on average, A is more informative than G for them, mostly due to the extremely low performance of G for *happy*, which gets confused with *neutral*. For FR specifically, A moderately surpasses F when recognizing *happy*. For NO→NO, A obtains the highest accuracy

TABLE 6.16: Emotion recognition results for video-based labels trained on the FRANCE (left) and WHOLE (right) training subsets under speech only or speech and silence instances, and evaluated on the FRANCE test subset under speech (train→test), reported as unweighted average accuracy ± SEM over 10 folds and three runs per fold. Bold: best accuracy per group. Underlined: best accuracy per training type. Italics: best accuracy overall.

	FRANCE → FRANCE				**WHOLE → FRANCE**			
Modality	**Neutral Accuracy**	**Happy Accuracy**	**Pensive Accuracy**	**Average Accuracy**	**Neutral Accuracy**	**Happy Accuracy**	**Pensive Accuracy**	**Average Accuracy**
Training on speech data:								
F	77.57 ± 2.0	51.59 ± 4.0	50.35 ± 3.7	59.84 ± 1.5	70.16 ± 2.6	66.28 ± 4.5	57.62 ± 4.0	64.69 ± 2.0
A	34.42 ± 1.2	**52.6 ± 4.8**	64.79 ± 1.9	**50.6 ± 1.6**	18.42 ± 1.4	***68.63 ± 3.6***	60.37 ± 2.0	49.14 ± 0.8
G	**45.16 ± 2.0**	34.75 ± 4.2	**67.29 ± 3.6**	49.07 ± 0.8	**46.25 ± 3.0**	44.46 ± 4.1	**69.48 ± 3.9**	**53.39 ± 1.5**
F+A	80.27 ± 2.0	***55.61 ± 4.1***	44.12 ± 4.4	60.0 ± 1.9	70.78 ± 2.4	**67.9 ± 4.0**	59.51 ± 4.1	66.07 ± 1.8
F+G	80.24 ± 1.8	50.04 ± 4.0	**57.48 ± 4.0**	62.59 ± 1.5	***72.17 ± 2.4***	65.3 ± 4.2	65.76 ± 3.6	67.74 ± 1.6
F+A+G	**81.3 ± 1.9**	53.7 ± 4.1	53.35 ± 4.8	**62.78 ± 1.9**	71.72 ± 2.3	67.4 ± 3.9	**66.98 ± 3.9**	***68.7 ± 1.6***
F+G_{vc}	**80.47 ± 1.8**	**51.62 ± 4.0**	**54.34 ± 3.8**	**62.14 ± 1.5**	**70.42 ± 2.4**	66.25 ± 4.2	**67.28 ± 3.2**	**67.98 ± 1.6**
F+G_{3dg}	77.99 ± 2.1	51.55 ± 4.0	52.97 ± 3.8	60.84 ± 1.5	69.96 ± 2.6	66.32 ± 4.2	61.73 ± 4.0	66.0 ± 1.8
F+G_{eye}	77.41 ± 2.1	50.87 ± 4.0	52.55 ± 3.8	60.28 ± 1.4	69.44 ± 2.5	65.78 ± 4.3	61.95 ± 3.7	65.72 ± 1.8
F+G_h	78.3 ± 2.0	50.49 ± 4.0	49.27 ± 3.7	59.35 ± 1.5	70.05 ± 2.6	**66.43 ± 4.1**	59.07 ± 4.0	65.18 ± 1.8
F+A+G_{vc}	***81.33 ± 1.9***	**55.13 ± 4.2**	**50.26 ± 4.5**	**62.24 ± 1.9**	**71.93 ± 2.2**	**67.93 ± 4.0**	**66.1 ± 3.6**	**68.65 ± 1.6**
F+A+G_{3dg}	80.09 ± 2.0	54.9 ± 4.1	46.96 ± 4.5	60.65 ± 1.9	70.38 ± 2.4	67.62 ± 4.0	62.45 ± 4.5	66.82 ± 1.8
F+A+G_{eye}	80.09 ± 1.9	54.19 ± 4.1	47.31 ± 4.5	60.53 ± 1.9	69.79 ± 2.4	67.5 ± 4.0	62.61 ± 4.2	66.63 ± 1.8
F+A+G_h	80.63 ± 1.9	54.8 ± 4.1	44.01 ± 4.4	59.81 ± 2.0	71.12 ± 2.4	67.84 ± 4.0	59.38 ± 4.3	66.11 ± 1.8
Training on all data (speech + silence):								
F	75.26 ± 2.1	52.45 ± 4.0	53.7 ± 3.8	60.47 ± 1.5	67.17 ± 2.7	67.7 ± 4.4	61.6 ± 3.8	65.49 ± 1.9
G	36.01 ± 1.6	40.98 ± 3.8	*71.09 ± 3.5*	49.36 ± 0.7	37.85 ± 3.0	51.68 ± 3.7	*71.46 ± 3.8*	53.67 ± 1.5
F+G	77.82 ± 1.9	51.07 ± 3.9	62.77 ± 4.1	*63.89 ± 1.6*	69.2 ± 2.5	66.85 ± 4.0	69.47 ± 3.6	68.5 ± 1.6
F+G_{vc}	**78.2 ± 1.8**	52.44 ± 4.0	**59.96 ± 3.8**	**63.53 ± 1.5**	**67.23 ± 2.4**	67.46 ± 4.1	**71.37 ± 3.0**	**68.68 ± 1.5**
F+G_{3dg}	75.1 ± 2.2	**52.68 ± 4.0**	57.34 ± 4.1	61.71 ± 1.5	67.13 ± 2.6	67.47 ± 4.1	65.69 ± 4.0	66.76 ± 1.8
F+G_{eye}	74.82 ± 2.2	52.3 ± 3.9	55.69 ± 3.9	60.94 ± 1.5	66.18 ± 2.6	67.2 ± 4.2	65.45 ± 3.6	66.27 ± 1.7
F+G_h	76.01 ± 2.0	51.9 ± 4.0	52.79 ± 3.9	60.24 ± 1.6	66.72 ± 2.6	**67.9 ± 4.0**	63.7 ± 3.7	66.11 ± 1.7

among all models for *pensive*, instead of G. Actually, for the latter, *happy* instances are mostly detected as *neutral*, showing the highest confusion among countries. Analyzing the confusion patterns for all datasets, we confirm that gaze cues are highly discriminative for *pensive* and audio cues are highly discriminative for *happy*.

Statistical tests show significant differences for the following cases. For WH: F vs A and F vs G (p=.015), and F vs F+G/A (p<.001) when training with speech data, and for all comparisons when training with all data (p<.001). For SP: all cases of G vs F+G (p<.01), and A vs F+A (p<.01) only for WH→SP training with speech data. For FR: when training on speech only, all F vs A and F vs G comparisons (p=.024 for FR→FR and p=.005 for WH→FR) and all G/A vs F+G/A comparisons (all p<.001 except A vs F+A for FR→FR, p=.024); when training on speech and silence, F vs G (p=0.013 for FR→FR, and p=.003 for WH→FR) and G vs F+G (p=.005 for FR→FR and p<.0001 for WH→FR). And for NO: for NO→NO, F vs G/A (p=.046 and p=.048, respectively) and G vs F+G (p=.008) when training with speech data, and G vs F+G (p=.011) when training with speech and silence; for WH→NO, F vs G and G vs F+G for both training types (p<.0001), F vs G (p<.0001), A vs G (p=.007) and A vs F+A (p=.001) when training with speech data.

TABLE 6.17: Emotion recognition results for video-based labels trained on the NORWAY (left) and WHOLE (right) training subsets under speech only or speech and silence instances, and evaluated on the NORWAY test subset under speech (train→test), reported as unweighted average accuracy ± SEM over 10 folds and three runs per fold. Bold: best accuracy per group. Underlined: best accuracy per training type. Italics: best accuracy overall.

	NORWAY → NORWAY				WHOLE → NORWAY			
Modality	**Neutral Accuracy**	**Happy Accuracy**	**Pensive Accuracy**	**Average Accuracy**	**Neutral Accuracy**	**Happy Accuracy**	**Pensive Accuracy**	**Average Accuracy**
Training on speech data:								
F	84.03 ± 1.6	49.73 ± 5.4	47.0 ± 3.8	60.26 ± 2.7	70.8 ± 1.8	68.08 ± 4.9	59.4 ± 4.0	66.09 ± 2.2
A	46.58 ± 1.2	***63.82 ± 4.0***	***68.32 ± 2.0***	**59.57 ± 1.4**	42.24 ± 1.6	**75.74 ± 2.8**	65.31 ± 1.6	**61.1 ± 1.1**
G	**69.65 ± 3.0**	21.41 ± 3.1	66.99 ± 4.2	52.68 ± 1.8	**42.86 ± 3.2**	53.86 ± 4.5	**68.46 ± 4.2**	55.06 ± 1.6
F+A	85.13 ± 1.3	**56.82 ± 4.9**	52.69 ± 3.8	64.88 ± 2.6	74.64 ± 1.5	***75.89 ± 3.8***	60.21 ± 4.0	70.25 ± 2.0
F+G	86.69 ± 1.1	47.62 ± 5.4	59.45 ± 4.1	64.59 ± 2.8	75.15 ± 1.5	67.25 ± 5.2	67.62 ± 4.4	70.01 ± 2.5
F+A+G	***87.77 ± 1.0***	55.13 ± 4.9	**62.63 ± 3.9**	***68.51 ± 2.7***	**76.7 ± 1.2**	75.34 ± 3.9	**70.69 ± 4.0**	***74.24 ± 2.2***
F+G_{vc}	84.6 ± 1.2	**50.38 ± 5.2**	**58.06 ± 4.4**	**64.35 ± 2.9**	**73.25 ± 1.5**	66.45 ± 5.2	**67.79 ± 4.3**	**69.16 ± 2.5**
F+G_{3dg}	85.04 ± 1.3	48.66 ± 5.4	51.42 ± 3.7	61.71 ± 2.7	71.4 ± 1.9	67.35 ± 4.9	63.43 ± 4.4	67.39 ± 2.3
F+G_{eye}	**85.93 ± 1.1**	47.84 ± 5.4	48.83 ± 3.5	60.87 ± 2.6	71.71 ± 1.6	67.57 ± 4.9	62.12 ± 4.1	67.13 ± 2.2
F+G_h	84.03 ± 1.4	50.15 ± 5.6	48.1 ± 4.0	60.76 ± 2.7	70.77 ± 1.8	**68.5 ± 5.1**	60.62 ± 4.1	66.63 ± 2.3
F+A+G_{vc}	86.78 ± 1.0	55.66 ± 5.0	**60.48 ± 4.1**	**67.64 ± 2.8**	***77.05 ± 1.2***	74.56 ± 3.8	**68.88 ± 4.1**	**73.49 ± 2.2**
F+A+G_{3dg}	**87.07 ± 1.1**	54.69 ± 4.8	56.22 ± 4.0	65.99 ± 2.7	76.8 ± 1.3	74.95 ± 3.6	65.14 ± 4.0	72.3 ± 2.0
F+A+G_{eye}	87.01 ± 0.9	54.59 ± 5.0	54.87 ± 3.6	65.49 ± 2.7	76.0 ± 1.2	**75.36 ± 3.7**	64.66 ± 3.7	72.01 ± 2.1
F+A+G_h	85.58 ± 1.2	**56.75 ± 4.9**	53.15 ± 4.0	65.16 ± 2.6	76.57 ± 1.3	75.33 ± 3.8	60.51 ± 4.0	70.8 ± 2.0
Training on all data (speech + silence):								
F	81.57 ± 1.8	53.71 ± 5.1	48.37 ± 3.8	61.22 ± 2.5	66.5 ± 1.9	69.13 ± 4.9	62.85 ± 4.0	66.16 ± 2.3
G	60.23 ± 3.4	37.93 ± 5.5	66.65 ± 4.2	54.93 ± 2.0	31.69 ± 3.0	62.77 ± 5.1	70.27 ± 4.1	54.91 ± 1.6
F+G	85.05 ± 1.2	52.31 ± 5.1	59.77 ± 4.0	65.71 ± 2.8	70.95 ± 1.8	69.25 ± 5.2	*71.45 ± 4.1*	70.55 ± 2.4
F+G_{vc}	82.04 ± 1.3	53.98 ± 5.0	**59.95 ± 4.3**	**65.32 ± 2.8**	**69.11 ± 1.6**	68.14 ± 5.0	**70.98 ± 4.2**	**69.41 ± 2.4**
F+G_{3dg}	82.42 ± 1.6	53.7 ± 5.0	53.69 ± 3.7	63.27 ± 2.6	67.04 ± 2.0	68.85 ± 5.0	66.96 ± 4.3	67.62 ± 2.3
F+G_{eye}	**83.16 ± 1.4**	53.61 ± 5.0	50.29 ± 3.5	62.35 ± 2.6	67.31 ± 1.8	69.59 ± 5.0	65.33 ± 4.1	67.41 ± 2.3
F+G_h	82.04 ± 1.7	**54.13 ± 5.1**	49.29 ± 4.0	61.82 ± 2.6	65.84 ± 1.9	**70.0 ± 5.2**	64.34 ± 4.1	66.73 ± 2.4

Multimodality

For WH, when training on speech data, we observe that adding A or G to F increases accuracy, and the highest is achieved by combining the three modalities, with F+A+G showing a 6.9% relative improvement over F alone. Class-wise, adding G substantially improves performance for *pensive* followed by *neutral*, while adding F has a more subtle effect. By contrast, *happy* appears to benefit from A and not G, but does so when combining the three modalities. Nevertheless, *happy* does benefit from G_s feature sets, outperforming both F and F+G. We observe similar trends when training on all speech and silence data.

Trends are overall maintained across countries, except for some differences. For instance, contrary to the other countries, we find that adding A hinders the recognition of *pensive* for FR→FR, and the accuracy increase on average is also minimal. Nonetheless, this anomaly gets corrected when training on WHOLE. In addition, contrary to the other countries, adding G and A to F obtains similar performance on average for NO, with F+A slightly outperforming F+G. However, it is with the combination of the three modalities that the highest performance is achieved.

Statistical tests confirm significant differences for the following pairwise comparisons when training on speech data: for WH, F vs F+G (p=.031), F vs F+A+G (p=.018), F+G vs F+A (p=.044), and F+A vs F+A+G (p=.015); for SP→SP, F+A vs F+A+G (p=.009); for WH→SP, F vs F+G (p=.004), F vs F+A (p=.015), F vs F+A+G

(p<.001), and F+A vs F+A+G (p=.003); for FR→FR, F vs F+G (p=.02), F+A vs F+A+G (p=.021); for WH→FR, F vs F+G (p=.024), F vs F+A+G (p=.007), F+G vs F+A+G (p=.044), F+A vs F+A+G (p=.005); for NO→NO, all pairwise comparisons (all p<.001 except F+G vs F+A+G, with p=.003); and for WH→NO, all pairwise comparisons (p<.001 except F vs F+A and F vs F+G vs F+A+G with p <.01). Significance results when training on speech and silence data: for WH, F vs F+G (p=.015); for WH→SP, F vs F+G (p=.001); for FR, F vs F+G (p=.009 for FR→FR, and p=.018 for WH→FR); and for NO, F vs F+G for all cases (p<.0001).

Eye and head feature subsets

Contrary to the audio-based scenario, none of the models with G_s subsets outperform multimodal G-based models (except G_{vc} for WH→FR when training with all data), suggesting that all feature subsets provide complementary information for the task. Nevertheless, some G_s models are found to be the top performers among all evaluated models for specific classes depending on the training regime. For instance, we observe that G_{vc} is almost consistently the top performer for *pensive* across countries among the different feature sets, with a large accuracy disparity between them. This can be attributed to its strong association with the characteristic behavior of this class, where eyes shifting away from the VC location is associated with thinking episodes. For SP, however, G_{eye} outperforms G_{vc}, suggesting that the richer information about dynamics provided by G_{eye} is important for this country, or that the VC location estimation has not been sufficiently precise. In general, G_h appears to be the least correlated with *pensive*, while G_{eye} and G_{3dg} fall within an intermediate range. Accuracy results for *neutral* also change depending on the feature set, although to a lesser extent than *pensive*, with G_{vc} being the most informative for WH and across most countries and G_h the least informative. Finally, accuracy results for *happy* are almost identical across feature sets. Interestingly, different variants of G_s marginally but consistently outperform (F+)G and F for this class for WH and all countries. For instance, for WH, we find that F+G_h is the top performer when training with speech and silence data. Contrary to other countries, for FR, G_{vc} is the most beneficial feature subset for *happy* when training with speech data, and G_{3dg} when training with speech and silence.

Statistical tests confirm statistically significant differences for the following pairwise comparisons. For WH: (F+A+)G_h vs (F+A+)G_{vc} (p=.015), and G vs all G_s subsets (p<.05) except G_{vc}. For SP: G_{vc} vs G_h when training on speech data (p=.037 and p=.048 without and with A, respectively, for SP→SP, and p<.01 for WH→SP); when training with speech and silence, F+G_{vc} vs F+G_h for WH→SP (p<.01), and most comparisons between G and G_s (p<.05), although for G and G_{vc} it is only for WH→SP (p=.03). 65% of the pairwise comparisons for FR show significant differences at various p-value levels (p<.05), including all (F+A+)G_{vc} vs (F+A+)G_h/G_{3dg}/G_{eye} comparisons. Finally, for NO, 75% of the pairwise comparisons are different at different p-value levels (p<.05), including all pairwise comparisons with G_{vc} feature set and the rest of the feature sets, but excluding G vs G_{vc} for NO→NO.

Comparison across countries

Figure 6.7a illustrates per-country average results for all modalities and the two training regimes, along with Tables 6.15, 6.16, and 6.17. In general, SP achieves the highest accuracy for all modality combinations, greatly benefiting from G and followed by A when added to F. It obtains the highest accuracy overall with F+A+G

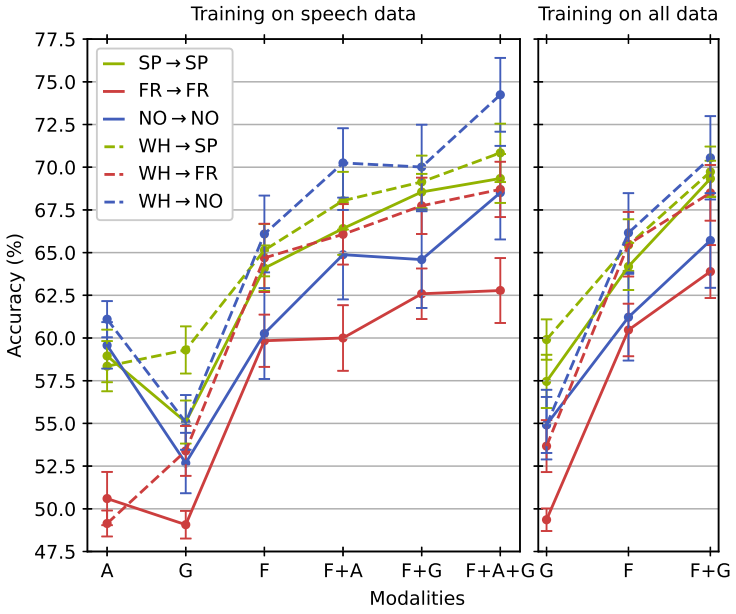
(69.34%) when training with speech, and with F+G (69.33%) when training with all data. NO achieves a similar accuracy with the trimodal model, although it benefits more from A than from G. By contrast, FR barely benefits from adding A, and repeatedly scores the lowest, despite having slightly more data than NO and almost equal performance on average with F. This might be partly caused by the difference in class proportions across countries, with *happy* being the most variant class (4.7% of the total data for FR, 0.8% for SP, and 2.7% for NO).

The difference in the distribution of audio features for FR (discussed in Section 6.5.3) is also noticeable here, with A alone obtaining the lowest accuracy for FR, and with a substantial difference compared to the other countries. Class-wise, adding A to F hurts *pensive* recognition for FR due to a high confusion between *neutral* and *pensive*, although their performance alone is better, and the highest with A. Regarding G, we observe the highest discriminative power for *pensive* with SP, with more elevated levels of *neutral-happy* confusion for the other two countries, thus scoring lower in the comparison.

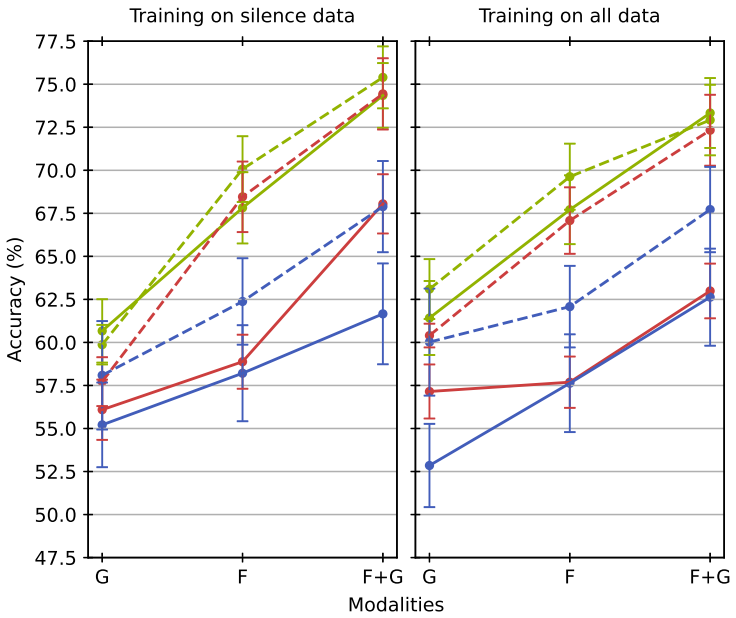
Expanding training data including other countries

Figure 6.7a, and Tables 6.15, 6.16, and 6.17 also illustrate the effect of training with the WH dataset instead of each country separately for the two training regimes. In general, adding more data increases accuracy on average for all cases except for A, for which accuracy is mostly maintained or slightly reduced. As can be seen, FR obtains the highest performance increase overall despite still scoring the lowest, and NO obtains the highest accuracy results overall, with F+A+G being the top performer (74.24%). SP and FR also achieve the highest accuracy with F+A+G. Models are slightly less stable when training on WH for F-based models, except for NO.

SP obtains the lowest gain, probably because it is the country with the most instances in WH for all classes except for *happy*. However, when we investigate class-wise trends, we observe an interesting difference. For NO and FR, the minority classes have their accuracy significantly increased, and the *neutral* accuracy decreased for all modality combinations, proving that the increase in variability and effective training data is beneficial for them to decrease confusion with *neutral*. By contrast, SP sees the *neutral* accuracy increase with all F-based models for all cases and with G only when training with speech data, while F and F+G maintain accuracy for *pensive* and decrease it for *happy*. However, performance increases when adding A for the minority classes, although A alone gets *happy* accuracy reduced, while with G alone they both improve. It is important to highlight that SP is the most unbalanced dataset over the three countries for both training regimes and, while its minority class, *happy*, is increased the most with respect to the other countries (473% when training with speech and 635% when training with speech and silence), this does not translate into higher accuracy for such class, suggesting that the higher number of samples does not provide the necessary variability to increase discriminative power. For silence-based evaluation (Section 6.5.5), *happy* accuracy does increase with F-based models when training on WH. Thus, we believe this difference in *happy* performance is due to the facial deformations caused by speaking being different across countries, greatly increasing variability. Although this is true for all countries, due to the small number of *happy* instances for SP, this distribution might be narrower than that of FR and NO. Thus, an increase in variability might be detrimental to performance. Then, when adding A to F, A helps discriminate better among classes that visually might be more similar. Regarding A alone,



(a) Testing on speech data



(b) Testing on silence data

FIGURE 6.7: Per-country video-based results under (a) speech or (b) silence, trained on SP, FR, NO, and WH training sets, and evaluated on SP, FR and NO test sets. Reported as unweighted average accuracy ± SEM over 10 folds and three runs per fold.

we mostly observe a redistribution of accuracies among classes, with an increase in *happy-neutral* confusion.

Statistical tests confirm significant differences for the following cases. For FR: when training with speech data, F+G (p=.027), F+A (p=.048), F+A+G (p=.028), F+G_{vc} (p=.03) and F+G_{vc} (p=.043), all F+A+G_s cases ($p<.05$); and when training with speech and silence data, F+G (p=.032), and all F+G_s cases (p=0.03). For NO: all models and cases at different p-value levels ($p<.01$) except for A and G. No significant differences are found for SP.

Expanding training data including silence instances

As can be seen in Table 6.14 for WH, training on all data (speech and silence instances) marginally but consistently improves performance on average, with the highest improvement obtained with G. Class-wise, confusion patterns reveal that the minority classes are less predicted as *neutral*, with a slight increase in confusion in the other direction in some cases. Nonetheless, the change is positive for *pensive* and *happy*. For G specifically, the *neutral-happy* confusion patterns are inverted.

The consistent improvement is also observed across countries, as depicted in Figure 6.7a and Tables, 6.15, 6.16, and 6.17, both when training per country and when training on WH, and the class-wise trends are generally maintained. As a matter of fact, per-class accuracies tend to be more balanced in this setting. This indicates that, by including training instances with no facial deformations caused by speaking, the models can pick up other cues that are consistent regardless of speaking status, which helps detect more actual *happy* and *pensive* instances. Specifically, *pensive* always obtains the highest accuracy overall, with a slight performance increase when training on WH with models including G. The other classes also increase their accuracy when training with all data, although the highest accuracy is obtained from including A, which can only be accomplished when training with speech instances. Contrary to the effect of adding more training data with cross-country samples, incorporating silence instances for SP→SP mainly causes an increase in variability for *neutral* instances, since the increase in *happy* and *pensive* instances is relatively low (95% increase for *neutral*, 37% *happy*, 7% *pensive*), and due to the subsampling technique only 15% of the *neutral* instances are used during training. By contrast, for WH→SP, the number of *neutral* instances is doubled (with similar relative increase percentages as with SP→SP), but only 7.8% of such instances are used. Unlike other countries, NO obtains the highest accuracy for *pensive* with F+G instead of with G for WH→NO. For F-based models, WH→NO training on speech and silence achieves the highest balance among class accuracies, while for G alone it is better when training on speech only, despite *happy* achieving lower accuracy. Nonetheless, if we are interested in being better at detecting the minority classes even with a slight increase in false positives, then training on speech and silence would be recommended.

Statistical tests confirm significant differences for the following cases. For FR→FR, F, F+G, F+G_{vc}, F+G_{3dg}, and F+G_{eye}, at various p-value levels ($p<.05$). For WH→FR, F, F+G, and all F+G_s variants, at various p-value levels ($p<.05$). For NO→NO, F+G ($p=.035$), F+G_{vc} ($p=.029$), F+G_{3dg} ($p=.008$), and F+G_{eye} ($p=.02$). No significant differences are found for WH→NO, WH, and SP.

6.5.5 Video-based emotion expression recognition under silence

Tables 6.18, 6.19, 6.20, and 6.21 show the results for video-based labels under silence trained and evaluated on WH, SP, FR, and NO, respectively, using two different training regimes: 1) training on samples where the user is not speaking (*silence data*); and 2) training on all samples irrespective of speaking status (*silence+speech*).

Main modality

For WH, results match the video-under-speech scenario on average and per class, with a decrease in stability. The *neutral-pensive* confusion also decreases. Country-wise, trends are similar to the video-under-speech scenario but differ from WH. For SP, *pensive* is the top performer (73.3% accuracy) followed closely by *neutral*, while

TABLE 6.18: Video-based results under silence on WHOLE, reported as unweighted average accuracy ± SEM over 10 folds and three runs per fold. Bold: best accuracy per group. Underlined: best accuracy per training type. Italics: best accuracy overall.

Modality	Neutral Accuracy	Happy Accuracy	Pensive Accuracy	Average Accuracy
Training on silence data:				
F	73.33 ± 1.8	*74.33 ± 3.2*	57.76 ± 2.8	68.47 ± 1.5
G	61.3 ± 1.6	48.15 ± 3.4	*74.09 ± 1.2*	61.18 ± 0.8
F+G	79.45 ± 1.3	72.76 ± 3.2	71.74 ± 1.8	*74.65 ± 1.2*
F+G_{vc}	77.49 ± 1.4	72.41 ± 3.5	**70.23 ± 1.8**	**73.38 ± 1.3**
F+G_{3dg}	**78.21 ± 1.3**	**73.32 ± 3.1**	65.72 ± 2.7	72.42 ± 1.3
F+G_{eye}	76.6 ± 1.3	73.3 ± 3.1	63.62 ± 3.2	71.17 ± 1.4
F+G_h	74.85 ± 1.7	73.31 ± 3.1	58.47 ± 2.6	68.87 ± 1.4
Training on all data (speech + silence):				
F	79.56 ± 1.4	72.3 ± 3.4	51.35 ± 2.8	67.74 ± 1.5
G	67.34 ± 1.4	43.54 ± 3.4	72.3 ± 1.5	61.06 ± 1.1
F+G	*85.23 ± 1.0*	71.35 ± 3.4	61.98 ± 1.8	72.85 ± 1.1
F+G_{vc}	**83.33 ± 1.1**	71.5 ± 3.4	**60.79 ± 1.9**	**71.88 ± 1.2**
F+G_{3dg}	82.7 ± 1.1	71.67 ± 3.4	55.7 ± 2.7	70.02 ± 1.3
F+G_{eye}	82.48 ± 1.1	71.55 ± 3.4	54.88 ± 2.8	69.63 ± 1.3
F+G_h	82.27 ± 1.2	**71.76 ± 3.3**	50.01 ± 2.7	68.01 ± 1.4

TABLE 6.19: Emotion recognition results for video-based labels trained on the SPAIN (left) and WHOLE (right) training subsets under silence only or speech and silence instances, and evaluated on the SPAIN test subset under silence (train→test), reported as unweighted average accuracy ± SEM over 10 folds and three runs per fold. Bold: best accuracy per group. Underlined: best accuracy per training type. Italics: best accuracy overall.

	SPAIN → SPAIN				WHOLE → SPAIN			
Modality	Neutral Accuracy	Happy Accuracy	Pensive Accuracy	Average Accuracy	Neutral Accuracy	Happy Accuracy	Pensive Accuracy	Average Accuracy
Training on silence data:								
F	69.42 ± 3.5	60.91 ± 4.9	73.3 ± 1.9	67.82 ± 2.1	77.2 ± 2.8	*67.55 ± 4.8*	65.57 ± 3.0	70.07 ± 1.9
G	66.46 ± 1.9	33.21 ± 5.7	80.01 ± 3.0	60.67 ± 1.8	70.72 ± 2.0	27.23 ± 4.1	*78.35 ± 3.2*	59.87 ± 1.2
F+G	80.61 ± 2.0	56.19 ± 5.7	*84.57 ± 1.1*	*74.35 ± 1.9*	84.92 ± 1.4	64.35 ± 5.2	75.85 ± 1.6	*75.4 ± 1.8*
F+G_{vc}	78.02 ± 2.1	60.72 ± 5.1	81.85 ± 1.3	**73.89 ± 1.9**	82.4 ± 1.6	65.14 ± 5.1	**74.27 ± 1.8**	74.27 ± 1.8
F+G_{3dg}	**78.27 ± 2.2**	56.51 ± 5.8	82.4 ± 1.4	72.97 ± 2.0	**84.81 ± 1.5**	64.79 ± 5.2	74.22 ± 1.8	**75.08 ± 1.8**
F+G_{eye}	77.97 ± 2.6	57.73 ± 5.7	**82.79 ± 1.6**	73.37 ± 1.9	82.31 ± 1.8	64.51 ± 5.2	72.93 ± 2.4	73.73 ± 1.8
F+G_h	70.54 ± 3.4	**61.3 ± 5.2**	73.31 ± 1.6	68.34 ± 2.1	79.11 ± 2.7	**66.1 ± 4.9**	65.51 ± 3.0	70.31 ± 1.9
Training on all data (speech + silence):								
F	77.84 ± 3.0	62.22 ± 5.0	63.18 ± 2.0	67.71 ± 2.0	83.54 ± 2.2	65.17 ± 5.2	59.59 ± 3.1	69.63 ± 1.9
G	67.14 ± 1.6	40.78 ± 6.1	75.08 ± 3.1	61.42 ± 2.2	76.29 ± 1.7	35.47 ± 5.6	75.01 ± 3.2	63.1 ± 1.7
F+G	*86.71 ± 1.5*	58.06 ± 5.6	73.55 ± 1.6	73.33 ± 2.0	*89.6 ± 1.2*	61.39 ± 6.1	66.72 ± 2.0	72.92 ± 2.0
F+G_{vc}	**84.19 ± 2.0**	***64.57 ± 4.4***	70.3 ± 1.9	**73.24 ± 1.9**	87.75 ± 1.4	**64.25 ± 5.2**	**65.62 ± 2.4**	**72.94 ± 1.9**
F+G_{3dg}	83.42 ± 2.0	61.29 ± 5.4	**70.44 ± 1.5**	72.03 ± 2.0	**88.05 ± 1.3**	62.99 ± 5.6	63.49 ± 2.7	72.02 ± 2.0
F+G_{eye}	82.73 ± 2.3	61.77 ± 5.4	69.77 ± 1.8	71.74 ± 1.9	87.26 ± 1.4	63.43 ± 5.3	64.36 ± 3.1	72.23 ± 2.0
F+G_h	80.63 ± 2.6	63.03 ± 4.7	61.16 ± 2.0	68.34 ± 1.9	86.82 ± 1.8	62.1 ± 6.1	57.31 ± 3.1	69.17 ± 2.1

for FR and NO, the top performer is *neutral* by a great margin (around 78-80% accuracy). For NO specifically, *pensive* obtains extremely low accuracy (38.3%). As usual,

TABLE 6.20: Emotion recognition results for video-based labels trained on the FRANCE (left) and WHOLE (right) training subsets under silence only or speech and silence instances, and evaluated on the FRANCE test subset under silence (train→test), reported as unweighted average accuracy ± SEM over 10 folds and three runs per fold. Bold: best accuracy per group. Underlined: best accuracy per training type. Italics: best accuracy overall.

	FRANCE → FRANCE				**WHOLE → FRANCE**			
Modality	**Neutral Accuracy**	**Happy Accuracy**	**Pensive Accuracy**	**Average Accuracy**	**Neutral Accuracy**	**Happy Accuracy**	**Pensive Accuracy**	**Average Accuracy**
Training on silence data:								
F	78.16 ± 2.4	*48.13 ± 5.0*	50.35 ± 3.2	58.88 ± 1.6	73.36 ± 2.4	*72.74 ± 4.6*	59.29 ± 2.4	68.46 ± 2.0
G	52.24 ± 2.7	36.49 ± 4.3	*79.54 ± 3.2*	56.09 ± 1.8	56.21 ± 3.0	40.36 ± 4.3	*76.58 ± 3.1*	57.72 ± 1.4
F+G	84.16 ± 1.8	47.39 ± 5.1	72.58 ± 2.2	68.05 ± 1.7	78.07 ± 2.0	70.85 ± 4.4	74.41 ± 2.8	*74.44 ± 2.1*
F+G_{vc}	81.21 ± 1.9	**47.94 ± 5.3**	**75.4 ± 2.0**	***68.18 ± 1.7***	76.18 ± 2.1	70.48 ± 4.8	**72.66 ± 3.0**	**73.11 ± 2.3**
F+G_{3dg}	**83.11 ± 2.1**	46.44 ± 5.1	62.58 ± 3.2	64.04 ± 1.6	**76.54 ± 2.0**	71.45 ± 4.3	67.49 ± 2.7	71.83 ± 1.8
F+G_{eye}	80.55 ± 2.4	47.17 ± 4.8	58.55 ± 3.0	62.09 ± 1.5	74.13 ± 2.2	71.39 ± 4.3	68.0 ± 2.1	71.18 ± 1.7
F+G_h	78.98 ± 2.6	47.04 ± 5.1	54.11 ± 2.7	60.04 ± 1.6	75.05 ± 2.3	**71.71 ± 4.4**	58.21 ± 2.3	68.32 ± 2.0
Training on all data (speech + silence):								
F	84.31 ± 1.9	47.34 ± 5.0	41.41 ± 3.3	57.69 ± 1.5	78.68 ± 2.2	70.71 ± 4.7	51.84 ± 2.3	67.08 ± 1.9
G	59.74 ± 2.6	36.14 ± 5.0	75.58 ± 3.0	57.15 ± 1.6	63.11 ± 3.1	42.56 ± 5.8	75.53 ± 2.8	60.4 ± 0.7
F+G	*89.57 ± 1.7*	45.74 ± 5.0	53.66 ± 3.3	62.99 ± 1.6	*82.88 ± 2.0*	69.87 ± 4.5	64.24 ± 2.8	72.33 ± 2.1
F+G_{vc}	**88.9 ± 1.7**	**47.81 ± 5.2**	**49.34 ± 3.5**	**62.02 ± 1.6**	**81.21 ± 1.9**	70.14 ± 4.6	**64.31 ± 3.2**	**71.89 ± 2.2**
F+G_{3dg}	86.64 ± 1.8	46.17 ± 5.1	48.96 ± 3.3	60.59 ± 1.5	80.19 ± 2.1	70.28 ± 4.5	56.41 ± 2.7	68.96 ± 1.8
F+G_{eye}	87.1 ± 1.8	45.06 ± 4.7	45.46 ± 3.3	59.2 ± 1.4	79.69 ± 2.1	70.04 ± 4.5	56.07 ± 2.2	68.6 ± 1.8
F+G_h	86.8 ± 1.8	46.44 ± 5.0	40.01 ± 3.2	57.75 ± 1.5	80.07 ± 2.1	**70.4 ± 4.4**	49.72 ± 2.4	66.73 ± 1.9

TABLE 6.21: Emotion recognition results for video-based labels trained on the NORWAY (left) and WHOLE (right) training subsets under silence only or speech and silence instances, and evaluated on the NORWAY test subset under silence (train→test), reported as unweighted average accuracy ± SEM over 10 folds and three runs per fold. Bold: best accuracy per group. Underlined: best accuracy per training type. Italics: best accuracy overall.

	NORWAY → NORWAY				**WHOLE → NORWAY**			
Modality	**Neutral Accuracy**	**Happy Accuracy**	**Pensive Accuracy**	**Average Accuracy**	**Neutral Accuracy**	**Happy Accuracy**	**Pensive Accuracy**	**Average Accuracy**
Training on silence data:								
F	80.91 ± 2.6	55.41 ± 4.9	38.3 ± 5.0	58.21 ± 2.8	64.02 ± 3.6	*65.08 ± 4.4*	58.04 ± 5.9	62.38 ± 2.5
G	65.2 ± 3.6	37.64 ± 5.3	62.78 ± 5.5	55.21 ± 2.5	47.0 ± 2.2	58.26 ± 6.0	69.0 ± 5.7	58.09 ± 3.2
F+G	84.78 ± 1.9	53.46 ± 4.7	46.75 ± 5.1	61.66 ± 2.9	69.08 ± 2.7	63.54 ± 5.2	*71.04 ± 5.4*	*67.89 ± 2.6*
F+G_{vc}	80.57 ± 2.3	53.29 ± 4.7	**54.1 ± 6.0**	***62.65 ± 2.9***	**67.43 ± 2.9**	63.57 ± 4.9	**69.79 ± 5.6**	**66.93 ± 2.7**
F+G_{3dg}	80.6 ± 2.4	***58.9 ± 4.4***	41.72 ± 5.1	60.41 ± 2.7	66.08 ± 3.1	63.27 ± 5.3	60.24 ± 5.5	63.2 ± 2.5
F+G_{eye}	**82.99 ± 2.3**	55.26 ± 5.0	38.39 ± 4.8	58.88 ± 2.9	67.23 ± 2.9	**64.81 ± 4.7**	55.84 ± 5.5	62.63 ± 2.5
F+G_h	78.75 ± 2.5	54.83 ± 4.6	40.51 ± 4.7	58.03 ± 2.6	64.51 ± 3.4	64.18 ± 4.6	59.3 ± 5.7	62.66 ± 2.4
Training on all data (speech + silence):								
F	85.62 ± 1.9	51.65 ± 4.9	35.61 ± 4.9	57.63 ± 2.8	71.36 ± 3.0	62.84 ± 4.7	52.04 ± 5.5	62.08 ± 2.4
G	71.26 ± 3.3	20.84 ± 5.1	*66.47 ± 6.2*	52.85 ± 2.4	53.96 ± 3.2	58.23 ± 4.8	67.88 ± 5.2	60.02 ± 3.1
F+G	88.33 ± 1.4	50.9 ± 5.0	48.67 ± 4.8	62.63 ± 2.8	*77.71 ± 2.0*	62.69 ± 4.9	62.77 ± 5.5	67.72 ± 2.5
F+G_{vc}	86.87 ± 1.4	50.51 ± 4.7	**49.48 ± 5.4**	**62.29 ± 2.9**	**75.36 ± 2.4**	62.1 ± 4.8	**63.83 ± 5.8**	**67.1 ± 2.6**
F+G_{3dg}	88.05 ± 1.5	51.28 ± 5.0	38.67 ± 5.0	59.33 ± 3.0	73.63 ± 2.5	62.48 ± 4.9	53.33 ± 5.4	63.15 ± 2.4
F+G_{eye}	***89.04 ± 1.4***	51.0 ± 5.0	35.48 ± 4.6	58.51 ± 2.9	74.86 ± 2.3	**63.0 ± 4.8**	50.34 ± 5.2	62.73 ± 2.3
F+G_h	85.59 ± 1.7	**51.82 ± 4.8**	38.89 ± 4.7	58.77 ± 2.7	74.03 ± 2.7	62.55 ± 5.0	51.51 ± 5.4	62.69 ± 2.4

the latter is mostly misclassified with *neutral*, but also with *happy* to a lesser extent for

such country, a behavior we had not observed until now. Compared to the video-under-speech scenario, the proportion of *neutral* instances under silence is higher across countries (around 93-98% depending on the country), which might explain the accuracy bias. For the other classes, however, accuracy results do not generally match class proportions, with *happy* corresponding only to 1.4% of the total sample size for WH but scoring very close to *pensive*.

For WH, the SD across folds is 8.5%, slightly higher than the speech scenario. The SD across runs has a fairly larger range, with an approximate mean of 0.1%.

Auxiliary modality

For WH, and following previous trends, G alone achieves the highest performance and discriminative power for *pensive*. The average accuracy for G gets the closest to F compared to audio and video-under-speech scenarios. G is more stable than F on average for *neutral* and *pensive*, while for *happy* is not.

These trends are also found across countries. More specifically, for SP, G outperforms F for *pensive* and is the top performer for such class for all training configurations except for SP→SP when training with silence data, for which F+G outperforms G for this class. By contrast, *happy* performance is quite low. For FR, the average performance difference between F and G is even smaller than for SP. For NO, G alone obtains the highest accuracy for *pensive* for all scenarios except for WH→NO training on silence, for which F+G slightly outperforms it. The difference between F and G on average is moderately larger than for FR but smaller than for SP. Interestingly, for NO, *pensive* is misclassified as *happy* at a similar proportion compared to the main modality (F).

We find statistically significant differences for the following cases. For SP: F vs F+G for SP→SP (p=.041 and p=.046 when training on silence, and speech and silence, respectively); for FR→FR, G vs F+G when training on silence (p=.009); for WH→FR, F vs G (p=.012) and G vs F+G (p<.001) when training with silence data, and G vs F+G (p=.008) when training on silence and speech. No significant differences are found for WH and NO.

Multimodality

Adding G to F consistently increases accuracy on average, following previous trends class-wise, i.e., increasing accuracy and stability for *neutral* and *pensive*, while not contributing for *happy*, for which F alone is usually the top performer. As a matter of fact, *pensive* is the most benefited class for WH and across countries, and *neutral* obtains the highest accuracy overall across training regimes.

However, for *happy*, at least one of the different F+G_s variants outperform F+G for all countries, and sometimes even F, which is similar to the video-under-speech evaluation scenario. For instance, F is surpassed by F+G_h when training with silence data and F+G_{vc} when training with silence and speech for SP→SP. For FR, G is only beneficial for *happy* with F+G_{vc} for FR→FR when training with speech and silence, which marginally outperforms F alone. For NO, F+G_{3dg} obtains the highest accuracy for *happy* for NO→NO, but F is the top performer for WH→NO.

For WH, no statistically significant differences are found, likely due to the high SEM of F alone, which decreases with multimodality. For SP, we find F vs F+G statistically different when WH→SP (p=.041). For FR, F+G surpasses F on average significantly (p=.001 for FR→FR, and p<.001 for WH→FR). And for NO, we find

statistically significant differences for all pairwise comparisons ($p<.01$) except on NO→NO training on silence data.

Eye and head feature subsets

Similarly to the video-under-speech evaluation, the G_s feature subsets do not outperform the complete G feature set for multimodal models on average (except F+G_{vc} for FR→FR and NO→NO training on silence data). The same is true for *pensive* and *neutral* (except F+G_{eye} for NO→NO for the latter). For *happy*, the reduced feature subsets modestly achieve a better result than F+G, while for all countries, the same is true compared to F, as previously commented. We also observe that, again, G_{vc} outperforms the other variants on average and for *pensive* by a large margin, followed by *neutral*. For SP, however, and in line with previous video-based results, F+G_{eye} scores higher than G_{vc} to recognize *pensive*, with the three gaze-based subsets obtaining very similar accuracy. Additionally, for *neutral*, the highest accuracy is obtained by F+G_{vc} or F+G_{3dg} for all cases except for NO→NO, where F+G_{eye} is the top performer, the importance of which slightly decreases when training on WHOLE.

For SP, the difference among the different gaze subset versions for *happy* increases with respect to WH. As a matter of fact, while the top-performer subset for *happy* is G_h when training with silence data, the top performer becomes G_{vc} when training with silence and speech data, obtaining the highest accuracy among all models for SP→SP. Although these two subsets obtain higher accuracy than F for SP→SP for *happy*, for WH→SP F alone is enough to recognize this emotion with the highest accuracy. For this class, G_{vc} is also the top performer for FR→FR, while for WH→FR, G_h takes over.

Statistical tests confirm significant differences for the following cases. For WH: G_{3dg} vs G_h ($p=.009$ when training with silence, and $p=.002$ when training with speech and silence), and G_{vc} vs G_h ($p=.037$ when training with speech and silence). For SP: for G_{vc} vs G_h when training with speech and silence data for SP→SP ($p=.013$) and WH→SP ($p=.041$), and when training on silence only ($p=.013$); G_{3dg} vs G_h for WH→SP when training on speech and silence data ($p=.041$), and G vs G_h also for WH→SP ($p=.041$ when training on silence only and $p=.046$ when training on speech and silence). For FR: 72.5% of all pairwise comparisons are significantly different at various p-value levels ($p<.05$), including G_h vs all other subsets for all configurations, and G vs all gaze feature subsets except for G_{vc}. And for NO: G vs G_h/G_{eye} when NO→NO, and G_{vc} vs the other feature subsets, and G vs G_h/G_{eye}/G_{3dg} when WH→NO, all at various p-value levels ($p<.05$).

Comparison across countries

Figure 6.7b depicts per-country average accuracy results, complementing Tables 6.19, 6.20, and 6.21. As can be seen, all countries benefit from multimodality, with FR benefiting the most when training on silence data. SP achieves the highest accuracy for all settings by a large margin, obtaining the best accuracy overall with F+G (74.35%) when training with silence data. Contrary to the video-under-speech evaluation, NO scores the lowest. With respect to the auxiliary modality G, its difference with respect to F is small for FR for both training regimes and for NO only when training with silence, as the performance of G decreases when training with all data for the latter country. Class-wise, trends vary compared to WH depending on the modality, as already discussed above. Again, one highlight is the low accuracy and stability for *pensive* with either modality for NO compared to

the other countries and classes, as well as its confusion with *happy*. In fact, NO is the country with the highest number of *pensive* instances (around 5.4% of the NO sample size), which might indicate a difference in the annotation procedure.

Expanding training data including other countries

Figure 6.7b and Tables 6.19, 6.20, and 6.21 also show the effect of training with WH. Similarly to the video-under-speech evaluation, SP barely benefits from adding cross-country data compared to the other countries (accuracy mostly increases when training on silence data only, except for G). Still, the highest accuracy is obtained by SP with F+G (75.4%) when training with silence. For FR, the highest performance increase is obtained by F, while for NO, G is the most beneficial modality.

Class-wise, we again observe differences between SP and the other two countries. For NO and FR, F-based models increase performance for *happy* and *pensive* and reduce it for *neutral*, while for SP, *neutral* and *happy* performance increases while *pensive* performance decreases. We conclude that *happy* is easier to recognize for F-based models in all countries when no facial deformations caused by speaking occur. However, for SP, we observe an increase in confusion between *neutral* and *pensive* when training with WH, which could be explained by a difference in user behavior or annotation procedure between SP and the other countries when users do not speak. This, in turn, might explain their difference in *neutral-pensive* ratios (61:1 for SP, 33:1 for FR, and 17:1 for NO), compared to their similarity when users speak (around 5:1 for all countries).

NO shows a different trend compared to the other countries for *pensive* with G when training with WH. More concretely, for SP and FR, *pensive* performance decreases when training with silence and it is maintained when training with speech and silence, while for NO, its performance increases for both cases. However, we again observe that this increase in performance also comes with a slight increase in confusion with *happy*. The performance decrease for *pensive* for FR and SP is opposed to the general increase observed for this class in the video-under-speech scenario.

For FR, statistical tests confirm significant differences for all models (p=[.01,.03] when training with silence and $p<.01$ when training with silence and speech) except G, and G_{vc} only when training with silence and data. For NO, only the increase for G when training on speech and silence is statistically significant (p=.03). No statistically significant differences are found for SP.

Expanding training data including speech instances

As shown in Table 6.18 for WH, training with speech and silence instances slightly reduces the performance of all models on average except for G, for which accuracy is maintained but stability decreases. Class-wise, *neutral* increases performance and *pensive* decreases. For F-based models, this effect is caused by an increase in confusion between *neutral* and the other two classes since the facial deformations added during training decrease discriminative power, while for G, confusion with *neutral* mostly increases for *happy*.

Figure 6.7b allows us to compare cross-country performance, complementing Tables 6.19, 6.20, and 6.21. As observed, this training regime slightly reduces the performance of all F-based models, except NO→NO with F+G. By contrast, performance is increased for all G models except for NO→NO. Class-wise, WH trends are maintained across countries for *neutral* and *pensive* for all models except for G with NO→NO, for which *pensive* increases accuracy. On the contrary, *happy* performance

is improved for SP→SP with F-based models, but its stability decreases, probably caused by the smaller number of *happy* instances for SP. Additionally, for SP and FR, *happy* also increases with G, showing similar behavior. By contrast, *happy* is extremely confused with *neutral* for NO when adding speech instances. The difference in overall behavior for NO is unclear. The general performance decrease of *pensive* could be caused by the high difference in the number of instances between speech and silence sets (808% increase for WH when including speech instances, and similarly high per country), causing the models to learn patterns more tailored to *pensive* episodes while speaking. This issue may also be causing part of the performance deterioration for *happy*, and it is the opposite effect observed for the video-under-speech scenario when training on all data. In general, and in contrast to the video-under-speech scenario, per-class accuracies tend to be more balanced when trained on silence only for F-based models, but also when trained on WH. For G, it is harder to categorize due to *pensive* being the main discriminated class. Nonetheless, this auxiliary modality tends to perform best on average when training on all data, indicating that it benefits from the added variability of different countries and speaking status, regardless of the evaluation scenario.

Statistical tests confirm significant differences for G_{3dg} ($p=.043$) for WH→SP, and for all FR comparisons at different p-value levels ($p<.05$) except for G. For NO and WH, we find no significant differences.

Effect of speaking status on video-based evaluation

For WH, the users' emotional expressions can be better recognized when they are not speaking with G and F-based models on average. Indeed, evaluating on silence instances with no facial deformations that add noise should increase the discriminative power for *happy* and *pensive*. However, G should not be directly affected by facial deformations, so this leads us to believe that gaze and head patterns may be correlated with speaking status beyond facial deformations, and that such patterns are more discriminative for silence instances. Country-wise, G also tends to work better when the user is silent. By contrast, trends differ for F-based models depending on the country and the training regime.

Class-wise, *neutral* always obtains the highest accuracy for silence samples trained on silence and speech per country, along with a higher number of false positives. *Happy* is easier to recognize during silence instances for SP and FR with F when training on WH with silent instances. For NO, *happy* is easier to recognize during speech instances with models that combine F and A, also when training with WH, highlighting again the importance of A for this country. For WH, both scenarios are equally performant. Finally, *pensive* is easier to discriminate when evaluating on silence instances for FR, SP, and WH, although it is difficult to determine which training regime is best due to fluctuating confusion patterns. By contrast, for NO, this class is easier to categorize when the user is speaking with F+G when training on WH with speech instances. Note that, with respect to sample size, both speech and silence subsets are given in around 1:1 ratio for *neutral* and 2:1 for *happy*, while for *pensive*, ratios range from 14:1 for SP to 4:1 for NO.

6.6 Discussion

In this section, we summarize the results obtained individually for each research question (Section 6.5.1), and discuss limitations and potential future work.

Q_1) Discriminative power of the main modality

We observed varying levels of confusion between the minority classes and the majority class for both label types, with the highest confusion being observed with *pensive*. We strongly believe that this is mostly caused by the FER model considering only spatial information, since this class is mostly characterized by specific dynamics. Consequently, it is the most benefited when F is combined with G, which does provide salient dynamics.

Nevertheless, the minority classes tend to be well discriminated against each other. Despite employing a balancing strategy, per-class accuracies seem to be associated with their corresponding sample sizes. It is important to note, however, that the majority class in both label types (*neutral* and *calm*) is characterized by a relatively lower intensity or absence of emotional expression and can be encountered in transitional phases between other emotional expressions. This poses a challenge in establishing clear category boundaries. All things considered, average accuracies are around 60-70%, which is a decent rate for a three-class unimodal classifier given the nature of our scenario. This could be further improved by leveraging temporal dependencies across time steps (Tzirakis et al., 2017; Filali Razzouki et al., 2023).

Q_2) Performance of auxiliary modalities compared to the main modality

This depends on the label type. For the audio-based scenario, G struggles to achieve sufficient accuracy. By contrast, F does a better job, usually recognizing *pleased* with similar or higher accuracy and discrimination power than A alone, particularly due to the similarity between *happy* and *pleased* facial features. Conversely, for the video-based scenario, both A and G achieve accuracies closer to the main modality, especially when evaluating on silence instances. In particular, G alone is able to recognize *pensive* with extremely high accuracy, always better than F alone. In addition, in some cases, A is capable of obtaining higher accuracies than F for *pensive* and *happy*, being especially informative for the latter.

Achieving such recognition rates using only the auxiliary modalities proves advantageous in various scenarios. For instance, network or sensor failures may produce potential asynchronies between audio and video streams or even data loss. Having channel-specific labels already allows for individual processing. Therefore, if one stream is affected or even deactivated to reduce processing times, the system can still maintain functionality by utilizing the auxiliary modalities. Another example involves privacy considerations, wherein the video modality may be intentionally altered or disabled to ensure anonymization. In the case of alteration, G could still be extracted; however, in the event of video deactivation, only A would remain useful, from which the important video-based events could still be recognized. Nonetheless, as a prospective direction, it is worth considering crossmodal training techniques (Abdou et al., 2022), which learn from multiple modalities at training time to improve single-modality recognition during inference.

Q_3) Multimodality

Multimodality has proven beneficial, provided that the modality that is added to the main modality provides discriminative information for a given class. This is the case when adding F to A for the audio-based scenario for the minority classes, and when adding A or G to F for the video-based scenario for all classes, except *happy* for G, although most (F+A+)G_s subsets outperform (F+A+)G for this class. Results also depend on the distribution of features. In addition, the number of network

parameters usually increases with the number of input features, which may lead to overfitting if not accounted for. We have not evaluated the contribution of the combination of the auxiliary modalities without adding the main modality, which we leave for future work.

The increase in accuracy when combining A and F is in line with the large audio-visual emotion recognition literature (Poria et al., 2017) and with the few works addressing our target population (Ma et al., 2019). Compared to A+F, it is difficult to contextualize the A/F+G results within the literature, as the conclusions depend on the scenario and features used, and most speech-video-gaze/head research employs the VAD model instead of a categorical one (O'Dwyer, Murray, and Flynn, 2018). As an example of discrete emotion recognition, our results resemble the findings of the aforementioned crossmodal work (Abdou et al., 2022), for which including gaze features also improved performance for video-test, while for audio-test only one of their gaze alternatives outperforms the no-gaze option. Class-wise, they also found that the improvement was higher for video than for audio when using gaze, and observed a similar gaze spatial distribution between *neutral* and *happy* instances.

Ideally, however, a multimodal system should effectively disregard irrelevant information from individual modalities, thereby preventing any detrimental impact on overall performance. In that sense, simple feature concatenation may not be a solution to this problem, as it may fail to capture the interactions between features adequately. In contrast to concatenation, attention-based methods are known to adaptively balance the contributions of different modalities (Guo, Wang, and Wang, 2019). Nonetheless, our preliminary experiments found no differences among fusion types. Although it is possible that both approaches do yield similar results in our scenario, 1) the stage-wise training, and 2) the fixed modality synchronization applied are factors that may have influenced the outcome. For the former, end-to-end training would allow the features of different modalities to jointly evolve from early network layers instead (Tzirakis et al., 2017). For the latter, a flexible temporal synchronization able to capture long-term crossmodal dependencies would compensate for the unaligned nature of communication (Tsai et al., 2019).

Q_4) Contribution of G_s features

Similarly to Q_2, this depends on the label type. For the audio-based scenario, the combination of G_s subsets provided redundant information and increased the effective number of training parameters, which negatively affected performance. All individual subsets outperformed the complete G feature set but with marginal differences among subsets, with G_{vc} and G_h usually standing out. In comparison, the combination of G_s subsets provided complimentary information for the video-based scenario, thus always obtaining the highest accuracy with the complete G set. On average, G_{vc} is always the top performer among subsets, mainly due to the relationship between *pensive* and gaze dynamics during thinking episodes, while G_h scores last. Indeed, if the setup is partly calibrated and/or the VC location is known, G_{vc} can be computed quickly and precisely from raw gaze estimates, and be used to infer the direction of overt attention and emotional expressions (*pensive* at least) with high accuracy. However, if the VC location is unknown or cannot be estimated properly, using this feature set can be error prone, as the VC location estimation can introduce noise. In such cases, G_{3dg} or G_{eye} can be used instead for emotion recognition.

Recent works have aimed to identify the most important gaze features for different emotion categories. For instance, Abdou et al. (2022) identified quartile and median statistics of gaze angles as important for recognizing *happy*. However, in

our case, G (or G_{3dg}/G_{eye} specifically) have been found not to contribute toward the recognition of this class (only in a handful of cases have they marginally outperformed F), suggesting that such findings depend on the considered emotion categories, setting, and/or age group. Further research is needed to identify which specific features of these G_s subsets are more informative for our scenario to minimize the number of features while maintaining informative power, both when using gaze and head information to infer emotional expressions individually or in combination with other modalities.

Q_5) Cross-country differences

There are numerous differences that have been previously discussed and will not be reiterated here, which have exposed disparities in how emotions are expressed and perceived across countries. Such differences in behavior and annotation are also reflected in the divergence of the proportions of per-class sample sizes and distribution in the feature space. Nonetheless, results are subject to the imbalance in the number of samples among countries. This causes NO, the country with the lowest sample size, to obtain lower performances in general.

Q_6) Multicountry training

In general, main modalities and multimodal combinations benefit from multicountry training. With respect to auxiliary modalities, however, only G benefits from this training regime for the video-based scenario, indicating that the additional features are not as transferable cross-country as the main modalities, or that there is not enough data to capture all possible behavior manifestations.

Performance gains come mainly from increased variability, which especially benefits the minority classes. However, the extent of these improvements does not correlate with the increase in sample size, since such increase is proportionally similar for the audio-based scenario than for the video-based scenario, but gains are higher for the latter. This can be attributed to: 1) the audio-based scenario having fewer samples than the video-based scenario; 2) A features already being more generic since they come from the large-scale WavLM model, thus more variability may not have a high impact on the results; 3) some acoustic features of emotional expressions being more language- or culture-specific and hence less transferable across countries than facial expressions of emotion, which is in line with previous studies (Riviello and Esposito, 2012). Note that no sampling strategy has been applied to balance the number of data samples across countries, hence having a bias toward SP, the country with the highest number of samples. Nonetheless, as found in our experiments, the effect of multicountry training will always depend on the feature distribution and the cultural similarities across the countries considered.

Q_7) Training with spoken and silent instances

Training on all data consistently but marginally improves performance when evaluating on spoken instances, increasing discriminative power for the minority classes. This is due to the fact that the added variability, without the noise produced by speaking, helps the models learn discriminative features that are less influenced by the speaking effect. Interestingly, though, the performance obtained with this training regime is on par with that obtained when combining F and A trained on spoken instances only. By contrast, when evaluating on silent instances, training with all

data produces the opposite effect, as the added variability makes the recognition task harder. Note that there is no sampling strategy to balance the number of spoken/silent data, hence having a slight bias toward spoken samples, primarily for the minority classes. Nonetheless, these findings indicate that the speaking status significantly influences the learning process and should be taken into account when devising solutions similar to ours.

Q_8) Speech vs silence performance

This is highly country-specific. In general, and related to Q_7, both speaking statuses achieve similar accuracy on average, but the minority classes tend to be better discriminated against when the user is not speaking, provided the model was trained on silent instances only (note that the original F features were trained on both types of instances, so differences in performance will come exclusively from the final models). This conclusion is intuitive for F-based models since facial deformations affect discriminative power. However, since this difference in performance is also observed for G, this leads us to believe that gaze and head patterns may be correlated with speaking status beyond facial deformations, with such patterns being more discriminative for silence instances. Nonetheless, if A is combined with F used for spoken instances, both spoken and silent instances can obtain comparable results.

Q_9) Audio vs video performance

Audio-based results are constrained by the limited amount of data. While this effect is balanced for A by the use of a pretrained large-scale model, which may also aid in generalization ability, auxiliary modalities are indeed impacted. By contrast, the video-based evaluation can leverage sample sizes two orders of magnitude higher, but with higher redundancy among samples. Overall, we observe a higher performance for video-based evaluation than for audio. This can be attributed to the differences in the number of samples. An additional reason could be that WavLM was trained on English speech; therefore, even if the extracted features are generic enough due to the large pretraining, they are not specialized to the languages considered in this work, and hence important subtleties might have been lost.

6.6.1 Limitations

Apart from the limitations raised above, we identify further general limitations of this work, which we comment on below.

The findings of this work are subject to the choice of methods, data imbalance handling, and synchronization and data annotation procedures employed. Similarly, results might vary when using all the annotated emotion categories. In particular, the video-based category *other* might provide interesting insights as it includes instances of multiple categories taking place simultaneously; thus, we leave it for future work. The analysis carried out cannot disentangle the effects of users' culture and raters' annotation skills; therefore, performance differences across countries cannot be attributed to differences in culture entirely.

With respect to gaze-related features in particular, results are contingent on the functionals considered, which were selected based on prior research and on their potential contribution to our scenario. Furthermore, gaze and head pose estimates are noisy. Since we do not have a ground-truth subset to measure accuracy, we are

unable to determine how the quality of the estimates affects the accuracy for the downstream task. In relation to the dataset employed for training the gaze estimation model, we have identified three primary factors that might contribute to a decrease in robustness. Firstly, as previously commented, the dataset lacks representation of older adults, potentially causing a domain shift when applied in our scenario. Secondly, the dataset features a discontinuous range of head poses; thus, accuracy may be lower for samples that fall outside the covered poses. Lastly, the dataset was recorded with a relatively constant head-screen distance, resulting in minimal vergence variation (vergence is possible correlated to the target). This lack of variation poses a challenge for scenarios where users may interact with agents at different distances, as in our case, leading to variations in vergence that the model has not been trained to account for.

Finally, future work would require a real-time evaluation of the system using the methodology presented to assess the effect of possible network latencies, and how emotion recognition would impact the user-VC interaction both quantitatively and qualitatively, including user studies.

6.7 Conclusions

In this chapter, we presented a comprehensive study on non-verbal emotion expression recognition in interactions between older adults and a simulated VC within the context of the EMPATHIC project. We also described the rationale for data collection and annotation procedure aimed at developing a computational approach that could leverage cues from audio and video channels separately. By analyzing the influence of different modalities, training approaches, and communication modes, this research aimed to shed light on some of the factors that affect the effectiveness of emotion recognition in this scenario. Our findings demonstrate that facial, speech, head, and gaze cues can contribute to the accurate recognition of the channel-specific emotional expressions considered with varying levels of discriminative power. As the evaluation was conducted in a subject-independent manner, these cues would prove even more valuable for a personalized online setup, in which the model could continuously learn from the user's behavior during the interaction. Furthermore, we determined that multicountry training can generally compensate for limited data from a particular country, thereby enhancing overall performance despite country-specific differences.

In particular, with respect to gaze-related features, we have found that the considered feature sets do not provide discriminative information for the audio-based emotion categories considered when used alone, but can help increase accuracy when added to speech and facial expressions. On the contrary, such gaze features do provide discriminative information for the video-based scenario, both individually and when added to the other modalities. In addition, the gaze feature subsets provide redundant information for the audio-based scenario, while for the video-based scenario, the subsets complement each other. This motivates future work to find better feature sets that minimize the number of features while maximizing informative and discriminative power.

The insights gained from this work are expected to contribute to the development of more accurate emotion recognition systems, and pave the way for improved VC experiences and personalized technologies catering to the emotional well-being of this age group.

Part III

Closing remarks

Chapter 7

Discussion and Conclusions

In this concluding chapter, we summarize the work done and conclusions reached from this thesis (Section 7.1). Furthermore, we discuss some observations made during the thesis period and propose prospective directions for the gaze estimation community (Section 7.2). We also introduce future research lines that we have initially explored, stemming from the present investigation (Section 7.3). Finally, given the risks of ubiquitous gaze estimation, we discuss the potential societal impact and ethical implications (Section 7.4).

7.1 Conclusions

In this thesis, we approached the problem of subject-independent, appearance-based gaze estimation through two different but complementary strategies: leveraging spatiotemporal and multimodal (from the same or different sensors) information. We explored these strategies for the first time for this task, with the ultimate goal of:

1. Increasing gaze estimation accuracy and sampling rate, which are important aspects for existing and emerging eye tracking approaches for improving robustness and applicability in different settings (addressed in Part I);

2. Promoting the use of gaze input to improve the performance of existing and emerging CV/ML applications, with the example of emotion recognition in a conversational HMI scenario focused on older adults (addressed in Part II).

In particular, we defined the following three research questions to be answered in this thesis:

RQ_1. *Is temporal information beneficial for appearance-based gaze estimation?*

RQ_2. *Can the fusion of different modalities or sensors improve appearance-based gaze estimation performance, in terms of accuracy and/or sampling rate?*

RQ_3. *Is gaze-related information beneficial for emotion recognition in older adults, either alone or in combination with other modalities?*

Given the remarkable success of DL in the past decade across countless applications, including gaze estimation, we built our approaches throughout the thesis on such a robust foundation. In particular, we mainly relied on similar convolutional-recurrent networks, which are capable of exploiting spatial information from CNNs and temporal information from recurrent networks (LSTMs or GRUs). For multimodality, we relied on feature-level and hybrid fusion to exploit the strengths of the different information sources. We summarize the findings for each part and discuss the main limitations associated with them below.

7.1.1 Part I: Methods

First, we explored the spatiotemporal (RQ_1) and multimodal (RQ_2) strategies for gaze estimation in a remote, off-the-shelf RGB camera-based scenario. This setting introduces some challenges. First, a gaze estimation approach applied to such setting is expected to work with the same accuracy for a variety of people's appearances, head poses, lighting conditions, etc. In addition, the usage of regular color cameras is characterized by a lower sampling rate and a lower-resolution image of the eye (compared to the near-eye cameras used in dedicated desktop or wearable systems), which impacts the amount of spatiotemporal information that can be acquired from the eye. The evaluation was performed on both a screen-target task, characterized by a restricted range of head poses, as in most prior literature, and a free-moving-target task, less frequently explored due to its associated complexities, with a wider range of head poses. We found that adding shape cues from extracted 3D facial landmarks to the appearance information obtained from face and eye images slightly regularizes (i.e., decreases SD) gaze estimates. Certainly, adding geometric constraints in the form of landmarks comes at a negligible cost since landmark estimation is a necessary preprocessing step in gaze estimation to locate the head and eyes in the 3D space. We also showed that our proposed approaches, without (i.e., static) and with temporal (i.e., dynamic) information, outperformed previous state-of-the-art static methods, aided by other design choices such as the pretrained backbone used. We encountered that our dynamic approach outperformed the static one for the free-moving-target task, but not the screen-target one, influenced by the target's slower velocity and smaller range of movement in the latter. Furthermore, for the free-moving-target task, results favored the static approach when the head was fixed, suggesting that useful temporal information primarily stems from head movements, at least for our particular setting.

Following such findings, we continued with a detailed investigation of the relevance of eye movements for spatiotemporal gaze estimation (RQ_1). To do so, we moved from the remote camera-based setting to a near-eye one with IR cameras mounted on an HMD-VR device, which provided us with higher-resolution, higher-frequency eye images and reduced head movement effects. In this setting, model- and feature-based approaches were still more prominent. However, using a DL-powered end-to-end appearance-based gaze estimation approach proved to be a feasible alternative. Although head pose variability is no longer an issue in this setting (from an appearance perspective, as head movements do alter eye gaze dynamics), other challenges arise, such as headset slippage, specular reflections caused by eyeglasses, and of course, differences in subject appearance and eye geometry. We evaluated our hypothesis on a larger dataset than in the previous evaluation, with a higher number of participants and variability with respect to ethnicities, age, gender, and eye accessories. Results confirmed the utility of temporal information, with the estimates of a dynamic approach being less noisy and better following the ground-truth gaze trajectories than a static counterpart. In addition, we discovered that accuracy was higher for image sequences featuring a fixation followed by a saccade, suggesting that the dynamic approach takes advantage of the stability of the first frames of the sequence to better distinguish between noisy and informative features for the task. We found two specific contributions with respect to conventional glint- or pupil-based approaches. First, the addition of temporal information proved especially beneficial for the vertical component of gaze, which can cause problems when the eyelid occludes the pupil for previous approaches. Second, we did not detect a decrease in accuracy for participants wearing glasses.

In the third study, we continued with a near-eye setting to focus on RQ_2. Contrary to the first study, where different modalities computed from the signal of a single sensor (RGB camera) regularized the obtained accuracy in a remote-camera-based setting, here we leveraged the signal from multiple sensors (camera and photosensors at a high sampling rate) to increase both accuracy and effective sampling rate in a near-eye, portable eye-tracker setting. High-speed eye tracking is demanded for applications that require precise detection of fast and small eye movements. There had been initial attempts to fuse sensors with different quality and sampling rates for increasing the sampling rate of the final system, but not with a DL-powered approach. In addition, no publicly available datasets that offered time-synchronized signals from multiple sensors were available, which allows for the comparison of fusion approaches as a function of the sampling rate. For this reason, we built a new synthetic dataset of time-synchronized camera-photosensor image pairs featuring variability in illumination, subject appearance (including pupil size), sensor locations, and gaze directions. The dataset featured the three basic eye movement types in a game-based, screen-target task. We evaluated different feature-level fusion strategies with increased complexity for single- (both sensors operating at the same sampling rate) and multirate (each sensor operating at a different rate) operation on this simulated setting, assuming perfectly synchronized sensors and zero system latency as a first baseline. We confirmed that camera-based gaze estimation is quite robust to different sources of variability, while photosensor-based estimation struggled with such variability, particularly with extreme changes in sensor location and rotation. Despite that, we found that adding photosensor information regularized camera-only results in the single-rate setting, similar to our first study. Furthermore, in the multirate setting, the multisensor approach was able to better follow the ground-truth traces and improve gaze estimation accuracy, especially during fast eye movements, in comparison to a camera-only gaze forecasting approach.

With the first two studies, we can answer RQ_1: we confirm that temporal information is indeed beneficial for appearance-based gaze estimation, being a promising cue to exploit for near-eye scenarios in general, and for remote-camera scenarios at least when free head movement is allowed. Furthermore, with the first and third studies, we can also positively confirm RQ_2, that is, that multimodal information can be useful for gaze estimation to improve its performance in the settings considered. This proves yet another promising cue to exploit with either single-sensor (camera) approaches to increase accuracy from image-based modalities, or with multisensor approaches to increase accuracy and/or sampling rate.

7.1.2 Part II: Applications

Progress in eye-tracking technology has consistently been driven by the opportunities it can unlock for various applications. In Part II, we focused on an emerging application that non-intrusive, remote-camera-based gaze estimation can facilitate: emotion expression recognition during the interaction with a VC, as a type of conversational HMI scenario (RQ_3). The application was specifically focused on older adults, an age group that is scarcely considered in affective computing research and almost non-existent in gaze estimation research. We investigated the contribution of gaze-related features for the emotion recognition task, both individually and in combination with other modalities, namely speech from audio and facial expressions for video, using feature-based and hybrid fusion. This application did not require high gaze estimation accuracy and sampling rate, as long as gaze behavior

and direction changes could be sufficiently captured, but it did require robustness against common confounding appearance factors. In particular, we studied features related to looking-at-VC instances, and 3D gaze, eye-in-head rotation, and head pose trajectories. The last three feature sources consisted of functionals used in previous literature computed from smoothed trajectory estimates. We found that our features could improve emotion recognition accuracy in a multimodal setting with different levels of discriminative power for the emotion categories considered. More specifically, gaze alone was not useful for recognizing emotion categories annotated from the audio channel, although it helped increase performance when combined with the other modalities. In addition, the groups of gaze-related features that we studied turned out to provide redundant information with respect to each other for the task. By contrast, for categories annotated from video, gaze features provided complementary information, and were capable of reaching accuracies close to those obtained with facial expression features. This indicated that the contributing feature groups and specific functionals depend on the emotion category to be recognized.

With that, we can positively answer RQ_3: gaze-related information is beneficial for emotion recognition in older adults with respect to the emotion categories considered. This outcome is even more notable given the relatively greater challenge associated with recognizing emotional expressions in conversational HMI scenarios. It contributes to the growing body of evidence supporting the significance of gaze (and head) input as valuable cues for such task.

7.1.3 Limitations

Certain caveats accompany the confirmation of our research questions:

- RQ_1 and partially RQ_2: There were no illumination or (substantial) camera location changes during the input sequences used for the three studies. Although we have seen that our dynamic models (CNN coupled with LSTMs or GRUs) can learn from temporal correlations and sequential information, we have not evaluated the effect of such changes on the final model performance.

- RQ_1: Eye and head movement dynamics are task-dependent. Our evaluation was based on three artificial tasks provided by existing datasets. Therefore, it is yet unclear how dynamic models would perform in more ecologically valid settings, and whether dynamic models trained on one task would perform equally well on other tasks. As we observed, for screen-target tasks that elicit very little head or eye movements in remote-camera settings, a static approach might provide similar performance to dynamic ones.

- RQ_2: Multisensor fusion has been evaluated on a perfectly synchronized setting without sensor noise. System latency and real-world perturbations may affect the results.

- RQ_3: Similarly, the emotion recognition evaluation was carried out in a semi-controlled WoZ paradigm. The performance of such a system deployed in an actual household remains uncertain.

- RQ_1, RQ_2, and RQ_3: The addition of temporal and multimodal (or multisensor) information may introduce additional computational complexity and latency in the final system. The decision to incorporate these features should be contingent upon the specifications of the target device where the solution would be deployed, as well as the intended applications of the solution.

7.2 Observations and prospective directions

In this section, we discuss important observations made during this research period, which are of significance to both the eye tracking and CV/DL gaze estimation communities. Along with such observations, we outline potential future research directions, some of which stem from our study's findings and limitations.

Performance metrics

We observed high variability in gaze estimation results throughout the three methodological studies, stemming from differences in subject appearance (e.g., ethnicity, gender, age, face accessories, pupil size), illumination, and sensor placement, as well as from extreme head poses (which mainly affects remote-camera scenarios) and gaze directions. A high proportion of the gaze estimation literature, in particular the one coming from a CV/DL perspective, relies excessively on the average angular error as a singular metric. However, taking this variability into account, we emphasize that the average gaze error is not an informative metric and should not be used exclusively. Instead, we reported and analyzed this variability by means of dispersion and uncertainty measures (SD and SEM), 1D and 2D graphs representing gaze error as a function of the different variability ranges, and later percentiles such as p95, which better represent the behavior of the estimation models for more complicated cases throughout the entire population. The effect of specific sources of subject appearance variability has not been possible to assess due to privacy considerations. However, when including sequential information, head, and eye movement dynamics are also subject-specific, which can affect performance regardless of subject appearance. The eye-tracking literature usually includes other measures in addition to spatial accuracy, such as spatial precision (how close different measurements of the same gaze direction are to each other), which are important for quantifying robustness (Aziz and Komogortsev, 2022). We argue that precision should also be a routine metric for gaze estimation approaches coming from a CV/DL perspective. Depending on the dataset used, it might be difficult to have the same ground truth gaze direction for two different samples of the same or different subjects under different conditions. However, these samples can be perturbed and such impact on performance assessed, as preliminarily explored by Xu et al. (2021). In addition, given the spread of the obtained results, we also state that gaze estimation should move away from deterministic estimates toward incorporating uncertainty during modeling and output.

Data collection during eye and head movements

Another important observation is the quality of gaze ground truth when using video sequences to leverage temporal information. Data acquisition is already complicated and cumbersome for static poses, and the difficulty increases exponentially when collecting data during eye movements due to involuntary eye movements and latency in the target-following process. Relying on an imperfect ground truth poses an upper bound on the accuracy of the system. Future work should aim to devise novel protocols to record accurate ground-truth data while performing eye movements, ideally mimicking real-world tasks. Given the difficulty, simulation is an existing alternative that currently poses a trade-off between accuracy and photorealism, but new approaches that enhance such photorealism are emerging every day (Nair et al., 2020; Wood et al., 2021; Chaudhary et al., 2022; Deng et al., 2023).

Generalization ability

Appearance-based gaze estimation approaches require large-scale datasets to achieve generalization. Despite current methodological advances, cross-dataset evaluation (i.e., training a model on one dataset and evaluating it on a different dataset) often obtains poor generalization results, as datasets are usually tight to specific setups, and camera and noise configurations. This affects particularly remote-camera scenarios that rely on off-the-shelf cameras, as near-eye approaches are usually linked to specific device configurations and are not expected to work directly for other devices. Domain adaptation approaches started to emerge to address these challenges (Shen, Komogortsev, and Talathi, 2020; Wang et al., 2022). Nevertheless, we argue that hybrid approaches, which combine model- and appearance-based approaches, have the potential to overcome generalization issues, not only cross-dataset but also cross-subject, and without the need for large amounts of data. Our first study was just an example of how geometric constraints can regularize results, but incorporating the vast knowledge of eye geometry and dynamics into appearance-based approaches is a potential key to success, as has already been shown in the literature (Funes Mora and Odobez, 2014; Wang, Zhao, and Ji, 2018; Wang, Su, and Ji, 2019; Kaur, Jindal, and Manduchi, 2022; Jin, Dai, and Nguyen, 2023). In addition, such knowledge may aid in providing a measure of gaze estimation uncertainty.

Demographic diversity in available datasets

Throughout the course of this thesis, we gained insight into the prevailing limitations of publicly available datasets, which predominantly cater to the young adult demographic, often ignoring significant age groups such as children and the elderly. These omissions can be attributed to a combination of factors, including privacy concerns (since these groups are vulnerable populations and children are particularly protected) and the inherent challenges in recruiting voluntary participants from these age brackets. In the context of gaze estimation, these demographic groups present unique challenges in acquiring accurate ground truth data. For example, young children may find it more challenging to follow instructions to focus on calibration points, while older adults may struggle to maintain a steady gaze over extended periods. These issues affect not only new appearance-based approaches but also conventional eye trackers. Given the importance of accurate gaze estimation as a biomarker for the early diagnosis of several conditions associated with these age groups, in addition to other applications, it is crucial to address these challenges through tailored data collection protocols and innovative methodologies to ensure equitable and reliable gaze estimation across diverse age demographics.

Further observations

The findings and limitations of our work motivate further research along different axes, of which we highlight three. First, due to the potential increase in computational complexity due to incorporating additional spatiotemporal and multimodal cues, future work should study novel ways of incorporating such useful cues while minimizing the computational footprint. Second, the substantial variations in accuracy across different gaze components, observed in Chapter 4, prompt the exploration of separate models for each gaze component, as opposed to the usual jointly trained methods. One way to achieve this is by employing a shared backbone and

training separate per-component heads. Lastly, the results of our gaze-related features applied to the emotion recognition task motivate further exploration to identify more efficient feature sets that reduce the feature count while preserving their informative capabilities.

7.3 Future research lines

Our journey through gaze estimation has offered valuable insights and opportunities. As we move forward, we look to a future where gaze technology becomes even more robust, accurate, and accessible. This vision motivates us to continue pushing the boundaries of what can be achieved, making gaze technology a practical tool for understanding and interaction. As such, before concluding, we outline below three specific research lines as potential avenues for further exploration, which we have preliminary explored during this period but are not part of the thesis. These research directions aim to address methodological challenges and practical applications, further advancing the field and contributing to its broader impact.

Self-supervised gaze estimation

The first line is related to the generalization issue and the difficult and error-prone data collection process. One prospective avenue for mitigating these challenges involves considering eye or face images that lack corresponding gaze ground-truth data. These images can be acquired with relative ease (compared to usual ground-truth gaze data collection), and can be leveraged by appearance-based (or hybrid) approaches to increase generalization ability across subjects of different ages, genders, or face/eye geometries, in addition to different head poses and illumination conditions. This can be achieved with self-supervised ML, which has recently gained prominence in enhancing generalization in a wide range of CV tasks (Jing and Tian, 2020). As such, self-supervised techniques (in addition to weakly- and unsupervised) have started to emerge for gaze estimation. Self-supervised learning aims at solving a pretext task to learn a useful representation, which is then used in downstream tasks via transfer learning. The common pretext task is to enforce consistency between features extracted from two differently transformed views of the same image, so as to learn invariant representations. These transformations usually include geometric and appearance variations. However, most mainstream self-supervised approaches are tailored to classification, and gaze estimation is a regression task; thus, applying a geometric transformation to an image will potentially change the associated ground truth. Consequently, this change must be taken into account.

Following this research gap, we conducted an initial exploration with a clustering-based self-supervised approach (Caron et al., 2020) adapted for remote camera-based full-face input and for regression, to learn informative representations for gaze estimation without gaze ground truth. The results were positive for within- and cross-dataset evaluation scenarios (Farkhondeh et al., 2022), which motivates future work in this direction. More concretely, our work only exploited 2D image geometric variations, which change the gaze direction with respect to the CCS in the 3D space. However, the 3D geometric relationship between eye rotation in the HCS and gaze direction in the CCS could be leveraged, as image rotations do not affect gaze direction in the HCS. Nonetheless, privacy and ethical concerns arise when using unlabeled face images for model training. It is of utmost importance to examine the source and variability of such images, to not perpetuate biases in the

data, and to corroborate the consent of the depicted subjects for the intended usage. When these aspects cannot be guaranteed, synthetic data and hybrid approaches might be more suitable candidates for promoting generalization.

Irregularly sampled data and uncertainty

The second line is related to the challenges associated with data that is irregularly acquired or lost from the sensors used in eye tracking. To be more attuned to real-world issues like non-synchronized sensors or system latency, techniques to deal with irregularly sampled time series such as neural ordinary differential equations (ODEs) are potential research directions. For instance, some works outside the gaze estimation field have already explored the transformation of the discrete-time dynamics of RNNs into continous-time-based using ODEs (Rubanova, Chen, and Duvenaud, 2019) to model transportation trajectories (Liang et al., 2021). This could be applied to gaze estimation to model gaze trajectories. With respect to sensor noise, state-space models like conventional Kalman filters (Kalman, 1960), which enable the modeling of measurement uncertainty, have been extensively used for sensor fusion. Since such filters require knowledge of the system dynamics beforehand (in this case, eye movement dynamics), recent works have powered Kalman filters with DL techniques, like RNNs, to model non-linear, time-varying system dynamics and noise sources for sensor fusion and regularization (Coskun et al., 2017; Hosseinyalamdary, 2018). Probabilistic recurrent state-space models using Gaussian processes and variational inference are also potential candidates for non-deterministic gaze estimation (Doerr et al., 2018).

Gaze in social interactions

The final line is related to the role of gaze in social interactions. As discussed in previous chapters, appearance-based gaze estimation facilitates the application of gaze estimation in scenarios where wearable eye trackers would be deemed intrusive, potentially disrupting natural interactions. Looking ahead, we would like to analyze and model social gaze (e.g., eye contact, averted gaze) and gaze dynamics in the context of dyadic and small group interactions, in addition to other communication modalities, to better understand social behavior during interactions. Furthermore, we would like to leverage gaze as an informative cue to predict intentions and future actions. The ultimate goal is to model human behavior through gaze and other communication channels to enhance the interaction capabilities of embodied agents. Socially interactive and intelligent systems must be endowed with excellent perception and reasoning capabilities to, among others, be able to understand the people and world around them, anticipate users' actions and intentions to actuate in a timely and appropriate manner, and tailor their behavior and communication style to the user and situation. This research would be enabled by the UDIVA dataset (*Understanding Dyadic Interactions using Video and Audio signals*), collected during the PhD period by our research group (Palmero et al., 2021a). UDIVA is a large-scale, non-acted dataset of zero- and previous-acquaintance face-to-face dyadic interactions, where participants perform collaborative and competitive tasks with different behavior elicitation. The dataset features a diverse population in terms of age, gender, culture, education, and personality, and was recorded using multiple cameras, microphones, and physiological sensors. Figure 7.1 depicts sample frames from the five tasks featured in the dataset. The last one, *Gaze*, was designed precisely to serve as ground truth for some gaze gestures and face modeling with varied head poses.

FIGURE 7.1: Sample frames from the five tasks included in the UDIVA dataset. From left to right: *Talk*, *Lego*, *Animals*, *Ghost*, and *Gaze*.

For such task, participants were asked to follow directions to look *at other's face*, *at static/moving object*, or *elsewhere*, while moving the head and eyes. The dataset provides a valuable foundation for advancing research in understanding and modeling social human behavior.

7.4 Ethical and societal implications

High-accuracy, high-speed eye tracking will boost a myriad of applications for good, ranging from medical diagnosis (Crutcher et al., 2009; Vidal et al., 2012; Termsarasab et al., 2015) to gaze-controlled devices for people with disabilities (Holmqvist, Thunberg, and Dahlstrand, 2018; Shafti, Orlov, and Faisal, 2019). However, as this technology gets integrated into more consumer devices, such as AR/VR headsets and glasses, virtual assistants, or cars with advanced driver assistance systems, the prospects of other potential applications may raise societal concerns about data privacy, security, and ethics (Liebling and Preibusch, 2014; Larsen et al., 2020). Indeed, gaze data can reveal highly sensitive, personal data, such as health data and protected attributes, and can also uniquely identify individuals (Kröger, Lutz, and Müller, 2020). Thus, devices equipped with eye-tracking capabilities have the ability to inadvertently capture a wealth of information beyond what a user intends or anticipates disclosing, leading to discrimination or other harms. As with other sensors and technologies, there is the possibility that these data are repurposed for deviant purposes rather than their original intended use. However, what distinguishes gaze behavior from other modalities like speech is that it is partly beyond volitional control, making it virtually impossible for users to exert intentional influence. And it does not only affect active users. For instance, when gaze-tracking technology is deployed on public displays, faces and gaze of passersby or bystanders (henceforth referred to as nonusers) who are not interacting with the display may also be unknowingly detected and processed. A similar issue arises for passengers of gaze-assisted cars. With headsets with front-facing cameras, nonusers' faces can also be recorded.

To mitigate such risks, there is an increasing trend in research programs aimed at developing privacy-preserving approaches for eye tracking, by means of perturbations to the input eye images (Chaudhary and Pelz, 2020) or estimated gaze (David-John et al., 2021), gatekeepers that aggregate the raw estimates into, e.g., areas of interest (David-John et al., 2021), or differential privacy mechanisms (Liu et al., 2019; Steil et al., 2019a). The content recorded by front-facing cameras can also be altered by means of new generative approaches like Stable Diffusion to deindentify nonusers (Kurzhals, 2023). What is more, headsets can integrate mechanical front-facing camera shutters that automatically close depending on the users' eye movements and scene context changes (Steil et al., 2019b). However, these systems are still not perfect and may negatively affect the accuracy of the downstream task. Developers and providers are also beginning to embrace the so-called privacy-enhancing technologies (Heurix et al., 2015). These are tools that embody core data protection

principles, emphasizing minimal collection and use of personal data, robust data security measures, and empowerment of individuals to protect their personally identifiable information. Furthermore, data protection laws such as the European Union and United Kingdom versions of the General Data Protection Regulation (GDPR)[29] are in place, which govern the way in which data controllers and processors can use, process, and store the data. To be GDPR compliant, the law requires, among others: consent from users to collect data, allow users to rectify or delete such data, transparent data processing, and use the collected data for the originally intended purposes only, since purpose determines the legal basis for data processing. However, these requirements present some challenges. For instance, it is clear that consent can be difficult to obtain for nonusers. These and other challenges are an active subject of concern in the eye-tracking community (Gressel et al., 2023), which calls for interdisciplinary action to consider technical, ethical, social, and legal aspects throughout the whole eye-tracking technology development and deployment cycle.

Nonetheless, the ubiquitous integration of eye tracking into society depends entirely on our collective understanding of the information we may inadvertently disclose through our gaze and the safeguards in place to prevent its misuse. Much like any other technological advancement, building social trust is key. As such, providers must endow users and nonusers with the ability to make informed decisions about the use of eye tracking and other facial processing technologies, inform them when eye tracking is active, and give them the choice to opt-out immediately. By fostering a culture of transparency, accountability, and respect for privacy, we can pave the way for the widespread and responsible adoption of this technology. As we move forward, a collaborative effort between technology providers, policymakers, and society at large can ensure that the benefits of eye tracking are harnessed while safeguarding individuals' rights and interests. In this light, the future of eye tracking holds great promise, offering valuable insights and applications while respecting the ethical considerations that underpin its use.

[29] EU-GDPR: https://gdpr-info.eu/. UK-GDPR: https://uk-gdpr.org/.

Bibliography

Abdou, Ahmed, Ekta Sood, Philipp Müller, and Andreas Bulling (2022). "Gaze-enhanced Crossmodal Embeddings for Emotion Recognition". In: *Proceedings of the ACM on Human-Computer Interaction* 6.ETRA, pp. 1–18 (cit. on pp. 83, 84, 121, 122).

Adams, Andra, Marwa Mahmoud, Tadas Baltrušaitis, and Peter Robinson (2015). "Decoupling facial expressions and head motions in complex emotions". In: *2015 International conference on affective computing and intelligent interaction (ACII)*. IEEE, pp. 274–280 (cit. on p. 83).

Agarwal, Pankaj K, Jiří Matoušek, and Subhash Suri (1992). "Farthest neighbors, maximum spanning trees and related problems in higher dimensions". In: *Computational Geometry* 1.4, pp. 189–201 (cit. on p. 59).

Alghowinem, Sharifa, Roland Goecke, Michael Wagner, Julien Epps, Matthew Hyett, Gordon Parker, and Michael Breakspear (2016). "Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors". In: *IEEE Transactions on Affective Computing* 9.4, pp. 478–490 (cit. on pp. 83, 84).

Alhargan, Ashwaq, Neil Cooke, and Tareq Binjammaz (2017). "Multimodal affect recognition in an interactive gaming environment using eye tracking and speech signals". In: *Proceedings of the 19th ACM international conference on multimodal interaction*, pp. 479–486 (cit. on p. 84).

Alos, Juan, Julien Boullié, Maria Inés Torres, Eneko Ruiz, Andoni Beristain Iraola, Jacobo López Fernández, Inaki Tellería, Janeth Carreno, Iker Garay, Arkaitz Carbajo, Amaia Santamaría, Urtzi Zubiate, Jon Ander Arzallus, Francisco Martínez, and Adriana Martínez (2022). "ORKESTA Comprehensive Solution for the Orchestration of Services and Soci-Sanitary Care at Home". In: *Proc. IberSPEECH 2022*, pp. 251–253 (cit. on p. 79).

Amorese, Terry, Claudia Greco, Marialucia Cuciniello, Carmela Buono, Cristina Palmero, Pau Buch-Cardona, Sergio Escalera, Maria Inés Torres, Gennaro Cordasco, and Anna Esposito (2022). "Using Eye Tracking to Investigate Interaction Between Humans and Virtual Agents". In: *IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, pp. 125–132 (cit. on p. 10).

Anderson, Nicola C, Evan F Risko, and Alan Kingstone (2016). "Motion influences gaze direction discrimination and disambiguates contradictory luminance cues". In: *Psychonomic bulletin & review* 23.3, pp. 817–823 (cit. on p. 32).

Anderson, Tim J and Michael R MacAskill (2013). "Eye movements in patients with neurodegenerative disorders". In: *Nature Reviews Neurology* 9.2, pp. 74–85 (cit. on p. 5).

Angelopoulos, Anastasios N, Julien NP Martel, Amit PS Kohli, Jorg Conradt, and Gordon Wetzstein (2021). "Event-Based Near-Eye Gaze Tracking Beyond 10,000 Hz." In: *IEEE Transactions on Visualization and Computer Graphics* (cit. on pp. 25, 54, 57).

Armstrong, Thomas and Bunmi O Olatunji (2012). "Eye tracking of attention in the affective disorders: A meta-analytic review and synthesis". In: *Clinical psychology review* 32.8, pp. 704–723 (cit. on p. 5).

Ashraf, Hajra, Mikael H Sodergren, Nabeel Merali, George Mylonas, Harsimrat Singh, and Ara Darzi (2018). "Eye-tracking technology in medical education: A systematic review". In: *Medical teacher* 40.1, pp. 62–69 (cit. on p. 5).

Atmaja, Bagus Tris, Akira Sasou, and Masato Akagi (2022). "Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion". In: *Speech Communication* 140, pp. 11–28 (cit. on pp. 82, 84).

Atrey, Pradeep K, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli (2010). "Multimodal fusion for multimedia analysis: a survey". In: *Multimedia systems* 16.6, pp. 345–379 (cit. on p. 57).

Aziz, Samantha and Oleg Komogortsev (2022). "An assessment of the eye tracking signal quality captured in the HoloLens 2". In: *2022 Symposium on eye tracking research and applications*, pp. 1–6 (cit. on p. 133).

Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli (2020). "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: *Advances in neural information processing systems* 33, pp. 12449–12460 (cit. on p. 81).

Bahill, A Terry, Michael R Clark, and Lawrence Stark (1975). "The main sequence, a tool for studying human eye movements". In: *Mathematical biosciences* 24.3-4, pp. 191–204 (cit. on p. 94).

Balim, Haldun, Seonwook Park, Xi Wang, Xucong Zhang, and Otmar Hilliges (2023). "EFE: End-to-end Frame-to-Gaze Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2687–2696 (cit. on p. 26).

Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency (2018). "Multimodal machine learning: A survey and taxonomy". In: *IEEE transactions on pattern analysis and machine intelligence* 41.2, pp. 423–443 (cit. on pp. 7, 57).

Baltrušaitis, Tadas, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency (2018). "Openface 2.0: Facial behavior analysis toolkit". In: *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, pp. 59–66 (cit. on pp. 26, 83).

Baluja, Shumeet and Dean Pomerleau (1993). "Non-intrusive gaze tracking using artificial neural networks". In: *Advances in Neural Information Processing Systems* 6 (cit. on p. 24).

Baranes, Adrien, Pierre-Yves Oudeyer, and Jacqueline Gottlieb (2015). "Eye movements reveal epistemic curiosity in human observers". In: *Vision research* 117, pp. 81–90 (cit. on p. 94).

Beltrán, Jessica, Mireya S García-Vázquez, Jenny Benois-Pineau, Luis Miguel Gutierrez-Robledo, and Jean-François Dartigues (2018). "Computational techniques for eye movements analysis towards supporting early diagnosis of Alzheimer's disease: a review". In: *Computational and mathematical methods in medicine* 2018 (cit. on p. 7).

Benjamini, Yoav, Abba M Krieger, and Daniel Yekutieli (2006). "Adaptive linear step-up procedures that control the false discovery rate". In: *Biometrika* 93.3, pp. 491–507 (cit. on p. 100).

Bentivoglio, Anna Rita, Susan B Bressman, Emanuele Cassetta, Donatella Carretta, Pietro Tonali, and Alberto Albanese (1997). "Analysis of blink rate patterns in normal subjects". In: *Movement disorders* 12.6, pp. 1028–1034 (cit. on p. 82).

Beymer, David and Myron Flickner (2003). "Eye gaze tracking using an active stereo head". In: *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.* Vol. 2. IEEE, pp. II–451 (cit. on p. 23).

Blanz, V and T Vetter (1999). "A Morphable Model for the Synthesis of 3D Faces". In: *26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1999)*. ACM Press, pp. 187–194 (cit. on p. 26).

Blignaut, Pieter and Tanya Beelders (2008). "The effect of fixational eye movements on fixation identification with a dispersion-based fixation detection algorithm". In: *Journal of eye movement research* 2.5 (cit. on p. 17).

Boateng, George and Tobias Kowatsch (2020). "Speech emotion recognition among elderly individuals using multimodal fusion and transfer learning". In: *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pp. 12–16 (cit. on p. 84).

Bouckaert, Remco R and Eibe Frank (2004). "Evaluating the replicability of significance tests for comparing learning algorithms". In: *PAKDD*. Vol. 3056. Springer, pp. 3–12 (cit. on p. 100).

Bowers, Norick R and Martina Poletti (2017). "Microsaccades during reading". In: *PloS one* 12.9, e0185180 (cit. on p. 19).

Bradley, Margaret M, Laura Miccoli, Miguel A Escrig, and Peter J Lang (2008). "The pupil as a measure of emotional arousal and autonomic activation". In: *Psychophysiology* 45.4, pp. 602–607 (cit. on p. 82).

Braff, David L (1993). "Information processing and attention dysfunctions in schizophrenia". In: *Schizophrenia bulletin* 19.2, pp. 233–259 (cit. on p. 5).

Brooks, Rechele and Andrew N Meltzoff (2005). "The development of gaze following and its relation to language". In: *Developmental science* 8.6, pp. 535–543 (cit. on p. 2).

Brunyé, Tad T, Trafton Drew, Donald L Weaver, and Joann G Elmore (2019). "A review of eye tracking for understanding and improving diagnostic interpretation". In: *Cognitive research: principles and implications* 4.1, pp. 1–16 (cit. on p. 5).

Bulat, Adrian and Georgios Tzimiropoulos (2017). "How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)". In: *International Conference on Computer Vision* (cit. on p. 38).

Bulling, Andreas, Jamie A Ward, Hans Gellersen, and Gerhard Tröster (2010). "Eye movement analysis for activity recognition using electrooculography". In: *IEEE transactions on pattern analysis and machine intelligence* 33.4, pp. 741–753 (cit. on p. 94).

Busso, Carlos, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth Narayanan (2007). "Rigid head motion in expressive speech animation: Analysis and synthesis". In: *IEEE transactions on audio, speech, and language processing* 15.3, pp. 1075–1086 (cit. on p. 83).

Caesar, Holger, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom (2020). "nuscenes: A multimodal dataset for autonomous driving". In: *Proc. IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631 (cit. on p. 57).

Callahan-Flintoft, Chloe, Christian Barentine, Jonathan Touryan, and Anthony J Ries (2021). "A case for studying naturalistic eye and head movements in virtual environments". In: *Frontiers in Psychology* 12, p. 650693 (cit. on p. 6).

Calvo, Rafael A and Sidney D'Mello (2010). "Affect detection: An interdisciplinary review of models, methods, and their applications". In: *IEEE Transactions on affective computing* 1.1, pp. 18–37 (cit. on p. 80).

Caron, Mathilde, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin (2020). "Unsupervised learning of visual features by contrasting cluster assignments". In: *Advances in neural information processing systems* 33, pp. 9912–9924 (cit. on p. 135).

Carvalho, Nicolas, Eric Laurent, Nicolas Noiret, Gilles Chopard, Emmanuel Haffen, Djamila Bennabi, and Pierre Vandel (2015). "Eye movement in unipolar and bipolar depression: A systematic review of the literature". In: *Frontiers in psychology* 6, p. 1809 (cit. on p. 5).

Castellano, Ginevra, Ana Paiva, Arvid Kappas, Ruth Aylett, Helen Hastie, Wolmet Barendregt, Fernando Nabais, and Susan Bull (2013). "Towards empathic virtual and robotic tutors". In: *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings 16*. Springer, pp. 733–736 (cit. on p. 31).

Cazzato, Dario, Marco Leo, Cosimo Distante, and Holger Voos (2020). "When i look into your eyes: A survey on computer vision contributions for human gaze estimation and tracking". In: *Sensors* 20.13, p. 3739 (cit. on pp. 22, 56).

Chakraborty, Rupayan, Meghna Pandharipande, and Sunil Kumar Kopparapu (2017). *Analyzing emotion in spontaneous speech*. Springer (cit. on p. 78).

Chamberlain, Ann Christine (1996). *Dual purkinje-image eyetracker*. US Naval Academy (cit. on p. 21).

Chaudhary, Aayush K, Nitinraj Nair, Reynold J Bailey, Jeff B Pelz, Sachin S Talathi, and Gabriel J Diaz (2022). "Temporal RIT-Eyes: From real infrared eye-images to synthetic sequences of gaze behavior". In: *IEEE Transactions on Visualization and Computer Graphics* 28.11, pp. 3948–3958 (cit. on pp. 57, 133).

Chaudhary, Aayush Kumar and Jeff B Pelz (2020). "Privacy-preserving eye videos using rubber sheet model". In: *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–5 (cit. on p. 137).

Chen, Changhao, Stefano Rosa, Yishu Miao, Chris Xiaoxuan Lu, Wei Wu, Andrew Markham, and Niki Trigoni (2019). "Selective sensor fusion for neural visual-inertial odometry". In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10542–10551 (cit. on pp. 54, 57).

Chen, Sanyuan, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei (July 2022). "Wavlm: Large-scale self-supervised pre-training for full stack speech processing". In: *IEEE Journal of Selected Topics in Signal Processing* 16.6, pp. 1505–1518 (cit. on pp. 81, 91).

Cheng, Yihua, Haofei Wang, Yiwei Bao, and Feng Lu (2021). "Appearance-based gaze estimation with deep learning: A review and benchmark". In: *arXiv preprint arXiv:2104.12668* (cit. on p. 22).

Chita-Tegmark, Meia (2016). "Social attention in ASD: A review and meta-analysis of eye-tracking studies". In: *Research in developmental disabilities* 48, pp. 79–93 (cit. on p. 5).

Cho, Hyunggi, Young-Woo Seo, BVK Vijaya Kumar, and Ragunathan Raj Rajkumar (2014). "A multi-sensor fusion system for moving object detection and tracking in urban driving environments". In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 1836–1843 (cit. on p. 54).

Chollet, François (2017). "Xception: Deep learning with depthwise separable convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258 (cit. on p. 93).

Chumachenko, Kateryna, Alexandros Iosifidis, and Moncef Gabbouj (2022). "Self-attention fusion for audiovisual emotion recognition with incomplete data". In: *26th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 2822–2828 (cit. on p. 84).

Chung, Junyoung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio (2014). "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *NIPS 2014 Workshop on Deep Learning, December 2014* (cit. on p. 37).

Clark, Eve V and Marisa Casillas (2015). "First language acquisition". In: *The Routledge handbook of linguistics*. Routledge, pp. 311–328 (cit. on p. 2).

Clay, Viviane, Peter König, and Sabine Koenig (2019). "Eye tracking in virtual reality". In: *Journal of Eye Movement Research* 12.1 (cit. on p. 5).

Cohn, Jeffrey F, Lawrence Ian Reed, Tsuyoshi Moriyama, Jing Xiao, Karen Schmidt, and Zara Ambadar (2004). "Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles". In: *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.* Pp. 129–135 (cit. on p. 84).

Collewijn, Han and Eileen Kowler (2008). "The significance of microsaccades for vision and oculomotor control". In: *Journal of Vision* 8.14, pp. 20–20 (cit. on p. 17).

Comaniciu, Dorin and Peter Meer (2002). "Mean shift: A robust approach toward feature space analysis". In: *IEEE Transactions on pattern analysis and machine intelligence* 24.5, pp. 603–619 (cit. on p. 95).

Conover, William Jay (1999). *Practical nonparametric statistics*. Vol. 350. john wiley & sons (cit. on pp. 40, 70).

Cooke, Neil James and Martin Russell (2008). "Gaze-contingent automatic speech recognition". In: *IET signal processing* 2.4, pp. 369–380 (cit. on p. 8).

Corneanu, Ciprian Adrian, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero (2016). "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications". In: *IEEE transactions on pattern analysis and machine intelligence* 38.8, pp. 1548–1568 (cit. on p. 82).

Cornsweet, Tom N and Hewitt D Crane (1973). "Accurate two-dimensional eye tracker using first and fourth Purkinje images". In: *JOSA* 63.8, pp. 921–928 (cit. on p. 19).

Cortacero, Kevin, Tobias Fischer, and Yiannis Demiris (2019). "RT-BENE: A dataset and baselines for real-time blink estimation in natural environments". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0 (cit. on p. 83).

Coskun, Huseyin, Felix Achilles, Robert DiPietro, Nassir Navab, and Federico Tombari (2017). "Long short-term memory kalman filters: Recurrent neural estimators for pose regularization". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5524–5532 (cit. on p. 136).

Cowen, Alan S and Dacher Keltner (2017). "Self-report captures 27 distinct categories of emotion bridged by continuous gradients". In: *Proceedings of the national academy of sciences* 114.38, E7900–E7909 (cit. on p. 86).

Crutcher, Michael D, Rose Calhoun-Haney, Cecelia M Manzanares, James J Lah, Allan I Levey, and Stuart M Zola (2009). "Eye tracking during a visual paired comparison task as a predictor of early dementia". In: *American Journal of Alzheimer's Disease & Other Dementias®* 24.3, pp. 258–266 (cit. on pp. 5, 137).

Dahlberg, Joakim (2010). "Eye tracking with eye glasses". MA thesis. Umea University, Sweden (cit. on p. 21).

Das, Julia, Lisa Graham, Rosie Morris, Gill Barry, Alan Godfrey, Richard Walker, and Samuel Stuart (2022). "Eye Movement in Neurological Disorders". In: *Eye Tracking: Background, Methods, and Applications*, pp. 185–205 (cit. on pp. 2, 5).

David-John, Brendan, Diane Hosfelt, Kevin Butler, and Eakta Jain (2021). "A privacy-preserving approach to streaming eye-tracking data". In: *IEEE Transactions on Visualization and Computer Graphics* 27.5, pp. 2555–2565 (cit. on p. 137).

De Lope, Javier and Manuel Graña (2023). "An ongoing review of speech emotion recognition". In: *Neurocomputing* 528, pp. 1–11 (cit. on p. 81).

De Velasco, Mikel (2023). "Analysis and Automatic Identification of Spontaneous Emotions in Speech from Human-Human and Human-Machine Communication". PhD thesis. Departamento de Electricidad y Electrónica. Universidad del País Vasco UPV/EHU (cit. on p. 81).

De Velasco, Mikel, Raquel Justo, and María Inés Torres (2022). "Automatic Identification of Emotional Information in Spanish TV Debates and Human-Machine Interactions". In: *Applied Sciences* 12.4 (cit. on pp. 78, 81, 82).

De Velasco Vázquez, Mikel, Raquel Justo Blanco, Asier López Zorrilla, and María Inés Torres Barañano (2023). "Analysis of Deep Learning-Based Decision-Making in an Emotional Spontaneous Speech Task". In: *Applied Sciences* (cit. on p. 81).

Delabarre, Edmund B (1898). "A method of recording eye-movements". In: *The American Journal of Psychology* 9.4, pp. 572–574 (cit. on p. 18).

Demaeght, Annebeth, Christina Miclau, Julia Hartmann, Janina Markwardt, and Oliver Korn (2022). "Multimodal emotion analysis of robotic assistance in elderly care". In: *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 230–236 (cit. on p. 79).

Deng, Haoping and Wangjiang Zhu (2017). "Monocular Free-head 3D Gaze Tracking with Deep Learning and Geometry Constraints". In: *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, pp. 3162–3171 (cit. on p. 33).

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (June 2009). "ImageNet: A large-scale hierarchical image database". In: *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE (cit. on p. 93).

Deng, Kangle, Gengshan Yang, Deva Ramanan, and Jun-Yan Zhu (2023). "3D-Aware Conditional Image Synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4434–4445 (cit. on p. 133).

Diefendorf, Allen Ross and Raymond Dodge (1908). "An experimental study of the ocular reactions of the insane from photographic records." In: *Brain* 31.3, pp. 451–489 (cit. on pp. 18, 20).

Ding, Yu, Lei Shi, and Zhigang Deng (2018). "Low-level characterization of expressive head motion through frequency domain analysis". In: *IEEE Transactions on Affective Computing* 11.3, pp. 405–418 (cit. on p. 83).

D'mello, Sidney K and Jacqueline Kory (2015). "A review and meta-analysis of multimodal affect detection systems". In: *ACM computing surveys (CSUR)* 47.3, pp. 1–36 (cit. on pp. 78, 84).

Dodge, Raymond (1926). "A pendulum-photochronograph." In: *Journal of Experimental Psychology* 9.2, p. 155 (cit. on p. 18).

Dodge, Raymond and Thomas Sparks Cline (1901). "The angle velocity of eye movements." In: *Psychological Review* 8.2, p. 145 (cit. on p. 18).

Doerr, Andreas, Christian Daniel, Martin Schiegg, Nguyen-Tuong Duy, Stefan Schaal, Marc Toussaint, and Trimpe Sebastian (2018). "Probabilistic Recurrent State-Space Models". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1280–1289 (cit. on p. 136).

Donahue, Jeffrey, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell (2015). "Long-term recurrent convolutional networks for visual recognition and description". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634 (cit. on p. 37).

Duchowski, Andrew T (2002). "A breadth-first survey of eye-tracking applications". In: *Behavior Research Methods, Instruments, & Computers* 34.4, pp. 455–470 (cit. on p. 5).

Duchowski, Andrew T. (2017). *Eye Tracking Methodology: Theory and Practice*. 3rd. Springer Publishing Company, Incorporated (cit. on pp. 4, 15, 17).

Duchowski, Andrew T, Nathan Cournia, and Hunter Murphy (2004). "Gaze-contingent displays: A review". In: *Cyberpsychology & behavior* 7.6, pp. 621–634 (cit. on p. 5).

Ebner, Natalie C, Michaela Riediger, and Ulman Lindenberger (2010). "FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation". In: *Behavior research methods* 42, pp. 351–362 (cit. on p. 79).

Ehinger, Benedikt V, Katharina Groß, Inga Ibs, and Peter König (2019). "A new comprehensive eye-tracking test battery concurrently evaluating the Pupil Labs glasses and the EyeLink 1000". In: *PeerJ* 7, e7086 (cit. on p. 20).

Ekman, Paul (1999). "Basic emotions". In: *Handbook of cognition and emotion* 98.45-60, p. 16 (cit. on pp. 80, 82, 88).

El Kaliouby, Rana and Peter Robinson (2005). "Real-time inference of complex mental states from facial expressions and head gestures". In: *Real-time vision for human-computer interaction*, pp. 181–200 (cit. on pp. 83, 89).

Emery, Nathan J (2000). "The eyes have it: the neuroethology, function and evolution of social gaze". In: *Neuroscience & biobehavioral reviews* 24.6, pp. 581–604 (cit. on p. 1).

Esposito, Anna (2009). "The Perceptual and Cognitive Role of Visual and Auditory Channels in Conveying Emotional Information". In: *Cognitive Computation* 1 (3), pp. 268–278 (cit. on p. 87).

Eyben, Florian, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, and Khiet P. Truong (2015). "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing". In: *IEEE transactions on affective computing* 7.2, pp. 190–202 (cit. on p. 81).

Eyben, Florian, Martin Wöllmer, Michel F Valstar, Hatice Gunes, Björn Schuller, and Maja Pantic (2011). "String-based audiovisual fusion of behavioural events for the assessment of dimensional affect". In: *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, pp. 322–329 (cit. on p. 83).

Falck-Ytter, Terje, Sven Bölte, and Gustaf Gredebäck (2013). "Eye tracking in early autism research". In: *Journal of neurodevelopmental disorders* 5.1, pp. 1–13 (cit. on p. 5).

Farkhondeh, Arya, Cristina Palmero, Simone Scardapane, and Sergio Escalera (2022). "Towards Self-Supervised Gaze Estimation". In: *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press (cit. on pp. 56, 135).

Fayyad, Jamil, Mohammad A Jaradat, Dominique Gruyer, and Homayoun Najjaran (2020). "Deep learning sensor fusion for autonomous vehicle perception and localization: A review". In: *Sensors* 20.15, p. 4220 (cit. on pp. 54, 57).

Feng, Di, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer (2020). "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges". In: *IEEE Transactions on Intelligent Transportation Systems* 22.3, pp. 1341–1360 (cit. on p. 57).

Ferhat, Onur and Fernando Vilariño (2016). "Low cost eye tracking". In: *Computational intelligence and neuroscience* 2016, p. 17 (cit. on p. 32).

Filali Razzouki, Anas, Laetitia Jeancolas, Graziella Mangone, Sara Sambin, Alizé Chalançon, Manon Gomes, Stéphane Lehéricy, Jean-Christophe Corvol, Marie Vidailhet, Isabelle Arnulf, Mounim A. El-Yacoubi, and Dijana Petrovska-Delacrétaz (2023). "Early-stage parkinson's disease detection based on action unit derivatives". In: *"Dispositifs biomédicaux et technologies numériques en santé ; des besoins aux usages" (Colloque JETSAN)*, pp. 1–8 (cit. on p. 121).

Fischer, Tobias, Hyung Jin Chang, and Yiannis Demiris (2018). "Rt-gene: Real-time eye gaze estimation in natural environments". In: *Proc. European Conference on Computer Vision*, pp. 334–352 (cit. on p. 56).

Fischler, Martin A and Robert C Bolles (1981). "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Communications of the ACM* 24.6, pp. 381–395 (cit. on p. 26).

Fitts, Paul M, Richard E Jones, and John L Milton (1950). "Eye movements of aircraft pilots during instrument-landing approaches." In: *Aeronautical Engineering Review* (cit. on p. 19).

Fletcher, Abbey, Stephen Dunne, and Joe Butler (2022). "A Brief History of Eye Movement Research". In: *Eye Tracking: Background, Methods, and Applications*. Ed. by Samuel Stuart. New York, NY: Springer US, pp. 15–29. ISBN: 978-1-0716-2391-6 (cit. on pp. 3, 19, 21).

Fogarty, Christine and John A Stern (1989). "Eye movements and blinks: their relationship to higher cognitive processes". In: *International journal of psychophysiology* 8.1, pp. 35–42 (cit. on p. 2).

Fuhl, Wolfgang and Enkelejda Kasneci (2021). "A multimodal eye movement dataset and a multimodal eye movement segmentation analysis". In: *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–7 (cit. on p. 57).

Funes-Mora, Kenneth A and Jean-Marc Odobez (2016). "Gaze Estimation in the 3D Space Using RGB-D Sensors: Towards Head-Pose and User Invariance". In: *International Journal of Computer Vision* 118, pp. 194–216 (cit. on pp. 7, 26, 33).

Funes Mora, Kenneth Alberto, Florent Monay, and Jean-Marc Odobez (2014a). "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras". In: *Proceedings of the Symposium on Eye Tracking Research and Applications*. ACM, pp. 255–258 (cit. on pp. 24, 32, 33, 39, 57).

— (2014b). *Eyediap database: Data description and gaze tracking evaluation benchmarks*. Tech. rep. Idiap-RR-08-2014. Idiap (cit. on p. 39).

Funes Mora, Kenneth Alberto and Jean-Marc Odobez (2014). "Geometric generative gaze estimation (g3e) for remote rgb-d cameras". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1773–1780 (cit. on p. 134).

Funke, Gregory, Eric Greenlee, Martha Carter, Allen Dukes, Rebecca Brown, and Lauren Menke (2016). "Which eye tracker is right for your research? performance evaluation of several cost variant eye trackers". In: *Proceedings of the Human Factors and Ergonomics Society annual meeting*. Vol. 60. 1. SAGE Publications Sage CA: Los Angeles, CA, pp. 1240–1244 (cit. on p. 21).

Fölster, Mara, Ursula Hess, and Katja Werheid (2014). "Facial age affects emotional expression decoding". In: *Frontiers in Psychology* 5 (cit. on p. 79).

Garde, Gonzalo, Andoni Larumbe-Bergera, Benoît Bossavit, Rafael Cabeza, Sonia Porta, and Arantxa Villanueva (2020). "Gaze estimation problem tackled through synthetic images". In: *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–5 (cit. on p. 55).

Gardner, Frances (2000). "Methodological issues in the direct observation of parent–child interaction: Do observational findings reflect the natural behavior of participants?" In: *Clinical child and family psychology review* 3, pp. 185–198 (cit. on p. 31).

Ghosh, Shreya, Abhinav Dhall, Munawar Hayat, Jarrod Knibbe, and Qiang Ji (2021). "Automatic gaze analysis: A survey of deep learning based approaches". In: *arXiv preprint arXiv:2108.05479* (cit. on pp. 4, 22, 83).

Glorot, Xavier and Yoshua Bengio (2010). "Understanding the difficulty of training deep feedforward neural networks". In: *Proc. 13th international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, pp. 249–256 (cit. on p. 68).

Gomez, Argenis Ramirez and Hans Gellersen (2018). "Smooth-i: smart re-calibration using smooth pursuit eye movements". In: *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pp. 1–5 (cit. on p. 23).

Gou, Chao, Yue Wu, Kang Wang, Kunfeng Wang, Fei-Yue Wang, and Qiang Ji (2017). "A joint cascaded framework for simultaneous eye detection and eye state estimation". In: *Pattern Recognition* 67, pp. 23–31 (cit. on p. 58).

Graham, Lisa, Julia Das, Jason Moore, Alan Godfrey, and Samuel Stuart (2022). "The Eyes as a Window to the Brain and Mind". In: *Eye Tracking: Background, Methods, and Applications*. Springer, pp. 1–14 (cit. on pp. 6, 17).

Graham, Reiko and Kevin S LaBar (2007). "Garner interference reveals dependencies between emotional expression and gaze in face perception." In: *Emotion* 7.2, p. 296 (cit. on p. 83).

— (2012). "Neurocognitive mechanisms of gaze-expression interactions in face processing and social attention". In: *Neuropsychologia* 50.5, pp. 553–566 (cit. on p. 82).

Graves, Alex (2012). "Long short-term memory". In: *Supervised sequence labelling with recurrent neural networks*, pp. 37–45 (cit. on p. 37).

Greco, Claudia, Carmela Buono, Pau Buch-Cardona, Gennaro Cordasco, Sergio Escalera, Anna Esposito, Anais Fernandez, Daria Kyslitska, Maria Stylianou Kornes, Cristina Palmero, Jofre Tenorio-Laranga, Anna Torp Johansen, and M. Inés Torres (2021). "Emotional Features of Interactions with Empathic Agents". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2168–2176 (cit. on p. 89).

Greff, Klaus, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber (2016). "LSTM: A search space odyssey". In: *IEEE transactions on neural networks and learning systems* 28.10, pp. 2222–2232 (cit. on p. 47).

Gressel, Céline, Rebekah Overdorf, Inken Hagenstedt, Murat Karaboga, Helmut Lurtz, Michael Raschke, and Andreas Bulling (2023). "Privacy-Aware Eye Tracking: Challenges and Future Directions". In: *IEEE Pervasive Computing* 22.1, pp. 95–102 (cit. on p. 138).

Griffith, Henry, Dmytro Katrychuk, and Oleg Komogortsev (2019). "Assessment of Shift-Invariant CNN Gaze Mappings for PS-OG Eye Movement Sensors". In: *Proc. IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0 (cit. on pp. 56, 67, 68).

Griffith, Henry, Dillon Lohr, Evgeny Abdulin, and Oleg Komogortsev (2021). "GazeBase, a large-scale, multi-stimulus, longitudinal eye movement dataset". In: *Scientific Data* 8.1, pp. 1–9 (cit. on pp. 55, 61).

Gross, Ralph, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker (2010). "Multi-pie". In: *Image and vision computing* 28.5, pp. 807–813 (cit. on p. 35).

Grossman, Gerard E, R John Leigh, Larry A Abel, Douglas J Lanska, and SE Thurston (1988). "Frequency and velocity of rotational head perturbations during locomotion". In: *Experimental brain research* 70.3, pp. 470–476 (cit. on p. 94).

Gudi, Amogh, Xin Li, and Jan van Gemert (2020). "Efficiency in real-time webcam gaze tracking". In: *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, pp. 529–543 (cit. on p. 25).

Guestrin, Elias Daniel and Moshe Eizenman (2006). "General theory of remote gaze estimation using the pupil center and corneal reflections". In: *IEEE Transactions on biomedical engineering* 53.6, pp. 1124–1133 (cit. on pp. 23, 48, 63).

Guillon, Quentin, Nouchine Hadjikhani, Sophie Baduel, and Bernadette Rogé (2014). "Visual social attention in autism spectrum disorder: Insights from eye tracking studies". In: *Neuroscience & Biobehavioral Reviews* 42, pp. 279–297 (cit. on p. 5).

Guitton, Daniel and Michel Volle (1987). "Gaze control in humans: eye-head coordination during orienting movements to targets within and beyond the oculomotor range". In: *Journal of neurophysiology* 58.3, pp. 427–459 (cit. on p. 83).

Gunes, Hatice and Maja Pantic (2010a). "Automatic, dimensional and continuous emotion recognition". In: *International Journal of Synthetic Emotions (IJSE)* 1.1, pp. 68–99 (cit. on pp. 80, 81).

— (2010b). "Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners". In: *Intelligent Virtual Agents: 10th International Conference, IVA 2010. Proceedings 10*. Springer, pp. 371–377 (cit. on p. 83).

Guo, Jianzhu, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li (2020). "Towards fast, accurate and stable 3d dense face alignment". In: *European Conference on Computer Vision*. Springer, pp. 152–168 (cit. on pp. 26, 92).

Guo, Wenzhong, Jianwen Wang, and Shiping Wang (2019). "Deep multimodal representation learning: A survey". In: *IEEE Access* 7, pp. 63373–63394 (cit. on pp. 7, 122).

Haith, Marshall M, Terry Bergman, and Michael J Moore (1977). "Eye contact and face scanning in early infancy". In: *Science* 198.4319, pp. 853–855 (cit. on p. 2).

Hammal, Zakia and Jeffrey F Cohn (2014). "Intra-and interpersonal functions of head motion in emotion communication". In: *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*, pp. 19–22 (cit. on p. 83).

Hammal, Zakia, Jeffrey F Cohn, and Daniel S Messinger (2015). "Head movement dynamics during play and perturbed mother-infant interaction". In: *IEEE transactions on affective computing* 6.4, pp. 361–370 (cit. on p. 83).

Hannula, Deborah E, Robert R Althoff, David E Warren, Lily Riggs, Neal J Cohen, and Jennifer D Ryan (2010). "Worth a glance: using eye movements to investigate the cognitive neuroscience of memory". In: *Frontiers in human neuroscience* 4, p. 166 (cit. on p. 5).

Hansen, Dan Witzner and Qiang Ji (2010). "In the eye of the beholder: A survey of models for eyes and gaze". In: *IEEE transactions on pattern analysis and machine intelligence* 32.3, pp. 478–500 (cit. on pp. 3, 16, 22, 31, 32, 56).

Hansen, Dan Witzner and Arthur EC Pece (2005). "Eye tracking in the wild". In: *Computer Vision and Image Understanding* 98.1, pp. 155–181 (cit. on p. 7).

Haro, Antonio, Myron Flickner, and Irfan Essa (2000). "Detecting and tracking eyes by using their physiological properties, dynamics, and appearance". In: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*. Vol. 1. IEEE, pp. 163–168 (cit. on p. 7).

Harston, J Alex and A Aldo Faisal (2022). "Methods and Models of Eye-Tracking in Natural Environments". In: *Eye Tracking: Background, Methods, and Applications*. Springer, pp. 49–68 (cit. on p. 6).

Hartridge, Hamilton and LC Thomson (1948). "Methods of investigating eye movements". In: *The British journal of ophthalmology* 32.9, p. 581 (cit. on p. 19).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proc. IEEE international conference on computer vision*, pp. 1026–1034 (cit. on p. 68).

— (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (cit. on pp. 47, 93).

Hess, Ursula, Reginald B Adams, and Robert E Kleck (2007). "Looking at you or looking elsewhere: The influence of head orientation on the signal value of emotional facial expressions". In: *Motivation and Emotion* 31, pp. 137–144 (cit. on p. 83).

Hessels, Roy S, Diederick C Niehorster, Gijs A Holleman, Jeroen S Benjamins, and Ignace TC Hooge (2020). *Wearable technology for "real-world research": realistic or not?* (Cit. on p. 6).

Heurix, Johannes, Peter Zimmermann, Thomas Neubauer, and Stefan Fenz (2015). "A taxonomy for privacy enhancing technologies". In: *Computers & Security* 53, pp. 1–17 (cit. on p. 137).

Ho, Simon, Tom Foulsham, and Alan Kingstone (2015). "Speaking and listening with the eyes: Gaze signaling during dyadic interactions". In: *PloS one* 10.8, e0136905 (cit. on p. 2).

Holland, Corey and Oleg V Komogortsev (2011). "Biometric identification via eye movement scanpaths in reading". In: *2011 International joint conference on biometrics (IJCB)*. IEEE, pp. 1–8 (cit. on p. 94).

Holmqvist, Eva, Gunilla Thunberg, and Marie P Dahlstrand (2018). "Gaze-controlled communication technology for children with severe multiple disabilities: Parents and professionals' perception of gains, obstacles, and prerequisites". In: *Assistive Technology* 30.4, pp. 201–208 (cit. on p. 137).

Holmqvist, Kenneth, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford (cit. on pp. 17, 54).

Hoppe, Sabrina, Tobias Loetscher, Stephanie A Morey, and Andreas Bulling (2018). "Eye movements during everyday behavior predict personality traits". In: *Frontiers in human neuroscience* 12, p. 105 (cit. on pp. 8, 94).

Hossain, Mir Rayat Imtiaz and James J Little (2018). "Exploiting temporal information for 3d human pose estimation". In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 68–84 (cit. on p. 7).

Hosseinyalamdary, Siavash (2018). "Deep Kalman filter: Simultaneous multi-sensor integration and modelling; A GNSS/IMU case study". In: *Sensors* 18.5, p. 1316 (cit. on p. 136).

Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed (2021). "Hubert: Self-supervised speech representation learning by masked prediction of hidden units". In:

IEEE/ACM Transactions on Audio, Speech, and Language Processing 29, pp. 3451–3460 (cit. on p. 81).

Hu, Ting, Xinyu Wang, and Haiming Xu (2022). "Eye-Tracking in Interpreting Studies: A Review of Four Decades of Empirical Studies". In: *Frontiers in Psychology* 13, p. 872247 (cit. on p. 5).

Huang, Kun-Yi, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su, and Yi-Hsuan Chen (2019). "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5866–5870 (cit. on p. 81).

Huang, Qiong, Ashok Veeraraghavan, and Ashutosh Sabharwal (2017). "TabletGaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets". In: *Machine Vision and Applications* 28.5-6, pp. 445–461 (cit. on p. 24).

Huber, Patrik, Guosheng Hu, Rafael Tena, Pouria Mortazavian, P Koppen, William J Christmas, Matthias Ratsch, and Josef Kittler (2016). "A multiresolution 3d morphable face model and fitting framework". In: *Proceedings of the 11th international joint conference on computer vision, imaging and computer graphics theory and applications* (cit. on p. 93).

Huey, Edmund B (1898). "Preliminary experiments in the physiology and psychology of reading". In: *The American Journal of Psychology* 9.4, pp. 575–586 (cit. on p. 18).

— (1900). "On the psychology and physiology of reading. I". In: *The American Journal of Psychology* 11.3, pp. 283–302 (cit. on p. 18).

Hutt, Stephen and Sidney K D'Mello (2022). "Evaluating Calibration-Free Webcam-Based Eye Tracking for Gaze-Based User Modeling". In: *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 224–235 (cit. on p. 7).

Hutt, Stephen, Aaron Wong, Alexandra Papoutsaki, Ryan S Baker, Joshua I Gold, and Caitlin Mills (2023). "Webcam-based eye tracking to detect mind wandering and comprehension errors". In: *Behavior Research Methods*, pp. 1–17 (cit. on p. 77).

Hutton, S. B. (2019). "Eye Tracking Methodology". In: *Eye Movement Research: An Introduction to its Scientific Foundations and Applications*. Ed. by Christoph Klein and Ulrich Ettinger. Cham: Springer International Publishing, pp. 277–308 (cit. on p. 20).

Ioffe, Sergey and Christian Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. pmlr, pp. 448–456 (cit. on p. 68).

Itier, Roxane J and Magali Batty (2009). "Neural bases of eye and gaze processing: the core of social cognition". In: *Neuroscience & Biobehavioral Reviews* 33.6, pp. 843–863 (cit. on p. 2).

Jacob, Robert JK (1991). "The use of eye movements in human-computer interaction techniques: what you look at is what you get". In: *ACM Transactions on Information Systems (TOIS)* 9.2, pp. 152–169 (cit. on p. 5).

Jacob, Robert JK and Keith S Karn (2003). "Eye tracking in human-computer interaction and usability research: Ready to deliver the promises". In: *The mind's eye*. Elsevier, pp. 573–605 (cit. on p. 5).

Jaimes, Alejandro and Nicu Sebe (2007). "Multimodal human–computer interaction: A survey". In: *Computer vision and image understanding* 108.1-2, pp. 116–134 (cit. on pp. 78, 90).

Jang, Young-Min, Rammohan Mallipeddi, Sangil Lee, Ho-Wan Kwak, and Minho Lee (2014). "Human intention recognition based on eyeball movement pattern and pupil size variation". In: *Neurocomputing* 128, pp. 421–432 (cit. on p. 8).

Jannat, Sk Rahatul and Shaun Canavan (2021). "Expression Recognition Across Age". In: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, pp. 1–5 (cit. on p. 85).

Javadi, Roya and Angelica Lim (2021). "The Many Faces of Anger: A Multicultural Video Dataset of Negative Emotions in the Wild (MFA-Wild)". In: *16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, pp. 01–08 (cit. on p. 84).

Jeni, László A and Jeffrey F Cohn (2016). "Person-independent 3d gaze estimation using face frontalization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 87–95 (cit. on p. 33).

Ji, Qiang and Zhiwei Zhu (2002). "Eye and gaze tracking for interactive graphic display". In: *Proceedings of the 2nd International Symposium on Smart graphics*, pp. 79–85 (cit. on p. 24).

Jin, Shiwei, Ji Dai, and Truong Nguyen (2023). "Kappa Angle Regression With Ocular Counter-Rolling Awareness for Gaze Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2658–2667 (cit. on p. 134).

Jin, Xin and Xiaoyang Tan (2017). "Face alignment in-the-wild: A survey". In: *Computer Vision and Image Understanding* 162, pp. 1–22 (cit. on pp. 32, 35).

Jindal, Swati and Roberto Manduchi (2023). "Contrastive Representation Learning for Gaze Estimation". In: *Proceedings of The 1st Gaze Meets ML workshop*. Ed. by Ismini Lourentzou, Joy Wu, Satyananda Kashyap, Alexandros Karargyris, Leo Anthony Celi, Ban Kawas, and Sachin Talathi. Vol. 210. Proceedings of Machine Learning Research. PMLR, pp. 37–49 (cit. on p. 25).

Jing, Longlong and Yingli Tian (2020). "Self-supervised visual feature learning with deep neural networks: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 43.11, pp. 4037–4058 (cit. on p. 135).

Jozefowicz, Rafal, Wojciech Zaremba, and Ilya Sutskever (2015). "An empirical exploration of recurrent network architectures". In: *International conference on machine learning*. PMLR, pp. 2342–2350 (cit. on p. 68).

Judd, Charles H, Cloyd N McAllister, and WM Steele (1905). "General introduction to a series of studies of eye movements by means of kinetoscopic photographs". In: *Psychological Review Monographs* 7.1, pp. 1–16 (cit. on p. 18).

Justo, Raquel, Leila Ben Letaifa, Cristina Palmero, Eduardo Gonzalez-Fraile, Anna Torp Johansen, Alain Vázquez, Gennaro Cordasco, Stephan Schlögl, Begoña Fernández-Ruanova, Micaela Silva, Sergio Escalera, Mikel deVelasco, Joffre Tenorio-Laranga, Anna Esposito, Maria Korsnes, and M. Inés Torres (2020). "Analysis of the interaction between elderly people and a simulated virtual coach". In: *Journal of Ambient Intelligence and Humanized Computing* 11, pp. 6125–6140 (cit. on pp. 10, 78, 79, 85).

Kalman, R. E. (Mar. 1960). "A New Approach to Linear Filtering and Prediction Problems". In: *Journal of Basic Engineering* 82.1, pp. 35–45 (cit. on p. 136).

Kanungo, Tapas, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu (2002). "An efficient k-means clustering algorithm: Analysis and implementation". In: *IEEE transactions on pattern analysis and machine intelligence* 24.7, pp. 881–892 (cit. on p. 59).

Kapitaniak, Bronisław, Marta Walczak, Marcin Kosobudzki, Zbigniew Jóźwiak, and Alicja Bortkiewicz (2015). "Application of eye-tracking in drivers testing: A review of research". In: *International journal of occupational medicine and environmental health* 28.6 (cit. on p. 5).

Kapoor, Ashish, Winslow Burleson, and Rosalind W Picard (2007). "Automatic prediction of frustration". In: *International journal of human-computer studies* 65.8, pp. 724–736 (cit. on pp. 83, 84).

Kar, Anuradha and Peter Corcoran (2017). "A Review and Analysis of Eye-Gaze Estimation Systems, Algorithms and Performance Evaluation Methods in Consumer Platforms". In: *IEEE Access* 5, pp. 16495–16519 (cit. on pp. 22, 32).

Karatekin, Canan (2007). "Eye tracking studies of normative and atypical development". In: *Developmental review* 27.3, pp. 283–348 (cit. on p. 5).

Karg, Michelle, Ali-Akbar Samadani, Rob Gorbet, Kolja Kühnlenz, Jesse Hoey, and Dana Kulić (2013). "Body movements for affective expression: A survey of automatic recognition and generation". In: *IEEE Transactions on Affective Computing* 4.4, pp. 341–359 (cit. on p. 83).

Kasprowski, Pawel (2022). "Eye Tracking Hardware: Past to Present, and Beyond". In: *Eye Tracking: Background, Methods, and Applications*. Springer, pp. 31–48 (cit. on p. 21).

Katrychuk, Dmytro, Henry Griffith, and Oleg Komogortsev (2020). "A Calibration Framework for Photosensor-based Eye-Tracking System". In: *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–5 (cit. on p. 56).

Kaur, Harsimran, Swati Jindal, and Roberto Manduchi (2022). "Rethinking model-based gaze estimation". In: *Proceedings of the ACM on computer graphics and interactive techniques* 5.2, pp. 1–17 (cit. on pp. 26, 134).

Kellnhofer, Petr, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba (2019). "Gaze360: Physically unconstrained gaze estimation in the wild". In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6912–6921 (cit. on pp. 45, 56).

Khan, Khalil, Rehan Ullah Khan, Riccardo Leonardi, Pierangelo Migliorati, and Sergio Benini (2021). "Head pose estimation: A survey of the last ten years". In: *Signal Processing: Image Communication* 99, p. 116479 (cit. on p. 83).

Khan, Muhammad Qasim and Sukhan Lee (2019). "Gaze and eye tracking: Techniques and applications in ADAS". In: *Sensors* 19.24, p. 5540 (cit. on p. 5).

Kim, Elizabeth S, Adam Naples, Giuliana Vaccarino Gearty, Quan Wang, Seth Wallace, Carla Wall, Jennifer Kowitt, Linda Friedlaender, Brian Reichow, Fred Volkmar, Frederick Shic, and Michael Perlmutter (2014). "Development of an untethered, mobile, low-cost head-mounted eye tracker". In: *Proceedings of the Symposium on Eye Tracking Research and Applications*, pp. 247–250 (cit. on p. 21).

Kim, Joohwan, Michael Stengel, Alexander Majercik, Shalini De Mello, David Dunn, Samuli Laine, Morgan McGuire, and David Luebke (2019). "Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation". In: *Proc. 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12 (cit. on pp. 24, 55, 57, 63, 68).

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (cit. on pp. 38, 48, 68).

Klein, Christoph and Ulrich Ettinger (2008). *A hundred years of eye movement research in psychiatry* (cit. on p. 2).

Ko, Byoung Chul (2018). "A brief review of facial emotion recognition based on visual information". In: *sensors* 18.2, p. 401 (cit. on p. 82).

Koch, Kristin, Judith McLean, Ronen Segev, Michael A Freed, Michael J Berry, Vijay Balasubramanian, and Peter Sterling (2006). "How much the eye tells the brain". In: *Current biology* 16.14, pp. 1428–1434 (cit. on p. 1).

Komogortsev, Oleg V, Denise V Gobert, Sampath Jayarathna, Do Hyong Koh, and Sandeep M Gowda (2010). "Standardization of automated analyses of oculomotor fixation and saccadic behaviors". In: *IEEE Transactions on Biomedical Engineering* 57.11, pp. 2635–2645 (cit. on pp. 46, 70).

Komogortsev, Oleg V and Alex Karpov (2013). "Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades". In: *Behavior research methods* 45.1, pp. 203–215 (cit. on p. 70).

Komogortsev, Oleg V, Young Sam Ryu, and Do Hyong Koh (2012). *Fast Target Selection via Saccade-driven Methods*. Tech. rep. TXSTATE-CS-TR-2012-6. Texas State University-San Marcos, Department of Computer Science. (cit. on p. 61).

Kossaifi, Jean, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller, Kam Star, Elnar Hajiyev, and Maja Pantic (2019). "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild". In: *IEEE transactions on pattern analysis and machine intelligence* 43.3, pp. 1022–1040 (cit. on pp. 79, 80, 87).

Kothari, Rakshit, Shalini De Mello, Umar Iqbal, Wonmin Byeon, Seonwook Park, and Jan Kautz (2021a). "Weakly-Supervised Physically Unconstrained Gaze Estimation". In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9980–9989 (cit. on pp. 25, 56).

Kothari, Rakshit, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B Pelz, and Gabriel J Diaz (2020). "Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities". In: *Scientific reports* 10.1, pp. 1–18 (cit. on p. 57).

Kothari, Rakshit S, Aayush K Chaudhary, Reynold J Bailey, Jeff B Pelz, and Gabriel J Diaz (2021b). "Ellseg: An ellipse segmentation framework for robust gaze tracking". In: *IEEE Transactions on Visualization and Computer Graphics* 27.5, pp. 2757–2767 (cit. on p. 22).

Krafka, Kyle, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba (2016). "Eye tracking for everyone". In: *Proc. IEEE conference on computer vision and pattern recognition*, pp. 2176–2184 (cit. on pp. 25, 32–34, 56).

Kredel, Ralf, Christian Vater, André Klostermann, and Ernst-Joachim Hossner (2017). "Eye-tracking technology and the dynamics of natural gaze behavior in sports: A systematic review of 40 years of research". In: *Frontiers in psychology* 8, p. 1845 (cit. on p. 5).

Kreibig, Sylvia D (2010). "Autonomic nervous system activity in emotion: A review". In: *Biological psychology* 84.3, pp. 394–421 (cit. on p. 82).

Kröger, Jacob Leon, Otto Hans-Martin Lutz, and Florian Müller (2020). "What does your gaze reveal about you? On the privacy implications of eye tracking". In: *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers 14*, pp. 226–241 (cit. on p. 137).

Kurzhals, Kuno (2023). "Privacy in Eye Tracking Research with Stable Diffusion". In: *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, pp. 1–7 (cit. on p. 137).

Lai, Meng-Lung, Meng-Jung Tsai, Fang-Ying Yang, Chung-Yuan Hsu, Tzu-Chien Liu, Silvia Wen-Yu Lee, Min-Hsien Lee, Guo-Li Chiou, Jyh-Chong Liang, and Chin-Chung Tsai (2013). "A review of using eye-tracking technology in exploring learning from 2000 to 2012". In: *Educational research review* 10, pp. 90–115 (cit. on p. 5).

LaLonde, Rodney, Dong Zhang, and Mubarak Shah (2018). "Clusternet: Detecting small objects in large scenes by exploiting spatio-temporal information". In:

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4003–4012 (cit. on p. 7).

Lamb, Maurice, Malin Brundin, Estela Perez Luque, and Erik Billing (2022). "Eye-tracking beyond peripersonal space in virtual reality: validation and best practices". In: *Frontiers in Virtual Reality* 3, p. 864653 (cit. on p. 6).

Lamb, Trevor D, Shaun P Collin, and Edward N Pugh Jr (2007). "Evolution of the vertebrate eye: opsins, photoreceptors, retina and eye cup". In: *Nature Reviews Neuroscience* 8.12, pp. 960–976 (cit. on p. 1).

Lang, Peter J, Margaret M Bradley, and Bruce N Cuthbert (1990). "Emotion, attention, and the startle reflex." In: *Psychological review* 97.3, p. 377 (cit. on p. 82).

Larrazabal, Agostina J, CE García Cena, and César Ernesto Martínez (2019). "Video-oculography eye tracking towards clinical applications: A review". In: *Computers in biology and medicine* 108, pp. 57–66 (cit. on p. 5).

Larsen, Ethan P, Jacob M Kolman, Faisal N Masud, and Farzan Sasangohar (2020). "Ethical considerations when using a mobile eye tracker in a patient-facing area: lessons from an intensive care unit observational protocol". In: *Ethics & Human Research* 42.6, pp. 2–13 (cit. on p. 137).

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *nature* 521.7553, pp. 436–444 (cit. on p. 3).

LeGrand, Yves and Sami G ElHage (2013). *Physiological optics*. Vol. 13. Springer (cit. on p. 22).

Leigh, R John and David S Zee (2015). *The neurology of eye movements*. Vol. 90. Oxford University Press, USA (cit. on pp. 1, 2, 5, 16, 17).

Lepetit, Vincent, Francesc Moreno-Noguer, and Pascal Fua (2009). "EPnP: An Accurate O(n) Solution to the PnP Problem". In: *International Journal Of Computer Vision* 81, pp. 155–166 (cit. on p. 93).

Levenson, Robert W, Laura L Carstensen, Wallace V Friesen, and Paul Ekman (1991). "Emotion, physiology, and expression in old age." In: *Psychology and aging* 6.1, p. 28 (cit. on p. 79).

Lhommet, Margaux and Stacy C Marsella (2014). "19 Expressing Emotion Through Posture and Gesture". In: *The Oxford handbook of affective computing*, p. 273 (cit. on p. 83).

Li, Benjamin J, Jeremy N Bailenson, Adam Pines, Walter J Greenleaf, and Leanne M Williams (2017). "A public database of immersive VR videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures". In: *Frontiers in psychology* 8, p. 2116 (cit. on p. 83).

Li, Richard, Eric Whitmire, Michael Stengel, Ben Boudaoud, Jan Kautz, David Luebke, Shwetak Patel, and Kaan Akşit (2020). "Optical Gaze Tracking with Spatially-Sparse Single-Pixel Detectors". In: *2020 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, pp. 117–126 (cit. on pp. 54–56).

Li, Shan and Weihong Deng (2020). "Deep facial expression recognition: A survey". In: *IEEE transactions on affective computing* 13.3, pp. 1195–1215 (cit. on p. 82).

Li, Tianxing, Qiang Liu, and Xia Zhou (2017). "Ultra-low power gaze tracking for virtual reality". In: *Proc. 15th ACM Conference on Embedded Network Sensor Systems*, pp. 1–14 (cit. on pp. 54, 56).

Liang, Paul Pu, Amir Zadeh, and Louis-Philippe Morency (2022). "Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions". In: *arXiv preprint arXiv:2209.03430* (cit. on p. 84).

Liang, Yuxuan, Kun Ouyang, Hanshu Yan, Yiwei Wang, Zekun Tong, and Roger Zimmermann (2021). "Modeling Trajectories with Neural Ordinary Differential

Equations." In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, pp. 1498–1504 (cit. on p. 136).

Liebling, Daniel J and Sören Preibusch (2014). "Privacy considerations for a pervasive eye tracking world". In: *Proc. 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pp. 1169–1177 (cit. on p. 137).

Lim, Jia Zheng, James Mountstephens, and Jason Teo (2020). "Emotion recognition using eye-tracking: taxonomy, review and current challenges". In: *Sensors* 20.8, p. 2384 (cit. on pp. 8, 77, 83).

Liu, Ao, Lirong Xia, Andrew Duchowski, Reynold Bailey, Kenneth Holmqvist, and Eakta Jain (2019). "Differential privacy for eye-tracking data". In: *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–10 (cit. on p. 137).

Liu, Chengjun and H. Wechsler (2002). "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition". In: *IEEE Transactions on Image Processing* 11.4, pp. 467–476 (cit. on p. 82).

Liu, Gang, Yuechen Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez (2018). "A differential approach for gaze estimation with calibration." In: *Proc. British Machine Vision Conference (BMVC)*. Vol. 2. 3, p. 6 (cit. on p. 25).

Liversedge, Simon P and John M Findlay (2000). "Saccadic eye movements and cognition". In: *Trends in cognitive sciences* 4.1, pp. 6–14 (cit. on p. 2).

Lopes, Nuno, André Silva, Salik Ram Khanal, Arsênio Reis, João Barroso, Vitor Filipe, and Jaime Sampaio (2018). "Facial emotion recognition in the elderly using a SVM classifier". In: *2nd International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW)*, pp. 1–5 (cit. on p. 79).

Lu, Feng, Takahiro Okabe, Yusuke Sugano, and Yoichi Sato (2011a). "A head pose-free approach for appearance-based gaze estimation." In: *Proc. British Machine Vision Conference (BMVC)*, pp. 1–11 (cit. on p. 33).

Lu, Feng, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato (2011b). "Inferring human gaze from appearance via adaptive linear regression". In: *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, pp. 153–160 (cit. on p. 24).

Luna-Jiménez, C., R. Kleinlein, D. Griol, Z. Callejas, J.M. Montero, and F. Fernández-Martínez (2022). "A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset". In: *Applied Sciences* 12 (327) (cit. on p. 81).

Ma, Kaixin, Xinyu Wang, Xinru Yang, Mingtong Zhang, Jeffrey M Girard, and Louis-Philippe Morency (2019). "ElderReact: a multimodal dataset for recognizing emotional response in aging adults". In: *2019 international conference on multimodal interaction*, pp. 349–357 (cit. on pp. 79, 84, 85, 122).

Mack, David J, Sandro Belfanti, and Urs Schwarz (2017). "The effect of sampling rate and lowpass filters on saccades–a modeling approach". In: *Behavior Research Methods* 49.6, pp. 2146–2162 (cit. on p. 61).

Magai, Carol, Nathan S Consedine, Yulia S Krivoshekova, Elizabeth Kudadjie-Gyamfi, and Renee McPherson (2006). "Emotion experience and expression across the adult life span: insights from a multimodal assessment study." In: *Psychology and aging* 21.2, p. 303 (cit. on p. 79).

Mahanama, Bhanuka, Yasith Jayawardana, Sundararaman Rengarajan, Gavindya Jayawardena, Leanne Chukoskie, Joseph Snider, and Sampath Jayarathna (2022). "Eye movement and pupil measures: A review". In: *frontiers in Computer Science* 3, p. 733531 (cit. on p. 21).

Majaranta, Päivi (2011). *Gaze interaction and applications of eye tracking: Advances in assistive technologies: Advances in assistive technologies*. IGI Global (cit. on p. 5).

Majaranta, Päivi and Andreas Bulling (2014). "Eye tracking and eye-based human–computer interaction". In: *Advances in physiological computing*. Springer, pp. 39–65 (cit. on p. 5).

Marandi, Ramtin Z and Parisa Gazerani (2019). "Aging and eye tracking: in the quest for objective biomarkers". In: *Future Neurology* 14.4, FNL33 (cit. on p. 5).

Marg, Elwin (1951). "Development of electro-oculography: Standing potential of the eye in registration of eye movement". In: *AMA archives of ophthalmology* 45.2, pp. 169–185 (cit. on pp. 3, 19).

Mariooryad, Soroosh and Carlos Busso (2015). "Facial expression recognition in the presence of speech using blind lexical compensation". In: *IEEE Transactions on Affective Computing* 7.4, pp. 346–359 (cit. on p. 78).

Massé, Benoît, Silèye Ba, and Radu Horaud (2017). "Tracking gaze and visual focus of attention of people involved in social interaction". In: *IEEE transactions on pattern analysis and machine intelligence* 40.11, pp. 2711–2724 (cit. on p. 6).

Massey Jr, Frank J (1951). "The Kolmogorov-Smirnov test for goodness of fit". In: *Journal of the American statistical Association* 46.253, pp. 68–78 (cit. on p. 51).

Mathôt, Sebastiaan (2018). "Pupillometry: Psychology, physiology, and function". In: *Journal of Cognition* 1.1 (cit. on p. 2).

Mavadati, S. Mohammad, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn (2013). "DISFA: A Spontaneous Facial Action Intensity Database". In: *IEEE Transactions on Affective Computing* 4.2, pp. 151–160 (cit. on p. 82).

McHugh, Mary L (2012). "Interrater reliability: the kappa statistic". In: *Biochemia medica* 22.3, pp. 276–282 (cit. on p. 88).

McKeown, Gary, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder (2011). "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent". In: *IEEE transactions on affective computing* 3.1, pp. 5–17 (cit. on p. 78).

McMurrough, Christopher D, Vangelis Metsis, Jonathan Rich, and Fillia Makedon (2012). "An eye tracking dataset for point of gaze detection". In: *Proc. Symposium on Eye Tracking Research and Applications*, pp. 305–308 (cit. on p. 57).

McTear, Michael, Kristiina Jokinen, Mirza Mohtashim Alam, Qasid Saleem, Giulio Napolitano, Florian Szczepaniak, Mossaab Hariz, Gérard Chollet, Christophe Lohr, Jérôme Boudy, Zohre Azimi, Sonja Dana Roelen, and Rainer Wieching (2023). "Interaction with a Virtual Coach for Active and Healthy Ageing". In: *Sensors* 23.5 (cit. on p. 79).

Medsker, Larry R and LC Jain (2001). "Recurrent neural networks". In: *Design and Applications* 5.64-67, p. 2 (cit. on p. 32).

Meißner, Martin and Josua Oll (2019). "The promise of eye-tracking methodology in organizational research: A taxonomy, review, and future avenues". In: *Organizational Research Methods* 22.2, pp. 590–617 (cit. on p. 5).

Mele, Maria Laura and Stefano Federici (2012). "Gaze and eye-tracking solutions for psychological research". In: *Cognitive processing* 13, pp. 261–265 (cit. on p. 5).

Mignault, Alain and Avi Chaudhuri (2003). "The many faces of a neutral face: Head tilt and perception of dominance and emotion". In: *Journal of nonverbal behavior* 27, pp. 111–132 (cit. on p. 83).

Milders, Maarten, Jari K Hietanen, Jukka M Leppänen, and Marc Braun (2011). "Detection of emotional faces is modulated by the direction of eye gaze." In: *Emotion* 11.6, p. 1456 (cit. on p. 83).

Model, Dmitri and Moshe Eizenman (2010). "An automatic personal calibration procedure for advanced gaze estimation systems". In: *IEEE Transactions on Biomedical Engineering* 57.5, pp. 1031–1039 (cit. on pp. 16, 22).

Mohammadi, M.R., E. Fatemizadeh, and M.H. Mahoor (July 2014). "PCA-based dictionary building for accurate facial expression recognition via sparse representation". In: *Journal of Visual Communication and Image Representation* 25.5, pp. 1082–1092 (cit. on p. 82).

Mollahosseini, Ali, David Chan, and Mohammad H Mahoor (2016). "Going deeper in facial expression recognition using deep neural networks". In: *IEEE Winter conference on applications of computer vision (WACV)*, pp. 1–10 (cit. on p. 82).

Mora, Kenneth Alberto Funes and Jean-Marc Odobez (2012). "Gaze estimation from multimodal kinect data". In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, pp. 25–30 (cit. on pp. 33, 41).

— (2013). "Person independent 3d gaze estimation from remote rgb-d cameras". In: *2013 IEEE International Conference on Image Processing*. IEEE, pp. 2787–2791 (cit. on p. 24).

Morais, Edmilson, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz (2022). "Speech Emotion Recognition Using Self-Supervised Features". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6922–6926 (cit. on p. 81).

MSC, IMO (2000). *Circ. 982 (2000) Guidelines on ergonomic criteria for bridge equipment and layout* (cit. on p. 39).

Nair, Nitinraj, Rakshit Kothari, Aayush K Chaudhary, Zhizhuo Yang, Gabriel J Diaz, Jeff B Pelz, and Reynold J Bailey (2020). "RIT-Eyes: Rendering of near-eye images for eye-tracking applications". In: *ACM Symposium on Applied Perception 2020*, pp. 1–9 (cit. on pp. 55, 133).

Newell, Alejandro, Kaiyu Yang, and Jia Deng (2016). "Stacked hourglass networks for human pose estimation". In: *European Conference on Computer Vision*. Springer, pp. 483–499 (cit. on p. 38).

Niehorster, Diederick C, Tim HW Cornelissen, Kenneth Holmqvist, Ignace TC Hooge, and Roy S Hessels (2018). "What to expect from your remote eye-tracker when participants are unrestrained". In: *Behavior research methods* 50, pp. 213–227 (cit. on p. 21).

Niehorster, Diederick C, Thiago Santini, Roy S Hessels, Ignace TC Hooge, Enkelejda Kasneci, and Marcus Nyström (2020). "The impact of slippage on the data quality of head-worn eye trackers". In: *Behavior Research Methods* 52.3, pp. 1140–1160 (cit. on p. 54).

O'Dwyer, Jonny (2019). "Speech, Head, and Eye-based Cues for Continuous Affect Prediction". In: *8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, pp. 16–20 (cit. on p. 84).

O'Dwyer, Jonny, Ronan Flynn, and Niall Murray (2017). "Continuous affect prediction using eye gaze and speech". In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, pp. 2001–2007 (cit. on pp. 83, 84).

O'Dwyer, Jonny, Niall Murray, and Ronan Flynn (2018). "Affective computing using speech and eye gaze: a review and bimodal system proposal for continuous affect prediction". In: *arXiv preprint arXiv:1805.06652* (cit. on pp. 77, 78, 83, 84, 122).

— (2019). "Eye-based Continuous Affect Prediction". In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pp. 137–143 (cit. on pp. 83, 84).

Oh, Catherine S, Jeremy N Bailenson, and Gregory F Welch (2018). "A systematic review of social presence: Definition, antecedents, and implications". In: *Frontiers in Robotics and AI* 5, p. 409295 (cit. on p. 5).

Olaso, J. M., M. García-Sebastián, A. López Zorrilla, M. Tainta, M. Ecay-Torres, M.I. Torres, and P. Martínez-Lage (2023). "The CITA GO-ON dialogue system: Mid-term Achievements". In: *Proceedings of the 16th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA (cit. on p. 79).

Olaso, Javier M., Alain Vázquez, Leila Ben Letaifa, Mikel de Velasco, Aymen Mtibaa, Mohamed Amine Hmani, Dijana Petrovska-Delacrétaz, Gérard Chollet, César Montenegro, Asier López-Zorrilla, Raquel Justo, Roberto Santana, Jofre Tenorio-Laranga, Eduardo González-Fraile, Begoña Fernández-Ruanova, Gennaro Cordasco, Anna Esposito, Kristin Beck Gjellesvik, Anna Torp Johansen, Maria Stylianou Kornes, Colin Pickard, Cornelius Glackin, Gary Cahalane, Pau Buch, Cristina Palmero, Sergio Escalera, Olga Gordeeva, Olivier Deroo, Anaïs Fernández, Daria Kyslitska, Jose Antonio Lozano, María Inés Torres, and Stephan Schlögl (2021). "The EMPATHIC Virtual Coach: A Demo". In: *Proceedings of the 2021 International Conference on Multimodal Interaction*. ICMI '21, 848–851 (cit. on pp. 79, 90).

Orquin, Jacob L and Simone Mueller Loose (2013). "Attention and choice: A review on eye movements in decision making". In: *Acta psychologica* 144.1, pp. 190–206 (cit. on p. 2).

O'Driscoll, Gillian A and Brandy L Callahan (2008). "Smooth pursuit in schizophrenia: a meta-analytic review of research since 1993". In: *Brain and cognition* 68.3, pp. 359–370 (cit. on p. 5).

Palinko, Oskar, Francesco Rea, Giulio Sandini, and Alessandra Sciutti (2016). "Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration". In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 5048–5054 (cit. on p. 5).

Palmero, Cristina, Elsbeth A van Dam, Sergio Escalera, Mike Kelia, Guido F Lichtert, Lucas PJJ Noldus, Andrew J Spink, and Astrid van Wieringen (2018a). "Automatic mutual gaze detection in face-to-face dyadic interaction videos". In: *Proceedings of Measuring Behavior*. Vol. 1, p. 2 (cit. on p. 8).

Palmero, Cristina, Mikel deVelasco, Mohamed Amine Hmani, Aymen Mtibaa, Leila Ben Letaifa, Pau Buch-Cardona, Raquel Justo, Terry Amorese, Eduardo González-Fraile, Begoña Fernández-Ruanova, Jofre Tenorio-Laranga, Anna Torp Johansen, Micaela Rodrigues da Silva, Liva Jenny Martinussen, Maria Stylianou Korsnes, Gennaro Cordasco, Anna Esposito, Mounim A. El-Yacoubi, Dijana Petrovska-Delacrétaz, M.Inés Torres, and Sergio Escalera (2023a). "Exploring Emotion Expression Recognition in Older Adults Interacting with a Virtual Coach". Under review (cit. on p. 10).

Palmero, Cristina, Oleg V Komogortsev, Sergio Escalera, and Sachin S Talathi (2023b). "Multi-Rate Sensor Fusion for Unconstrained Near-Eye Gaze Estimation". In: *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, pp. 1–8 (cit. on p. 9).

Palmero, Cristina, Oleg V Komogortsev, and Sachin S Talathi (2020). "Benefits of temporal information for appearance-based gaze estimation". In: *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–5 (cit. on pp. 9, 10, 56, 67).

Palmero, Cristina, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera (2018b). "Recurrent CNN for 3D Gaze Estimation using Appearance and Shape Cues". In: *Proc. British Machine Vision Conference (BMVC)* (cit. on pp. 8, 45, 56).

Palmero, Cristina, Javier Selva, Sorina Smeureanu, Julio Junior, CS Jacques, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, et al. (2021a). "Context-aware personality inference in dyadic scenarios: Introducing

the udiva dataset". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1–12 (cit. on p. 136).

Palmero, Cristina, Abhishek Sharma, Karsten Behrendt, Kapil Krishnakumar, Oleg V Komogortsev, and Sachin S Talathi (2020). "Openeds2020: Open eyes dataset". In: *arXiv preprint arXiv:2005.03876* (cit. on pp. 13, 48).

— (2021b). "OpenEDS2020 Challenge on Gaze Tracking for VR: Dataset and Results". In: *Sensors* 21.14, p. 4769 (cit. on pp. 13, 48, 56, 57, 67).

Pan, Bing, Helene A Hembrooke, Geri K Gay, Laura A Granka, Matthew K Feusner, and Jill K Newman (2004). "The determinants of web page viewing behavior: an eye-tracking study". In: *Proceedings of the 2004 symposium on Eye tracking research & applications*, pp. 147–154 (cit. on p. 5).

Panda, Renato, Ricardo Manuel Malheiro, and Rui Pedro Paiva (2020). "Audio Features for Music Emotion Recognition: a Survey". In: *IEEE Transactions on Affective Computing*, pp. 1–1 (cit. on p. 81).

Park, Seonwook, Emre Aksan, Xucong Zhang, and Otmar Hilliges (2020). "Towards end-to-end video-based eye-tracking". In: *Proc. European Conference on Computer Vision*. Springer, pp. 747–763 (cit. on p. 56).

Park, Seonwook, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz (2019). "Few-shot adaptive gaze estimation". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9368–9377 (cit. on p. 25).

Park, Seonwook, Xucong Zhang, Andreas Bulling, and Otmar Hilliges (2018). "Learning to find eye region landmarks for remote gaze estimation in unconstrained settings". In: *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–10 (cit. on p. 58).

Parker, A. (2004). *In The Blink Of An Eye: How Vision Sparked The Big Bang Of Evolution*. Basic Books. ISBN: 9780465054381 (cit. on p. 1).

Parkhi, O. M., A. Vedaldi, and A. Zisserman (2015). "Deep Face Recognition". In: *British Machine Vision Conference* (cit. on pp. 37, 38).

Pastor, Miguel, Dayana Ribas, Alfonso Ortega, Antonio Miguel, and EDUARDO LLEIDA SOLANO (2022). "Cross-Corpus Speech Emotion Recognition with HuBERT Self-Supervised Representation". In: *IberSPEECH 2022* (cit. on p. 81).

Patel, Saumil S, Joseph Jankovic, Ashley J Hood, Cameron B Jeter, and Anne B Sereno (2012). "Reflexive and volitional saccades: biomarkers of Huntington disease severity and progression". In: *Journal of the neurological sciences* 313.1-2, pp. 35–41 (cit. on p. 5).

Patney, Anjul, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn (2016). "Towards foveated rendering for gaze-tracked virtual reality". In: *ACM Transactions on Graphics (TOG)* 35.6, p. 179 (cit. on pp. 5, 54).

Pavlidis, George Th (1985). "Eye movements in dyslexia: their diagnostic significance". In: *Journal of learning disabilities* 18.1, pp. 42–50 (cit. on p. 5).

Peißl, Sylvia, Christopher D Wickens, and Rithi Baruah (2018). "Eye-tracking measures in aviation: A selective literature review". In: *The International Journal of Aerospace Psychology* 28.3-4, pp. 98–112 (cit. on p. 5).

Płużyczka, Monika (2018). "The first hundred years: A history of eye tracking as a research method". In: *Applied Linguistics Papers* 25/4, pp. 101–116 (cit. on pp. 3, 18, 19, 21).

Poole, Alex and Linden J Ball (2006). "Eye tracking in HCI and usability research". In: *Encyclopedia of human computer interaction*. IGI global, pp. 211–219 (cit. on p. 5).

Poria, Soujanya, Erik Cambria, Rajiv Bajpai, and Amir Hussain (2017). "A review of affective computing: From unimodal analysis to multimodal fusion". In: *Information fusion* 37, pp. 98–125 (cit. on pp. 78, 84, 122).

Porta, Sonia, Benoit Bossavit, Rafael Cabeza, Andoni Larumbe-Bergera, Gonzalo Garde, and Arantxa Villanueva (2019). "U2Eyes: a binocular dataset for eye tracking and gaze estimation". In: *Proc. IEEE/CVF International Conference on Computer Vision Workshops* (cit. on pp. 57, 58, 63).

Pouget, P (2015). "The cortex is in overall control of 'voluntary'eye movement". In: *Eye* 29.2, pp. 241–245 (cit. on p. 1).

Purves, Dale, George J Augustine, David Fitzpatrick, Lawrence C Katz, Anthony-Samuel LaMantia, James O McNamara, and S Mark Williams (2001). "Types of Eye Movements and Their Functions". In: *Neuroscience. 2nd edition*. Sinauer Associates (cit. on p. 16).

Rahal, Rima-Maria and Susann Fiedler (2019). "Understanding cognitive and affective mechanisms in social psychology through eye-tracking". In: *Journal of Experimental Social Psychology* 85, p. 103842 (cit. on p. 5).

Rahman, Wasifur, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque (2020). "Integrating multimodal information in large pretrained transformers". In: *Proceedings of the conference. Association for Computational Linguistics. Meeting*. Vol. 2020. NIH Public Access, p. 2359 (cit. on p. 65).

Rajan, Vandana, Alessio Brutti, and Andrea Cavallaro (2022). "Is cross-attention preferable to self-attention for multi-modal emotion recognition?" In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4693–4697 (cit. on p. 98).

Ramachandram, Dhanesh and Graham W Taylor (2017). "Deep multimodal learning: A survey on recent advances and trends". In: *IEEE signal processing magazine* 34.6, pp. 96–108 (cit. on p. 57).

Ramirez, Geovany A, Tadas Baltrušaitis, and Louis-Philippe Morency (2011). "Modeling latent discriminative dynamic of multi-dimensional affective signals". In: *Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II*. Springer, pp. 396–406 (cit. on pp. 83, 84).

Ramzan, Muhammad, Hikmat Ullah Khan, Shahid Mahmood Awan, Amina Ismail, Mahwish Ilyas, and Ahsan Mahmood (2019). "A survey on state-of-the-art drowsiness detection techniques". In: *IEEE Access* 7, pp. 61904–61919 (cit. on p. 5).

Rayner, Keith (1998). "Eye movements in reading and information processing: 20 years of research." In: *Psychological bulletin* 124.3, p. 372 (cit. on p. 2).

Raynowska, Jenelle, John-Ross Rizzo, Janet C Rucker, Weiwei Dai, Joel Birkemeier, Julian Hershowitz, Ivan Selesnick, Laura J Balcer, Steven L Galetta, and Todd Hudson (2018). "Validity of low-resolution eye-tracking to assess eye movements during a rapid number naming task: performance of the eyetribe eye tracker". In: *Brain injury* 32.2, pp. 200–208 (cit. on p. 61).

Rigas, Ioannis, Hayes Raffle, and Oleg V. Komogortsev (2017). "Hybrid ps-v technique: A novel sensor fusion approach for fast mobile eye-tracking with sensor-shift aware correction". In: *IEEE Sensors Journal* 17.24, pp. 8356–8366 (cit. on pp. 7, 54, 57, 58).

Rigas, Ioannis, Hayes Raffle, and Oleg V Komogortsev (2018). "Photosensor oculography: survey and parametric analysis of designs using model-based simulation". In: *IEEE Transactions on Human-Machine Systems* 48.6, pp. 670–681 (cit. on pp. 54, 56).

Riviello, Maria Teresa and Anna Esposito (2012). "A cross-cultural study on the effectiveness of visual and vocal channels in transmitting dynamic emotional information". In: *Acta Polytechnica Hungarica* 9.1, pp. 157–170 (cit. on pp. 87, 123).

Robinson, David A. (1963). "A Method of Measuring Eye Movemnent Using a Scieral Search Coil in a Magnetic Field". In: *IEEE Transactions on Bio-medical Electronics* 10.4, pp. 137–145 (cit. on pp. 3, 19).

Robinson, David A (1968). "The oculomotor control system: A review". In: *Proceedings of the IEEE* 56.6, pp. 1032–1049 (cit. on p. 16).

Rohrschneider, Klaus (2004). "Determination of the location of the fovea on the fundus". In: *Investigative ophthalmology & visual science* 45.9, pp. 3257–3258 (cit. on p. 15).

Rommelse, Nanda NJ, Stefan Van der Stigchel, and Joseph A Sergeant (2008). "A review on eye movement studies in childhood and adolescent psychiatry". In: *Brain and cognition* 68.3, pp. 391–414 (cit. on p. 2).

Rouast, Philipp V, Marc TP Adam, and Raymond Chiong (2019). "Deep learning for human affect recognition: Insights and new developments". In: *IEEE Transactions on Affective Computing* 12.2, pp. 524–543 (cit. on pp. 78, 84).

Rubanova, Yulia, Ricky TQ Chen, and David K Duvenaud (2019). "Latent ordinary differential equations for irregularly-sampled time series". In: *Advances in neural information processing systems* 32 (cit. on p. 136).

Russell, James A (1980). "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6, p. 1161 (cit. on pp. 80, 92).

Russell, James A, Jo-Anne Bachorowski, and José-Miguel Fernández-Dols (2003). "Facial and vocal expressions of emotion". In: *Annual review of psychology* 54.1, pp. 329–349 (cit. on p. 87).

Salehinejad, Hojjat, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee (2017). "Recent advances in recurrent neural networks". In: *arXiv preprint arXiv:1801.01078* (cit. on p. 32).

Samanta, Atanu and Tanaya Guha (2017). "On the role of head motion in affective expression". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2886–2890 (cit. on p. 83).

— (2020). "Emotion sensing from head motion capture". In: *IEEE Sensors Journal* 21.4, pp. 5035–5043 (cit. on pp. 83, 94).

Sand, KM, A Midelfart, L Thomassen, A Melms, H Wilhelm, and JM Hoff (2013). "Visual impairment in stroke patients–a review". In: *Acta Neurologica Scandinavica* 127, pp. 52–56 (cit. on p. 5).

Santini, Thiago, Diederick C Niehorster, and Enkelejda Kasneci (2019). "Get a grip: Slippage-robust and glint-free gaze estimation for real-time pervasive head-mounted eye tracking". In: *ACM symposium on eye tracking research and applications*, pp. 1–10 (cit. on pp. 24, 54).

Saxe, Andrew M, James L McClelland, and Surya Ganguli (2014). "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks". In: *Proceedings of the International Conference on Learning Represenatations 2014*. International Conference on Learning Represenatations 2014 (cit. on p. 68).

Schlögl, Stephan, Gavi Doherty, and Saturnino Luz (2015). "Wizard of oz experimentation for language technology applications: challenges and tools." In: *Interact Comput* 27.6, pp. 592–615 (cit. on p. 85).

Schuller, Björn, Anton Batliner, Stefan Steidl, and Dino Seppi (2011). "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge". In: *Speech communication* 53.9-10, pp. 1062–1087 (cit. on p. 80).

Schuller, Björn, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim (2013). "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism". In: *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France* (cit. on p. 81).

Schuller, Björn, Felix Weninger, Yue Zhang, Fabien Ringeval, Anton Batliner, Stefan Steidl, Florian Eyben, Erik Marchi, Alessandro Vinciarelli, Klaus Scherer, Mohamed Chetouani, and Marcello Mortillaro (2019). "Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge". In: *Computer Speech & Language* 53, pp. 156–180 (cit. on p. 78).

Schuller, Björn W (2018). "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends". In: *Communications of the ACM* 61.5, pp. 90–99 (cit. on p. 84).

Schuller, Björn W, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, Harald Baumeister, Alexis Deighton MacIntyre, and Simone Hantke (2020). "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks". In: *Interspeech 2020* (cit. on p. 79).

Sesma, Laura, Arantxa Villanueva, and Rafael Cabeza (2012). "Evaluation of pupil center-eye corner vector for gaze estimation using a web cam". In: *Proceedings of the symposium on eye tracking research and applications*, pp. 217–220 (cit. on p. 23).

Sewell, Weston and Oleg Komogortsev (2010). "Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network". In: *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pp. 3739–3744 (cit. on p. 25).

Shafti, Ali, Pavel Orlov, and A Aldo Faisal (2019). "Gaze-based, context-aware robotic system for assisted reaching and grasping". In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 863–869 (cit. on p. 137).

Shan, Caifeng, Shaogang Gong, and Peter W. McOwan (May 2009). "Facial expression recognition based on Local Binary Patterns: A comprehensive study". In: *Image and Vision Computing* 27.6, pp. 803–816 (cit. on p. 82).

Shannon, Claude Elwood (1949). "Communication in the presence of noise". In: *Proceedings of the IRE* 37.1, pp. 10–21 (cit. on p. 63).

Shehu, Ibrahim Shehi, Yafei Wang, Athuman Mohamed Athuman, and Xianping Fu (2021). "Remote eye gaze tracking research: a comparative evaluation on past and recent progress". In: *Electronics* 10.24, p. 3165 (cit. on pp. 7, 22, 31).

Shen, Yiru, Oleg Komogortsev, and Sachin S Talathi (2020). "Domain Adaptation for Eye Segmentation". In: *Proc. European Conference on Computer Vision*. Springer, pp. 555–569 (cit. on p. 134).

Shepherd, Stephen V (2010). "Following gaze: gaze-following behavior as a window into social cognition". In: *Frontiers in integrative neuroscience* 4, p. 5 (cit. on p. 2).

Shih, Sheng-Wen and Jin Liu (2004). "A novel approach to 3-D gaze tracking using stereo cameras". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 34.1, pp. 234–245 (cit. on p. 23).

Shin, Y, HW Lim, MH Kang, M Seong, H Cho, and JH Kim (2016). "Normal range of eye movement and its relationship to age". In: *Acta Ophthalmologica* 94 (cit. on p. 94).

Siegfried, Rémy and Jean-Marc Odobez (2021). "Visual focus of attention estimation in 3d scene with an arbitrary number of targets". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3153–3161 (cit. on p. 6).

Simon, Gérard (1975). "14.3. On the theory of visual perception of Kepler and Descartes: reflections on the role of mechanism in the birth of modern science". In: *Vistas in Astronomy* 18, pp. 825–832 (cit. on p. 17).

Singh, Youddha Beer and Shivani Goel (2022). "A systematic literature review of speech emotion recognition approaches". In: *Neurocomputing* 492, pp. 245–263 (cit. on p. 81).

Smith, Brian A, Qi Yin, Steven K Feiner, and Shree K Nayar (2013). "Gaze locking: passive eye contact detection for human-object interaction". In: *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, pp. 271–280 (cit. on p. 33).

Sogancioglu, Gizem, Oxana Verkholyak, Heysem Kaya, Dmitrii Fedotov, Tobias Cadée, Albert Ali Salah, and Alexey Karpov (2020). "Is Everything Fine, Grandma? Acoustic and Linguistic Modeling for Robust Elderly Speech Emotion Recognition." In: *INTERSPEECH*, pp. 2097–2101 (cit. on p. 84).

Soleymani, Mohammad, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic (2011). "A multimodal database for affect recognition and implicit tagging". In: *IEEE transactions on affective computing* 3.1, pp. 42–55 (cit. on p. 78).

Spector, Robert H (1990). "The pupils". In: *Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition* (cit. on p. 82).

Sreevidya, P, S Veni, and OVR Murthy (2022). "Elder emotion classification through multimodal fusion of intermediate layers and cross-modal transfer learning". In: *Signal, Image and Video Processing* 16.5, pp. 1281–1288 (cit. on p. 85).

Steil, Julian, Inken Hagestedt, Michael Xuelin Huang, and Andreas Bulling (2019a). "Privacy-aware eye tracking using differential privacy". In: *ACM Symposium on Eye Tracking Research and Applications*, pp. 1–9 (cit. on p. 137).

Steil, Julian, Marion Koelle, Wilko Heuten, Susanne Boll, and Andreas Bulling (2019b). "Privaceye: privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features". In: *Proceedings of the 11th ACM symposium on eye tracking research & applications*, pp. 1–10 (cit. on p. 137).

Steininger, Silke, Florian Schiel, Olga Dioubina, and S Raubold (2002). "Development of user-state conventions for the multimodal corpus in smartkom". In: *Proc. Workshop on Multimodal Resources and Multimodal Systems Evaluation*. Citeseer, pp. 33–37 (cit. on p. 89).

Stone, Scott A, Quinn A Boser, T Riley Dawson, Albert H Vette, Jacqueline S Hebert, Patrick M Pilarski, and Craig S Chapman (2022). "Generating accurate 3D gaze vectors using synchronized eye tracking and motion capture". In: *Behavior Research Methods*, pp. 1–14 (cit. on p. 26).

Sugano, Yusuke, Yasuyuki Matsushita, and Yoichi Sato (2013). "Appearance-based gaze estimation using visual saliency". In: *IEEE transactions on pattern analysis and machine intelligence* 35.2, pp. 329–341 (cit. on p. 24).

— (2014). "Learning-by-synthesis for appearance-based 3d gaze estimation". In: *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, pp. 1821–1828 (cit. on pp. 24, 33, 35, 42).

Sugano, Yusuke, Xucong Zhang, and Andreas Bulling (2016). "Aggregaze: Collective estimation of audience attention on public displays". In: *Proceedings of the 29th*

Annual Symposium on User Interface Software and Technology, pp. 821–831 (cit. on p. 31).

Sweeney, John A, Yukari Takarae, Carol Macmillan, Beatriz Luna, and Nancy J Minshew (2004). "Eye movements in neurodevelopmental disorders". In: *Current opinion in neurology* 17.1, pp. 37–42 (cit. on p. 5).

Tan, Kar-Han, David J Kriegman, and Narendra Ahuja (2002). "Appearance-based eye gaze estimation". In: *Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on*. IEEE, pp. 191–195 (cit. on p. 24).

Tatler, Benjamin W, Nicholas J Wade, Hoi Kwan, John M Findlay, and Boris M Velichkovsky (2010). "Yarbus, eye movements, and vision". In: *i-Perception* 1.1, pp. 7–27 (cit. on p. 3).

Termsarasab, Pichet, Thananan Thammongkolchai, Janet C Rucker, and Steven J Frucht (2015). "The diagnostic value of saccades in movement disorder patients: a practical guide and review". In: *Journal of clinical movement disorders* 2.1, pp. 1–10 (cit. on p. 137).

Tonsen, Marc, Julian Steil, Yusuke Sugano, and Andreas Bulling (2017). "Invisibleeye: Mobile eye tracking using multiple low-resolution cameras and learning-based gaze estimation". In: *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3, pp. 1–21 (cit. on pp. 56, 59).

Torres, M. I., J. M. Olaso, C. Montenegro, R. Santana, A. Vázquez, R. Justo, J. A. Lozano, S. Schlögl, G. Chollet, N. Dugan, M. Irvine, N. Glackin, C. Pickard, A. Esposito, G. Cordasco, A. Troncone, D. Petrovska-Delacretaz, A. Mtibaa, M. A. Hmani, M. S. Korsnes, L. J. Martinussen, S. Escalera, C. Palmero Cantariño, O. Deroo, O. Gordeeva, J. Tenorio-Laranga, E. Gonzalez-Fraile, B. Fernandez-Ruanova, and A. Gonzalez-Pinto (2019). "The EMPATHIC Project: Mid-Term Achievements". In: *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA '19. Rhodes, Greece, 629–638 (cit. on p. 79).

Tsai, Yao-Hung Hubert, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov (2019). "Multimodal transformer for unaligned multimodal language sequences". In: *Proceedings of the conference. Association for Computational Linguistics. Meeting*. Vol. 2019. NIH Public Access, p. 6558 (cit. on pp. 84, 122).

Tzirakis, Panagiotis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou (2017). "End-to-end multimodal emotion recognition using deep neural networks". In: *IEEE Journal of selected topics in signal processing* 11.8, pp. 1301–1309 (cit. on pp. 121, 122).

Valliappan, Nachiappan, Na Dai, Ethan Steinberg, Junfeng He, Kantwon Rogers, Venky Ramachandran, Pingmei Xu, Mina Shojaeizadeh, Li Guo, Kai Kohlhoff, and Vidhya Navalpakkam (2020). "Accelerating eye movement research via accurate and affordable smartphone eye tracking". In: *Nature communications* 11.1, p. 4553 (cit. on p. 7).

Valstar, Michel, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic (2014). "AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge". In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. AVEC '14. New York, NY, USA: ACM, pp. 3–10 (cit. on p. 81).

Van Essen, David C (2003). "Organization of visual areas in macaque and human cerebral cortex". In: *The visual neurosciences* 1, pp. 507–521 (cit. on p. 1).

Van Huynh, Thong, Hyung-Jeong Yang, Guee-Sang Lee, Soo-Hyung Kim, and In-Seop Na (2019). "Emotion recognition by integrating eye movement analysis and

facial expression model". In: *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*, pp. 166–169 (cit. on p. 83).

Vidal, Mélodie, Jayson Turner, Andreas Bulling, and Hans Gellersen (2012). "Wearable eye tracking for mental health monitoring". In: *Computer Communications* 35.11, pp. 1306–1311 (cit. on p. 137).

Villanueva, Arantxa and Rafael Cabeza (2008). "A novel gaze estimation system with one calibration point". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 38.4, pp. 1123–1138 (cit. on p. 23).

Vorontsov, Eugene, Chiheb Trabelsi, Samuel Kadoury, and Chris Pal (2017). "On orthogonality and learning recurrent networks with long term dependencies". In: *International Conference on Machine Learning*. PMLR, pp. 3570–3578 (cit. on p. 68).

Vázquez, Alain, Asier López Zorrilla, Javier Mikel Olaso, and María Inés Torres (2023). "Dialogue Management and Language Generation for a Robust Conversational Virtual Coach: Validation and User Study". In: *Sensors* 23.3 (cit. on pp. 78, 90).

Wade, Nicholas J (2010). "Pioneers of eye movement research". In: *i-Perception* 1.2, pp. 33–68 (cit. on pp. 2, 17).

— (2015). "How were eye movements recorded before yarbus?" In: *Perception* 44.8-9, pp. 851–883 (cit. on p. 17).

Wade, Nicholas J and Benjamin W Tatler (2008). "Did Javal measure eye movements during reading?" In: *Journal of Eye Movement Research* 2.5 (cit. on p. 18).

Wagner, Johannes, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller (2023). "Dawn of the transformer era in speech emotion recognition: closing the valence gap". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (cit. on p. 81).

Wang, Chengyi, Yu Wu, Yao Qian, Kenichi Kumatani, Shujie Liu, Furu Wei, Michael Zeng, and Xuedong Huang (2021a). "Unispeech: Unified speech representation learning with labeled and unlabeled data". In: *International Conference on Machine Learning*. PMLR, pp. 10937–10947 (cit. on p. 81).

Wang, Jianyou, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding, and Vahid Tarokh (2020). "Speech Emotion Recognition with Dual-Sequence LSTM Architecture". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6474–6478 (cit. on p. 81).

Wang, Kang and Qiang Ji (2017). "Real Time Eye Gaze Tracking with 3D Deformable Eye-Face Model". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1003–1011 (cit. on p. 23).

Wang, Kang, Hui Su, and Qiang Ji (2019). "Neuro-inspired eye tracking with eye movement dynamics". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9831–9840 (cit. on pp. 18, 45, 134).

Wang, Kang, Rui Zhao, and Qiang Ji (2018). "A hierarchical generative model for eye image synthesis and eye gaze estimation". In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 440–448 (cit. on pp. 55, 134).

Wang, Kunxia, Ning An, Bing Nan Li, Yanyong Zhang, and Lian Li (2015). "Speech emotion recognition using Fourier parameters". In: *IEEE Transactions on affective computing* 6.1, pp. 69–75 (cit. on p. 79).

Wang, Kunxia, Qingli Zhang, and Siyu Liao (2014). "A database of elderly emotional speech". In: *Proc. Int. Symp. Signal Process. Biomed. Eng Informat*, pp. 549–553 (cit. on p. 79).

Wang, Kunxia, ZongBao Zhu, Shidong Wang, Xiao Sun, and Lian Li (2016). "A database for emotional interactions of the elderly". In: *IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pp. 1–6 (cit. on p. 79).

Wang, Ning, Wengang Zhou, Jie Wang, and Houqiang Li (2021b). "Transformer meets tracker: Exploiting temporal context for robust visual tracking". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1571–1580 (cit. on p. 7).

Wang, Yaoming, Yangzhou Jiang, Jin Li, Bingbing Ni, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Teng Li (2022). "Contrastive regression for domain adaptation on gaze estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19376–19385 (cit. on p. 134).

Wedel, Michel and Rik Pieters (2017). "A review of eye-tracking research in marketing". In: *Review of marketing research*, pp. 123–147 (cit. on p. 5).

Williams, Oliver, Andrew Blake, and Roberto Cipolla (2006). "Sparse and Semi-supervised Visual Mapping with the Sˆ 3GP". In: *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. Vol. 1. IEEE, pp. 230–237 (cit. on p. 24).

Winkler, Stefan and Ramanathan Subramanian (2013). "Overview of eye tracking datasets". In: *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, pp. 212–217 (cit. on p. 57).

Wisiecka, Katarzyna, Krzysztof Krejtz, Izabela Krejtz, Damian Sromek, Adam Cellary, Beata Lewandowska, and Andrew Duchowski (2022). "Comparison of webcam and remote eye tracking". In: *2022 Symposium on Eye Tracking Research and Applications*, pp. 1–7 (cit. on p. 7).

Wollaston, William Hyde (1824). "XIII. On the apparent direction of eyes in a portrait". In: *Philosophical Transactions of the Royal Society of London* 114, pp. 247–256 (cit. on pp. 34, 35).

Wood, Erroll, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton (2021). "Fake it till you make it: face analysis in the wild using synthetic data alone". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3681–3691 (cit. on p. 133).

Wood, Erroll, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling (2016a). "A 3D Morphable Eye Region Model for Gaze Estimation". In: *Proc. European Conference on Computer Vision*, pp. 297–313 (cit. on pp. 33, 54, 58).

— (2016b). "Learning an appearance-based gaze estimator from one million synthesised images". In: *ACM Symp. Eye Tracking Research and Applications*, pp. 131–138 (cit. on pp. 24, 33, 54, 55, 58).

Wood, Erroll, Tadas Baltrušaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling (2015). "Rendering of Eyes for Eye-Shape Registration and Gaze Estimation". In: *Proc. IEEE International Conference on Computer Vision*, pp. 3756–3764 (cit. on pp. 33, 54, 58).

Wu, Chung-Hsien, Jen-Chun Lin, and Wen-Li Wei (2014). "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies". In: *APSIPA transactions on signal and information processing* 3, e12 (cit. on p. 84).

Wu, Suowei, Zhengyin Du, Weixin Li, Di Huang, and Yunhong Wang (2019). "Continuous emotion recognition in videos by fusing facial expression, head pose and eye gaze". In: *2019 International Conference on Multimodal Interaction*, pp. 40–48 (cit. on p. 84).

Xiong, Xuehan, Zicheng Liu, Qin Cai, and Zhengyou Zhang (2014). "Eye gaze tracking using an RGBD camera: a comparison with a RGB solution". In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pp. 1113–1121 (cit. on p. 7).

Xu, Mingjie, Haofei Wang, Yunfei Liu, and Feng Lu (2021). "Vulnerability of appearance-based gaze estimation". In: *arXiv preprint arXiv:2103.13134* (cit. on p. 133).

Xue, Tong, Abdallah El Ali, Gangyi Ding, and Pablo Cesar (2021). "Investigating the relationship between momentary emotion self-reports and head and eye movements in hmd-based 360 vr video watching". In: *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pp. 1–8 (cit. on p. 83).

Yang, Tao, Zeyun Yang, Guangzheng Xu, Duoling Gao, Ziheng Zhang, Hui Wang, Shiyu Liu, Linfeng Han, Zhixin Zhu, Yang Tian, Yuqi Huang, Lei Zhao, Kui Zhong, Bolin Shi, Juan Li, Shimin Fu, Peipeng Liang, Michael J. Banissy, and Pei Sun (2020). "Tsinghua facial expression database–A database of facial expressions in Chinese young and older women and men: Development and validation". In: *PloS one* 15.4 (cit. on p. 79).

Yarbus, Alfred L (1967). "Eye movements during perception of complex objects". In: *Eye movements and vision*, pp. 171–211 (cit. on pp. 2, 18–20).

Yiu, Yuk-Hoi, Moustafa Aboulatta, Theresa Raiser, Leoni Ophey, Virginia L Flanagin, Peter Zu Eulenburg, and Seyed-Ahmad Ahmadi (2019). "DeepVOG: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning". In: *Journal of neuroscience methods* 324, p. 108307 (cit. on pp. 4, 22, 56).

Young, Laurence R and David Sheena (1975). "Survey of eye movement recording methods". In: *Behavior research methods & instrumentation* 7.5, pp. 397–429 (cit. on pp. 19, 20).

Yu, Yu and Jean-Marc Odobez (2020). "Unsupervised representation learning for gaze estimation". In: *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7314–7324 (cit. on pp. 25, 56).

Zangemeister, Wolfgang H. and Lawrence Stark (1981). "Active head rotations and eye-head coordination". In: *Annals of the New York Academy of Sciences* 374.1, 540–559 (cit. on p. 83).

Zemblys, Raimondas and Oleg Komogortsev (2018). "Making stand-alone PS-OG technology tolerant to the equipment shifts". In: *Proc. 7th Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction*, pp. 1–9 (cit. on pp. 56, 58).

Zeng, Zhihong, M Pantic, GI Roisman, and TS Huang (2009). "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1.31, pp. 39–58 (cit. on pp. 79, 84).

Zhang, Shifeng, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li (2017a). "Faceboxes: A CPU Real-time Face Detector with High Accuracy". In: *IEEE International Joint Conference on Biometrics* (cit. on p. 92).

Zhang, Xucong, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges (2020). "ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation". In: *Proc. European Conference on Computer Vision*. Springer, pp. 365–381 (cit. on pp. 25, 57, 93).

Zhang, Xucong, Yusuke Sugano, and Andreas Bulling (2018). "Revisiting data normalization for appearance-based gaze estimation". In: *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pp. 1–9 (cit. on pp. 35, 42).

— (2019). "Evaluation of appearance-based methods and implications for gaze-based applications". In: *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–13 (cit. on p. 56).

Zhang, Xucong, Yusuke Sugano, Mario Fritz, and Andreas Bulling (2015). "Appearance-based gaze estimation in the wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4511–4520 (cit. on pp. 4, 24, 31, 33, 35).

— (2017b). "It's written all over your face: Full-face appearance-based gaze estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 51–60 (cit. on pp. 32–34, 41, 56).

— (2017c). "Mpiigaze: Real-world dataset and deep appearance-based gaze estimation". In: *IEEE transactions on pattern analysis and machine intelligence* 41.1, pp. 162–175 (cit. on pp. 24, 57).

Zhou, Xiaolong, Jianing Lin, Jiaqi Jiang, and Shengyong Chen (2019). "Learning A 3D Gaze Estimator with Improved Itracker Combined with Bidirectional LSTM". In: *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, pp. 850–855 (cit. on p. 45).

Zhu, Zhiwei, Kikuo Fujimura, and Qiang Ji (2002). "Real-time eye detection and tracking under various light conditions". In: *Proceedings of the 2002 symposium on Eye tracking research & applications*, pp. 139–144 (cit. on p. 24).

Zhu, Zhiwei and Qiang Ji (2005). "Eye gaze tracking under natural head movements". In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE, pp. 918–923 (cit. on p. 24).

— (2007). "Novel eye gaze tracking techniques under natural head movement". In: *IEEE Transactions on biomedical engineering* 54.12, pp. 2246–2260 (cit. on p. 23).