EPFL

Thèse n° 8509

Modeling and Inferring Attention between Humans or for Human-Robot Interactions

Présentée le 2 décembre 2021

Faculté des sciences et techniques de l'ingénieur
Laboratoire de l'IDIAP
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Rémy Alain SIEGFRIED

Acceptée sur proposition du jury

Dr D. Gillet, président du jury
Dr J.-M. Odobez, directeur de thèse
Prof. E. Kasneci, rapporteuse
Prof. Y. Demiris, rapporteur
Prof. P. Dillenbourg, rapporteur

École
polytechnique
fédérale
de Lausanne

2021

He (Hari Seldon) smiled and said, “Can you read, Raych?”
Raych spat again. “Nab! Who wants to read?”
“Can you use a computer?”
“A talking computer? Sure. Anyone can.”
— I. Asimov, Prelude to Foundation, 1988

To my beloved fiancee, my parents, my brother, my friends
and to everyone who contributed to make me who I am and teached me what I know...

Acknowledgements

The beginning of a thesis, like many other experiences in life, starts when someone gives you an opportunity. I started this unexpected journey in November 2016, when, after some interviews, Jean-Marc offered me to start a PhD and thus become a researcher. He introduced me to research and coached me in reading, reviewing, writing, and presenting papers, pointing out when, after several rounds of rewriting, my sentences no longer made sense. In addition, he always emphasized critical thinking and understanding of what we were doing. So thank you Jean-Marc that you gave me this opportunity and for guiding me through these doctoral studies.

During this four and a half years, I learned that time is precious, that at some point we cannot do everything we have in mind, and that choosing among ideas is as crucial as it is difficult. Thus, I would like to thank Dr. D. Gillet, Prof. E. Kasneci, Prof. Y. Demiris, and Prof. P. Dillenbourg for the valuable time they spent reading and listening to me, and for the interesting discussions that occured during my oral exam.

Also, this work would not have been possible without funding, so I would like to thank the European Union for its Horizon 2020 research and innovation program (grant n° 688147, MuMMER project) which funded most of my research, as well as the Idiap Research Institute for their "Valais-Wallis Ambition" initiative, which allowed more flexibility in the organization of my thesis work.

Then, I would like to thank colleagues I worked with, for the help they gave me and the things I learned from them. A special thanks to Olivier, who gave me a lot of technical support and helped me to further improve my programming skills. I also had the opportunity to work with several people throughout my PhD and I would like to thank Kenneth, Yu, Skanda, Bozorgmehr, and Michael for what I learned from them and the exciting discussions we had.

Next, I would like to thank the members of the Perception and Activity Understanding group for our technical discussions in the group meetings, for their feedbacks on my work, and for the good time we had together. So thanks to Michael, Nam, Yu, Weipeng, Angel, Gang, Yuanzhouhan, Nicolas, Bozorgmehr, Mattia, Marco, Samy, and Anshul. A special thanks to Angel, who did his PhD at the same time as me. It was nice to have a companion along the way.

Working at Idiap also allowed me to meet nice and interesting people, that I would like to thank for the time I spent with them. I would like to thank especially Sandrine for sharing with me her peculiar world vision, for our philosophical sharings, and for her support; Noémie for guiding

me on the path of research, for our interesting discussions, and her always frank and sincere point of view; Emmanuel for his creative mind, his optimism, and his "brainstorming way" to approach discussions; and Christian for his sanely sarcastic vision of things, his enthusiasm for technique, and his sense of humor. Also, a big thank to Bozorgmehr, Nicolas, Thibault, Hakan, Olivia, Adrian, and Samuel for the nice discussions we shared.

Finally, I would like to thank all the people who surrounded me during this work. I would like to take this opportunity to thank my parents for their unconditional love and support, for the solution-oriented mindset they teached me, and for providing me with the necessary logistics to follow my path. I thank my brother for our complicity and the precious time we have spent and continue to spend together. I also thank my friends for being there, reminding me that life goes on despite the difficulties.

My last thanks will be for Ludmilla, my beloved fiancée. She was my "Sam" when I climbed my "Mount Doom": unable to carry my burden but always there to carry me, always there when I thought I was alone. We are the "super cool team" and we will always be.

Lausanne, 15 July 2021 Rémy

Abstract

More and more intelligent systems have to interact with humans. In order to communicate efficiently, these systems need to perceive and understand us. A key factor of communication is the people's visual focus of attention (VFOA), which is useful to estimate addressees and engagement among others. It is also strongly related to the gaze, its continuous counterpart, whose analysis allows to estimate high-level features, e.g. confidence and tiredness. Beyond communication, interesting statistics can be derived from eye movements themselves. For instance, fixation duration and blink rate were shown to be related to mental health. Thus, VFOA, gaze, and eye movements estimation have great potential in a wide range of fields, like human behaviors analysis, human-computer interactions, psychiatric diagnosis, and so on. However, despite recent improvements in sensors and methods, the precise tracking of people's gaze and VFOA remains difficult without using intrusive sensors or constraining people's movement, which do not suit applications where users behaving naturally are recorded by remote sensors, like in many human-robot interactions or psychological studies, for example.

This thesis introduces new approaches to improve gaze and VFOA estimation from videos recorded by remote sensors which could be embedded on robots or be part of a room monitoring system for example. It proposes an unsupervised and online method to calibrate gaze trackers from attention priors used in conversation and manipulation, removing the need for a dedicated calibration session. Also, it proposes a method to estimate the VFOA in setups with an arbitrary and dynamic number of visual targets by using a fixed-sized representing for all the subject and context-related features. Finally, it proposes an eye movements recognition to detect saccades and blinks in eye image sequences without the need for accurate gaze traces.

These approaches were validated through experiments on human conversation and object manipulation recordings. Overall, by focusing on the improvement of VFOA and gaze estimation usability, this thesis attempts to make a step toward the democratization of these methods and their application to weakly constrained setups relying on cheap sensing devices.

Keywords: remote sensors, human interactions, visual focus of attention, 3D gaze, eye movements, gaze calibration, appearance-based methods.

Résumé

De plus en plus de systèmes intelligents sont tenus d'interagir avec des humains. Afin de communiquer efficacement, ces systèmes doivent donc nous percevoir et nous comprendre. Un facteur clé de la communication réside dans la focalisation visuelle de l'attention (FVA), qui est utile pour savoir par exemple à qui s'adresse une personne et son niveau d'engagement. Elle est également fortement liée au regard, son pendant continu, dont l'analyse permet d'estimer des caractéristiques de haut niveau, comme la confiance en soi et la fatigue. Au-delà de l'aspect communicationnel, des statistiques intéressantes peuvent être inférées des mouvements oculaires. Par exemple, il a été démontré que la durée de fixation du regard et le taux de clignement des yeux peuvent être reliés à la santé mentale. Ainsi, l'estimation de la FVA, du regard et des mouvements oculaires présente un grand potentiel dans un large éventail de domaines, comme l'analyse des comportements humains, les interactions homme-machine et le diagnostic psychiatrique. Cependant, malgré les récents progrès en matière de capteurs et de méthodologie, le suivi précis du regard et de la FVA des personnes reste difficile sans utiliser des capteurs intrusifs et sans contraindre les mouvements des personnes. Ces restrictions ne conviennent pas aux applications où les utilisateurs se comportent naturellement et sont enregistrés par des capteurs à distance, comme dans de nombreuses interactions homme-robot ou lors d'études psychologiques, par exemple.

Cette thèse introduit de nouvelles approches pour améliorer l'estimation du regard et de la FVA à partir de vidéos enregistrées par des capteurs à distance qui pourraient être embarqués sur des robots ou faire partie d'un système de surveillance d'une pièce par exemple. Elle propose une méthode non supervisée et en ligne pour calibrer des traqueurs de regard à partir d'aprioris sur l'attention visuelle étudiés dans des contexts de conversations et de manipulations d'objets, éliminant le besoin d'une session de calibration dédiée. Elle propose également une méthode pour estimer la FVA dans des situations avec un nombre arbitraire et dynamique de cibles visuelles, en utilisant une représentation de taille fixe pour toutes les caractéristiques liées au sujet et au contexte. Enfin, elle propose une reconnaissance des mouvements oculaires pour détecter les saccades et les clignements dans des séquences d'images oculaires sans avoir besoin d'estimer précisement le regard.

Ces approches ont été validées par des expériences sur des enregistrements de conversations humaines et de manipulations d'objets. Globalement, en se concentrant sur l'amélioration de l'utilisabilité de l'estimation de la FVA et du regard, cette thèse tente de faire un pas en avant vers la démocratisation de ces méthodes et leur application à des environments faiblement

contraints et à des capteurs peu coûteux.

Mots-clés : capteurs à distance, interactions humaines, cible de l'attention visuelle, regard 3D, mouvements oculaires, calibration du regard, méthodes basées sur l'apparance.

Contents

List of Figures

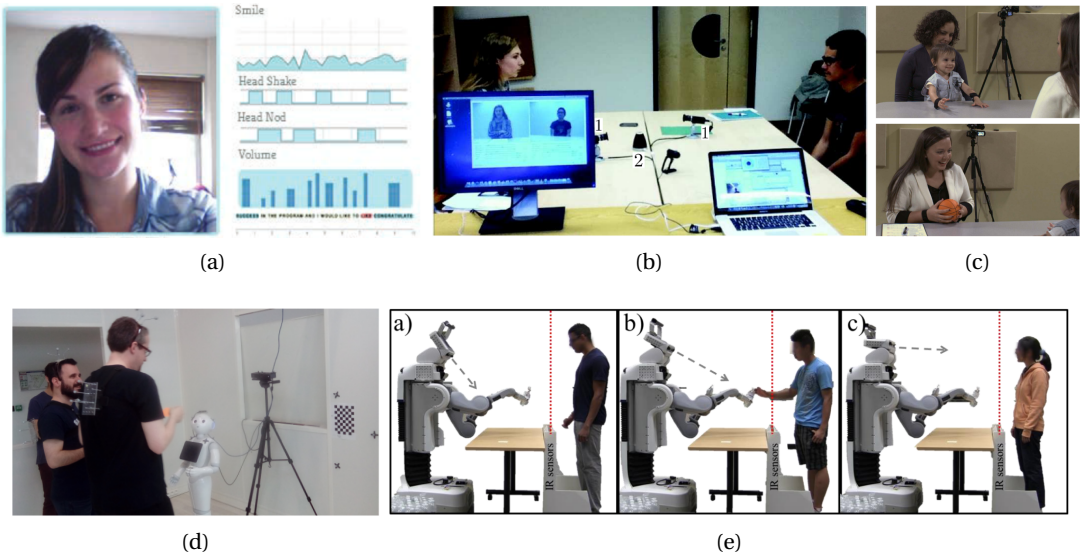
List of Tables

1 Introduction

The advances in computer science and hardware allow intelligent systems to understand us better and better and the design of systems that interact naturally with humans, like robots or intelligent personal assistants, lately grew in interest. Indeed, these systems present many advantages. Using an intuitive interface, i.e. natural human communication, decreases the amount of skill required for the user and thus increases the accessibility to technology, e.g. for elderly people. Moreover, a better interpretation of the user's instructions and an increased autonomy would reduce the required amount of supervision, letting the user focus on other tasks, like in the case of driving assistance. Generally speaking, the automatic detection and understanding of human behaviors has a great potential in a wide range of applications.

Nevertheless, implementing such systems is still challenging. Their perception abilities are usually limited to verbal components, although human communication also relies heavily on non-verbal components. These last years have seen great progress in computer vision, whether it is in object detection (Redmon et al., 2016), people detection (Cao et al., 2019), or human features extraction, like 3D body pose (Martínez-González et al., 2020a,b), 3D head pose (Chen et al., 2011; Yu et al., 2018a), gaze direction (Funes Mora and Odobez, 2014; Liu et al., 2020), etc. Applying these technologies to intelligent systems would allow them to better understand humans and thus to better meet user's needs.

In this thesis, we focus on estimating the visual focus of attention (VFOA) of people and related features like gaze, its continuous counterpart, and eye movements, their related underlying physical behavior. Indeed gaze and VFOA play a major role in human-human interactions (HHI) (Gatica-Perez et al., 2014; Kendon, 1967), as they are involved in communication for floor control management (Bohus and Horvitz, 2010; Oertel et al., 2015) or signaling addressees (Jayagopi et al., 2013) and engagement (Bohus and Horvitz, 2009). They are also related to higher-level constructs, like thought process (e.g. cognitive load), personality traits, like conscientiousness (Batrinca et al., 2011), leadership (Sanchez-Cortes et al., 2013), and dominance (Hung et al., 2011), or the state of mind, like stress (Huang et al., 2016) and distraction (Sigari et al., 2014).



(a) (b) (c)

(d) (e)

Figure 1.1 – Examples of applications that can profit from VFOA, gaze, and/or eye movements estimation. (a) Feedback from an automated conversation coach (Hoque et al., 2013). (b) Monitored interview setup that allows the nonverbal behavior from an automated conversation coach (Nguyen et al., 2014). (c) Adult-children dyadic interaction setup for child actions recognition (Rehg et al., 2013). (d) A conversational robot in interaction with multiple users (Canévet et al., 2020). (e) Different gaze strategies during handovers (left to right: none, looks at the object, looks at the user) (Moon et al., 2014).

In the remaining part of this section, we first present some application cases for VFOA and gaze estimation, as well as eye movement recognition. Then, we introduce our study case and the tasks we will address in this thesis. Finally, we state the goal of this thesis and present our contributions toward this goal.

1.1 Application examples

The capacity to estimate gaze and VFOA has great potential in many fields, as shown by the following examples. Some of them are illustrated in Fig. 1.1.

1.1.1 Human-human interaction (HHI) analysis

There is a growing interest in collaboration between social and computer sciences regarding social interaction analyses (Reiter-Palmon et al., 2017) and, as introduced before, attention is central in human communication. Indeed, experts in human behaviors have a high interest in the study of gaze behavior in a lot of different scenarios (Kleinke, 1986). Usually, when conducting studies, the gaze direction is manually annotated which is time-consuming and annotations consist in overall directions rather than in precise directions, as it is difficult for

people to predict the exact direction of the gaze. Thus, a lot of studies could benefit from automatic gaze estimation (Funes Mora and Odobez, 2014), as it would significantly reduce the cost in time and manpower of gaze annotations (Frauendorfer et al., 2014).

Also, interpersonal communication skills are critical in many situations, like in job interviews (Nguyen et al., 2014) or for professions involving interaction with customers (Sundaram and Webster, 2000) or patients. In such cases, there is an interest in behavioral training systems (Hoque et al., 2013; Muralidhar et al., 2016), that could also provide feedback about gaze behavior.

1.1.2 Psychiatric diagnosis

Psychology research showed that gaze is related to cognitive ability (Knapp et al., 2013) and can, to some extent, reflect the mental state. Indeed several gaze characteristics were shown to be influenced by cognitive processes, which makes them potential markers to monitor mental health (Vidal et al., 2012).

For example, studies showed that blink rate (Zhou et al., 2015), eye contact frequency (Scherer et al., 2014), and glance duration at positive stimuli (Isaac et al., 2014; Duque and Vázquez, 2015) are significantly altered in patients with depression. Thus, detecting gaze patterns and eye behaviors using low-end sensors could help to monitor health out of the lab, e.g. at home.

Another example is the support of the Autism Spectrum Disorders (ASD) diagnosis. The high prevalence of ASD (Baio, 2018) compared to the limited diagnosis resources (Ning et al., 2019) raised the interest for assistive technology, which could, among others, perform automatic diagnosis or at least automatic detection of social interaction patterns. Indeed, research showed that children with ASD look less at other people's eyes (Boraston and Blakemore, 2007) and monitor their gaze during joint attention tasks differently than typically developing children (Anzalone et al., 2019). Detecting such behaviors could help experts by making a first rough classification of the patients or, when the diagnosis is made from a video, to select the most relevant interaction part to assess.

1.1.3 Human-computer interaction (HCI)

One of the most famous applications in HCI is the use of eye movements and gaze tracking as an input medium for computers and other devices like tablets (Holland et al., 2013) or smartphones (Krafka et al., 2016). Although it can be used by any user, these technologies was successfully implemented to allow people whose disabilities prevent the use of traditional input methods to use computers (Lupu et al., 2013; Grauman et al., 2003).

Another application is the detection of the user's state, which allows the system to inform the user and/or to adapt to the situation. For example, the detection of drowsiness (Soukupova and Cech, 2016; Dreißig et al., 2020) or inattention (Fletcher and Zelinsky, 2009) of car drivers

could help to reduce the frequency of car accidents.

Also, gaze information can be used to adapt computers' processes, like in foveated rendering applied in virtual reality (Patney et al., 2016), which allows a significant display speedup by rendering fewer details in the peripheral field of view of the user.

1.1.4 Human-robot interaction (HRI)

There is an increasing interest in robots that communicate with users (Bennewitz et al., 2007; Foster et al., 2012) and the most obvious cases are conversational robots. Indeed, analyzing and synthesizing gaze behaviors can improve HRI by both having robots better understand human intentions and making their actions more natural in a social context, allowing them to appropriately answer humans (Moon et al., 2014) or improve the perception toward the robot(Andrist et al., 2014). For example, a conversational robot could understand when a question is addressed to it and when it can answer the user, or when someone wants to interact (Bohus and Horvitz, 2009) Also, it would allow it to better estimate users' addressees and disambiguate the situation when multiple users are interacting with the robot.

Furthermore, learning by demonstration could also benefit from attention estimation, as people's gaze often reveals the importance of a motion and they rarely looks at objects that are not relevant to the current task (Hayhoe and Ballard, 2005). For example, when people move a glass of water: they will look directly to the destination point if it is empty, but look at it during the whole movement if it is full. This difference is significant of the need to not spill the water and the motion will be highly affected by this constraint (slower and more careful motions). Thus, detecting these differences could help a robot to focus on important parts of the motion to learn. Also, the perception and propoer usage of the gaze was shown to improve human robot collaboration during handovers (Strabala et al., 2013; Moon et al., 2014).

1.2 Challenges

Despite the interest in gaze estimation, scene understanding and the perception of humans by intelligent systems are far from being solved. Indeed, although human feature extraction like body pose and gaze estimation improved a lot in the last years, intelligent systems are usually used in setups that make the integration of such technology difficult as several challenges remain to be addressed.

- **Remote and embedded sensors.** The usage of wearable, intrusive, or constraining sensors is difficult as it makes the interaction less natural. For example, it would be difficult to convince users to put dedicated head-mounted eye trackers before interacting with an information robot in a shopping mall. Also, in psychological studies and assessments, constraining the patient's movements might alter their behaviors and bias the interaction. Thus, they should rely on embedded and remote sensors, which produce

noisier data at a lower sampling rate compared to laboratory equipment and usually provide only partial scene information. The proper position of these sensors is also a problem faced in most interaction setups, as the viewpoint will affect the system's sensing capabilities.
- **Dynamic environment.** These systems are often exposed to highly dynamic environments, whether due to movements of the user, like in surveillance applications, or of the system itself, e.g. in robotic cases, or due to unexpected external events that affect the interaction.
- **Environment diversity.** The diversity of encountered environments requires robust and/or adaptive perception and behavior design, which can be challenging to achieve.
- **Real-time computing.** All the computation must be done in real-time, despite the usually limited available computational power.

VFOA and gaze estimation are also challenging in themselves.

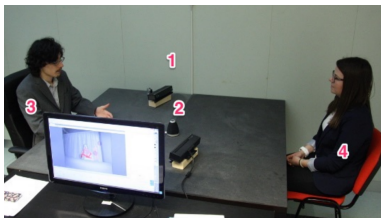
- **Scene monitoring.** VFOA estimation relies on both the gaze direction of the user but also on scene information, like the positions of potential visual targets or other contextual information. Such scene monitoring can be difficult to get depending on the setup.
- **Low resolution image.** Accurate gaze estimation is difficult when applied to pictures and videos recorded through remote sensors, because the distance between the sensor and the user leads to lower eye image resolution compared to high end gaze tracking devices (e.g. wearables). For example, the eye of a person recorded at 2m by an HD camera will appear with a size of approximately 40x60 pixels.
- **Dynamic environment.** The unconstraint behavior of the user leads to a big variance in illumination, head pose, and facial expression, which modify the eyes' appearance and thus perturbates the estimation.
- **User diversity.** The difference of appearance between users and the invisble anatomical differences in eye characteristics are usually addressed by a calibrating the gaze tracker, i.e. by adapting the gaze estimation model to the user. However, the nature of the interaction often makes it difficult to get a proper calibration session,

All these challenges make difficult the transfer of current human sensing technologies from the laboratory to the "real world". Thus, there is a need for improvement in gaze and attention estimation methods accuracy, e.g. by limiting their sensibility to data quality and increasing their robustness, but also in usability by designing more flexible and generalizable models, so that they can adapt to new users and environments.

1.3 Study case

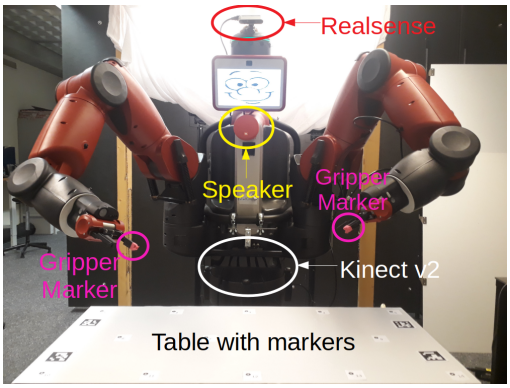
1.3.1 Setups of interest

The data, setups, and method involved in our work were chosen/designed to address some of the above challenges. From a setup perspective, we focused on situations where people



(a) Dyadic interaction.

(b) Multi-party meeting.



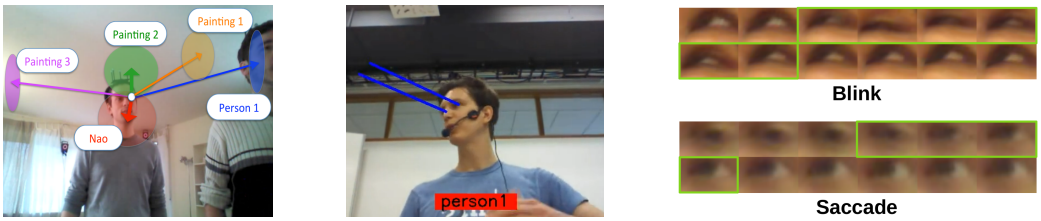
(c) HRI.

Figure 1.2 – Setups of interest for the present work. (a) Dyadic interaction recorded by two Kinect2 sensors and a microphone array in the center of the table. (b) Four-party meeting recorded by four Kinect and four GoPro cameras as well as lapel microphones. (c) HRI setup for object manipulation (user point of view), where a Baxter collaborative robot is equipped with two additional RGB-D cameras to record the user and the workspace.

are not constraint in their behavior, i.e. where they are free in the way they perform the task. Moreover, we rely on consumer cameras used remotely. Although they are not embedded sensors strictly speaking, they have similar image resolutions, sampling rates, and operating distances. From the methodology point of view, we built our work on a previously developed approach that leverages depth information to robustly estimate head pose, gaze, and attention. Furthermore, we focused on developing methods that can adapt to new setups without the need for supervision.

However, to frame our study and focus on the above points, we had to let aside other aspects. We did not focus on computing speed and we did not systematically check the computational load of the proposed methods. Moreover, we worked on data recorded in monitored environments using multiple sensors (mainly cameras and microphones), so we did not have to interpret the whole scene with a single partial point of view, letting aside the question of the scene monitoring.

Fig. 1.2 presents the setups used in this work (more details in Chapter 4). In all three cases, each participant performs the given task, i.e. participating in a conversation or manipulating objects, while the scene is monitored by one RGB-D camera per person, an additional camera recording manipulation actions when needed, and audio sensors to get voice activity (lapel microphones or microphone array). Lapel microphones are not remote sensors, but they could be reasonably replaced with microphone arrays (He et al., 2018a,b). The availability of depth information combined with the known position of cameras relatively to each other allow to build a complete 3D scene representation and thus to leverage geometrical reasoning



(a) VFOA (b) 3D gaze (c) eye movements

Figure 1.3 – Features of interest for the present work. (a) Given a subject, estimate his VFOA implies a categorical decision among potential visual targets. (b) Given a subject, estimate his 3D gaze implies the estimation of a 3D vector. (c) Given a sequence of eye images, eye movements recognition is performed by segmenting and labelling the action.

in our methods.

1.3.2 Addressed tasks

In this work, we focus on the estimation of three features, namely VFOA, gaze, and eye movements, that are illustrated in Fig. 1.3 and described below.

Visual focus of attention (VFOA). The VFOA is a categorical value designing the target of the visual attention of someone, i.e. the person or object that this person is looking at. It is usually estimated from the gaze direction (Funes Mora and Odobez, 2016; Salam et al., 2016) or from the head pose (Otsuka et al., 2006; Ba and Odobez, 2008a), when the former is not available. In both cases, it requires to get the position of the potential visual targets and to make a decision among them to set the VFOA. There are moments when a person is not attending to anything in paticular, typically while thinking. Such case is called an aversion and the VFOA is then undefined.

3D gaze. In this work, when we talk about gaze, we always refer to the 3D gaze (unless otherwise specified) which is defined as a direction in the 3D space and can thus be represented either by a 3D vector (gaze vector) or by 2 angles (gaze angles). It is estimated from eye images, so each eye have ist own gaze, which theoretically converge and intersect at the point of gaze. However in practice, intersection is not guaranteed due to estimation noise and inaccuracy, so they are usually combined in a single gaze for the person, e.g. by averaging left and right gazes or by considering only the most visible eye to avoid perturbation due to partial eye image.

Eye movements. The eye movements are categorical values describing actions of the eye ball and, unlike VFOA and gaze, they are dynamical in nature and only make sense over time. Beyond their intrinsic utility, recognizing eye movements can help to better estimate the gaze (Feng et al., 2011; Wang et al., 2019). For the same reasons, we are also interested in detecting blinks, even if they are note eye movements strictly speaking. Most current eye movements are:

- fixation: action of the eye looking at the same point over time;
- saccade: action of the eye moving to another fixation point;
- pursuit: action of the eye following a smoothly moving target;
- blink: action of closing and opening the eye.

There are several other more subtle eye movements like post-saccadic oscillations or micro-saccades among others, but the temporal and spatial resolution of the system studied in this work do not allow to distinguish them.

1.4 Goal and contributions

The main objective of this thesis is to improve the accuracy and the usability of gaze and VFOA estimation in weakly constrained settings relying on consumer sensors. To fulfil this goal, we made the following contributions.

1.4.1 Unsupervised and online context-based gaze calibration.

We proposed a method that exploits context knowledge to provide a weak estimation of the VFOA and exploit this estimation to perform a person-specific calibration of gaze estimators in an unsupervised and online way. Especially, we showed how task contextual attention priors can be used to gather reference gaze samples, which is a cumbersome process otherwise. Also, we proposed a robust estimation framework to exploit these weak labels for the estimation of the calibration model parameters. Finally, we demonstrated the applicability of this approach on two HHI and one HRI settings, namely dyadic interaction, multi-party meeting and object manipulation. Experiments on three datasets validated our approach, providing insights on the effectiveness of the priors and on the impact of different calibration models, in particular the usefulness of taking head pose into account.

This work has resulted in several publications (Siegfried and Odobez, 2017; Siegfried et al., 2017; Siegfried and Odobez, 2021a), with the last one being under publication process. Also, it has been the occasion of collaborations in the fields of dyadic interactions analysis (Muralidhar et al., 2018) and HRI (Siegfried et al., 2020; Foster et al., 2019).

1.4.2 Deep VFOA estimation in 3D scenes with an arbitrary number of targets.

We proposed a novel deep learning method that encodes all VFOA relevant features, i.e. gaze direction, head pose, subject speaking status, target directions, and target speaking status, as a fixed number of 2D maps, regardless of the number of targets. Thus, unlike previous approaches that train a model for a specific setup and using a fixed number of interacting people, the model learned by this method can be used on new setups with a different geometry and a different number of targets without needing to be retrained. Experiments on two datasets demonstrated that our method can be trained in a cross-dataset fashion without loss in VFOA

accuracy compared to intra-dataset training.

This work was published in Siegfried and Odobez (2021b).

1.4.3 Eye movements recognition from videos.

We proposed a method based on computer vision and deep learning in order to detect fixations, saccades, and blinks in natural interaction videos recorded with remote sensors. Unlike traditional methods, by directly processing eye images instead of analyzing gaze traces, this method allows to recognize eye movements even in absence of a precise gaze tracker. Although the overall performances indicate that detecting eye movements in low-sampling rate data acquired with remote sensors in natural conditions remains a challenging task, experiments on subjects participating naturally in a conversation demonstrated the benefit of our approach compared to previous methods, including deep learning models applied to gaze outputs.

This work was published in Siegfried et al. (2019).

1.5 Thesis plan

First, in **Chapter 2**, we summarize the main VFOA and gaze estimation approaches, as well as common gaze calibration methods. We then discuss the literatures about eye movements recognition, and position our work toward the state-of-the-art. Then, in (**Chapter 3**, we present the robust 3D gaze estimation framework, on which we built our improvements. We explain how head pose, gaze and finally VFOA are sequentially estimated and discuss the advantages and limitation of this method. In **Chapter 4**, we present the datasets we use in our diverse experiments. We also discuss the annotation of eye related features and its challenges and present our contributions in terms of data collection. Consecutively, we proposed three methods to address some of the challenges discussed above:

- In **Chapter 5**, we introduce an approach to perform online and unsupervised gaze estimation calibration and circumvent the traditional calibration session by leveraging a context-based weak VFOA estimation. We also present experiments on three different datasets, figuring dyadic interactions, four-party meetings, and objects manipulation.
- In **Chapter 6**, we address the flexibility of current VFOA estimation methods and introduce a deep learning approach to estimate the VFOA of users in a setup with an arbitrary number of targets. We also present a promising cross-dataset experiment that validate the generalization capabilities of the proposed approach.
- In **Chapter 7**, we address the recognition of eye movements when precise gaze traces are not available and introduce a convolutional neural network to perform image-based saccades and blinks recognition.

Finally, in **Chapter 8**, we briefly summarize the works and contributions we made through this thesis and discuss the limitations and perspectives of the proposed approaches.

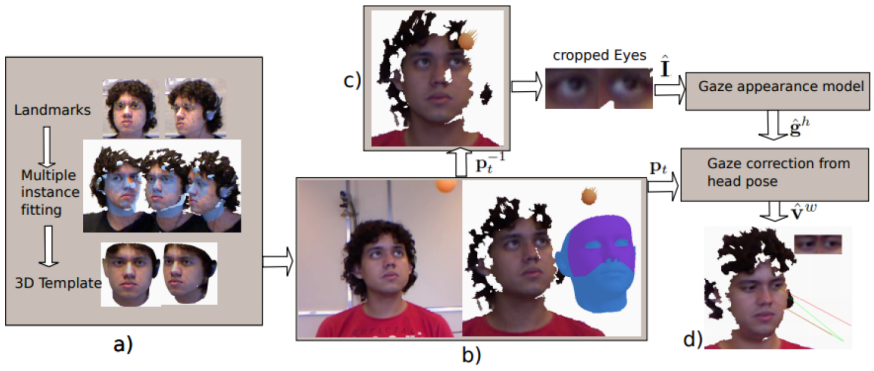
2 Related works

In this chapter, we provide a literature review of works that are relevant to the three tasks defined in Chapter 1, namely VFOA estimation, gaze estimation, and eye movements recognition. As VFOA and eye movements often involve gaze, we first review gaze estimation methods and especially appearance-based approaches, which are the most promising ones for remote gaze estimation nowadays, as well as gaze estimation calibration, also known as few-shot learning in the deep learning literature. Then, we review VFOA estimation methods and mainly works on conversation setups, as this problem was mainly treated in the field of HHI analysis. Next, we review eye movements recognition, focusing on methods based on gaze traces as they are the most common ones and highlighting the differences between the traditional study case and ours. Finally, we summarize the state of the art, as well as its limitations regarding these three problems, and present to which extent they are addressed by our contributions.

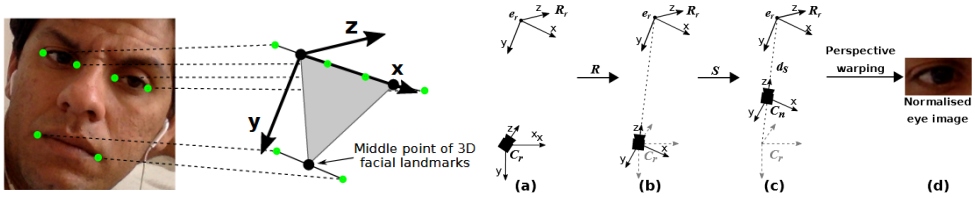
2.1 Gaze estimation

Gaze estimation includes many different problems and several review papers exists on the topic (Hansen and Ji, 2009; Chennamma and Yuan, 2013; Kar and Corcoran, 2017; Cheng et al., 2021). The difficulty of gaze estimation was traditionally addressed by using expensive specialized hardware or highly controlled scenarios (Guestrin and Eizenman, 2006). Still today, best performances are achieved by using head-mounted devices (Mayberry et al., 2014; Mansouryar et al., 2016), which remains invasive, or active gaze trackers leveraging corneal reflection (Morimoto and Mimica, 2005; Huang et al., 2014), whose usage is limited to screen-based setups where the user is relatively near to the sensor (less than 1 meter).

However, recent progress in technology (e.g. sensors) coupled with computer vision and machine learning techniques have opened the way to remote gaze estimation, which is a promising approach to democratize gaze tracking to more open and less contrained setups. As already mentioned in Chapter 1, we focus on remote 3D gaze estimation using consumer cameras and in the following, we review methods that address this specific gaze estimation problem.



(a) Gaze estimation framework proposed by Funes Mora and Odobez (2012).



(b) Eye normalization as presented by Zhang et al. (2017b).

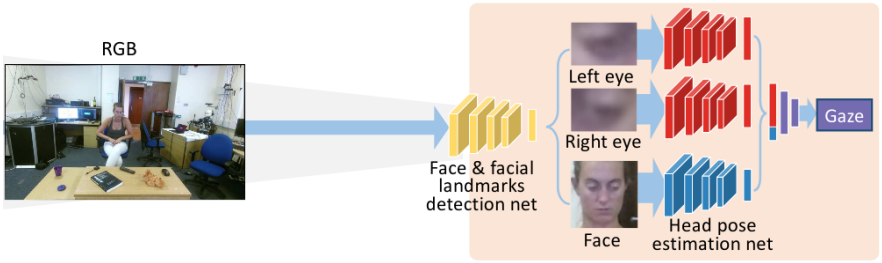
Figure 2.1 – Examples of eye rectification for appearance-based gaze estimation. Funes Mora and Odobez (2016) (a) rely on depth information to fit a 3DMM of the face, so they can rotate and project the textured mesh to get a frontal face image (step c), from which eyes can be cropped. Zhang et al. (2017b) (b) warp the eye image to a normalized view of the eye using the head pose and camera intrinsic parameters. Both approaches require an estimate of the head pose.

2.1.1 Appearance-based for remote gaze estimation

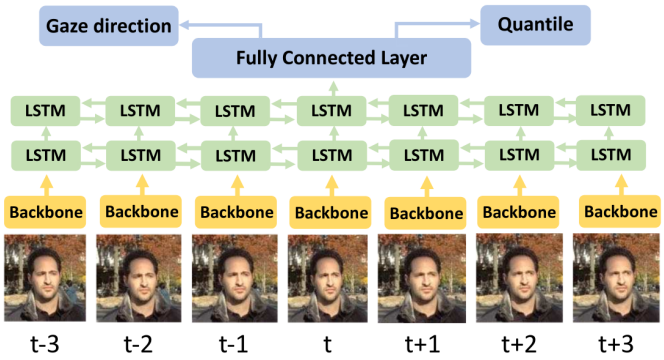
Computer vision based gaze estimation methods can be classified into two categories: geometric-based methods and appearance-based methods (Hansen and Ji, 2009). Geometric-based methods work by extracting local face and eye features and by mapping them into gaze cues. Features extraction either rely on IR illumination (corneal reflection) (Huang et al., 2014) or on RGB images of the eye (e.g. for pupil center (Jianfeng and Shigang, 2014)). However, these methods require high-resolution images, limiting user mobility, can only handle limited head pose variations, and are mainly targetting screen-gazing applications.

By directly learning a mapping from the eye image to the gaze parameters and avoiding feature tracking, appearance-based methods are more appropriate to handle lower image resolution and to be applied in HRI and HHI. The classical approach consists in computing visual features, e.g. Histogram of Oriented Gradients (Martinez et al., 2012), and estimating gaze using a learned regression model, e.g. Support Vector Regression (Martinez et al., 2012) or Random Forest (Sugano et al., 2014).

These methods either assume a relatively static head pose and train a user-specific gaze



(a) Gaze estimation network proposed by Fischer et al. (2018).



(b) Gaze estimation network proposed by Kellnhofer et al. (2019).

Figure 2.2 – Examples of deep learning methods for appearance-based 3D gaze estimation, i.e. with eye and/or face images as input and 3D gaze direction as output. Fischer et al. (2018) (a) use face and both eyes images (automatically extracted here). Kellnhofer et al. (2019) (b) propose an LSTM based architecture on top of a backbone network (ResNet-18). Also, they output a confidence measure in addition to the gaze estimate.

appearance model for this pose or they rely on some form of head pose dependent image rectification to crop the eye image in a canonical reference frame (Valenti et al., 2011; Funes Mora and Odobez, 2012; Li and Busso, 2014; Sugano et al., 2014), as shown in Fig. 2.1. For instance, Sugano et al. (2014) proposed a method that uses head pose and camera intrinsics parameters to warp the eye image so that it appears as seen from the front and at a normalized distance of the camera. Initially proposed to generate synthetic eye images, it was later used for normalization purposes (Zhang et al., 2017b). Authors in Funes Mora and Odobez (2012) alleviate this need by leveraging precise head pose estimation from RGB and depth data with a 3D Morphable Model (3DMM) facial mesh fitted online to compute a frontal face (and eye) image allowing to train an appearance-based gaze model on eyes with a canonical viewpoint. Still, they report a significantly higher error for the pose and person invariant model (i.e. handling an unknown user) compared to person-specific gaze model ($6° - 12°$ vs $2° - 6°$).

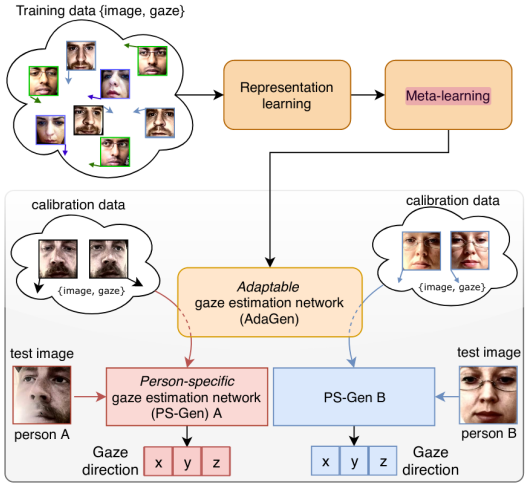
Deep learning approaches

Lately, appearance-based gaze estimation gained popularity with the growing number of gaze estimation dataset, like Columbia Gaze (Smith et al., 2013), UT Multiview (Sugano et al., 2014), EYEDIAP (Funes Mora et al., 2014), MPIIGaze (Zhang et al., 2017b), RT-GENE (Fischer et al., 2018), or ETH-XGaze (Zhang et al., 2020) allowing to take advantage of recent advances in deep learning, which led to better performances (Zhang et al., 2015; Krafka et al., 2016). Appearance-based methods are now considered more effective than their geometric-based counterparts (Zhang et al., 2019b).

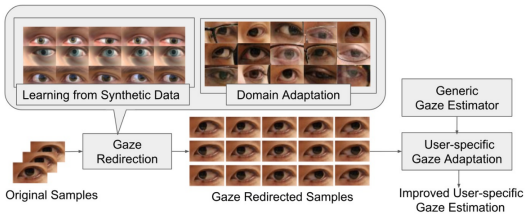
Zhang et al. (2015) proposed the first deep learning method for gaze estimation that is based on a shallow Convolutional Neural Network (CNN). They updated it later (Zhang et al., 2017b) by using a deeper network based on the VGG-16 architecture (Simonyan and Zisserman, 2014). It takes as input an eye image and outputs the 3D gaze direction, by using 13 convolutional layers, concatenating the head pose to the feature vector, and applying finally three fully connected layers. This approach was shown to beat state-of-the-art methods, even in cross-dataset settings. Later, Chen and Shi (2018) proposed to use dilated convolutions instead of max-pooling layers to avoid the strong reduction of image resolution that such pooling layers produce and guarantee that higher-level filters use the same pixels as input every time. Indeed, max pooling is convenient to reduce the spatial resolution of an image at the cost of losing some spatial information.

From that point, several works tried to improve gaze estimation by taking more information into account. For example, Zhang et al. (2017a) proposed regressing the gaze direction from the face image directly, using an attention mechanism so that the network learns what is relevant in the face. Studying the attention masks reveal that the network focused on eyes location but also other parts of the face. Other approaches used multiple inputs like both eyes, like in Cheng et al. (2018) where authors introduce an asymmetric loss to weight the contribution of both eyes and get a single output gaze vector. Also, some works combined gaze estimation with facial landmarks estimation in multitask frameworks to improve and/or fasten the learning (Fischer et al., 2018; Yu et al., 2018b). For example, Fischer et al. (2018) used both eyes and the face image altogether in a multi-task cascaded CNN, which first estimates facial landmarks position to align the face image given a canonical position, then estimate both head pose and gaze from face and eye images respectively, and finally combine them to get the final gaze direction (see Fig. 2.2a).

Another complementary research axis is the estimation of gaze in videos, which allows to leverage temporal information. For instance, Palmero et al. (2018) relied on a multi-stream CNN to jointly model face image, eyes image, and facial landmarks and feed the resulting feature vector to a Recurrent Neural Network to take into account gaze dynamic. Likewise, Kellnhofer et al. (2019) addressed the gaze estimation from people in the wild, by processing the whole face image through a pre-trained ResNet-18 network and feeding the results in a bi-directional Long Short-Term Memory (LSTM) layer. Interestingly, they also propose to use the



(a) Overview of the method proposed by Park et al. (2019).



(b) User-specific data augmentation scheme based on a redirection network as proposed by Yu et al. (2019).

Figure 2.3 – Examples of approaches training generalizable gaze estimation models, i.e. models that can be adapted to a new user with a few calibration samples. Park et al. (2019) (a) use meta-learning to train a model that can be easily fine-tuned. Yu and Odobez (2020) (b) propose to train a gaze redirection network, first from synthetic data and then from real data, using Cycle GAN and a frozen gaze estimator network. The trained network is then used to generate new calibration samples from a few ones, expanding in this way the training set used for user model adaptation.

pinball function as loss to learn the confidence of the gaze prediction (see Fig. 2.2b). However, as mentioned by Park et al. (2020), all these approaches were developed and evaluated on datasets where people are asked to look at some predefined points, leading to artificial gaze behaviors and a relatively low variance in facial expressions. In their work, they collected the EVE dataset, consisting of people looking at a movie without a specific task.

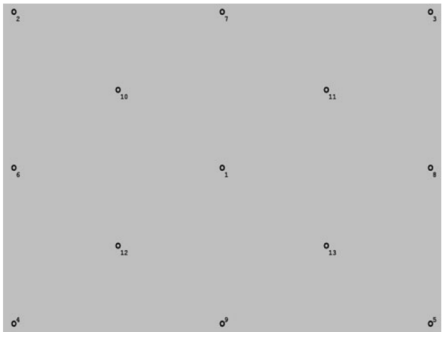
2.1.2 Gaze estimation models adaptation an calibration

Nevertheless, one important challenge in appearance-based gaze estimation is the performance drop in cross-dataset conditions, i.e. when such a method is trained on a reference dataset and used on data coming from a setup running in the wild. Such situations are common as gaze estimators often operate in environments where it is difficult to gather ground truth gaze points. Unfortunately, current methods have errors around $8-10°$ on cross-dataset evaluations conducted on existing gaze datasets, using for example the VGG 16 model presented above (Zhang et al., 2017b), or even more modern models, like an hourglass network with model-based gaze estimation (Park et al., 2018). Zhang et al. (2020) also made cross-dataset experiments using different public datasets, and found that the test accuracy depends a lot on the training and testing datasets, but remains higher than $10°$ in most cases. Such performances are far below those achieved in single dataset cross-subject situations ($3-5°$),

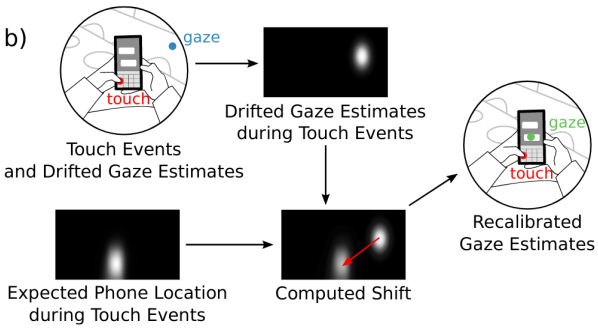
showing that it is still difficult to train models that generalize well to other setups.

To improve the results of appearance-based methods, the learned model is often adapted to the test domain. Indeed, difference in illumination and camera point of view can lead to a decrease in performance when passing from a setup to another. In this case, one can adapt the model to the specificities of the current setup, in which case we speak of domain adaptation. Also, people exhibit a large range of variations in eye characteristics like for instance the difference between the optical and visual axis which can not be measured from image only. Thus, another way to improve results is to adapt the model to each user, which is called calibration or few-shot learning. In general, calibration procedures rely on the collection of gaze direction for known target points (i.e. calibration or reference points) which can be exploited in several ways. A first classical approach is to learn a regression from the gaze output to a better gaze estimation. Even simple correction models, like linear regression, were shown to already significantly improve the results (Liu et al., 2020). Also, intermediate features, like the internal feature vectors of a neural network, can be used instead of the gaze output (Krafka et al., 2016). Another approach is to fine-tune the gaze estimation neural network model itself, by retraining part of it (Masko, 2017). Recent work focused on gaze models that can adapt without additional learning. For example, calibration parameters can be included as input to the gaze estimator (Linden et al., 2019), or the gaze model can be directly trained to estimate the difference in gaze between a new eye image and a reference one (Liu et al., 2020).

The difficulty to adapt gaze estimation models to new users and setups and the relatively small size of available datasets create interest in semi- and unsupervised learning. Indeed, a way to improve performances is to increase the generalization capabilities of the trained models, so that they be can more easily adapted to new users and setups (see Fig. 2.3). User-specific network adaptation relies on calibration training samples and collecting more samples leads to better adaptation. As it can be cumbersome to collect a lot of them, Yu et al. (2019) design a gaze redirection network, which takes an eye image and a numerical gaze shift as input and generates a new eye image with a shifted gaze direction. This network is first trained using eye synthesis, allowing to synthesize the same eye but with different (known) gaze directions. Then, domain adaptation, i.e. from synthetic to real, is performed using self-supervised learning by leveraging a loss based on Cycle GAN (Zhu et al., 2017) and a frozen gaze estimation error. Finally, the redirection network can be used to generate a big calibration set from a few calibration samples. Another example of a user adaptation model is the method of Park et al. (2019). Authors first train a robust gaze representation network using a disentangling autoencoder and then use meta-learning[1] to train a generic gaze estimation network, taking as input the aforementioned gaze representation, that can be adapted to new users.



(a) Typical calibration target positions, as presented in Nystrom et al. (2013).



(b) Overview of the methods proposed by Muller et al. (2019).

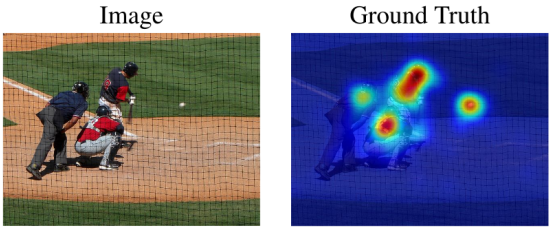
Figure 2.4 – Examples of traditional and unsupervised calibration samples collection. In traditional calibration (a), the user is asked to look sequentially at a set of points. Muller et al. (2019) (b) propose an automatic (re-) calibration method, assuming that the user is looking at his phone when he performs a touch. At this specific moment, the gaze error is known and can be used to recalibrate the gaze estimation model.

2.1.3 Calibration samples collection

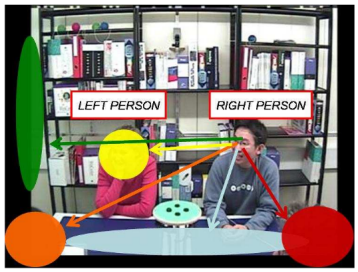
While improving gaze calibration for appearance-based methods recently grew in interest, less attention was paid to how to collect the required calibration samples. Traditional screen-based gaze trackers rely on a 5, 9, or 13 points calibration procedure (see Fig. 2.4a) consisting of asking users to look at some pre-defined points (Nystrom et al., 2013). This has a rather poor usability (Morimoto and Mimica, 2005) and the number of collectable calibration point is limited by the calibration session duration, which has an impact on user effort and fatigue (Holland et al., 2013). Thus further researches tried to overcome this limitation. Pursuit calibration (Santini et al., 2017) was explored to automatically detect when the user looks at a moving target by observing the correlation between the gaze and the target trajectories, which makes the calibration more robust to interruptions and does not need user contribution. This approach is less tedious and more flexible (Pfeuffer et al., 2013), but requires full target control, making it unsuitable for settings without screens.

Another approach to collect calibration samples without a dedicated calibration procedure with user collaboration relies on context knowledge, like environment, user's field of view, and/or the performed task, to infer highly probable gaze targets and use them as calibration points. For example, Sugano et al. (2015) showed that it was shown possible to calibrate gaze estimation methods using gaze and mouse operations coordination. When a mouse click occurs, their system gathers the eye image and the mouse position, which is assumed to be the gaze target. The gathered points are used to calibrate the gaze estimation methods. Similarly, Pi and Shi (2019) make the hypothesis that people look at the center of the buttons when they

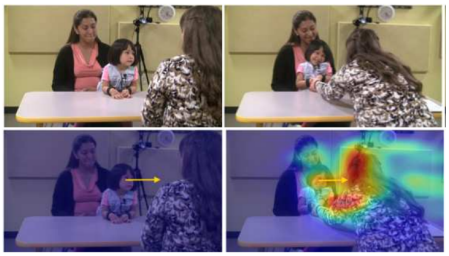
[1] A method that given pairs of calibration-evaluation sets minimizes the evaluation loss of a model after calibration.



(a) Saliency estimation (Pan et al., 2017).



(b) VFOA estimation in 3D scene (Ba and Odobez, 2008a).



(c) 2D VFOA estimation (Chong et al., 2018).

Figure 2.5 – Different VFOA estimation tasks. Saliency estimation (a) focuses on the visual stimulus and predicts what is salient for an external observer. VFOA estimation in 3D scenes (b) is the classification of a subject's VFOA among a set of visual targets which are not necessarily in the image (their direction are represented as color arrows and blobs here). 2D VFOA estimation (c) is an hybrid task, where the subject (the child in this case) is present in the image but attention is predicted in the image itself, regardless of target positions and image context. In this example, the gaze is also predicted (yellow arrow).

are using dwell-time based typing on a virtual keyboard. Thus, calibration points are found by storing the gaze estimate and the center of the activated virtual keys. Also, Muller et al. (2019) leverage the access to the subject's field of view, made available my a head-mounted eye tracker equipped with an egocentric camera, i.e. a camera that records what the user is seeing. When the user makes a touch on his smartphone, the gaze estimation and the position of the smartphone in the egocentric camera image are stored and used as calibration samples (see Fig. 2.4b). These approaches are promising, as they allow recalibrating the gaze tracker at any time during usage, without interruption or the need for specific actions of the user. However, these studies focused mainly on screen-based gaze trackers and head-mounted devices, so whether this paradigm hold in the case of 3D remote sensors setups, which present more variability in terms of head pose and gaze, remains to be investigated.

2.2 VFOA estimation

Attention is a complex behavior, which covers all factors that influence selection mechanisms in perception (Borji and Itti, 2012). In the case of vision, we usually refer to the visual focus of attention (VFOA), i.e. the object, person, or region to which a person attends, as a proxy

for the attention. In that context, attention is defined by what a person actually perceives and is usually described as driven by two different mechanisms: the top-down attention is conscious and related to the task while the bottom-up attention is unconscious and related to the saliency of our perception field (Katsuki and Constantinidis, 2014). As we usually look straight to the target of our visual attention to exploit the higher acuity of our central vision, the VFOA is strongly related to gaze. However, its dependency to the visual stimulus distinguishes it from the gaze, in the sense that we usually estimate gaze from the head and eye information regardless of the environment. VFOA estimation problems can be separated into three main categories (see Fig. 2.5) that we shortly describe here, before focusing on the VFOA estimation in 3D scenes.

Saliency estimation. Saliency estimation addresses the detection of what is salient in a visual content, i.e. what part of the stimulus will attract the visual attention (Borji and Itti, 2012). These models usually take as input an image or a video and output the probability that an observer will look at each pixel in the form of a saliency map. Traditional methods used low-level features extraction, like colors, intensity, and orientations (Itti et al., 1998), taking inspiration of human bottom-up attention, but deep learning now allows to process the entire input image directly (Li and Yu, 2016; Cornia et al., 2016; Pan et al., 2017). Parallelly, some works attempted to predict the scanpath, i.e. the fixation sequence, of a potential observer for a given image, e.g. by detecting and inhibiting the most salient part of the image in turn (Itti et al., 1998) or leveraging inverse reinforcement learning (Yang et al., 2020). One example of application field is autonomous driving. Indeed, several works attempt to predict car drivers' VFOA (Palazzi et al., 2018; Xia et al., 2018) to better understand and model drivers' decision making.

VFOA estimation in 3D scenes. VFOA estimation in 3D scenes addresses the classification of the VFOA among a set of visual targets, which are not always present in the subject's image but whose positions are known. The decision is made by extracting features, e.g. body pose, head pose, or gaze direction, from the subject's image, but also sometimes from other modalities, like utterances, events in the environment, and so on. This task and the different approaches to handle it are discussed further below.

2D VFOA estimation. A more recent, and somehow hybrid, problem is the recognition of the general attention of a person in an image without information about the 3D scene or the image's surroundings. In other terms, the goal is to predict the attention target and/or the gaze direction of a given person in the image. As such, this task lies in the middle of VFOA, gaze, and saliency estimation. As an example, Chong et al. (2018) proposed a CNN that takes the image and the cropped head image of the person of interest as inputs and outputs the person's gaze as well as a heatmap showing where the person is looking in the image (fixation likelihood). They later improved their model by incorporating the position of the person in the image and adding temporal information (Chong et al., 2020). Another example is the work of Marin-Jimenez et al. (2019) which focused on predicting if two people are looking at each

other. In this case, the head images of both persons are fed in a network together with the scene image and the output is binary.

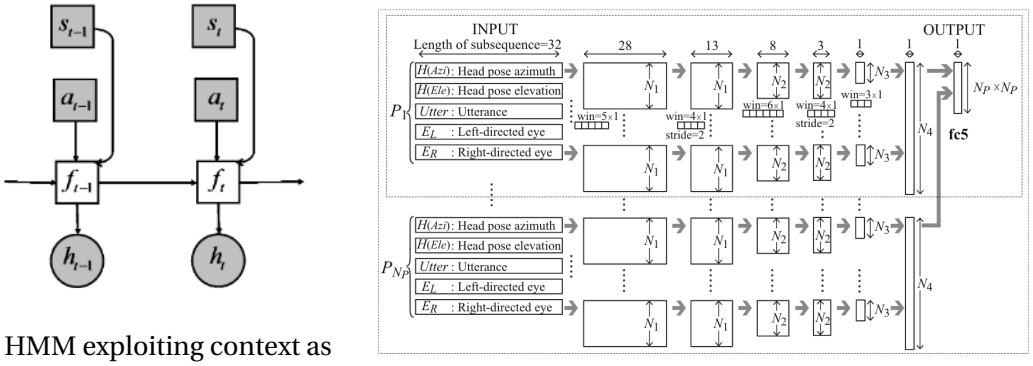
2.2.1 VFOA estimation in 3D scenes

Back when accurate gaze trackers were not available, VFOA in 3D scenes was inferred from head pose (Stiefelhagen et al., 2002), using inference mechanisms like Gaussian Mixture Models and Hidden Markov Models (HMM) (Ba and Odobez, 2008a). However, head pose only approaches are limited because head-gaze coordination highly depends on the subject (Sidenmark and Gellersen, 2019a) and the same head pose can be related to several targets. Also, the head does not always move during gaze shifts, e.g. when people are looking at a target and perform a short gaze aversion before looking back to the target. Reciprocally, the head sometimes move without the gaze, typically during head gestures like nods and shakes.

Two main approaches were proposed to improve the VFOA estimation, namely interaction models and individual-centered models. Interaction models estimate the VFOA of a person (or the joint VFOA of several) by clustering person and context-related features (Otsuka et al., 2006; Ba and Odobez, 2008b; Otsuka et al., 2018; Bai et al., 2019). This approach was studied a lot in conversation setups, where VFOA has a particular importance as it allows to understand the conversation dynamic and infer interesting behaviors. Individual-centered models rely on geometrical reasoning to exploit a gaze estimate, despite its relative inaccuracy, by comparing it to the positions of the different visual targets (Funes Mora et al., 2013; Yücel et al., 2013; Salam et al., 2016). Indeed, there are setups where the distance between likely targets is enough to allow estimating the VFOA from a coarse gaze direction.

Interaction models

VFOA estimation is a well-known problem in HHI and was studied a lot in the context of multi-party conversations, sometimes involving artificial agents. To overcome the head pose ambiguity, these researches focused on developing more complex interaction models, e.g. leveraging Dynamic Bayesian Networks, that model the joint VFOA of all participants by taking into account more features and contextual information. Otsuka et al. (2006) added the conversation regime, i.e. the distinction between monologue, dialogue, or other, as a hidden variable and the utterance, i.e. does each participant speaks or not, as an observation. Gorga and Otsuka (2010) propose to also use a rough gaze estimation, classified in left, right, or center. Other works build a richer model by adding contextual information. Ba and Odobez (2008b), working on presentation meetings, used slide activity as an observation (see Fig.2.6a), and Sheikhi and Odobez (2015) added discussion topic, as well as robot utterance and addressee in the context of HRI. To further model the scene, several works (Ba et al., 2009; Sheikhi and Odobez, 2015) introduced the target positions in their model and implicitly estimate the gaze direction.



(a) HMM exploiting context as proposed by Ba and Odobez (2008b).

(b) CNN proposed by Otsuka et al. (2018).

Figure 2.6 – Examples of VFOA estimation methods based on interaction models. Ba and Odobez (2008b) (a) proposed a HMM that inferring the VFOA of all the participants f_t from the observed head poses h_t, utterances $\tilde{s}_t$, and slide activity a_t. Otsuka et al. (2018) (b) proposed a CNN which take 32 frames as input and fuse the features of all the participants step by step.

Leveraging progress in computation and gaze estimation, recent works proposed more complex models, like deep learning ones (Zhang et al., 2019a; Otsuka et al., 2018), that use gaze along the head pose to predict the VFOA. For instance, Otsuka et al. (2018) proposed to use a multi-stream CNN, to combine gaze (left, center, or right), head pose, and speaking status information (see Fig.2.6b). The different features are processed independently, using 32 consecutive frames each time and the resulting feature vectors are first combined for each person before. The final VFOA prediction is taken from the combination of all participants' features. Bai et al. (2019) use the same features, but estimated the continuous gaze and extracted the speaking status, from visual information only, using the mouth landmarks positions over time. They proposed to use a "network" of classifiers (e.g. Random Forest) which propagates information both across participants and across time. These methods were shown more accurate than their Bayesian counterparts. Indeed, they provide more efficient ways to fuse the different features and learn more complex temporal and inter-modality relations, at the cost of interpretability and requirement for larger training datasets.

The main advantage of the methods based on interaction models is that the VFOA of each participant is estimated using an important amount of features, allowing to exploit co-occurrences of behaviors. Also, they allow estimating the VFOA even when the subject and target positions are not known or when these features can not be expressed in the same referential. However, an important drawback is that they train setup-specific models, which must be re-trained to be applied to a new setup and do not handle moving targets.

Individual-centered models

Individual-centered models rely on the estimation of the subject's 3D head pose and gaze direction to know if the subject is looking at a target given its position. Indeed, with the

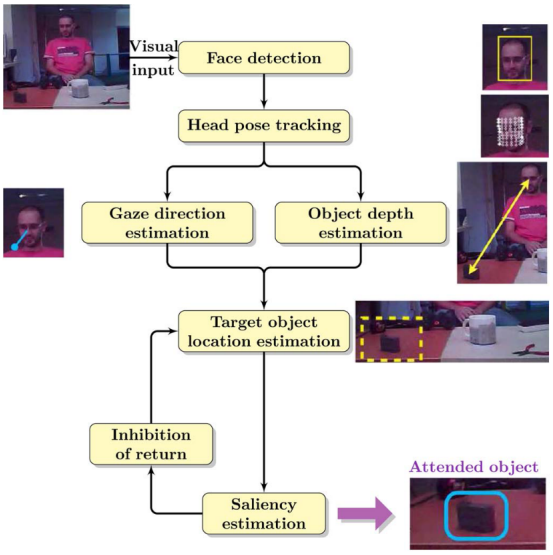
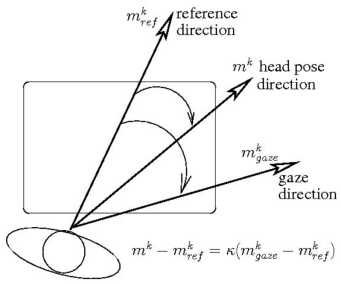


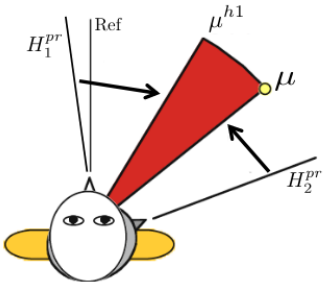
Figure 2.7 – Example of a VFOA estimation method using an individual-centered model. Yücel et al. (2013) (a) proposed a framework to sequentially estimate head pose, gaze, and VFOA using saliency estimation to get the position of the targets.

improvement of gaze estimation, even simple frame-based geometrical models were shown efficient to estimate VFOA. Thus, a common framework consists in sequentially estimating head pose, gaze, and finally VFOA (Funes Mora et al., 2013; Yücel et al., 2013; Asteriadis et al., 2014). For example, Funes Mora et al. (2013) used a threshold over the cosine distance between the subject's gaze and the subject-to-target vector, where the target positions are obtained through scene monitoring. Similarly, Yücel et al. (2013) compared the gaze to the position of objects obtained by a saliency model (see Fig. 2.7a).

Several works even relied on weak gaze proxies, like people's torso orientation (Foster et al., 2012), sometimes combined with the head pose (Salam et al., 2016), or head pose alone (Voit and Stiefelhagen, 2008; Ba and Odobez, 2011; Sheikhi and Odobez, 2015). Ba and Odobez (2011) proposed a cognitive model that estimate the most likely head pose direction given the target position, based on cognitive science findings about gaze shift dynamics (see Fig. 2.8a). In their case, they found that this geometric model performs as good as a data driven trained model. Later, Sheikhi and Odobez (2015) improved this model by taking into account the "midline effect" (see Fig. 2.8b). Indeed, the proportional model usually holds when the gaze moves away from the body orientation, with respect to the previous head pose, but we tend to move less the head when the gaze is going in the other direction. The challenge lies in the estimation of the body orientation, which is needed by the model. Before deep learning methods, it was difficult to estimate it in a precise manner. Nevertheless, Sheikhi and Odobez (2015) proposed to estimate it as the average of the head pose over the previous 40 seconds, which was shown to be effective in situations where people were standing in front of a robot. Overall these approaches work when the inaccuracy and noise of the gaze proxy (or gaze model) remain smaller than the distance between targets, so their performances depends on



(a) Proportional model used by Ba and Odobez (2011).



(b) Midline effect model used by Sheikhi and Odobez (2015).

Figure 2.8 – Examples of geometric model to estimate the head pose likelihood given the target position. Ba and Odobez (2011) (a) used simple geometric models to estimate the most likely head pose direction given the target position. Sheikhi and Odobez (2015) (b) proposed a more complex model based on the "midline effect", which takes into account the previous position of the head. Considering a visual target direction μ, the most likely head pose direction to look at this target is μ^{h1} (as in the simple geometric case) if the previous pose was below or to the left (e.g. H_1^{pr}). However, it is μ if the previous head pose was beyond μ (like H_2^{pr}) and the person needs to rotate the head back to look at the visual target. Note that in this work, an estimate of the body orientation was also proposed, as it is required by the model to define the midline (Ref).

the setup.

Another approach consists in estimating the VFOA of a person directly from its appearance. For instance, Dong et al. (2009) proposed a Dynamic Bayesian Network, which infers the VFOA from the face appearance and the position of the face. In short, a set of face clusters modeled by Principal Component Analysis subspaces are learned for different head pose orientations and used to infer the VFOA from the face image. Another example is the work of Duffner and Garcia (2016), where a particle filter is used to perform face tracking while training an HMM that infer VFOA. After some time, the HMM model is frozen and further exploited.

These methods usually involve simple computation, at least for the VFOA estimation itself, as they rely on simple rules or criteria. Also, they allow to use the same model on different setups and can handle an arbitrary number of people in the scene at the cost of computation power, as each person must be processed individually. However, they are sensitive to gaze estimate noise and inaccuracy and are usually more demanding in terms of input features. Indeed, the target positions must be known, they usually require head pose and gaze estimation, and all the features must be expressed in the same frame, which often requires scene monitoring.

2.3 Eye movements recognition

Besides visual attention, eye movements and blinks are also related to high-level cognitive processes, like drowsiness and cognitive load, as presented in Chapter 1. Furthermore, their

recognition can help to improve gaze estimation. Some works showed that taking into account the gaze dynamic by modeling the underlying eye movements improved the gaze estimation performances, even if eye movements themselves are not explicitly estimated. For example, Feng et al. (2011) proposed an HMM with three states (fixation, saccade, and smooth pursuit), to smooth the gaze estimated by a commercial eye tracker. Recently, Wang et al. (2019) adapted this method to appearance-based gaze estimation and made the model explicitly learn and exploit the duration of each eye movement type. Moreover, appearance-based gaze estimation methods usually do not check that the eye is opened, leading to irrelevant estimations when the eye is closed (Cortacero et al., 2019). It could lead to silent failures of methods relying on the gaze estimate, which can be better handled if blinks are detected.

Eye movements are defined as the voluntary or involuntary movements of the eyes that help in acquiring, fixating, and tracking visual stimuli. Mulvey (2011) described 9 types of eye movements:

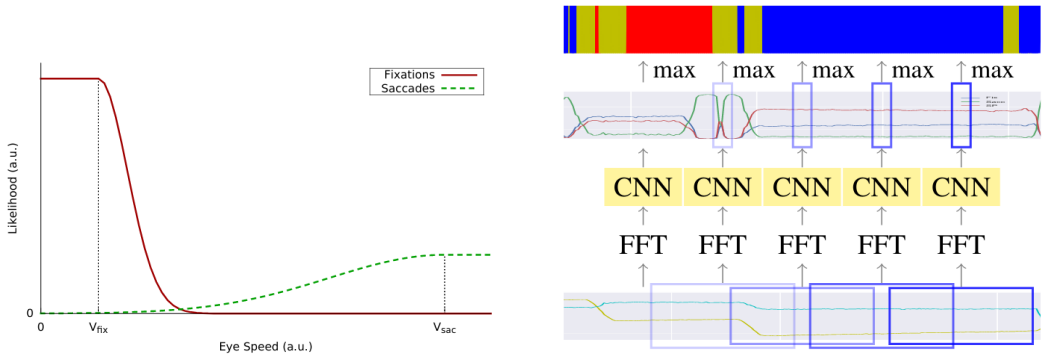
- fixation: the eye moves less than 1°, for at least 100ms;
- saccade: movement to new areas of the visual field, greater than 2 degrees;
- microsaccade: very small movements which occur irregularly during a saccade;
- smooth pursuit: movement of the eye to follow a moving target with the same velocity and trajectory as the target;
- vergence: movement of the eyes inward, i.e. in opposite directions, to offset retinal disparity for close objects;
- vestibular ocular reflex: movements to maintain the point of regard during head movements;
- optokinetic reflex: saccade-smooth pursuit movements to focus on moving scenes
- accommodation: changes in the shape of the lens to focus light from objects at varying distances;
- pupil dilation: changes in the size of the aperture in the iris to maintain optimal light levels inside the eye.

However, several of these movements are too small to be observed from a consumer camera at operating distance. Also, visual targets do not move a lot in the considered datasets (see Chapter 4), so the number of smooth pursuits is too small to allow proper training and evaluation. Thus, we focus on the recognition of the two main eye movements, namely fixations and saccades, as well as blink detection.

2.3.1 Fixations and saccades recognition

A lot of methods were proposed to recognize fixations and saccades. They usually take as input the gaze trace, i.e. the 2D coordinates of the point of gaze (PoG) over time, which is obtained by computing the intersection of the gaze and a 2D surface, like a screen.

The common approach for eye movement recognition is to define features to perform rule-



(a) Fixation and saccade likelihoods as defined by Santini et al. (2016).

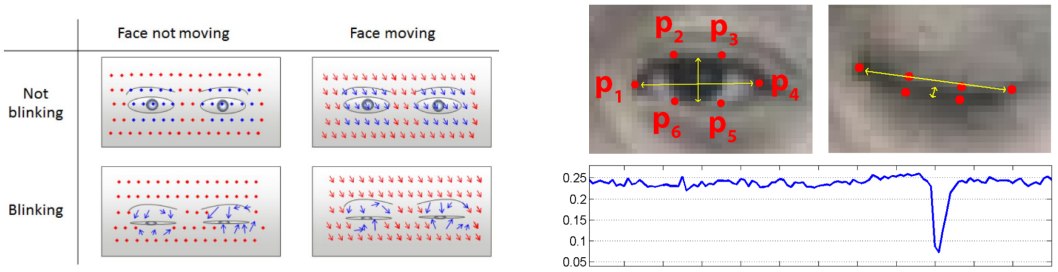
(b) Network architecture proposed by Hoppe and Bulling (2016).

Figure 2.9 – Examples of eye movement recognition methods. Santini et al. (2016) (a) proposed a Bayesian Decision Theory method, whose parameters are learned online. Here we see the likelihood of a PoG being a fixation or a saccade given the velocity. Hoppe and Bulling (2016) (b) proposed a CNN which processes the Fast Fourier Transform of the PoG signal.

based classification (review by Andersson et al. (2017)). Binary classification between fixations and saccades was first done by thresholding the PoG velocity or dispersion, i.e. the maximal distance between PoGs in a time window (Salvucci and Goldberg, 2000). Then, more complex features and/or rules were proposed to improve recognition (Veneri et al., 2011; Larsson et al., 2013; Hessels et al., 2017). For instance, Larsson et al. (2013) proposed a method that first detects saccades in a frame-based manner using the gaze acceleration and then refines the saccades boundaries by checking three consistency criteria. The main limitations of this approach are the need to carefully design features and rules, with the risk to overfit the method to a specific problem, and the assumption of high-quality data in terms of sampling rate and gaze accuracy (Zemblys et al., 2016).

To better model the eye dynamic, other works investigated probabilistic methods, whose parameters can be learned instead of being manually chosen. For example, Salvucci and Goldberg (2000) proposed a two-state HMM, one for fixations and one for saccades, which use the PoG velocity as observation. A similar idea was used by Tafaj et al. (2012), who proposed a Bayesian Mixture Model based on the distance between consecutive PoGs. The mixture consists of two Gaussians, like the above HMM, but this method allows to train the parameters online. It was shown to provide a good recognition rate on data recorded with a low sampling rate (25 fps). Finally, Santini et al. (2016) proposed a method based on Bayesian Decision Theory to also recognize smooth pursuits. The likelihood of their model is computed from the PoG velocity for fixations and saccades (see Fig. 2.9a), and a motion-related feature for smooth pursuits. This model allows recognizing these three eye movements in a fully probabilistic and online manner.

Other machine learning methods have been proposed to incorporate more features or leverage deep learning methods (Anantrasirichai et al., 2016). For instance, Zemblys et al. (2016) used a



(a) Eyelids vs face motion as illustrated by Mohanakrishnan et al. (2013).

(b) Eye aspect ratio as defined by Soukupova and Cech (2016).

Figure 2.10 – Examples of features used in blink detection methods. Mohanakrishnan et al. (2013) (a) proposed to compare the motion in the eye region and the whole face. Soukupova and Cech (2016) (b) relied on the eye aspect ratio (height over width), which is expected to decrease when the eye closes.

Random Forest applied to 14 features, using basic ones, like velocity and dispersion, together with complex rule-based criteria. Also, Hoppe and Bulling (2016) introduced a CNN that classifies the Fast Fourier Transform of the PoG signal in fixations, saccades, and smooth pursuits (see Fig. 2.9b). Later, human-level performances were achieved by Bellet et al. (2019), who proposed a U-Net inspired network to directly process the PoG signal. They also took time into account, by using 179 frames equally distributed in the past and the future as input, which corresponds to a 180-360 ms time window in their experiments, depending on the dataset.

A limitation of all these methods regarding our study case is that they use clean data usually recorded at a high sampling rate by powerful, but invasive, eye trackers. Indeed, low sampling rate signals were shown to be significantly more difficult to work with (Zemblys et al., 2016). Some methods still achieved good performances using extracted gaze from mobile eye trackers data at 25-30 fps (Tafaj et al., 2012; Santini et al., 2016; Anantrasirichai et al., 2016). However, their estimation of gaze mostly relies on eye images with higher resolution than remote sensors can provide.

2.3.2 Blink detection

In PoG-based approaches, blinks are often removed manually (Anantrasirichai et al., 2016; Pekkanen and Lappi, 2017; Santini et al., 2016) or by the eye tracker, as they lead to data loss (Larsson et al., 2013), but these methods are not suited for the use of remote sensors in dynamic interactions.

Blink detection from eye image was first addressed by the analysis of hand-crafted features extracted from eye images. Some methods relied on eye image characteristics, e.g. color (Panning et al., 2011). Other methods relied on template matching: a score is given to the input image based on its similarity to a set of pre-recorded opened eye images and the decision is taken by thresholding this score (Chau and Betke, 2005). However, these methods struggle

to detect blinks when the face is moving (Mohanakrishnan et al., 2013). To overcome this limitation, motion-based approaches were investigated, by relying on optical flow (Divjak and Bischof, 2009) or motion vectors deduced from features tracking (Mohanakrishnan et al., 2013). These methods work by detecting the fast vertical motion in the eye image and distinguishing it from the overall face movements (see Fig. 2.10a).

The improvement of facial landmarks detection made it easier to measure features like eye closure or eye aspect ratio. For example, Soukupova and Cech (2016) relied on the eye aspect ratio, i.e. height over width, using the 6 eye landmarks detected by a facial landmarks detector (see Fig. 2.10b). The status of the eye, i.e. opened or closed, was then estimated from the eye aspect ratio over time using a classifier, like a Support Vector Model or an HMM. Still, several challenges remain because blinks are similar to short downward glances from an eyelid pattern point of view. Also, blink behavior varies across people and over time due to eye fatigue (Baccour et al., 2019).

By using deep learning, recent methods attempted to skip the feature extraction step entirely. Cortacero et al. (2019) introduced the RT-BENE dataset by annotating a gaze estimation dataset (RT-GENE) for blinks. Doing so allows training models that jointly estimate gaze estimation and blinks. They also proposed a set of CNN baselines, which take both eye images as input and estimate the blink probability. Anas et al. (2017) also proposed a CNN to classify eye images between three classes, namely *opened*, *closed*, and *partially closed*. They showed that their method reached very high precision and recall even in cross-datasets experiments.

2.4 Conclusion

As shown in this chapter, a lot of work was done on the three tasks addressed in this thesis. In the three cases, the evolution from the analysis of hand-crafted features to the end-to-end processing of sensor signals highly increased the achieved performances. Indeed, deep learning methods, like CNN, are now the baseline in all the presented tasks. Nevertheless, as methods become more effective, the research interest shifts to more challenging tasks and scenarios, and remote sensing, among other problems, remains challenging. In the following, we highlight some of the state-of-the-art limitations and how our contributions address them.

Gaze estimation

Sec. 2.1 pointed the success of appearance-based gaze estimation in remote sensing cases, especially using deep learning methods which takes into account more and more information, e.g. face image, head pose, and gaze dynamics. Also, we observed the difficulty of adapting a model to new users and setups. This is addressed by few-shot learning, but the way to collect calibration samples remains unclear, despite being a critical step in the use of gaze estimation in real-time applications. As shown in Sec. 2.1.3 several methods were proposed to collect these samples in an unsupervised fashion, but they rely on ego-centric view or displayed

information, which are not available in remote sensing setups.

Moreover, we observed the multiplication of gaze estimation datasets, i.e. datasets providing eyes or faces image along with gaze ground truth, with a growing number of data points and increasing variability in illumination conditions, head pose, and so on. However, they are usually recorded while the user artifically looks at a sequence of points or passively observes a visual stimulus. Indeed, it is very difficult to get gaze ground truth when people are acting naturally and use actively their gaze to perform a task, like in conversation for example. Thus, there is a lack of data to evaluate gaze tracking in real application situations.

In Chapter 5, we address the unsupervised and online collection of calibration samples by using attention-based priors. Also, we annotated two datasets for VFOA and collected a third one (see Chapter 4) to evaluate remote gaze estimation and our calibration scheme during real and unconstrained interactions.

VFOA estimation

In Sec. 2.2, we showed the trade-off between interaction and individual-centered models. While the later ones are often simplistic and miss interesting contextual features that could improve the VFOA estimation, interaction models lack flexibility. They have the advantage of representing all participants' behaviors together to take into account dependencies between them but such models are in almost all cases trained on a single specific setup. Indeed, a defined number of static visual targets are usually assumed and 3D scene representation elements, like 3D positions, are sometimes absent, leading the model to learn setup-specific feature clustering rather than geometrical reasoning. Thus, they can not generalize to unseen setups with a different number of people or a different geometry without being retrained.

In Chapter 6, we propose a method that allows encoding contextual information while keeping the flexibility of the individual-centered models. The idea is to express the subject's and the contextual features in the same referential (the body frame of the subject) and to reformat them as a fixed number of 2D maps, one for each feature type. Doing so allows to train a single model for an arbitrary number of subjects and targets. Also, a trained model can be applied in new setups without retraining.

Eye movement recognition

In Sec. 2.3, we reviewed methods to perform fixation and saccade recognition on one side and blink detection on the other. Both tasks were successfully addressed, reaching human-level performances for the first and very high precision and recall for the second. However, the presented fixation and saccade recognition methods can not be applied to our study case, because of the important difference in data quality. Indeed, fixation and saccade recognition is usually performed from an accurate gaze signal, which is difficult to obtain from appearance-based gaze estimation.

In Chapter 7, we propose a deep learning method to estimate fixations, saccades, and blinks directly from the eye image stream. Experiments are conducted on remote sensors data recorded at 30 fps and show promising results despite the difficulty of the task.

3 Background methods

In this thesis, we study 3D scenes and we need to extract features, like people's position, head pose, and gaze. In this chapter, we present the baseline framework on top of which we made experiments and built improvements. From a data point of view, we consider that the RGB-D video of the subject and 3D target positions are available. Note that the workflow presented below is usually repeated on all the frames of a video but in this chapter, we omitted time indexes for readability.

3.1 Overview

In general, we used the head pose invariant gaze estimation framework proposed in Funes Mora and Odobez (2012), which is summarized in Fig. 3.1. In short, it takes as input the data provided by an RGB-D camera, which can be seen as textured 3D meshes once both RGB and depth camera are calibrated. Given a head pose 3D model, it then estimates at each timestamp the head pose which is used to rectify the face region into a frontal view so that normalized eye images can be cropped and processed through an appearance-based gaze estimator. The resulting gaze estimate in the frontal head reference is finally combined with the head pose to deliver an estimate of the gaze in the 3D space. Note that the normalization of eye images through the face frontalization allowed thanks to the availability of the depth information and the head pose estimation makes the gaze estimation more robust to head pose variations, a robustness which swould be difficult to obtain by directly estimating the gaze from the RGB image (Funes Mora and Odobez, 2016).

In our work, we updated this framework by using more recent methods for head pose and gaze estimation and added VFOA classification on the top. Also, we assume a calibrated setup, so that observations (i.e. positions and vectors) can be expressed either in the world coordinate system (WCS), the camera coordinate systems, or the head coordinate system (HCS) of the subject once his head pose is estimated. It is useful to process the observations of several cameras together.

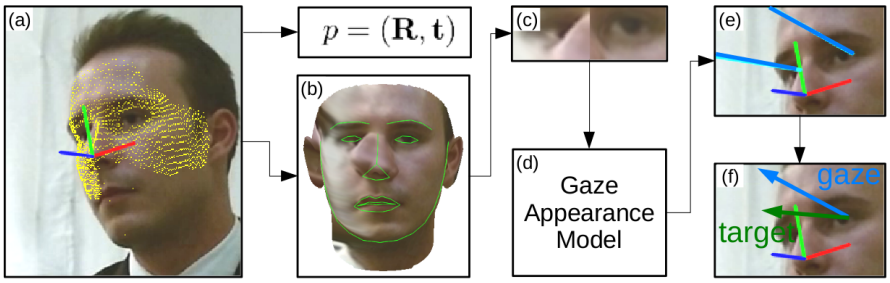


Figure 3.1 – Gaze and VFOA estimation framework as proposed by Funes Mora and Odobez (2012). a) Head pose estimation $\mathbf{p} = (\mathbf{R}, \mathbf{t})$ from RGB-D video stream. b) Frontalized face image computation using the head pose to rotate and project the textured mesh (depth + RGB image). c) Eye images cropping using a landmarks detector. d) Gaze angles $\mathbf{g}$ estimation in the head coordinate system from the eye images. e) 3D gaze estimation $\mathbf{v_g}$ by combining gaze angles and head pose. f) VFOA f estimation based on the angular distances e between the gaze vector $\mathbf{v_g}$ and the vectors pointing to targets $\mathbf{v}_j$.

3.2 Head pose estimation

In this work, we define the head pose $\mathbf{p}$ as the rigid transform $(\mathbf{R}, \mathbf{t})$ that maps a Head Coordinate System (HCS) rigidly attached to the subject's head to the Camera Coordinate System (CCS) attached to the camera recording the subject. [1] In other words, we want to estimate the position and orientation of the subject's head with respect to the camera.

To do that, we relied on the Headfusion method (Yu et al., 2018a) which is summarized in Fig. 3.2. The main head pose estimation is performed by minimizing the alignment error between a 3D head model and the observed depth data using an Iterative Closest Point (ICP) algorithm. The pose used as initialization of this iterative optimization process is obtained from the RGB stream, either using an RGB landmarks detector whose outputs are mapped in the 3D space leveraging the RGB-D calibration (at the tracking start), or by updating the estimation of the previous frame using a robust KLT tracker (following steps). Parallelly, this method adapts the 3D head model to the current subject by automatically fitting both a 3D Morphable Model (3DMM) (Paysan et al., 2009) of the face and a raw 3D representation of the head.

Yu et al. (2018a) reported an average error of 2° to 5° depending on the dataset. Also, this method was shown to be more robust to large head poses variations that might occur in recordings of people behaving freely in natural interaction settings, compared to 2D landmarks-based methods or methods solely relying on the 3DMM representation. Moreover, both the 3DMM and the 3D head representation are built online without any manual intervention, so this method can be applied to any subject. For these reasons, it satisfies our need for a robust method that can be applied to new setups without requiring supervision. Finally, as the semantic meaning of the 3DMM's vertices is known, this method allows estimating the position

[1] Note that the head pose angles expressed in the CCS can be directly computed from $\mathbf{R}$.

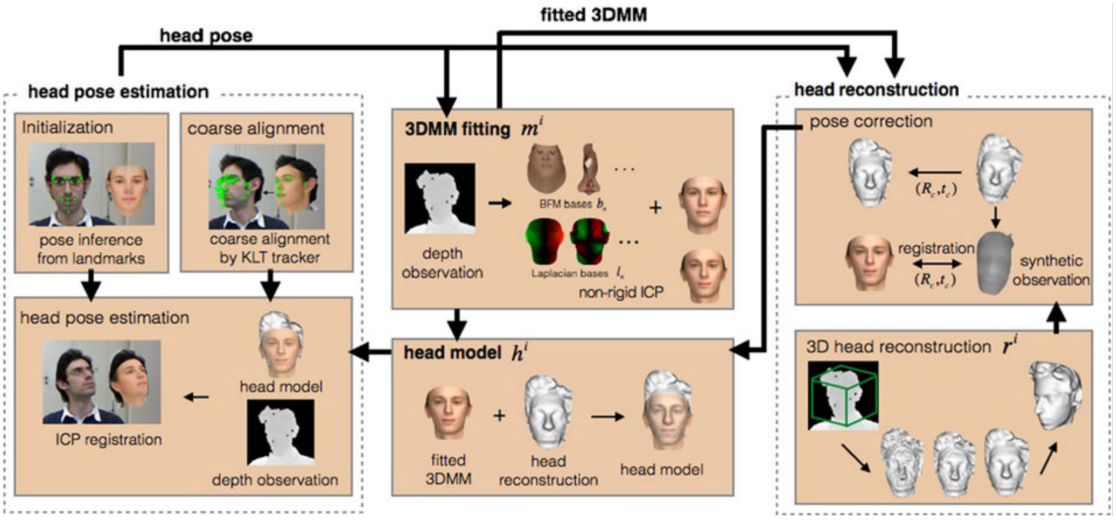


Figure 3.2 – Headfusion framework (Yu et al., 2018a). The head pose estimation module aligns the 3D head model h^i and the observed depth data using a point-to-plane ICP method. This head pose model h^i is composed of a 3DMM and a full 3D face model which are fitted (3DMM, middle block on the top) to the data or reconstructed (full face model, right block) by aggregating pose rectified depth data over time into a full 3D head representation r^i.

of the face features, e.g. the eyes.

3.3 Normalized eye image extraction

The next step of the framewok (Fig. 3.1) consists of the acquisition of the eye images (36x60 pixels for each eye) so that we can estimate gaze direction from them. To do that, we first use the head pose to rotate the textured mesh obtained from RGB and depth image and get a frontal view of the subject's face, which also normalized the size of the face across people. Once the face is frontalized, eyes can be localized and eye images can be cropped for further processing. Here we describe these two steps.

3.3.1 Face frontalization

The rectification of the face texture to a canonical head pose is a key step for head pose invariance and it is done as follows. The textured 3D mesh (i.e., a mesh where each 3D point is associated with an RGB color) of the face image, is rendered after the application of the rigid transformation $\mathbf{p}^{-1} = (\mathbf{R}^T, -\mathbf{R}^T\mathbf{t})$, i.e. the inverse of the estimated head pose, generating a frontal-looking face image (Fig. 3.1b). In our case, the textured 3D mesh is a template-based mesh resulting from the mapping of the RGB texture to the fitted person-specific 3DMM. This template approach depends on the 3DMM fitting quality, compared to using depth mesh directly, but it provides a smoother surface for the rectification and frontal rendering.

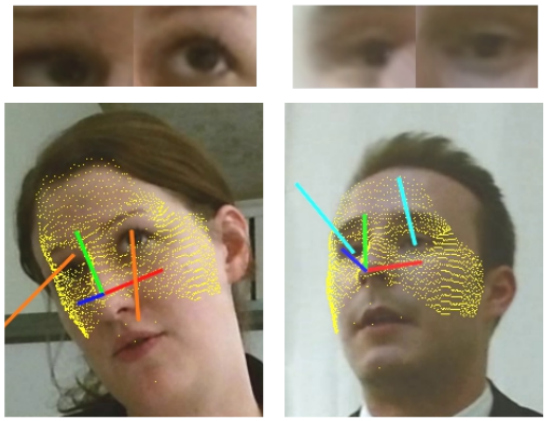


Figure 3.3 – Examples of eye cropping using 3DMM semantic, illustrating the importance of eye localization in the frontal face image (left: wrong, right: correct). Yellow dots represent the 3DMM mesh, the red-green-blue coordinate system represents the head pose, and the blue and orange arrays represent the estimated gaze direction originating from the estimated 3D eyeball centers. The cropped frontalized eye images are shown on the top of each result.

3.3.2 Eye images cropping

In Funes Mora and Odobez (2016), eye localization is performed by exploiting the semantic information of the 3DMM. However, the accuracy of the eye positions obtained through 3DMM semantic is directly affected by the head pose estimation and 3DMM fitting qualities, which can suffer some errors, either because the 3DMM features are not rich enough, or due to an inaccurate fitting (e.g. if the subject is only seen from a highly non-frontal pose during most of the video). Fig. 3.3 shows an example of misalignment due to a slightly inaccurate 3DMM fit, cropped eyes are lower and to the right, resulting in an underestimated gaze pitch. To address this issue, in Funes Mora and Odobez (2016), the eyes region window is further improved by using a person-specific mapping learned from a few samples.

In our work, we simply localized the eyes by applying a 2D landmarks detector (King, 2009) directly on the frontal face image, which led to more robust and stable results without the need to collect person-specific samples. Given an estimate of the eye corner locations, we simply compute the alignment translation that minimizes the discrepancy between the aligned corners (after translation) and their expected location in the canonical frame which was used to train the appearance model. Despite landmarks not being always accurate typically because the frontal face suffers some deformations compared to a real face or the landmark location of the eye varying slightly in function of eye closure, it was shown to improve gaze estimation accuracy to a reasonable accuracy (Siegfried et al., 2017).

Finally, we end up with two 36x60 pixels eye images, with normalized eye corners position, size, and orientation, which makes it easier to learn an appearance-based gaze estimation model.

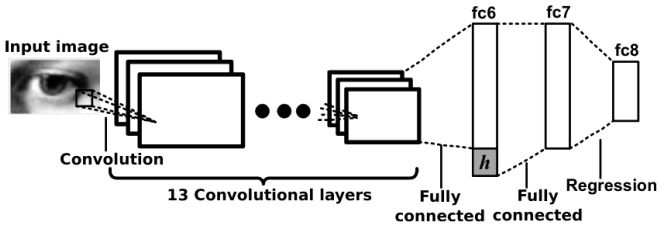


Figure 3.4 – Architecture of GazeNet (Zhang et al., 2017b). The network is based on VGG16. It estimates the 2D gaze angles from an eye image and the associated 2D head angles.

3.4 Gaze estimation

In order to estimate the 2D gaze angles $\mathbf{g} = (\phi_{\mathbf{g}}, \theta_{\mathbf{g}})$ from eye images, we rely on GazeNet (Zhang et al., 2017b), as it was shown efficient on state-of-the-art datasets in cross-subject and cross-dataset settings. It is based on the *VGG16* architecture and takes head pose angles into account, as shown in Fig. 3.4.

Unlike the presented head pose estimation method, deep learning methods need to be trained with a significant amount of data. However, our goal is to be able to adapt our gaze and attention estimation method to new setups with minimal supervision and as the datasets used in this thesis (see Chapter 4) do not contain enough gaze-annotated frames and diversity of target positions for training the network, we relied on an external dataset to train it. We used the Eyediap dataset (Funes Mora et al., 2014) as it provides depth data, allowing the usage of similar normalized eye image extraction, reducing the gap between the training and testing domains. We trained the network on all Eyediap subjects using the "floating target" setting with mobile head pose and used the resulting model for all our experiments.

3.5 VFOA estimation

Finally, we can compute the gaze vector $\mathbf{v_g}$ from the gaze angles in HCS $\mathbf{g}$ and the head pose $\mathbf{p}$:

$$\mathbf{v_g} = \mathbf{R} \cdot \mathsf{V}(\mathbf{g}). \tag{3.1}$$

Comparing it with the directions of the potential visual targets $\mathbf{v}_j$ derived from the target and subject's eye positions, it is possible to classify the subject's VFOA f. In our work, we used two different methods. In Chapter 5, we use a simple geometric model with a distance threshold as the single parameter, which allows its application on any gaze data without training. However, it is a rather coarse estimation, which assumes homogeneity in all directions and unbiased gaze estimation. Thus, in Chapter 6, where we study VFOA estimation more closely, we relied on a probabilistic model, which overcomes these limitations in cases where some training data are available. The details of each method is provided in the chapters where they are used.

Note that in all cases, we relied on the gaze values obtained from the closest eye to the camera

as gaze information, since it is usually the most visible and thus less prone to occlusions and deformations from the rotation, and results in more precise and stable estimations.

3.6 Conclusion

Using the framework presented in this chapter, we can now estimate the head pose and the gaze of a subject from an RGB-D recording, without the need for additional training. On top of that, we can classify the subject's VFOA given the 3D positions (expressed in the subject's HCS) of the potential visual targets. However, the performances of such a model applied in a new setup and with a new user are usually limited. In the following, we present methods to refine its gaze and VFOA estimation, using calibration (Chapter 5) or a more complex VFOA estimation method (Chapter 6).

4 Datasets

In this chapter, we introduce the datasets that we used in our different experiments including our contributions in terms of data collection and annotations. As mentioned in Chapter 1, we are interested in situations where people are performing a task in natural conditions and as we aim for 3D scene understanding, this requires access to the 3D position of people and objects. Also, as seen in Chapter 2, gaze estimation from remote sensors is a challenging task, so we imposed some constraints to maximize the chances of getting acceptable gaze estimates. We focus on datasets where:

- people are recorded by RGB-D cameras, so we can use the framework presented in Chapter 3;
- each person in the scene is recorded by a camera where the eyes (and pupils) are normally visible;
- the spatial and temporal calibrations are available, so we can get the relative 3D positions of the different people and objects in the scene which is required for VFOA estimation.

4.1 UBImpressed dataset

4.1.1 Data

The UBImpressed dataset (Muralidhar et al., 2016) consists of 330 short dyadic interactions (five to ten minutes) in which a participant interacts with an actor in two different scenarios. Both setups are presented in Fig. 4.1. From a gaze perspective, this dataset presents a situation where a unique important target is present, i.e. the other person. The short interaction and the formal setup favor mutual gazes, so it is interesting to study people's behaviors with little perturbations as people focus on the interaction.

In the *Interviews* scenario, the applicant and the interviewer (actor) are sitting in front of each other at a distance of two meters. Since this social interaction is rather formal, people remain still most of the time and exhibit constrained behaviors. In the *Desk* scenario, a receptionist

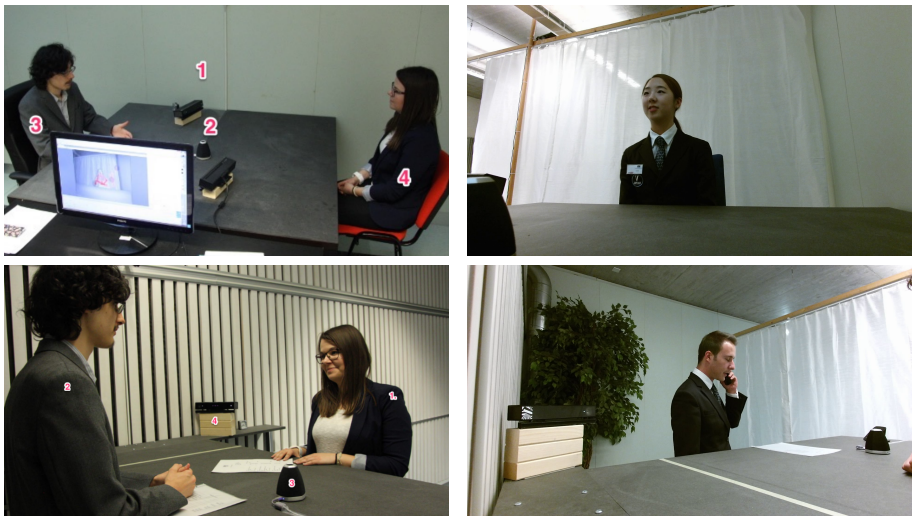


Figure 4.1 – UBImpressed dataset. *Interviews* (up) and *Desk* (bottom) setups (left) and examples of recording (right).

must deal with the questions and complaints of a client (actor), with moments where the receptionist uses the phone or discusses a bill on the desk with the client. This setup favors a more open and animated type of communication and presents a higher variety of gaze behaviors as well as body and head movements since people are standing.

Videos were acquired with Kinect 2 sensors (RGB-D, HD color images, 30 fps) placed on the table, recording each participant from the side (around 45°), as shown in Fig. 4.1. The dataset itself provides utterance annotations. Indeed, the sound was synchronously recorded by a microphone array that automatically detects the beginning and end of each participant's utterance, thus indicating who is talking in each video frame. Moreover, spatial calibration between cameras is provided as a rigid transform from the camera frame to a World Coordinate System (WCS), allowing to project 3D data from a camera frame to the other. Finally, temporal calibration is given as the time difference between the start of the two videos.

4.1.2 Contributions

VFOA annotations As annotating attention target is time-consuming, we focused on optimizing data usability. We started from the typical experimental protocol for gaze estimation, which consists in training/calibrating the evaluated method at the beginning of the interaction and evaluating it later. Thus, we annotated the first minute of each considered video and 5 additional segments of 10 seconds, separated by 50 seconds of not annotated video (so 6 segments in total). Thus, we ensure enough data that includes dynamic behaviors for training and sample diversity for testing, while avoiding annotating the whole videos. The per-frame annotation indicated whether the subject was blinking or not, and in the latter case, whether he was looking at the other person ("gazing" label) or not. As attention shifts are a

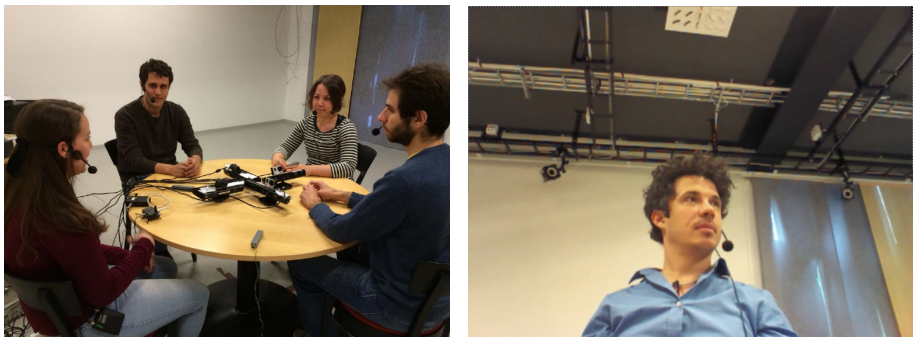


Figure 4.2 – KTH-Idiap dataset. Whole setup(left) and example of recording (right).

rather low-frequency behavior we annotated one frame over three, i.e. each 100ms, so we gain time on annotations. Afterward, we interpolated the annotations, so that every frame in the considered segments got a label. We annotated 4 *Interviews* and 4 *Desk* sessions (16 videos in total). Ignoring the blinking frames, we ended up with 42'000 annotated frames with 52% of them having the "gazing" label (hence 48% with the aversion label).

4.2 KTH-Idiap group-interviewing corpus

4.2.1 Data

This dataset (Oertel et al., 2014) consists of five one-hour four-party meetings in which three students present their Ph.D. projects to an interviewer who leads the discussion. The setup is presented in Fig. 4.2. From a gaze perspective, this dataset presents a situation where three important targets are present, favoring gaze shift between targets as the speaker changes. Also, more roles can be identified, as people will sometimes be side-listeners rather than the main addressee, potentially increasing the diversity of behaviors compared to dyadic interaction.

Compared to UBImpressed, it comprises a more relaxed type of social interaction. The alternation of monologues, dialogues, and animated discussions as well as the presence of three other people (i.e. potential visual targets) makes the interaction highly dynamic and generates a high variability of head gestures, facial expressions, and gaze patterns. Each session is organized into five parts: (1) students are left alone, (2) each student presents himself, (3) each student gives an elevator-pitch for his project, (4) each student discuss the impact of his project on society, (5) students discuss a joined research project. In our experiments, we consider the four last sections, ignoring the first section where there is less and more occasional interaction.

Videos were acquired with Kinect sensors (RGB-D, VGA color images, 30 fps) placed on the table at around 0.8 meters in front of each participant. As participants were wearing lapel microphones, the dataset provides automatic utterance estimation using sound intensity (Oertel et al., 2014). Also, spatial and temporal calibrations were provided in the same format as for

the UBImpressed dataset.

4.2.2 Contributions

VFOA annotations. Following the same protocol as in the UBImpressed dataset, VFOA was manually annotated on the first minute (i.e. first minute of the second part of the meeting), and on nine additional 30 seconds segments spread on the entire video, to catch different situations like monologues, dialogues, attentive listening, moments of aversion, etc. In these segments, one frame over three (i.e. every 100 ms) was annotated whether the participant was blinking, looking at another person (left, facing, right), or looking elsewhere (aversion case). Overall, it represents 110 minutes of annotation (180'000 frames excluding blinking but including interpolated labels) in which 19% are aversion and 81% are looking at another person.

Eye movements annotations. Following the same scheme as for VFOA, all videos were annotated for eye movements. We considered 5 classes:

- fixation;
- blink during a fixation (fix-blink);
- saccade;
- blink during a saccade (sac-blink);
- unknown.

The distinction between the blinks happening during fixation and saccade seems important to us, as eye behaviors are quite different in terms of gaze signal and eye appearance. Despite eye movements being faster events than attention shifts, we kept annotating frames each 100 ms. We ended up with a total of 180'000 annotated frames excluding *unknown* frames but including interpolated labels (81% of fixation, 13% of saccades, 4% of blinks during fixations, and 2% of blinks during saccades), with a predictable over-representation of fixation.

4.3 ManiGaze dataset

The two previous datasets concern conversation, but we are interested in studying attention in other kinds of setups too. It has been shown that vision is intimately bound up with the control of purposeful actions (Land, 2009). Especially, objects manipulation, in which the role of gaze has also been well studied (Admoni and Scassellati, 2017), is another interesting task for VFOA estimation, which is commonly encountered in human behaviors analysis and HRI.

The importance of gaze has been demonstrated during object manipulations, as the proactive use of the gaze informs about the intention and anticipation of the actor while its reactive use enlightens particular attention (Bader et al., 2009). Also, a strong correlation was found between eyes and hands movements, e.g. during pick and place actions (Johansson et al.,

2001; Sidenmark and Lundström, 2019). However, most studies in this domain relied on either hand-coded gaze events or the use of intrusive sensors like chin-rests (Johansson et al., 2001) or head-mounted devices (Newman et al., 2018), which limit the usability of gaze estimation in real applications. Thus, we are interested in investigating the use of remote sensors during object manipulation.

Nevertheless, such datasets are difficult to find. Newman et al. (2018) introduced the HARMONIC dataset to study collaboration between a user and a 6 degrees of freedom robot to pick marshmallows. This dataset provides an accurate gaze signal estimated by a head-mounted eye tracker together with the egocentric view of the user and a scene point of view, which is very interesting from a gaze estimation and analysis perspective. However, the user does not perform specifics gestures or actions, as the task consists in piloting the robot arm with a joystick. Also, Azagra et al. (2017) collected the MHRI dataset, which consist in the recording of people presenting objects to a Baxter robot, by pointing it, picking it to show it to the robot, or describing its position relatively to other objects. This dataset provides two RGB-D recordings, one focused on the user and the other on the table, plus an HD recording of the robot's point of view, together with sound. However, this dataset was meant to learn object models and thus does not provide the user's gaze or VFOA ground truth.

Therefore, to evaluate remote gaze estimation during object manipulation tasks, we collected and made public the ManiGaze dataset, providing both artificial scenarios where the user gaze is constrained, which makes it easy to get gaze ground truth, e.g. looking at a specific point on a table, and more natural ones, where the user is asked to freely perform actions, like picking an object.

4.3.1 Contributions

This dataset was recorded in the frame of and HRI project, whose goal is to investigate the social structure of the learning by demonstration process. This project proposed to rely on natural interactions for skill learning, involving queries about the skills and answers, and demonstrations made by both the human and the robot to show what it has learned.

In this context, the ManiGaze dataset was collected to benchmark gaze estimation in an HRI setup with a collaborative robot, with object manipulation tasks in mind. The dataset is described in Siegfried et al. (2020). In terms of contribution, B. Aminian designed the experiments, built the setup, organized the data collection sessions, and collected the data. I contributed by organizing the collected data, writing and publishing the related paper, and making the data publicly available.

4.3.2 Setup and Calibration

The physical setup is illustrated in Fig. 4.3. It consists of a Baxter robot separated from the participant by a table on which the manipulation tasks take place, similarly to the MHRI

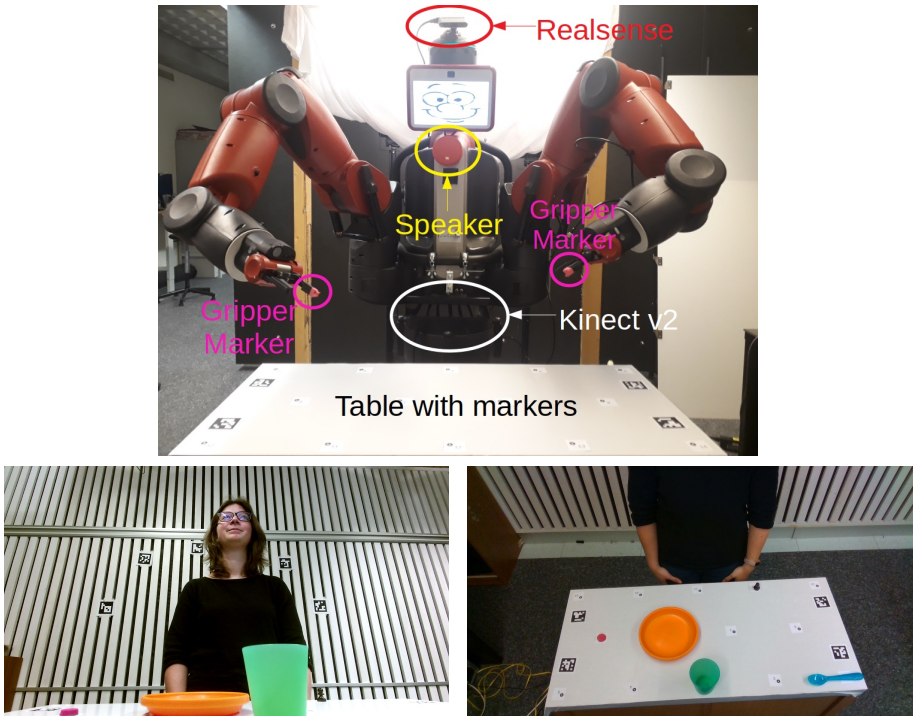


Figure 4.3 – ManiGaze dataset experimental setup from the point of view of the user (top), the field of view of the Kinect v2 (bottom left), and of the Intel RealSense (bottom right).

dataset (Azagra et al., 2017). On this table, 14 markers (numbered black dots) were placed on three rows in a triangular mesh pattern with a distance of around 20 cm between two rows and two markers within the same row. The distance between two markers of the same row is 20 cm and there is 20 cm between rows.

Three modalities, namely color video, depth video, and audio, were recorded by 2 RGB-D cameras and a microphone. Finding a good place to sense the gaze of the participants was difficult because of potential targets position (objects on the table, robot arms, and robot head) which were very spread in space. The best location we found was at the height of the table, and we used a Kinect v2 sensor due to its depth accuracy and large field of view. The low camera angle is unusual compared to classical gaze estimation datasets, but it can be cumbersome to get ideal sensing conditions (i.e. frontal view) in real-life setup, as sensors can not reasonably be placed in the workspace (here: the table). A second RGB-D camera (an Intel RealSense D435) was also placed on the robot head to record the table and participant hands from above. In addition to the sensor data, several features related to the interaction like mouse clicks (see below), robot speech, or robot arms position were recorded.

Wizard-of-Oz approach. The recordings were made using a Wizard-of-Oz approach. All the robot's behaviors were generated in advance, and the experimenter triggered them sequentially when it was appropriate. Also, at any time, the participant could ask to repeat the previous instruction. At the beginning of the experiment, the robot introduced himself and asked a

few questions to make the participant used to the robot (e.g. "how are you today?", "Nice to meet you, what is your name?"). Then it guided the participant through the whole set of experiments using voice to encourage interactions and natural behaviors. Randomness was introduced in the robot utterances ("look at X", "Can you look at X", "Now, look at X") and in the feedbacks ("ok", "congratulations", "well done") to make the interaction more natural. The robot also randomly asked to look at it to break the task monotony and gather gaze points to the robot. Note that participants were not told to look at a specific point on the robot, but the majority looked naturally to the "robot face". The result was qualitatively satisfactory: participants tend to ignore the experimenter, speaking naturally to the robot and turning the head toward it when the robot was speaking.

4.3.3 Recorded sessions

The experiment was organized into 4 sessions, going from the most artificial to the most natural, and the participant received some basic instructions before the start.

Markers on the Table Targets (MT). The participant stood approximately 1 meter in front of the robot and had the computer mouse in hand. The robot asked the participant to look at a numbered marker on the table. The participant would then press the mouse button upon gazing at the target marker without blinking. After feedback, the procedure would repeat, with the robot designating another marker until the participant looked at each marker. The participants' gaze fixation locations and their occurrence time were deduced from the mouse events and the target marker position. This session can be used to evaluate gaze estimation and calibration algorithms, as the markers build a dense and regular grid distributed on the whole manipulation space.

End-effector Targets (ET). This session is similar to the MT one except that the participants are asked to continuously look at one of the robot end-effectors and to press the mouse button when it stops moving (37 positions) while looking at it without blinking. Each robot's arm is used to cover half the space, to avoid participant's face occlusion. This session provides additional targets not limited to the manipulation area (the table) to evaluate gaze estimation. In our work, we focus on the fixations indicated by the mouse events, but data were continuously recorded, so the smooth pursuits performed by the participants could also be exploited in further studies.

Object Manipulation (OM). In this session, objects of different sizes and shapes (a backgammon pawn, a chess pawn, a spoon, a glass, and a plate) are initially placed on markers on the table. The robot asks the participant to perform pick and place actions to move a designated object to a designated numbered marker or on/in another object. Sometimes the robot also asked questions related to the actual position of the objects. As pick and place actions of the participant are guided by the robot instructions and objects are always placed on markers, the positions of the objects are known at each robot's occurrence (i.e. before and after each

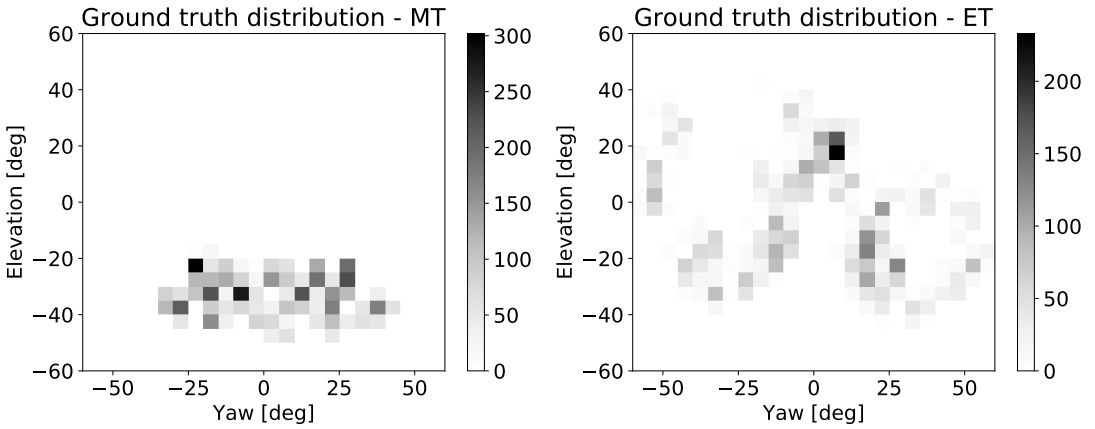


Figure 4.4 – Gaze histogram for MT and ET sessions. (Reference frame: Kinect v2 camera frame)

pick and place action). However, we do not know the gaze direction ground truth during this session, so it is best suited to study gaze behavior during pick and place actions (pick and place moments were annotated) using a gaze tracker that was evaluated and calibrated on the previous sessions. Indeed, it is less controlled than previous ones and as a result, participants act more naturally.

Set a Table (ST). In this final session, the participant is asked to set a table while explaining to the robot how to do it without referring to the markers on the table. The robot does not act or speak and the participant is free to choose how to proceed in order to encourage natural eye-hand coordination (e.g. usage of both hands at the same time, faster transitions between objects, and so on).

4.3.4 Data, annotations, and statistics

The multimodal dataset we collected involved 16 participants (24-36 years old, 4 women / 12 men, 6 glasses wearers). Both raw videos and annotations are provided publicly, i.e. both RGB-D video streams, gaze annotations for the ET and MT sessions, transcript of robot's speech, and grasp/release moments for the OM session.

The number of collected ground truth gaze points depends on the session and participant, as it is related to the duration of mouse events duration. We obtained an average of 807 labeled frames ($std = 350$, $min = 257$, $max = 1498$) by participant for the MT session (14 different targets) and 337 labeled frames ($std = 316$, $min = 146$, $max = 1381$) by participant for the ET session (37 different targets).

The gaze ground truth distributions for MT and ET sessions (see Fig. 4.4) have little overlap, as the MT session makes the subject look only at the table and the other makes him look at targets in a broader space. Also, because the position of the subjects' head is not constrained, each

target point leads to a lot of slightly different angles, providing more gaze angles variability than the defined number of targets. It should be noted that the range of elevation angles in the MT session (-50 to -20 degrees) represents a challenge, as gaze estimation is usually less accurate when people look down (Kellnhofer et al., 2019).

Regarding the manipulation session (OM), we manually annotated grasp and release events and obtained 11 annotated grasp events and the same amount of annotated release events for each participant.

4.4 Conclusion

In this chapter, we presented the dataset that we will be using to evaluate our methods in the following chapters and our contributions in terms of data collection and annotations.

Regarding conversation setups, we presented two datasets. The first one consists in dyadic interactions taking place in a formal setting as participants' skills are put to the test by actors. There is also an interesting difference between the *Interviews* and *Desk* scenarios, as the latter is much more dynamic in terms of people's positions, as people are standing and moving, and gaze behaviors, as they are sometimes discussing a paper placed on the desk. The second one consists in four-party meetings, where one participant leads the discussion. The setting is more relaxed and the interaction is much longer, so we can expect a higher diversity of behaviors. Also, there are more people involved in the interaction, which makes the interaction more complex.

Regarding manipulation setups, we presented the ManiGaze dataset, which was collected to provide a gaze estimation benchmark for HRI involving object manipulations. In this case, all the sensors are embedded on the robot and the camera recording the subject provides a challenging point of view, as it was not possible to place it in the manipulation workspace, where the participants are looking most of the time. Also, pick and place actions were recorded, which allows studying eye-hand coordination.

5 Unsupervised context-based gaze calibration

In this chapter, our goal is to improve gaze and VFOA estimation by building user-specific models, investigating weak conversation or manipulation attention prior along with robust estimation to perform online and unsupervised gaze calibration.

The idea of collecting samples based on attention priors was initially presented at the 2017 Communication by Gaze Interaction (COGAIN) Symposium (Siegfried and Odobez, 2017). The robust calibration parameter estimation and deeper analyses about the calibration effects were published in the Proceedings of 2017 ACM International Conference on Multimodal Interaction (ICMI) (Siegfried et al., 2017). Later, we further extended this work to more complex calibration models and different setups. Finally, the work presented in this chapter was accepted for publication in the ACM Transactions on Multimedia Computing, Communications, and Applications journal (TOMM) (Siegfried and Odobez, 2021a).

This chapter is structured as follows. First, we present our intuitions and discuss our contributions (Sec. 5.1). Second, we present the proposed method (Sec. 5.2). Third, we provide our experimental results on weak VFOA labeling (Sec. 5.3) as well as offline, and online calibration (Sec. 5.4 and 5.5). Finally, we discuss the limitations of this work and future work (Sec. 5.6).

5.1 Introduction

5.1.1 Motivation

Perceiving the gaze of others is a difficult task even for humans. Indeed, we tend to underestimate the amplitude of the gaze when people look aside (Kluttz et al., 2009) and to overestimate it when people look at a target that is far from us (Masame, 1990). Furthermore, we are biased by features like head orientation (Wollaston effect) (Kluttz et al., 2009) and nose direction (Langton et al., 2004). However, while we encounter difficulties at estimating gaze direction, we are good at exploiting context information to estimate people's visual focus of attention (VFOA), i.e. "what they are looking at", as illustrated in Fig. 5.1. Indeed, we can

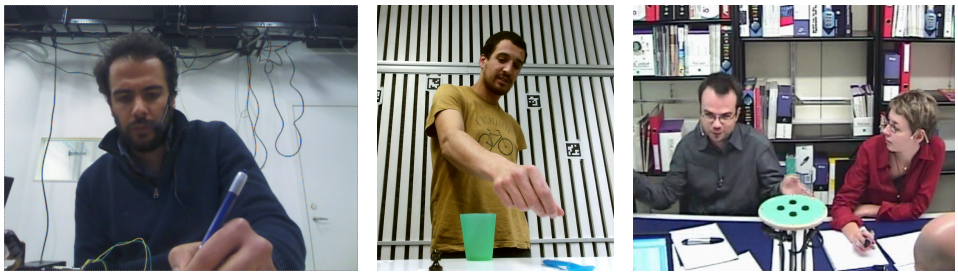


Figure 5.1 – Examples of people's gaze during a task. One can reasonably guess what they are looking at, even if it is difficult to see eyes (the word he is writing, the object that will be grasped, the speaking person).

recognize which objects are relevant visual targets in a given situation and use this information as a prior to indirectly better infer people's gaze direction.

Algorithms do not suffer the same biases as us but they have their own challenges, as discussed in Sec. 1.2, which make gaze estimation with remote sensors difficult. To overcome these difficulties, gaze estimation models are usually adapted to new users and settings through calibration, which highly increases their performances. However, as seen in Sec. 2.1.3, it usually requires dedicated calibration sessions which are cumbersome to conduct in dynamic and open settings. Our goal is to avoid such sessions, by exploiting human attention properties to automatically collect calibration points during the system exploitation.

As systems interacting with people in natural conditions using remote sensors rarely have access to people's field of view, which precludes the use of saliency as context. However, as they are usually designed for a set of specific tasks, they can leverage top-down context-based priors to estimate people's VFOA. Examples of priors include:

- conversation: people usually look at the other speaking participants of the discussion;
- manipulation: upon object manipulation, eye and hands movements are coordinated (Johansson et al., 2001);
- driving: people often look at the future path (Lappi et al., 2013);
- web browsing: the gaze is strongly related to the cursor position (Chen et al., 2001).

In this work, we will investigate the first two contexts as prior for calibration data collection, as they are the most common cases in HHI and HRI setups. Note that context information was already shown useful to improve VFOA estimation (Sheikhi and Odobez, 2015; Ba and Odobez, 2011), but in this work, we also propose a method to exploit them as calibration points to improve gaze estimation.

5.1.2 Approach summary and contributions

Our approach for unsupervised user-specific gaze calibration is summarized in Fig. 5.2. In essence, it comprises two main steps. The first one corresponds to our aim and consists in

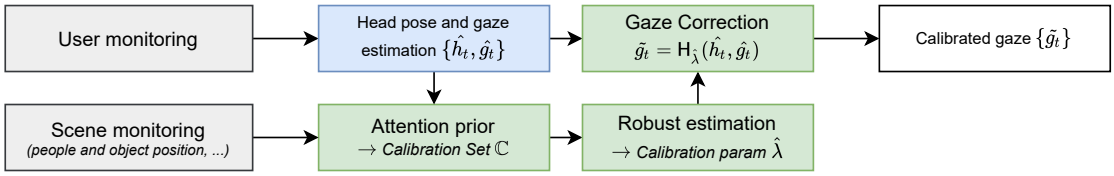


Figure 5.2 – Calibration framework. An initial head pose and gaze estimations $\{\hat{\mathbf{h}}_t, \hat{\mathbf{g}}_t\}$ are extracted from the video of the user for each frame t (blue block). In parallel, the calibration set $\mathbb{C}$ is automatically built relying on the attention prior and the scene monitoring information, and the parameters λ of a predefined calibration model H_λ are robustly estimated (green blocks), model which is then used to infer the calibrated gaze $\tilde{\mathbf{g}}_t$.

using contextual attention prior to collect calibration points without the conscious help of the user. Indeed, when some VFOA targets become more important according to the context, so does some gaze directions. Secondly, we propose a robust estimation framework to use these calibration samples to estimate the parameters of a pre-selected calibration model. Indeed, since such attention behaviors reflect tendencies more than strict rules (e.g. people do avert their gaze during a conversation or briefly take their eyes off the road while driving), the resulting VFOA labels are prone to error and a method for filtering outliers is needed to obtain reliable calibration parameters. In this context, our contributions can be summarized as:

- we investigate the use of context-based attention prior for collecting gaze calibration samples;
- we propose a robust estimation framework to exploit these samples for unsupervised user-specific gaze calibration, studying different calibration models;
- we propose an online approach to adapt the calibration to the local context;
- we study the application of these methods with two main priors and setups, namely conversation and manipulation.

Experiments on three datasets demonstrate the validity of our approach, providing insight on the validity of the prior and on their impact on calibration. Note that the resulting user-specific calibration procedure can be applied on top of any remote gaze estimator to improve gaze estimation.

5.2 Method

5.2.1 Approach overview

Our overall approach is shown in Fig. 5.2. We assume that we are given a calibrated setup and scene where the subject is monitored using an RGB-D video stream and that the 3D positions of the potential VFOA targets, as well as the contextual information (utterances, pick-and-place manipulation actions), are also extracted and monitored from this stream or

other sensors (RGB-D cameras, microphones). In the following we assume that all variables are expressed and processed in the Head Coordinate System (HCS), which is attached to the subject's head.

The gaze correction workflow is as follows. At time t, a first estimation $\hat{\mathbf{g}}_t = (\hat{\phi}_{\mathbf{g},t}\hat{\theta}_{\mathbf{g},t}) = \mathsf{G}(I_{RGBD})$ of the gaze yaw and pitch angles is computed from the visual data I_{RGBD} (RGB and depth) using an off-the-shelves gaze estimation module which additionally provides an estimate of the head pose $\hat{\mathbf{h}}_t = (\hat{\phi}_{\mathbf{h},t}\hat{\theta}_{\mathbf{h},t})$. The gaze estimate is then updated using a calibration function H according to:

$$\tilde{\mathbf{g}}_t = \mathsf{H}_\lambda(\hat{\mathbf{g}}_t, \hat{\mathbf{h}}_t). \tag{5.1}$$

As can be seen, the head pose is exploited in the correction as gaze errors might depend on it (more in Sec. 5.2.6). The goal is then to estimate online the parameters λ of this model using weakly labeled data collected over time thanks to the use of prior VFOA models.

5.2.2 Problem formalization

More precisely, the learning process is defined as follows. We assume that at any instant t, the set of relevant objects and people in the scene that the subject can look at is defined by $\mathbb{T}$. Then, through time, a set $\mathbb{C}$ of calibration samples is collected

$$\mathbb{C} = \left\{(t_i, j_i), i = 1 \dots N_{\mathbb{C}}\right\}, \tag{5.2}$$

where each pair i comprises a time instant t and a target index $j \in \mathbb{T}$. Given our calibrated set-up, the set $\mathbb{C}$ can then be transformed into an actual set $\mathbb{F}$ of gaze calibration samples according to:

$$\mathbb{F} = \left\{(\hat{\mathbf{g}}_t, \mathbf{g}_t), \text{ with } \mathbf{g}_t = \mathsf{V}^{-1}\left(\frac{x_{j,t} - e_t}{\|x_{j,t} - e_t\|}\right), \forall (t, j) \in \mathbb{C}\right\} \tag{5.3}$$

where $\hat{\mathbf{g}}_t$ is the estimated gaze, $\mathbf{g}_t$ is the angles associated to the gaze direction $(x_{j,t} - e_t)$ corresponding to looking at the target j at time t, where e_t and $x_{j,t}$ denotes the 3D positions of the eye and of the target j at time t. The function V is a function that transforms a 2D gaze angle representation into the corresponding 3D gaze direction vector in HCS, and V^{-1} is its inverse.

Given such a calibration set, the goal is then to estimate the parameters of the calibration model H_λ by optimizing an error function E (λ,$\mathbb{C}$) based on the residual discrepancy between the corrected gaze $\tilde{\mathbf{g}}$ and the calibration gaze $\mathbf{g}$:

$$r_t(\lambda) = \tilde{\mathbf{g}}_t - \mathbf{g}_t = \mathsf{H}_\lambda(\hat{\mathbf{g}}_t, \hat{\mathbf{h}}_t) - \mathbf{g}_t. \tag{5.4}$$

Since the calibrations points are not certain, we rely on robust estimation criterion (Least Median of Square, i.e. LMedS) to first filter out outliers, and then apply a regularized Least Mean Square optimization (LMS) on the remaining calibration points. In the following, we describe the main elements of this method:

- target position acquistion (Sec. 5.2.3);
- gaze estimation G from RGB-D images (Sec. 5.2.4);
- gaze calibration set $\mathbb{C}$ collection using different weak VFOA labeling schemes (Sec. 5.2.5);
- calibration function H_λ (Sec. 5.2.6);
- robust estimation of the calibration parameters λ (Sec. 5.2.7);
- online estimation of λ (sec. 5.2.8).

5.2.3 3D target positions aquisition

To estimate the VFOA of a subject, we need to know the potential visual targets in the scene and where they are. This requirement is intrinsic to the task, and not specific to our approach. In this work, we consider each target $x_{t,j}$ as a single 3D point in space, approximating the position of the whole target. In the case of people (interaction scenes), we use the nose tip as an approximation for the face, as people usually look at the eyes/mouth region when talking to each other. As mention in Sec. 5.2.1, we assume a scene calibrated setup. The nose tip is thus extracted by applying the Headfusion method (see above) on the video of people, and mapping its estimated 3D position in the subject's HCS. While this approximation induces a bias as we do not know where people are looking precisely, it is small given the distance between people (e.g. the nose-to-eye distance is around 3cm corresponding to 1.7° at 1m and people are usually further than that) and compared to the errors made by the gaze system in such distance sensing free-moving settings. In the case of objects (manipulation), we use the position of the marker below the object, which is provided by the dataset.

5.2.4 Gaze and VFOA estimation

Gaze estimation

The gaze estimation function G is the method described in Chapter 3 (see Fig. 3.1). As a reminder, it first estimates the head pose from the depth data and uses it to rotate the textured mesh obtained from RGB and depth image to get a frontal view of the subject's face. The eye images are then cropped from the frontal face image and fed into a pre-trained network that outputs the gaze angles in the HCS. The final 3D gaze estimation is computed by combining the gaze angles with the head pose.

VFOA estimation

Knowing the gaze direction, the 3D position of the subject's eye and the 3D positions of the visual targets (people or objects), the VFOA of the subject can be decided using a simple geometrical model. More precisely, we rely on the gaze values obtained from the closest eye to the camera as gaze information, since it is usually the most visible and thus less prone to occlusions and deformations from the rotation, and results in more precise and stable estimations. Regarding VFOA, we use as a gaze distance to target j the angular difference

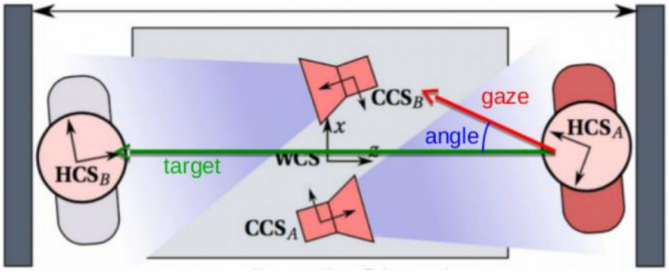


Figure 5.3 – Geometrical estimation of the VFOA in a dyadic interaction example. The angle between the subject's gaze direction and the target's direction is compared to a threshold to decide if the subject looks at the target or not.

between the gaze vector $\mathsf{V}(\tilde{\mathbf{g}}_t)$ and the unitary vector that goes from the subject's eye to the target j (see Fig. 5.3):

$$\kappa_{j,t} = \arccos\left(\mathsf{V}(\tilde{\mathbf{g}}_t) \cdot \frac{(x_{t,j} - e_t)}{||(x_{t,j} - e_t)||}\right). \tag{5.5}$$

Then, the VFOA at time t, denoted by f_t, is defined as the relevant target that is closest (according to κ) to the gaze direction and for which κ is smaller than a threshold κ_τ. Otherwise, if this last condition is not met, we decide that the VFOA is *aversion*, i.e. that the subject looks far away from any known target. Formally:

$$f_t = \begin{cases} j_{\min} & \text{if } j_{\min} = \text{argmin}_j(\kappa_{j,t}) \text{ and } \kappa_{j_{\min},t} < \kappa_\tau, \\ aversion & \text{otherwise.} \end{cases} \tag{5.6}$$

To set the threshold κ_τ, we must account for both uncertainties in the gaze direction estimates and the fact that visual targets are usually not a single point in space. As a typical value, we use $\kappa_\tau = 10°$, which is the angular size of an object of 35cm located at 2m away from the camera.

5.2.5 Calibration sets from VFOA prior

The calibration set $\mathbb{C}$ should ideally be built by collecting gaze samples for which the visual target of the user is known. However, as we do not have access to the actual VFOA, we aim to rely on the context to select gaze samples for which the probability of the VFOA is high knowing this context. To achieve this, we rely on rule-based weak VFOA labeling leveraging the knowledge of highly probable human gaze behavior associated with activities and tasks performed by the user, and propose three VFOA priors that can be used for that purpose: one that applies to conversations, one that applies to manipulation tasks, and a simple geometric prior that can be combined with the two others to improve their accuracy. We present them below. Their validity (statistics) will be studied in Sec. 5.3. It should be noted that our approach can be exploited in other situations and context involving such tasks (conversation, manipulation) or be adapted to other tasks in which another suitable VFOA context prior can be defined.

Conversation prior

During conversations, people unconsciously coordinate their behaviors to smooth the interaction. For example, turn-taking regulation requires some signaling which is mainly performed through gaze (Bohus and Horvitz, 2010). In particular, studies on social interactions showed that listeners are likely to look at the person being listened to (88% of the time) (Admoni and Scassellati, 2017). Even in presence of a slide presentation, it was shown that in four-party meetings people look 45% of the time at other people and that there are 5 to 8 times more chances to look at the speaker than to another listener (Ba and Odobez, 2011). Also, speakers are often looking at their addressee, i.e. the target of their speech (77% of the time) (Admoni and Scassellati, 2017). There is thus a clear relation between the speaking status and the VFOA of people that we can use to estimate weak VFOA labels and thus gather likely calibration points.

As identifying the addressee of a speaker is a rather difficult task, and speakers usually present highly dynamic head and gaze movements, we focus on exploiting the VFOAs of listeners. Thus, as conversation prior, we propose to use speakers as weak VFOA labels. Accordingly, the calibration frames set can be written as

$$\mathbb{C}_{conv} = \left\{(t, j) \mid j \in \mathbb{S}_t \text{ and } |\mathbb{S}_t| = 1\right\} \tag{5.7}$$

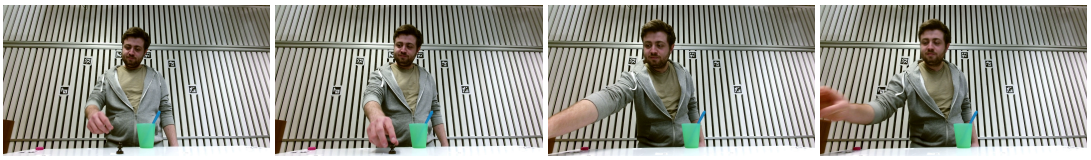
where $\mathbb{S}_t$ denotes the set of speaking targets at the instant t and $|\mathbb{S}_t|$ is its cardinality, i.e. the number of speakers.

Note that it has been shown that listeners tend to look at the speaker with a higher probability at some specific points (for example during turn changes (Oertel et al., 2013)). However, in our data, we haven't found that this was necessarily the case. Using these stronger constraints on context results in gathering fewer calibration points, which increases the risk for the model to be badly estimated if a subject presents an atypical behavior, so we did not use them.

Manipulation prior

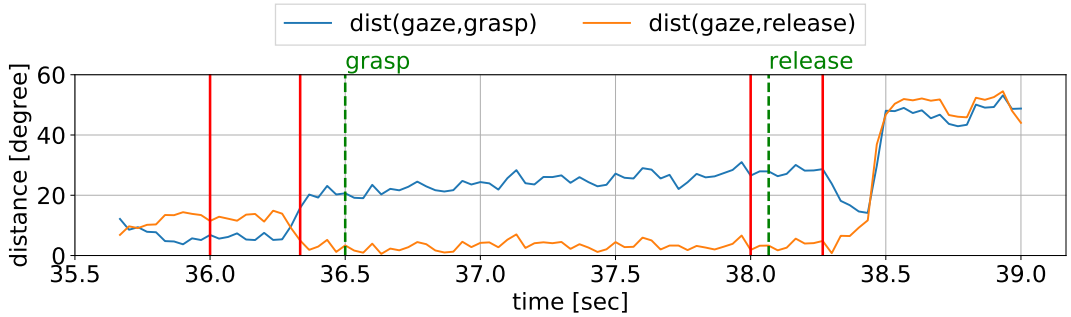
It has been shown that vision is intimately bound up with the control of purposeful actions (Land, 2009). During object manipulation, gaze typically reaches the object before any movement of the hands has started (Admoni and Scassellati, 2017) and upon reaching the object with the hand, people usually already anticipate the next action, for example by looking at where the object will be placed (Johansson et al., 2001). The reliability of this behavior makes it a good candidate for unsupervised gaze calibration (Sidenmark and Lundström, 2019).

More precisely, in pick-and-place actions, the gaze behavior is correlated to the hand touching instants (when the hand touches the object and when it releases it) and not necessarily to the start of the object motion. These two moments of interest were studied in (Johansson et al., 2001), where it was shown that people look at the origin and destination positions of the



a) Kinect2 recording at moments of interest.

b) Intel RealSense recording at moments of interest.



c) Cosine distance between gaze and grasping/releasing locations directions in the user's Head Coordinate System (HCS).

Figure 5.4 – Qualitative (a,b) and quantitative (c) gaze behavior during pick-and-place actions. Red vertical lines in the plot indicate when Kinect2 and Intel RealSense pictures were taken. Green vertical dashed lines in the plot indicate when actions occur (object grasp and release).

picked and placed object with very high confidence (> 90%) in a time window going from 1 to 0.6 second before the hand touching instants.

Fig. 5.4 shows the typical gaze pattern used by a subject during a pick-and-place action. The plot displays the cosine distance between the gaze and grasping/releasing location. The used gaze signal was calibrated on the ManiGaze dataset's MT session of the same subject and achieved a mean angular error of 4°. Despite gaze signal noise and relative inaccuracy, we can see when the user looks at the grasping location (first plateau around 36 seconds) and at the releasing location (second plateau from 36.3 to 38.4 seconds). It is interesting to see that the user anticipates the action by already looking at the releasing location before he grasped the object (second picture on each row). This anticipation is not seen at the release (fourth picture on each row), probably because there is no successive action to perform.

Thus, if the grasping or a release of a pick or place action of an object j occurs at time t', then we can infer the VFOA target as j with high confidence in a time window $\mathcal{W}_{t'} = \{t' - 1s, \ldots, t' - 0.6s\}$ of $0.4s$ duration, and use them for calibration. Accordingly, if we denote

by $M_{j,t'} \in \{grasp, release, other\}$ the action performed by the user on the object j at time t' and by $\mathbb{M}_j = \{t' \mid M_{j,t'} \in \{grasp, release\}\}$ the set of frame events where a given object j is grasped or released, then the calibration set can be derived as

$$\mathbb{C}_{mani} = \left\{(t, j) | t' \in \mathbb{M}_j \text{ and } t \in \mathscr{W}_{t'}\right\}. \tag{5.8}$$

Physical constraints

The proposed method estimates the head pose with high accuracy (2° to 5° errors as reported by Yu et al. (2018a)), so we can rely on it to constrain the possible gaze direction of the subject. Indeed, the maximal range of eye movement is 45° sideward and downward, and 30° upward (Lee et al., 2019). However, it was reported that looking further than 30° on the side is already uncomfortable (Smith et al., 2013). Thus, we can collect the set of objects that a person can be looking at as a weak calibration set. Practically, we compute the absolute difference between the gaze angles $(\phi_{j,t}, \theta_{j,t})$ which have to be used to look at the target j at time t, and the head pose angles $(\phi_{hp,t}, \theta_{hp,t})$, which are always zero in the head coordinate system. This difference is compared to a threshold vector (τ_ϕ, τ_θ), which represents the maximum tolerated eye rotation on each axis. Accordingly, we can define the calibration set for physical constraints (pc) as:

$$\mathbb{C}_{pc} = \left\{(t, j) \mid |\phi_{j,t} - \phi_{hp,t}| < \tau_\phi, |\theta_{j,t} - \theta_{hp,t}| < \tau_\theta\right\}. \tag{5.9}$$

These constraints only provide a rough estimation of potential VFOA, since it only checks whether objects are within the viewing frustum of the person. Thus, in practice, we use this condition together with the conversation or manipulation prior to increase their accuracy, as these constraints allow to remove obvious outliers (e.g. a target is speaking, but the user is looking far from it) from the calibration set. In this work, we use symmetrical constraints and thresholds for simplicity and to avoid overfitting this prior on our data. Also, note that the use of a robust estimator in our fitting scheme makes the approach resilient to these threshold values. Nevertheless, following (Sidenmark and Gellersen, 2019b), we could most probably use an asymmetrical prior for the pitch (i.e. define a calibration point as respecting $\tau_\theta^{down} < \theta_{j,t} - \theta_{hp,t} < \tau_\theta^{up}$ for pitch), since downward gaze movements are more common than upwards ones, and use different thresholds for the two angular directions. Further unsupervised estimation of these thresholds for different users could also be investigated, given the variety of behaviors between head-movers and eye-moves, but this would make the calibration more complex. We leave this for further research.

5.2.6 Calibration models

Selecting a good calibration model is important. While more complex ones may lead to better accuracies if learned with enough, diverse, and reliable data, they are also more easily prone to overfitting or outliers. In this work, we rely on rather simple models to avoid these issues,

and consider three variations of a calibration model whose general form is:

$$\mathsf{H}_\lambda(\mathbf{h},\mathbf{g}) = A\mathbf{g} + B\mathbf{h} + c, \tag{5.10}$$

$$\text{with } \mathbf{g} = \begin{bmatrix} g_\phi & g_\theta \end{bmatrix}^T, \mathbf{h} = \begin{bmatrix} h_\phi & h_\theta \end{bmatrix}^T, \text{ and } c = \begin{bmatrix} \gamma_\phi & \gamma_\theta \end{bmatrix}^T,$$

where A and B are 2×2 matrices. Below we describe these three variations, namely a constant, a linear, and a linear with head pose mapping.

Constant model

A simple model is to consider an offset depending on the user. It is especially useful when all reference points are near each other, like in dyadic interactions where the only available target is the other person. In such a condition, considering a constant bias is a simple way to avoid overfitting and better generalize to the remaining space. Also, it requires a smaller number of calibration points compared to more complex models. In this case, the A and B matrix and the set of parameters are:

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \text{ and } \lambda = \{\gamma_\phi, \gamma_\theta\}. \tag{5.11}$$

Linear model

In practice, we observe that large gaze values are often underestimated, which can be compensated by scaling the gaze estimation. Also, when more calibration frames are available and that targets are spread in space, more complex models can improve the calibration efficiency. To take this into account, we propose to use a linear model. In that case, we have:

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}, B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \text{ and } \lambda = \{a_1, \ldots, a_4, \gamma_\phi, \gamma_\theta\}. \tag{5.12}$$

Linear model with head pose

Gaze shifts usually involve both head pose and gaze-in-the-eye coordinated motion (Sheikhi and Odobez, 2015). We may thus expect that gaze estimation errors will also be related to head pose. For instance, looking to the side will induce both head pose and gaze shifts, with the latter potentially being underestimated depending on the camera point of view. Thus, we propose to improve the calibration model by taking into account the head pose. In that case, we have:

$$A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}, B = \begin{bmatrix} b_1 & b_2 \\ b_3 & b_4 \end{bmatrix}, \lambda = \{a_1, \ldots, a_4, b_1, \ldots, b_4, \gamma_\phi, \gamma_\theta\}. \tag{5.13}$$

5.2.7 Robust estimation

Algorithm 1: Robust calibration parameters estimation.

Data: $\mathbb{C} = \{(t, T)\}$
Result: $\hat{\lambda}$
$\lambda_{\text{Med}}, \hat{q} \leftarrow NULL, \infty$
for $i = 0$ *to* $maxIter$ **do**
  $\mathbb{C}' \leftarrow$ random_sampling($\mathbb{C}, n$) `// n=1 (cst), n=3 (lin), or n=5 (lin_h)`
  $\lambda_{\text{prop}} \leftarrow$ regression($\mathbb{C}'$)
  $\mathbb{R} \leftarrow$ compute_residuals($\mathbb{C}, \lambda_{\text{prop}}$)
  $q \leftarrow$ compute_median($\mathbb{R}$)
  if $q < \hat{q}$ **then**
    $\lambda_{\text{Med}}, \hat{q} \leftarrow \lambda_{\text{prop}}, q$
$\mathbb{C}'' \leftarrow$ filter($\lambda_{\text{Med}}, \mathbb{C}$)
$\hat{\lambda} \leftarrow$ ridge_regression($\mathbb{C}''$)

Because the calibration set is based on weak VFOA labeling, it will contain erroneous calibration points (i.e. with wrong labels, hence with wrong gaze references), which can greatly affect the gaze correction estimation quality. A typical example is when a subject looking to another person makes a short aversion by lowering his gaze. In multi-party conversations, it can also occur at turn changes: sometime the next speaker begins to talk but subject did not already switch his VFOA, leading to weak VFOA labeling errors. To handle this, we resort to the robust estimation paradigm by treating bad calibration points as outliers. In practice, this is achieved by applying a Least Median of Square estimator to filter out these outliers, and then applying Ridge regression on the remaining points. Algorithm 1 describes the whole process.

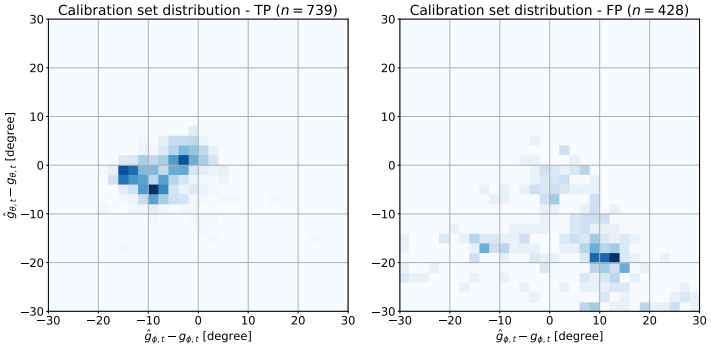
Least Median Square (LMedS) estimation

It is defined as the parameters optimizing the median of the residuals (defined in Equ. 5.4):

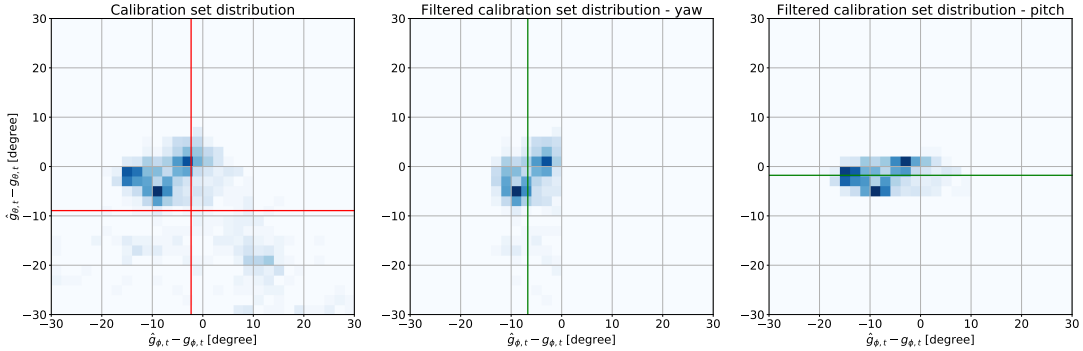
$$\lambda_{\text{Med}} = \underset{\lambda}{\operatorname{argmin}}\, \mathsf{E}_{Med}(\lambda, \mathbb{C}) \text{ with } \mathsf{E}_{Med}(\lambda, \mathbb{C}) = med\{\|r_i(\lambda)\|^2\}. \tag{5.14}$$

The advantage of the LMedS is that it has a breakdown point ϵ^* of 50%, meaning that even with 49% of outliers in the data, the estimator will not take an arbitrarily large value (Rousseeuw, 1984). However, it has a lower efficiency than the mean in terms of convergence to the optimum parameter value. This is why we use it for filtering outliers and then apply Least-Square (LS) for final estimation.

As the LMedS does not have an analytical solution, we rely on an iterative approach. At each iteration, a subset of samples is randomly selected and used to computed (in a LS sense) a parameter proposal λ_{prop}, which is then used to evaluate the median error $\mathsf{E}_{Med}(\lambda_{\text{prop}}, \mathbb{C})$. After a given number of iteration, the proposal minimizing E_{Med} is chosen as the estimate



a) Distribution of true positive (TP, left) and false positive (FP, right) in the calibration set.



b) Calibration parameters estimation without (left) and with (middle and right) LMedS filtering.

Figure 5.5 – Distribution of the calibration samples gathered using conversation prior (only annotated ones). Note that here, we plotted the difference between the gaze $\hat{g}_t$ and the target direction g_t. Vertical and horizontal lines indicate calibration parameters (constant model) estimated using the whole calibration set (red lines, left) or the filtered one (green lines, middle and right).

λ_{Med} and used to remove from the calibration set $\mathbb{C}$ half of the points which have the largest residuals.

Fig. 5.5 shows an example of LMedS filtering effect, using the constant model (without regularization) and the conversation prior on one subject in the UBImpressed dataset. Before filtering, outliers were miss-labeled as "looking at the target", as seen in Fig. 5.5a, and affects the calibration parameters estimation, especially in yaw, as seen in the Fig. 5.5b. Indeed, the computed bias (c parameter) represents the whole distribution instead of focusing on the distribution peak. It is solved by minimizing the median of the squared residuals instead, which is looking for parameters that minimize the residual of half the calibration samples only, ignoring potential outliers. From a geometrical point of view, it can be interpreted as choosing the 50% of the calibration samples that are the most packed in the calibration samples space to perform the estimation.

Ridge regression

The final calibration parameters $\hat{\lambda}$ can be estimated from the remaining samples using standard least-squares. However, as will be discussed in the results, instabilities and overfitting might occur due to the low variabilities of some observations; in particular when using the linear model, training samples may not span a large enough interval in yaw or pitch depending on the scenario and data, and the scaling might quickly depart from 1.

To handle this issue, we can use the Tikhonov regularization to penalize deviation from prior values of the parameters (1 for gaze scaling, 0 for all others). Introducing μ_0 as prior values, and Λ its precision matrix, the solution is given by:

$$\hat{\lambda} = (X^T X + \Lambda)^{-1}(X^T Y + \Lambda\mu_0), \tag{5.15}$$

where the specific form of X and Y depends of the used model. For instance, when using the linear model with head pose of Sec. 5.2.6 for which parameter estimation can be conducted separately for the ϕ and θ axes, we have when estimating the yaw parameters $\lambda_\phi = \{\alpha_{11}, \alpha_{12}, \beta_{11}, \beta_{12}, \gamma_\phi\}$:[1]

$$X = \begin{bmatrix} \hat{\mathbf{g}}_{\phi,1} & \hat{\mathbf{g}}_{\theta,1} & \hat{\mathbf{h}}_{\phi,1} & \hat{\mathbf{h}}_{\theta,1} & 1 \\ \dots & \dots & \dots & \dots & \dots \\ \hat{\mathbf{g}}_{\phi,N} & \hat{\mathbf{g}}_{\theta,N} & \hat{\mathbf{h}}_{\phi,N} & \hat{\mathbf{h}}_{\theta,N} & 1 \end{bmatrix}, \; Y = \begin{bmatrix} \mathbf{g}_{\phi,1} \\ \dots \\ \mathbf{g}_{\phi,N} \end{bmatrix}. \tag{5.16}$$

5.2.8 Offline and Online calibration

Until now, we discussed how to select calibration points and how to estimate the calibration parameters from them. However, though this may impact calibration and the performance we did not discuss when we should collect them. In general, as summarized below, we can identify two main strategies depending on the application.

- ***Offline.*** Some applications rely on a posteriori analysis of data, like in social interaction studies. In that case, the whole video can be used for collecting samples based on the prior, and the learned model is then applied to it.
- ***Online.*** Other applications require real-time gaze estimation, like in HRI. In that case, calibration samples are collected as they arrive, and model parameters are constantly updated as new data comes in so that corrected gaze predictions can be directly exploited.

Data updates

Regarding the *Online* case, we need to define how past information is accumulated and updated. The first aspect to take into account is that a minimal number of points is needed to

[1] Remind that the calibration pairs are $(\hat{\mathbf{g}}_t, \mathbf{g}_t)$.

average the gaze estimation noise and ensure some data diversity, but using too much data can also be computationally expensive; in addition, it can be good to forget old calibration points for long sessions where the gaze error potentially drifts.

To account for this, we simply defined $[N_{\mathbb{C}}^{min}, N_{\mathbb{C}}^{max}]$ as the interval for the calibration set size. Secondly, we need a strategy on how to update data points when the maximum number $N_{\mathbb{C}}^{max}$ is reached. As new data points can be somehow correlated (e.g. when interacting mainly with one person in a given time period) whereas older samples may contain valuable information (e.g. in terms of diversity, like resulting from having interacted with several people in the past), one interesting strategy is to select the data to be replaced in the calibration set at random. In that case, it can be shown that the probability for a sample (t, j) to remain in the calibration set is exponentially decreasing with the number of updates:

$$p\left((t,j) \in \mathbb{C}_k\right) = \left(\frac{N_{\mathbb{C}}^{max} - 1}{N_{\mathbb{C}}^{max}}\right)^{N_{upd}(t,k)}, \tag{5.17}$$

where $\mathbb{C}_k$ denotes the calibration set at time k, t denotes the time index when the calibration sample was collected, and $N_{upd}(t,k)$ indicates the number of collected samples between time t and k. Such a strategy has shown to be very effective in background subtraction tasks, where the goal is to model the distribution of past color values over time using a non-parametric approach.

5.2.9 Implementation details

In all experiments, the iterations number for the LMedS was set to 500, and physical constraints thresholds τ_ϕ and τ_θ (see Equ. 5.9) were set to 30°. Finally, when Ridge regularization was used, the penalization (i.e. the prior precision) was set to 10^4 for all parameters except for the translation ones, which are not penalized.

Also, we noticed that the calibration reacts sometimes poorly when the numbers of points per target are unbalanced. Indeed, in such cases, the robust parameter estimation will ignore the part of the space where there are less points. To avoid that, in practice, we used a weighted median in the LMedS procedure, so that each point labeled as "looking to target j" has a weight of:

$$w_j = \frac{1}{J \cdot N_{\mathbb{C}}^j}, \tag{5.18}$$

where J denotes the total number of targets and $N_{\mathbb{C}}^j$ the number of points in the calibration set that are labeled as "looking to target j".

Table 5.1 – Annotation statistics: number of subjects used in this work, annotation segments description, number of considered visual targets and average number of annotation per subject.

Dataset	Setting	Subjects nb	Annotations	Targets nb	Average nb of *target* annotations	Average nb of *aversion* annotations
UBImpressed	Desk	8 (4 sessions)	first minute + 5x 10sec	1	1525	1415
UBImpressed	Interviews	8 (4 sessions)	first minute + 5x 10sec	1	1856	1235
KTH-Idiap	Meeting	20 (5 sessions)	first minute + 9x 30sec	3	7327	1683
ManiGaze	MT	16	1 time each target	14	368	0
ManiGaze	ET	16	1 time each target	37[1]	337	0
ManiGaze	OM	16	special (see text)	22[1]	0[2]	0

1: not concurrent targets
2: only movements were annotated, not gaze

5.3 Weak labeling evaluation

We evaluate in this section the capacity of the VFOA priors at collecting good calibration sets. To do that, we use the annotated data in the three datasets presented in Chapter 4. Tab. 5.1 summarizes the amount of available samples. Key quality factors are: precision, which measures correct "looking at target j" labels in the calibration set, and size, which measures the amount and diversity of calibration points. Indeed, there is a trade-off between both since adding stricter constraints tends to increase the precision but also reduces the number of points accepted in the calibration set.

5.3.1 Conversation prior

Fig. 5.6 shows the precision of the "looking at target j" label and the ratio of frame fulfilling the prior over the physical constraints threshold: with no assumption, or assuming that the person is speaking (*spk*) or/and adding physical constraints (*pc*). In both datasets, adding assumptions (*spk*, *pc*) effectively increases the labeling precision, reaching values above 0.75. This increase is more significant in the KTH-Idiap dataset (+0.4), where more people targets are in competition. The *pc* condition also increases the probability to find a good calibration set with correct VFOA labels in all conditions. However, as seen in the bottom plots of Fig. 5.6, it has an impact on the number of candidate points for the calibration set. For example, setting a threshold to 25° in addition to the conversation prior in the KTH-Idiap dataset reduces by half the number of available frames.

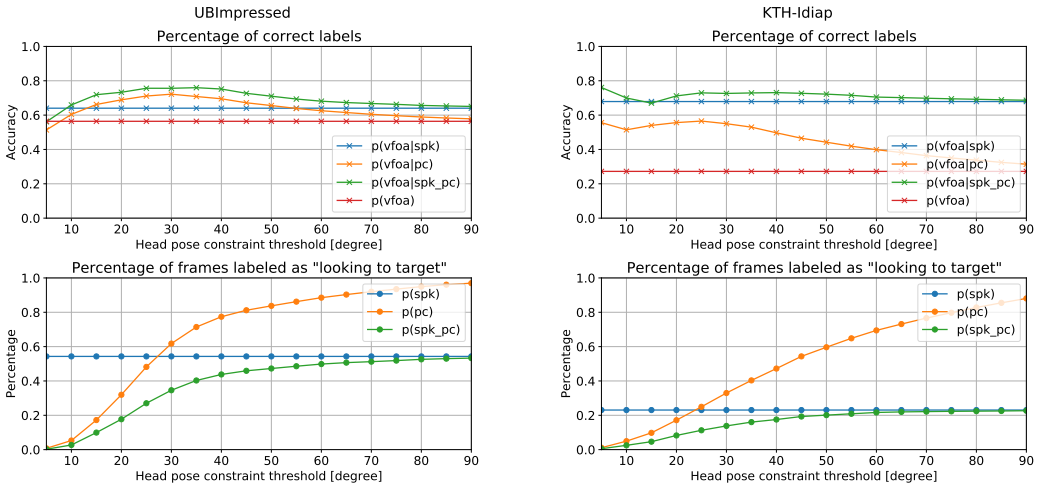


Figure 5.6 – Conversation prior statistics for each dataset. Top: weak VFOA labeling precision, i.e. probability that a subject looks at a given target (*vfoa*) during a *spk* event (the target is speaking) and/or a *pc* event (physical constraints are satisfied) for different values of τ_ϕ and τ_θ. Bottom: ratio of *spk* and/or *pc* events in data, i.e. the ratio of calibration points that would be gathered with a given prior and threshold over the total number of frames. For example, considering the UBImpressed dataset, using the conversation prior together with the physical constraint and setting a threshold to 35°, the proposed method will gather 4 points over 10 in the calibration set and 3 points of these 4 will be well labeled in average.

5.3.2 Manipulation prior

To assess the manipulation prior, we annotated frame by frame whether the subject looks at the origin/destination of a picked object in a 2.5 seconds time window around the grasping and releasing moments. Fig. 5.7 presents the resulting probabilities of looking at the target computed from activities of 4 subjects (44 grasps and 44 releases in total). It confirms that subjects indeed look with a high probability at the position where the grasp or the release will occur, as was shown ba Johansson et al. (2001). In our case, a maximum can be seen around 0.5 seconds before the action. Also, the fixation duration is shorter for the grasping as people anticipate the next action with the eye before the hand has finished the current action. Considering that we use a robust estimator which can theoretically tolerate up to 50% of errors, we could set a $[-1,0]$ time window for grasps and $[-1.2,0.4]$ for releases to gather more calibration points without affecting the results. Note however that in our experiments, as our goal is not to overfit our current setup, rather than using these statistics, we used the $[-1,-0.6]$ second interval before the grasp or release which was introduced by Johansson et al. (2001) as a prior to collect calibration samples, as described in Sec. 5.2.5.

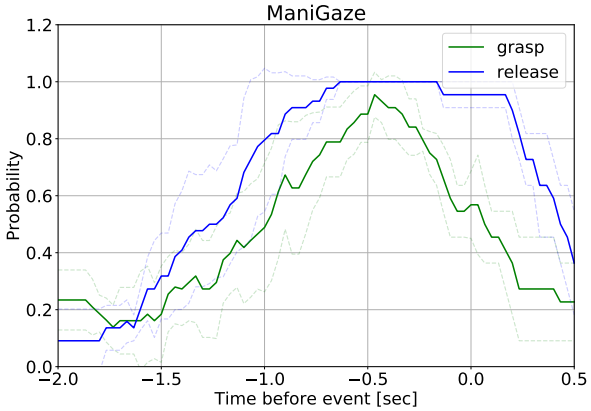


Figure 5.7 – Manipulation prior statistics. Probability that the subject looks at the initial (for grasps) or final (for releases) object position during a pick-and-place action in a time window around the grasp/release frame, computed from 4 ManiGaze dataset's subjects (OM session). Full line: mean. Dashed line: standard deviation.

5.4 Offline calibration experiments

First, we compare the gaze estimation performances using the *Offline* calibration protocol (see Sec. 5.2.8), i.e. using all the available data for both calibration and evaluation.

5.4.1 Experimental protocol

Performance measures

We are interested in two aspects: the accuracy of the gaze estimation itself and the accuracy of the subsequent VFOA estimation. Thus, we defined two performance metrics, namely the gaze angular error and the VFOA accuracy:

$$angErr = \frac{1}{T}\sum_{t=1}^{T} \arccos(\mathsf{V}(\mathbf{g}_t) \cdot \mathsf{V}(\tilde{\mathbf{g}}_t)), \qquad vfoaAcc = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\tilde{f}_t = f_t}, \tag{5.19}$$

where $\mathsf{V}(\mathbf{g}_t)$ is the ground truth gaze vector, and f_t is the ground truth VFOA. These metrics were computed for each subject and then averaged by dataset.

Note that because of the target position approximation (see Sec. 5.2.3) that affects $\mathbf{g}_t$, the angular error suffers a bias and can not reach 0. Indeed, people can look at different parts of the target region (e.g. face) without us knowing it. Still, we consider that it is a useful metric to compare methods, as this bias will be relatively small on average (around 1 to 1.5° maximum), much smaller than the errors of the studied systems (above 4 degrees).

Table 5.2 – Mean angular errors (in degrees) and mean VFOA accuracy (in percent) across subjects for the conversation datasets with calibration based on the entire session (*Offline*). Here we compare the raw gaze estimation without calibration (**Baseline**), the calibration using the manual VFOA annotations (**Oracle**), and the unsupervised calibration using the conversation prior and physical constraints (**Prior**).

	UBImpressed		KTH-Idiap	
Method	*angErr*	*vfoaAcc*	*angErr*	*vfoaAcc*
Baseline	9.19	0.77	14.52	0.40
Supervised (Oracle)				
Cst	6.00	0.88	12.38	0.62
LinGaze	4.95	0.83	8.64	0.73
LinGazeReg	5.00	0.88	8.45	0.76
LinHeadGazeReg	4.11	0.87	7.01	0.82
Unsupervised (Prior)				
Cst	8.29	0.82	12.67	0.56
LinGaze	6.78	0.72	9.30	0.70
LinGazeReg	6.52	0.76	9.21	0.71
LinHeadGazeReg	3.80	0.68	8.51	0.72

Evaluated models

We considered several calibration models (see Sec. 5.2.6) for our experiments:

- **Cst**: constant model without Ridge regularization;
- **LinGaze**: linear model without Ridge regularization;
- **LinGazeReg**: linear model with Ridge regularization;
- **LinHeadGazeReg**: linear model with gaze and head pose and Ridge regularization.

Calibration set

Two cases were considered. First (**Oracle** case), to evaluate the maximum achievable performance of the proposed method and calibration models we consider using the manual VFOA annotations to build the calibration set, as if the weak VFOA labeling process was perfect. Second, to evaluate our approach (**Prior** case), the calibration set was built as the intersection between the calibration sets defined by the main task prior (either conversation or manipulation) and physical constraints. In other words, $\mathbb{C} = \mathbb{C}_{conv} \cap \mathbb{C}_{pc}$ for conversation, and $\mathbb{C} = \mathbb{C}_{manip} \cap \mathbb{C}_{pc}$ for manipulation.

5.4.2 Results using Conversation prior

Tab. 5.2 presents the average of the angular error and VFOA accuracy across subjects in the conversation datasets for different models and calibration types.

Baseline results

The raw gaze estimate presents high angular errors compared to those reported in more traditional screen-based setups (9.19° and 14.52°). They can be in great part explained by the experimental setup (see Fig. 4.1 and 4.2). On both datasets, the subjects are seen from an unusually low angle. They are relatively far from the camera and the distance is changing, as people tend to lean back in their chair or lean on the table (KTH-Idiap) or move closer or away from the desk (UBImpressed), which creates variation in the eye image resolution and illumination. There are high head pose variations, especially in the KTH-Idiap dataset where subjects must significantly turn the head to look at other people. In such a situation, it is difficult to get a high gaze estimation accuracy, especially since the gaze estimator was not trained on these datasets.

Supervised calibration

As expected, it improves results for both metrics, and in general, more complex models achieve better results. **LinGaze** is an exception, as it is worse than the constant one regarding *vfoaAcc* on UBImpressed, which is mainly due to ill-conditioning: in this dataset, visual targets are not well distributed in the 3D space (there is a single target without much 3D change of relative position), so that the linear parameters are not well constrained. Hence, this issue is solved by regularization, so the **LinGazeReg** model maintains the performance of the constant model on the UBImpressed data (*vfoaAcc* = 0.88) while improving results on the KTH-Idiap dataset (*vfoaAcc* moving from 0.62 to 0.76). Finally, the **LinHeadGazeReg** model provides the overall best results, indicating that there can be a dependency between gaze estimation errors and head pose, especially when targets are distributed in space, and that it can be exploited for improvements.

Looking at the UBImpressed results, one can see that improvements in angular error and VFOA accuracy are not really correlated. This is due to two points. First, the two metrics do not involve the same set of frames. In particular, the VFOA accuracy takes into account frames where the subject does not look at a target and which are thus not considered neither for calibration nor for angular error evaluation. Secondly, as in UBImpressed the calibration data derives from a single target in the 3D space, the gaze correction may mainly consist of mapping gaze to a constant value, which would optimize calibration and angular error evaluation, but can be problematic when distinguishing between looking at the target or not. Hence, to some extent, *vfoaAcc* allows us to check if the calibration generalizes well to other points in the gaze space and detect overfitting.

Unsupervised calibration

On UBImpressed, only the **Cst** model improves the baseline VFOA accuracy (*vfoaAcc* of 0.82 instead of 0.77). Although more complex models reduce angular error, they do not necessarily

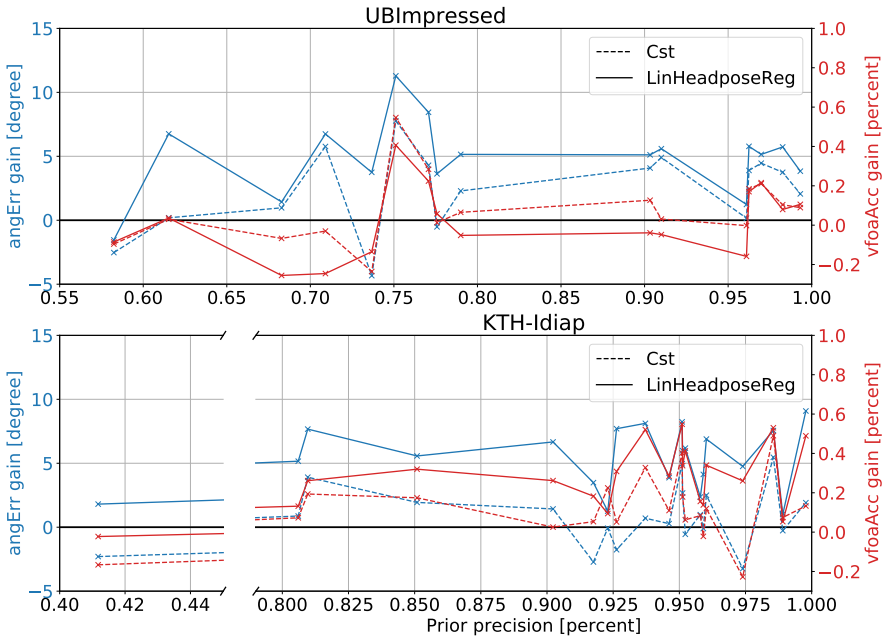


Figure 5.8 – Performance gain after calibration over the weak VFOA labeling precision for each subject using the **Cst** (dashed line) and **LinHeadGazeReg** (plain line) models.

improve *vfoaAcc*, as discussed above. This was expected: as we deal with one target in the 3D space, a translation is sufficient to correct the gaze, and the regularisation is not sufficient to handle the ill-defined problem for more complex models due to label noise, thus leading to the overfitting revealed by the drop in VFOA accuracy.

For the KTH-Idiap dataset, the unsupervised calibration consistently improves the results. The **LinGazeReg** shows its usefulness when dealing with more targets spread in front of the subject. However, adding head pose in the model has less impact than in the supervised case, probably because it is more sensible to the calibration set quality.

Overall, while there is an expected drop in performances compared to the supervised approach, unsupervised calibration improves baseline results given that the right correction model is selected.

Impact of the weak labeling accuracy

Previous results validate our approach globally, but we noticed that the weak VFOA labeling precision varies among subjects, which could imply that the proposed approach does not work for all subjects, potentially failing when this precision is too low. Fig. 5.8 displays the gain (or loss) for both metrics over the weak VFOA labeling precision (i.e. ratio of correctly labeled points in the calibration set), using either the **Cst** or **LinHeadGazeReg** models. We can first notice that the precision of the weak labeling is good, being more than 0.6 for all but two subjects in two different datasets. Secondly, the **LinHeadGazeReg** model significantly

improves the results for most subjects in the KTH-Idiap dataset but has a mixed effect on the UBImpressed ones, where it reduces the gaze error at the cost of the VFOA accuracy, as already known from the global results. Finally, our intuition was that a lower weak VFOA labeling precision would lead to lower performances but there is almost no correlation between the weak labeling precision and performance gain, demonstrating the robustness of the proposed method.

Accounting for gaze cues in weak labeling

Looking at some videos, we noticed that the proposed physical constraints between head pose and visual targets might be too loose and not be sufficient. Indeed, for example, people sometimes avert their gaze from the speaker without moving much the head, and the physical constraints will not filter those frames which will still be labeled as "looking to the target". Reversely, they may filter out valuable samples involving large head poses (and thus large gazes in the head coordinate system) when people look at targets on their sides, which may explain the smaller contributions of head poses on gaze correction in the unsupervised case compared to the supervised case.

One way to handle this consists of applying the constraints directly on the gaze estimates from the baseline. Indeed, although not calibrated, they are not completely wrong and the error is usually less than 20 degrees. Running the same experiments as before with this constraint has been shown to well improve the results on UBImpressed (with for instance $angErr = 7.08°$ and $vfoaAcc = 0.85$ for the **Cst** model, or $angErr = 4.12°$ and $vfoaAcc = 0.79$ for the **LinHeadGazeReg** model) but did not make much difference on the KTH-Idiap dataset.

5.4.3 Results using manipulation prior

We applied the proposed method to the ManiGaze dataset, estimating the calibration parameters on a given session (MT, ET, or OM) and evaluating them on another one (MT or ET). Tab. 5.3 presents the obtained results. Note that computing *vfoaAcc* for the ET session makes little sense, as only one visual target is present at a time, so we estimated it only for the MT session. Moreover, aversions were not annotated, so we used a very high threshold ($\tau = 90°$), so that aversion is never predicted.

We evaluated supervised calibration in two fashion: first calibrating and evaluating the gaze estimation on the same session (intra-session) and then calibrating the gaze estimation on one session and evaluating it on the other (cross-session). The former setup shows the best achievable performances, while the latter represents a more reasonable experiment for testing generalization.

Table 5.3 – ManiGaze dataset. Mean angular errors (in degrees) and VFOA accuracy across subjects. Here we evaluate the raw gaze estimation without calibration (**Baseline**), supervised calibration using the same session VFOA annotations (intra-session), supervised calibration using VFOA annotations of the other session (cross-session), and unsupervised calibration using the manipulation prior (**Prior**) applied on the OM session.

	ManiGaze-MT		ManiGaze-ET
Method	*angErr*	*vfoaAcc*	*angErr*
Baseline	18.82	0.21	16.27
Supervised intra-session			
Cst	6.28	0.64	8.26
LinGaze	5.47	0.67	7.17
LinGazeReg	5.66	0.67	7.24
LinHeadGazeReg	4.38	0.77	6.27
Supervised cross-session			
Cst	7.79	0.56	10.06
LinGaze	8.16	0.52	14.92
LinGazeReg	7.57	0.59	10.12
LinHeadGazeReg	7.21	0.62	9.87
Unsupervised (Prior), calibrated on OM session			
Cst	8.86	0.41	11.91
LinGaze	8.25	0.42	17.52
LinGazeReg	8.10	0.45	12.80
LinHeadGazeReg	8.62	0.40	15.06

Baseline results

As before, the error is quite high (16.27° and 18.82°). It can be in great part explained by the challenging experimental setup (see Fig. 4.3). The camera captures subjects from a very unusual point of view, which differs from data used to train the gaze estimation model. Also, the VFOA accuracy is very low, which can be explained by the challenging task: there are many more targets (14) in smaller visual space compared to the conversation settings, and they are close to each other (less than 20cm, gaze difference of around 8°).

Supervised calibration results

Supervised calibration consistently improves gaze and VFOA estimation by a large margin. As expected, results are better for the intra-session calibration. In the cross-session case, calibrating on ET and applying it to MT works better than the reverse. This was expected, as the visual location of targets in the ET session span a space that somehow comprises the one in the MT session. The MT-to-ET calibration highlights the difficulty to generalize calibration parameters to other parts of the space, but it still achieves much better results than the baseline

Table 5.4 – Mean angular errors (in degrees) and mean VFOA accuracy across subjects using adaptive calibration (*Online*) on the conversation datasets. Here we compare three maximal calibration set sizes (100, 1000, ∞), as well as two calibration models (**Cst** and **LinHeadGazeReg**).

			UBImpressed		KTH-Idiap	
Method	$N_{\mathbb{C}}^{min}$	$N_{\mathbb{C}}^{max}$	*angErr*	*vfoaAcc*	*angErr*	*vfoaAcc*
Baseline	-	-	9.19	0.77	14.54	0.40
Cst	10	100	6.60	0.84	10.36	0.67
Cst	10	1000	6.18	0.89	11.74	0.62
Cst	10	∞	6.17	0.90	12.86	0.57
LinHeadGazeReg	10	100	6.21	0.84	9.42	0.71
LinHeadGazeReg	10	1000	4.68	0.82	9.29	0.72
LinHeadGazeReg	10	∞	4.25	0.78	10.19	0.67

Unsupervised calibration results

The unsupervised cross-session calibration does not reach supervised performances but beats the baseline by a large margin. For the MT session, the results are not far from the supervised cross-session case in terms of angular error, which is promising. However, the gain in VFOA accuracy is half as good compared to the supervised case. This is due to the proximity of the visual targets: a small performance loss in angular error has a high impact on VFOA accuracy. On ET, while the **Cst** model is close to the supervised cross-session case (11.91° compared to 10.06°), more complex models seem to be more sensitive to the quality of the calibration set, resulting in much lower performances.

Generally, the rather small improvement of the **LinGazeReg** model compared to the **Cst** one shows that the calibration consists mainly in correcting a bias not depending on the gaze direction. Adding head pose has a strong effect when the calibration is based on ground truth gaze points, but it seems to be more sensitive to the quality of calibration set, as shown in the unsupervised case where it is not better than the **Cst** model.

5.5 Online calibration experiments

In this section, we evaluate the *Online* calibration method proposed in Sec. 5.2.8 on the UBImpressed and KTH-Idiap datasets. We let aside the ManiGaze dataset, as the sessions presenting natural interactions (OM and ST) do not provide ground truth gaze and thus do not allow a proper evaluation.

5.5.1 Experimental protocol

We use the same metrics and parameters as for the *Offline* experiments. We compared the **Cst** and **LinHeadGazeReg** calibration models with three maximal calibration set sizes (100, 1000,

and ∞) to study the impact of taking past information into account.

In practice, reference points are accumulated from the start of the video and the gaze estimation is calibrated when there are at least $N_{\mathbb{C}}^{min}$ points. When the calibration set is bigger than $N_{\mathbb{C}}^{max}$, a sample is discarded and replaced at random. Although not reported here, we tested other updating strategies, like discarding the oldest sample in the calibration set or setting a time constraint rather than a number of samples, but those strategies were slightly worse. Note that to allow comparison with the *Offline* results, the *Online* method is evaluated on the same test set as previously.

5.5.2 Results

Tab. 5.4 presents the obtained results. Overall performances are equivalent to or better than the *Offline* experiments. While the *Online* calibration is getting less information than in the *Offline* case, it allows adapting the gaze correction to the local context, which seems effective.

We observe the same distinction in model performance than before: the **Cst** works better on the UBImpressed dataset and the **LinHeadGazeReg** performs better on the KTH-Idiap one.

Regarding the calibration set size, using 1000 samples, i.e. at least 30 seconds, works better given that the model is appropriate for the dataset (**Cst** model for UBImpressed and **LinHeadGazeReg** one for KTH-Idiap) Indeed, there is a trade-off between getting enough context to mitigate the labeling noise, but still being able to adapt to local errors.

Also, despite its simplicity, the **Cst** model achieves better performances than in the *Offline* case (VFOA accuracy of 0.89 and 0.62 for $N_{\mathbb{C}}^{max} = 1000$ compared to 0.82 and 0.56). It can be explained by the nature of the task, as in the *Online* case, we search for calibration parameters that fit the local conditions around the evaluated point and not the whole interaction as in the *Offline* experiments. It tends to show that the gaze estimation error evolves over time, and the proposed online calibration framework allows taking it into account, which compensates for the simplicity of the **Cst** model.

5.6 Discussion

In the experiments above, we showed that a pre-trained gaze estimation model can be calibrated in an unsupervised fashion using context information. Also, taking head pose into account has a positive effect during supervised calibration, but not in unsupervised calibration, showing that it makes the calibration more sensitive to weak VFOA labeling errors. Moreover, the constant model reaches good performances despite its simplicity, performing almost as good as more complex models when the calibration parameters are adapted over time using local information. Overall, we showed the importance to adapt the calibration model to the setup, as the best model depends on the diversity of the available calibration samples.

Limitations

One limitation of this work is the simplicity of the proposed VFOA estimation method for collecting calibration data. Hard thresholds were used, like for physical constraints or manipulation priors time-window, and independent frame decisions were taken. A potential improvement could be to use some weighting and/or sampling strategies, by exploiting the probability of looking at a target or by using a subset of the frames coming from the same glance and/or utterance as they are likely to be similar. For instance, when a meeting participant is taking a long speaking turn, only a subset of frames of that turn (for other participants) could be used for calibration. Furthermore, learning better VFOA prior models would be interesting, for instance by using more cues (e.g. gaze/head pose of a speaker) or incorporating temporal filtering. Also, note that in this work, VFOA inference was only based on gaze and targets' positions, although it could itself benefit from using the VFOA priors. This was done on purpose here, as we focused on measuring the geometric impact of calibration on VFOA inference, not obtaining the best accuracy.

Another limitation of the proposed approach is the potentially low diversity of the provided calibration points. Indeed, depending on the application, only a few different targets might be available, which impacts the complexity of the model that can be used, as shown in the UBImpressed setting. Note however that this might not be a problem, since we have shown that a simple translation model which adapts to the local context can do well. As time passes, targets will eventually move and/or the user will change his pose increasing diversity, but it is not guaranteed. In this regard, a potential way to increase the calibration sample diversity could be to track object or sound sources and using some sort of bottom-up scene saliency or other forms of priors, but more experiments would be needed to validate such an approach.

Another concern is the need for contextual information in addition to the target 3D positions needed for infering the VFOA: speaking status, or user's actions (grasping, releasing objects). These are usually available in many research studies and applications, especially in HHI (Ba and Odobez, 2011; Müller et al., 2018; Otsuka et al., 2018; Bai et al., 2019) and HRI (Sheikhi and Odobez, 2015), where detecting people and object positions, speakers, or actions performed by people are also desired for other goals beyond gaze estimation (e.g. conversation analysis). However, in other applications (e.g. internet video analysis), the extraction of 3D positions or contextual information and the accounting for the potential uncertainties would require further investigation.

Future work

Future work could consist in investigating temporal gaze and/or VFOA models, which could provide more reliable and smoother estimates to improve both the collection of data points and the VFOA inference. Also, meta-learning gained popularity recently and fine-tuning a learned neural network model was shown more effective than using an additional regressor (Yu et al., 2019). It would be interesting to see if these approaches could benefit from the pro-

posed calibration samples selection. Regarding experiments on object manipulation, object detection and gesture recognition remain to be automated. It will add some inaccuracy in the calibration samples, but our experiments showed the robustness of the proposed approach which relied on prior timing statistics which are different than what was observed on our data (see discussion in Sec. 5.3.2). Finally, it would interesting to see how different prior could be combined, e.g. in setups where people discuss while performing another task. However, different tasks require a different level of visual attention (e.g. people can discuss without looking at each other, but not grasp an object without looking at it), so prior combination may not be straightforward.

5.7 Conclusion

In this chapter, we proposed an unsupervised user-specific calibration method for remote gaze estimators, relying on context-based weak VFOA labeling. We evaluated the empirical validity of this weak VFOA estimation and proposed a robust calibration parameter estimation to exploit it to calibrate a pre-trained remote gaze estimator. We evaluated the proposed calibration method on two popular HHI and HRI settings, namely conversation and manipulation, and showed its effectiveness at improving gaze and VFOA estimations, both in supervised and unsupervised setups. Finally, we proposed a framework to apply the unsupervised calibration in an online fashion, achieving similar or better results than in the offline experiments, depending on the dataset.

6 Visual Focus of Attention Estimation with an Arbitrary Number of Targets

The previous chapter focused on collecting calibration samples in an unsupervised way to improve gaze estimation through calibration. However, our end goal is often to infer the visual focus of attention (VFOA) of people (see Fig. 6.1). Although using a gaze-based geometrical model allows us to estimate the VFOA from basic information, we have shown in Sec. 2.2.1 that taking contextual information into account can be useful to improve the VFOA estimation accuracy. However, existing approaches mainly focus on static setups with a pre-defined number of people, although, in several applications, it is often necessary to handle an arbitrary and dynamic number of targets and/or to apply a model to a different setup, because of a lack of VFOA annotation in the setup of interest for instance.

In this chapter, we address the estimation of the VFOA in 3D scenes with an arbitrary number of targets and propose cross-datasets experiments to study our model and its generalization properties. We rely on a feature normalization method that represents all the subject and context-related features in a fixed number of 2D maps, allowing us to tackle the above challenges. The work in this chapter was published in the proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (GAZE2021 workshop) (Siegfried and Odobez, 2021b).

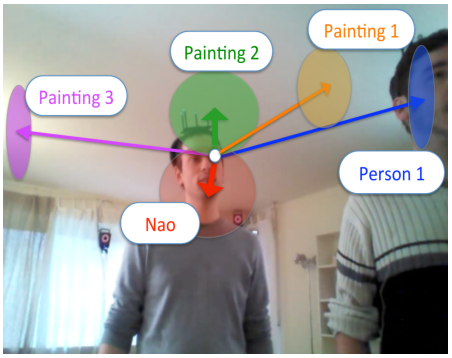


Figure 6.1 – VFOA estimation illustration (Vernissage corpus (Jayagopi et al., 2013)). Given a subject, estimate his VFOA implies a categorical decision among potential visual targets.

This chapter is organized as follows. First, we introduce our motivations and contributions (Sec. 6.1). Then, we recall some related works that share elements with our method (Sec. 6.2). Next, we present the proposed method, i.e. our feature normalization process and our VFOA network (Sec. 6.3). Finally, we provide our ablation study and cross-datasets experiments results (Sec. 6.4), before concluding this chapter (Sec. 6.5).

6.1 Introduction

As discussed in Sec. 2.2.1, VFOA estimation in 3D scenes is a challenging task as it relies on many features like the head pose and gaze of the person, and also scene information to know the positions of the person and VFOA targets or other potential contextual information (e.g. who is speaking), which can be difficult to estimate accurately depending on the setup. In particular, despite recent progress, the accuracy of gaze estimation is often limited when recording naturally acting people with remote sensors due to low image resolution and high variability in appearance (pose, eye), in particular for large head poses.

To overcome these difficulties, principled methods have been proposed to integrate different cues in the VFOA inference to make it more robust to noise in feature estimation. However, these methods usually rely on predefined target sets leading to fixed input/output sizes and inference structures, which constraints their usage to a single setup with a fixed number of targets. Thus, these models are usually trained for a specific setup and should be retrained to be applied to another one, which limits their usability as VFOA annotations can be difficult or expensive to gather. Having a model that can generalize to new setups and situations (geometry, number of people, and salient objects) would be useful.

To work towards this goal, we propose to reformulate the problem. The main idea is to reformat the input features into several 2D maps associated with the subject's field of view, allowing to encode all inputs and contextual cues (4 maps for head pose, gaze direction, directions of speaking sources, and person speaking status) within a single referential, as well as providing as input a visual saliency 2D map encoding an arbitrary number of candidate VFOA targets. This leads to a fixed number of maps that can be stacked as a tensor and processed as an image to produce a 2D map of VFOA direction probabilities, whose maximum is used in the final VFOA classification (details in Sec. 6.3). This has four advantages:

- it normalizes the inputs, removing spatial and feature dependencies to the camera view points and more generally the setup;
- it allows to consider an arbitrary number of targets and to encode all contextual cues in the same referential;
- it makes the input more suited to be processed by a Convolutional Neural Network (CNN), as images naturally encode proximity in space and channels;
- it makes data augmentation easier (an important step for generalizing to other situations), as targets can easily be be added and removed or context be modified during training.

Our contribution is thus a novel method to estimate the VFOA of a subject given an arbitrary number of targets and contextual cues, which combines the above advantages and allows the application of a learned model to different setups. We evaluated this method and the impact of the different input features on two publicly available datasets, namely UBImpressed and KTH-Idiap (see Chapter 4), including convincing cross-datasets/setup experiments.

6.2 Related works

As shown in Sec. 2.2.1, although VFOA models were improved using new methods and additional contextual features with the aim of modeling conversations, little work was done to improve their flexibility. Indeed, interaction models have the advantage of representing all participants' behaviours together to take into account dependencies between them and they were shown to be effective at learning models for VFOA estimation in setups for which annotations are available. However, such models are often trained on a single specific setup, in which a defined number of static visual targets are usually assumed and 3D scene representation is sometimes absent, leading the model to learn setup specific feature clustering rather than geometrical reasoning, so it can not generalize to unseen setups with a different number of people or a different geometry and can not handle people that join or leave the conversation.

Marin-Jimenez et al. (2019) addressed this issue by learning two-person VFOA predictors (do they look at each other or not), but their work focused on 2D VFOA estimation. Here we propose a new input data format that is more naturally processed by CNNs and allows the same trained model to handle an arbitrary number of targets as well as different setups. In this regard, unlike interaction models, the proposed method estimates the VFOA of each person individually (rather than jointly) using 2D maps while still encoding the conversation features from all subject as well as scene information.

Representing the gaze direction as an image was already studied in the context of gaze refinement in a screen-based setting, Park et al. (2020) proposed a method to refine the gaze estimation using the screen image content, i.e. leveraging saliency. They first estimated the 3D gaze from the eye images and then projected it on the screen, where the visual stimulus is displayed, to get the point of gaze (PoG). The PoG is then represented as a black image with the same resolution as the screen and a 2D Gaussian centered on the PoG estimate. This gaze image is then stacked with the actual displayed image on the screen and processed by a neural network which refines the PoG estimate. In this work, we extend this idea by representing all subject's and scene's features as images (i.e. 2D maps) and by considering the whole 3D field of view of the subject. In contrast to Park et al. (2020), we do not have access to what the subject sees or the gaze ground truth, so we must work in a virtual field of view that we fill with the scene information. Also, we must handle aversion cases without knowing where the subject is effectively looking.

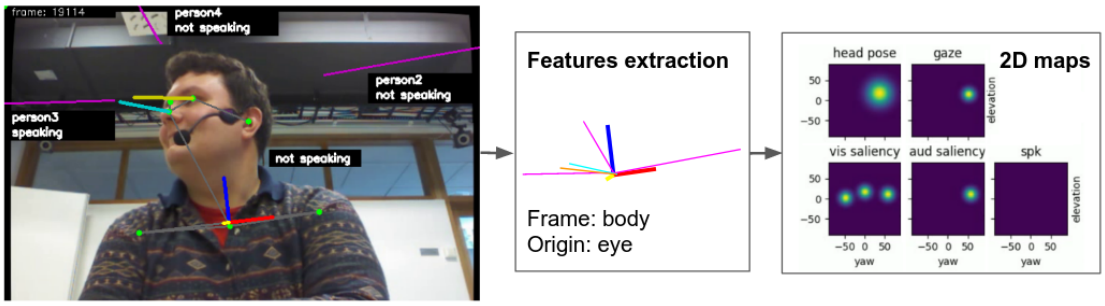


Figure 6.2 – Features extraction. Head pose (cyan) and gaze direction (yellow) are estimated from the RGB-D video while target directions (purple) and people speaking status are recovered from scene monitoring. Then, all these features are expressed in the body frame (red-blue-yellow) and represented as yaw and pitch angles. Finally, they are encoded into a fixed number (i.e. five) 2D maps.

6.3 Method

The proposed method is presented in Fig. 6.2 and 6.3. It can be divided into three main parts:

1. the extraction of the required features (head pose, gaze direction, target directions, and speaking status) from the input video and scene information;
2. the translation of these features into a fixed number of 2D maps;
3. the estimation of the VFOA from the 2D maps.

They are described in more detail below.

6.3.1 Features extraction

We consider the case where the scene is monitored, i.e. that the 3D positions and the speaking status (binary) of each person involved in the interaction are available. Our goal is to express all features (head pose, gaze, and target directions) as yaw and pitch angles in a frame associated with the body orientation. In this way, the representation can potentially exploit coordination patterns between the body, the head pose, and the gaze, and allows to normalize the data independently of the camera position. Target directions are easily computed as the eye-to-target (most forward of both eyes) vector and then translated to angles, while head pose and gaze direction are extracted from RGB-D video using the method presented in Chapter 3.

Body frame estimation

The 3D positions of the subject's joints are extracted by combining the body 2D keypoints from OpenPose Cao et al. (2019) and the depth provided by RGB-D cameras. To catch the orientation of the subject, the body frame is built using the vector going from the right to the left shoulder and the vertical axis of the camera frame. The latter axis was selected since available videos

only provide upper body views of the people, so the hip keypoints are not available and the estimation of the vertical axis of the body is then difficult. We consider it a minor drawback, as most of the body rotations are done in the yaw direction in our conversation scenarios. Note that this method is sensitive to occlusions, but the cases where the hands of the subject hide his shoulders are rare in the considered datasets.

6.3.2 2D feature maps

The proposed method takes five types of features as input: head pose angles, gaze direction angles, speaking status of the subject, the directions of potential VFOA targets (i.e. visual saliency), and the directions associated with speaking targets (i.e. audio saliency). To bring all these features in the same space and allow an arbitrary number of targets to be represented, we set in place two main elements. First, as explained earlier, all directions, whether from the scene (target directions) or from the human subject (i.e. head pose, gaze) have been expressed in the same reference frame (the subject's body frame). Secondly, each input feature type is encoded as a 2D map with a resolution of 180x180 pixels, in which each pixel represents an angle of 1 degree in both yaw and pitch axes. As a result, the 2D maps represent a unified view of the gazing activity and scene information in front of the person.

To generate these maps (see one example of such maps in Fig. 6.2), we proceed as follows. If $\{p_i, i = 1,\dots,N\}$ denotes the set of directions to be encoded in the map D, we simply place 2D Gaussians of covariance Σ^m at each direction p_i to provide information regarding this direction while taking into account the estimation noise or the size of targets. More formally:

$$D(p) \propto \max_{i=1,\dots,N} \mathcal{N}(p - p_i; \Sigma^m) \tag{6.1}$$

Using this process, the head pose map is created by using as p_i the (single) head pose direction, the gaze map is built using the gaze direction, the video saliency map is produced using as p_i the set of potential VFOA target directions (people in the conversation in our case), and the audio saliency map using the directions of people who are talking. The map associated with the speaking status is the exception. As it is not associated with any direction, we chose to fill it with its value, i.e. it is full of ones when the subject speaks and full of zeros otherwise.

Finally, all the five above maps are gathered into a single tensor, so that it can be processed as a 5-channel input by a CNN.

6.3.3 VFOA network and classification

Architecture

Our network's architecture is similar to the one in Park et al. (2020) and is a kind of hourglass network with the following elements (see Fig. 6.3).

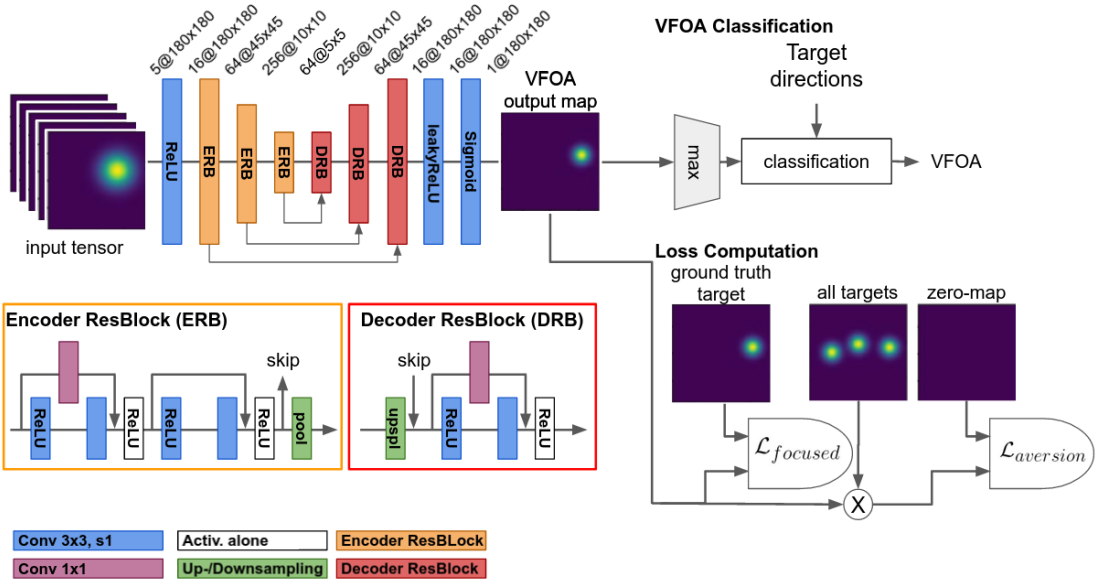


Figure 6.3 – VFOA network architecture and loss computation. The 5D feature maps tensor is processed by a king of hourglass network, i.e. a CNN with residual blocks and skip connections, to generate a VFOA probability map. The network mainly consists of three down- and upsampling layers, which shrinks the image spatial resolution down to 5x5. The maximum of the outptut VFOA probability map is used to derive the actual VFOA (top right). Network layer types are represented by colors (bottom left). To compute the loss (bottom right), we distinguish two cases. When the target looks at a target (focused), we compare the output map with a map containing only the position of the ground truth target. Otherwise (aversion), we compare it with an empty map but use the saliency map as weight to penalize only the position near the targets.

- An initial 3x3 convolutional layer.
- Three downsampling layers consisting of two residual blocks, with the first one performing the channel number update. After the two residual blocks, a copy of the feature vector is stored for later usage (skip connection) and it is down-sampled by a max-pooling layer.
- Three upsampling layers consisting of one residual block each. Before the residual block, the input feature vector is upsampled using bilinear interpolation and concatenated to the stored element of the same size (skip connection).
- Two final 3x3 convolutional layers ending up with sigmoid activation.

Every convolutional layer is followed by an instance normalization layer and all activations are Rectified Linear Units (ReLU), except the last one which is a Sigmoid. Note that while the feature map resolution decreases relatively fast, experiments with more than 3 layers did not improve the results.

VFOA classification

The predicted VFOA map is transformed into yaw and pitch angles by taking the angle coordinates of the map's maximal value. Then, VFOA classification is performed using the angular distance (i.e. cosine distance) to each target: it is an aversion if the minimal angular distance to all targets is above a given threshold κ_τ and the nearest target otherwise. In other words, we use the geometrical model presented in the last chapter (Sec. 5.2.4) and use the output of the VFOA network as if it was a gaze estimate.

Training

The VFOA map network is trained frame by frame using the binary cross-entropy (BCE) loss distinguishing two cases, as shown in the bottom right of Fig. 6.3. When the subject is looking at another person (focused), we want the output VFOA map to fit the target position, and the loss is the BCE between the output map M_{out} and a target map M_{tar} featuring only the ground truth VFOA target. In case of aversion, we want the output VFOA map values to be low where there are targets, so the loss is the BCE between a zero map 0_{map} and the output map M_{out} masked by the visual saliency map M_{vsal} to only keep VFOA outputs close from targets (and thus remove outputs far from targets which can be considered as valid outputs). So, we have:

$$\mathscr{L}_{VFOAmap} = F_{gt} \cdot \mathscr{L}_{focused} + (1 - F_{gt}) \cdot \mathscr{L}_{aversion}, \tag{6.2}$$

$$\mathscr{L}_{focused} = \mathbf{BCE}(M_{out}, M_{tar}), \tag{6.3}$$

$$\mathscr{L}_{aversion} = \mathbf{BCE}(\frac{M_{vsal}}{max(M_{vsal})} \cdot M_{out}, 0_{map}), \tag{6.4}$$

where F_{gt} is a binary indicator equal to 0 if the ground truth VFOA is "aversion" and 1 otherwise.

Data augmentation

To increase the generalization abilities of the trained model, we use several data augmentation strategies:

- target removal: a random number of targets (between 0 and the total number of targets minus 1) are removed from visual and audio saliency maps. If a removed target corresponds to the VFOA ground truth, the label is turned to "aversion";
- target addition: a random number of fake targets (between 0 and 2) are added to the visual saliency map. Their locations are sampled using the mean of the real target positions and a variance scaled by 1.5 in the yaw direction. Each fake target can also appear on the audio saliency map, as if it was speaking, with a probability of 0.5;
- global noise: random white noise ($\sigma = 5°$) is added to all angles (head pose, gaze, and target positions).

- feature noise: random white noise ($\sigma = 2°$) is added to each angle separately (head pose, gaze, and target positions).

When data augmentation is used, these four strategies are applied to each samples, meaning that the number of training samples does not increase.

6.4 Experiments

6.4.1 Baseline model

We looked for a baseline that is comparable to our method in terms of input features and usability, i.e. a method allowing to predict the VFOA for an arbitrary number of targets without retraining. However, to the best of our knowledge, previous state-of-the-art methods focused on predicting VFOA in scenarios involving a fixed number of people in the conversation and within a fixed setting (Ba and Odobez, 2011; Otsuka et al., 2018; Bai et al., 2019). No cross-dataset experiments were performed. As discussed in Sec. 2.2.1, applying these methods to a new setting with a different geometry or number of participants is not trivial without retraining the model from scratch, since these models do not introduce explicit spatial relationships, and they do not handle moving people.

Thus, we propose to use as baseline a strong multimodal statistical binary classifier predicting the probability that a subject looks to a target given the target direction, the subject's gaze and head pose, and the speaking status of all persons in the scene. Doing so allows this classifier to be applied to any number of potentially moving targets without retraining or fine-tuning and uses the same features as the proposed method.

More formally, let us denote F_j the subject's focus status towards target j (1 when being *focused* on target, and 0 otherwise) $\mathbf{g}$ his/her gaze direction, $\mathbf{h}$ his/her head pose direction, $\mathbf{t}_j$ the direction of target j, and s the speaking status of the scene, defined as the combination of three speaking status $s = (s_{subject}, s_{target}, s_{other})$ and can thus take 8 values. For example, it can take values such has $(0,0,0)$ (nobody speaks) or $(0,1,1)$ i.e. the subject does not speak, but the target and at least another person speak. Note that $\mathbf{g}$, $\mathbf{h}$ and $\mathbf{t}_j$ are all expressed as 2D angles. With this notation, we define the probability of the subject' focus status as follows:

$$p(F_j|\mathbf{Z}_j, s) \propto p(\mathbf{Z}_j|F_j)p(F_j|s), \tag{6.5}$$

$$\text{with } \mathbf{Z}_j = [\mathbf{g}-\mathbf{t}_j, \mathbf{h}-\mathbf{t}_j]^T \tag{6.6}$$

where we made the assumption that the gaze and head pose distances to target $\mathbf{g}-\mathbf{t}_j$ and $\mathbf{h}-\mathbf{t}_j$ do not depend on the speaking status. We then further define the likelihood $p(\mathbf{Z}_j|F_j)$ as a multivariate Gaussian for each possible value of F_j. It can formally be written as:

$$p(\mathbf{Z}_j|F_j) = F_j \cdot \mathcal{N}(\mathbf{Z}_j; 0, \Sigma^{foc}) + (1-F_j) \cdot \mathcal{N}(\mathbf{Z}_j; 0, \Sigma^{unf}). \tag{6.7}$$

Regarding the prior $p(F_j|s)$, it is defined as a categorical distribution over the eight speaking

status.

Finally, at inference, to make a decision we first find the target $\hat{j}$ for which the likelihood $p(F_{\hat{j}}|\mathbf{Z}_{\hat{j}}, s)$ of looking at this target is maximal. Then, if

$$p(F_{\hat{j}} = foc|\mathbf{Z}_{\hat{j}}, s) > p(F_{\hat{j}} = unfoc|\mathbf{Z}_{\hat{j}}, s),$$

$\hat{j}$ is set as the subject's VFOA, otherwise it is defined as being *aversion*.

In our experiments, the parameters Σ^{foc}, Σ^{unf}, and the prior distribution $p(F_j|s)$ are learned from a training set.

6.4.2 Experimental Protocol

Performance measure

To compare methods, we report the mean and standard deviation of VFOA classification accuracy of the subjects, where the accuracy per subjects is computed as:

$$vfoaAcc = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\hat{f}_t = f_t}, \tag{6.8}$$

where $\hat{f}_t$ is the estimated VFOA and f_t the ground truth. Moreover, we report the mean of VFOA classification macro F1-score of the subjects, to ensure that the model does not exploit classes' unbalance to reach a good accuracy.

Experimental protocol

Regarding the protocol, our method and the baseline are first evaluated on both datasets separately with a leave-one-out protocol, reporting the average of the mean VFOA accuracy computed on each subject. For *cross-datasets* experiments, a single model is trained on one of the datasets and evaluated on the second dataset without any adaptation, and we also compute the mean VFOA accuracy for each subject in the test dataset and report their average. Finally, our method is evaluated by training and testing a model on both datasets together (*all-datasets* experiments), for which we use a 4-fold cross-validation protocol, with 1 KTH-Idiap, 1 UBImpressed 'Interviews', and 1 UBImpressed 'Desk' sessions per fold. We compute the mean VFOA accuracy for each subject and report their average by dataset to allow comparison with other experiments.

Parameters

We fixed the VFOA classification threshold κ_τ (see Sec. 6.3.3) to 10°, which corresponds to a tolerance of 35cm at 2 meters. In addition, in our experiments, to produce the feature maps,

Table 6.1 – VFOA estimation accuracy mean, standard deviation, and macro F1-score across subjects. (Abbreviations: 'h' stands for head pose, 'g' for gaze direction, 'vsal' for visual saliency, 'asal' for audio saliency, and 'spk' for subject's speaking status).

Method	**UBImpressed**		**KTH-Idiap**	
	vfoaAcc	$F1$	*vfoaAcc*	$F1$
a) Overall results				
baseline	0.84 ± 0.13	0.80	0.80 ± 0.11	0.74
VFOAmap Net	0.85 ± 0.13	0.82	0.81 ± 0.15	0.75
VFOAmap Net + *dataAug*	0.82 ± 0.12	0.78	0.82 ± 0.15	0.74
b) Input ablation study				
headGaze (h-g)	0.80 ± 0.15	0.78	0.60 ± 0.17	0.56
onlyScene (vsal-asal-spk)	0.60 ± 0.14	0.37	0.63 ± 0.15	0.51
noGaze (h-vsal-asal-spk)	0.67 ± 0.12	0.55	0.73 ± 0.14	0.59
noHead (g-vsal-asal-spk)	0.83 ± 0.12	0.79	0.82 ± 0.15	0.74
noAudio (h-g-vsal)	0.88 ± 0.10	0.80	0.78 ± 0.14	0.70
c) Cross-datasets evaluation				
baseline	0.71 ± 0.11	0.58	0.62 ± 0.15	0.56
VFOAmap Net	0.74 ± 0.14	0.65	0.70 ± 0.15	0.61
VFOAmap Net + *dataAug*	0.85 ± 0.12	0.82	0.79 ± 0.13	0.71
c) All-datasets evaluation				
baseline	0.80 ± 0.10	0.77	0.82 ± 0.12	0.75
VFOAmap Net	0.85 ± 0.09	0.85	0.85 ± 0.17	0.75
VFOAmap Net + *dataAug*	0.87 ± 0.10	0.83	0.85 ± 0.17	0.77

we used an isotropic Gaussian kernel Σ^m with a standard deviation of 10° for all maps, except for the head pose map where we used 20° which better encompasses the range from the head pose where the gaze can be.

6.4.3 Results

Intra-datasets evaluation

Intra-dataset results are reported in Tab. 6.1a. Given the difficulty of the task, we can see that the multimodal baseline already produces very good results on both datasets. Looking at the standard deviation, we can also notice an important differences between subjects, which remains in all experiments. In intra-dataset experiments, the proposed method achieves marginally better than the baseline. Also, the F1-score, which puts more weight on incorrectly classified cases compared to accuracy, shows a similar trend, showing that performances are not only due to more *target* and fewer *aversion* predictions but to good recognition of all classes. The data augmentation does not help here, which is probably due to the amount of available data compared to the relatively low target positions variance in the datasets, and the actual potential overfitting when conducting such intra-dataset experiments.

Input ablation study

In the proposed approach, the input consists of five 2D maps and we are interested in testing the contribution of the different features to the overall performance. To do so, we removed some maps to see how it affects performance. We tested five combinations of inputs, and results are given in Tab. 6.1b.

Results confirm that VFOA estimation benefits both from subject features and scene information, as experiments with only head and gaze (*headGaze*) or scene cues (*onlyScene*, an experiment which allows to check the impact of only prior on results), do not reach the performance of the proposed approach. In addition, while adding head pose improves the performances of *onlyScene* (*noGaze* experiments), it is not as strong as adding gaze alone (see *noHead*) which almost reaches the performance of the proposed approach (*VFOAmap Net*), indicating that in our data, the head pose does not contribute much when the gaze is available. Finally, audio information (subject's speaking status and audio saliency) seems to be more relevant in the multi-target case of the KTH-Idiap dataset (compare *noAudio* to *VFOAmap Net*), which is expected as in such a case, the tendency of looking at the speaker can help to solve ambiguities.

These results show that the proposed method mainly exploits gaze and visual saliency when they are available, but that all inputs contribute to robustness (even if they are redundant, e.g. head pose).

Cross-datasets evaluation

Tab. 6.1c reports the resulting VFOA accuracy when the model is trained on the other dataset (i.e. trained on KTH-Idiap and evaluated on UBImpressed and vice-versa). These clearly demonstrate the generalization capabilities of the proposed method. In particular, while the results achieved only with the available training data are below the intra-dataset results by 10%, using data augmentation (*dataAug*) allows to close the gap and to achieve results as good as if the method was trained on the dataset itself. For comparison, the baseline's accuracy decreases of respectively 13% and 18%, which shows that generalizing from a dataset to another is not trivial.

All-datasets evaluation

In our case, training on both UBImpressed and KTH-Idiap together (see Tab. 6.1d) slightly improves the performances compared to cross-datasets experiments (+2% and +6% with data augmentation). Also, data augmentation only marginally improves the results, probably for the same reasons as in the intra-dataset case. These results, which are the best among our experiments, show the advantage of the proposed method that can successfully train a single model using several datasets with different setups and target numbers.

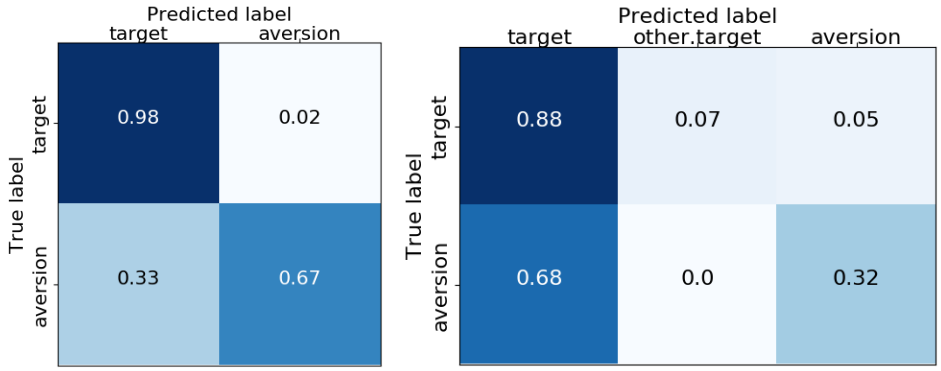


Figure 6.4 – Confusion matrices for UBImpressed (left) and KTH-Idiap (right) datasets after cross-dataset training with data augmentation.

Confusion matrices

Fig. 6.4 shows the confusion matrices for the models trained in a cross-dataset fashion with data augmentation. We did not report the confusion matrices for the intra-dataset experiments as they are very similar to these. In the KTh-Idiap case, computing a confusion matrix is difficult as when the VFOA ground truth is a target, the network can output *aversion* (false negative), the correct target (true positive), or another target. We fixed the latter case by adding an *other target* column in the confusion matrix.

Looking at the resulting matrices, the proposed method has more difficulties to detect aversions, as it achieves a better *target* recall (0.98 for UBImpressed and 0.88 for KTH-Idiap) compared to *aversion* recall (0.67 and 0.32 recall respectively).

Looking at the KTH-Idiap dataset, most of the errors come from the model predicting a *target* instead of *aversion*. This and the very low *aversion* recall can be explained by the class imbalance, as only 17% of VFOA are aversions. Nevertheless, class balancing strategies might not be desired, as this imbalance is a characteristic of multi-party meetings and from an application viewpoint, there is no obvious reason to favor recall over accuracy. The low aversion recall might also be explained by the defined loss, which does not set a precise prediction target in aversion cases. It should be noted that we reported only the confusion matrices for a VFOA threshold κ_τ of 10°, without searching to maximize the recall.

Accuracy versus VFOA classification threshold

In the above experiments, we set the VFOA classification threshold κ_τ to an arbitrary value of 10°. Figure 6.5 shows the impact of this parameter on the accuracy, precision, and recall of the *aversion* class. All these metrics were computed for each subject, and we report their average for each value of κ_τ.

Overall, the accuracy is maximal in a region between 5° and 15°, which is probably due to the

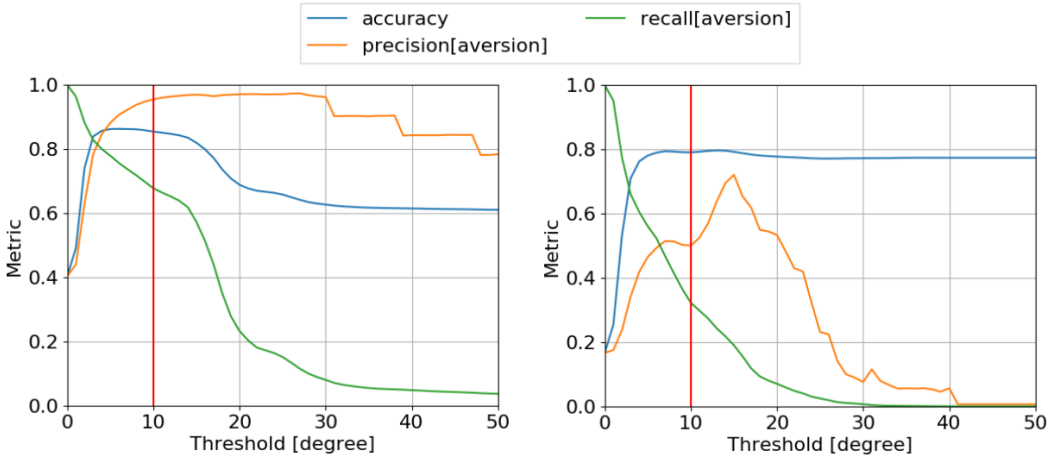


Figure 6.5 – Accuracy, as well as the precision and recall curves of the *aversion* class against the VFOA classification threshold for UBImpressed (left) and KTH-Idiap (right) datasets, after cross-dataset training with data augmentation. The red vertical line indicates the default value of $\kappa_\tau = 10°$ used in our experiments.

choice of the Gaussian kernel's standard deviation. Also, one can see that we could increase *aversion* recall without losing accuracy by choosing a smaller threshold.

Looking at the KTH-Idiap case, the accuracy peak is smaller and increasing the threshold makes the accuracy saturate toward a value of 0.77, which is near to that *target* class ratio in the dataset. This may suggest that the network's good score is particularly due to its ability to chose between the different targets. Still, when the threshold is around 5°, both accuracy and *aversion* recall are above 0.60, showing that the network is somehow able to distinguish aversion from target.

6.5 Conclusion

In this chapter, we proposed a deep learning based method that estimates VFOA estimation from visual and audio features encoded as 2D maps, which provides setup normalization and allows to consider an arbitrary number of targets. The presented experiments show the usefulness of five exploited features, namely head pose, gaze direction, visual saliency (i.e. potential targets direction), audio saliency (i.e. speaking potential targets direction), and subject's speaking status, even if head pose can be omitted in the presence of gaze. Especially, the proposed method was shown successful in cross-datasets experiments, which is a promising step to estimate VFOA in new setups without needing to retrain or fine-tune the model.

The main limitation of the proposed method is the need for the 3D position and speaking status of all participants in the scene, as already discussed in Sec. 5.6. Future work would consist in testing this method on other datasets with even more intra-setup variance in terms

of target position and number. Indeed, in both presented datasets, the number of targets does not change during the interaction, even if cross-dataset results are promising. Also, the proposed network could be enhanced with temporal information, using recurrent layers for example, or by adding other maps like encoding the gaze of the targets. Finally, it would be interesting to see if this approach could be applied to different tasks, like gaze refinement or gaze synthesis.

7 Eye movements recognition from remote sensors in videos

Beyond the sheer instantaneous estimation of the gaze direction, recognizing eye movements can be useful for a range of applications like in mental health assessment, control interfaces, or gaze analysis. Such a task is often addressed by postprocessing gaze traces, as shown in Sec. 2.3. However, remote sensors usually provide low resolution and low sampling rate eye image streams, which makes it difficult to recognize eye movements. Furthermore, fixations and saccades are usually estimated from the gaze signal and separately from blinks.

In this chapter, we address the combined recognition of fixations, saccades, and blinks from videos with a normal sampling rate (30 Hz) and low-resolution eye images (36x60 pixels) using a Convolutional Neural Network (CNN) that takes 9 consecutive eye images as input. The work in this chapter was published in the Proceedings of the 2019 ACM Symposium on Eye Tracking Research and Applications (ETRA) (Siegfried et al., 2019).

This chapter is organized as follows. First, we recall our motivations and present our contributions (Sec. 7.1). Then, we present the proposed method (Sec. 7.2), the proposed evaluation protocol (Sec. 7.3), and the obtained results (Sec. 7.4). Finally, we discuss the limitations of this work (Sec. 7.5).

7.1 Introduction

Previous eye movements recognition approaches mainly relied on common infrared-based sensors like Eyelink 1000, iView X or Tobi TX300, which are rather expensive, often require calibration and can restrain user movements (head pose, headbox size) or be quite invasive (need to wear goggles). These conditions might not be a problem for applications like medical exams or neurological investigation. However, they make difficult the application of eye movements recognition for gaze analytics at large scales in fields like driving assistance, conversational agents or sociological studies, where we want users to act naturally without head-mounted devices, constrained head pose or the need for user-specific calibration.

Computer vision technologies are best suited for such applications, as they adapt to cheaper

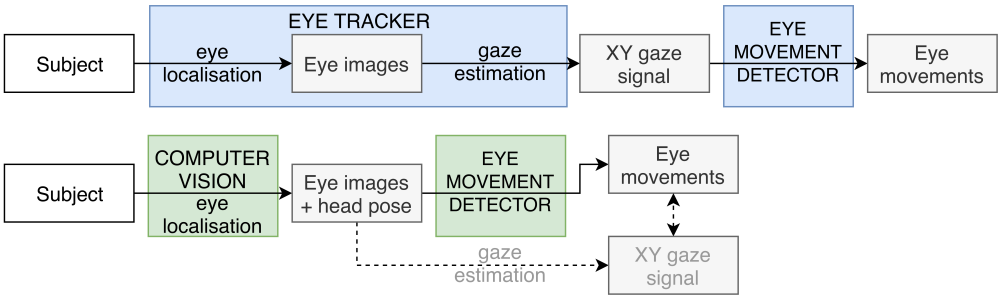


Figure 7.1 – Common workflow when using an eye tracker system and proposed workflow. Usually (top), an eye tracker provides a gaze signal, which is then processed to recognize eye movements. In the proposed approach (bottom), we process the eye images directly, so that eye movements can be recognized without the need for a gaze signal. Note that, as shown, gaze direction can still be estimated from the eye images and used along with the eye movements for further analysis. However, this is not used in this work.

sensors, to a larger head pose diversity and to larger spaces. However, as seen in Sec. 1.2, they have their own drawbacks: eye images have lower resolution, sampling rates are limited by the sensors used, and natural conditions introduce higher variabilities (e.g. head pose, illuminations, and dynamics of eye movements). Although promising works have been achieved in particular thanks to deep learning techniques, the extracted gaze signals remain noisier and less reliable than with dedicated IR eye trackers, making previous methods for eye movement recognition less suitable.

The method we propose in this chapter detects eye movements from the streams of eye images and head pose information, as presented in Fig. 7.1. Processing the raw signal allows to leverage computer vision and machine learning techniques to distinguish nuisance elements like low-quality data, illumination factors, eye shape variations, or bad eye alignments which might be responsible for noisy and unstable gaze outputs, from the information (e.g. pupil motion) useful for the classification of eye movements. We evaluate this method on the KTH-Idiap dataset (see Sec. 4.2). The nature of the recorded signals forces us to focus on macro movements like fixation and saccade, as post-saccadic oscillations are too subtle to be observed in this kind of data. Note that we could theoretically also recognize smooth pursuits, i.e. when a person looks at a moving target, but visual targets do not move a lot in the considered datasets, so the number of smooth pursuits is too small to allow proper training and evaluation. Our approach also allows to directly detect blink in addition to fixation and saccade, which might also be useful for behavior analysis (e.g. to evaluate light comfort in office space, fatigue state while driving, cognitive load, etc.).

7.2 Method

The workflow is presented in Fig. 7.1, and comprises two main steps. The first one performs an accurate head tracking followed by head frontalization and eye image cropping. The

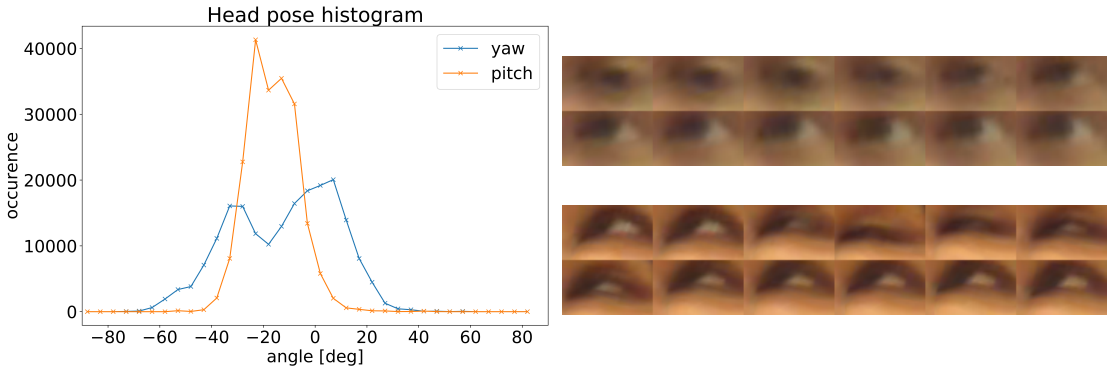


Figure 7.2 – Histogram of estimated head pose (left) and examples of eye image sequences (right).

frontalization reduces the variability of the eye appearance due to the head pose orientation. The second step consists of eye movement recognition.

The eye movement recognition is based on the processing of normalized eye images sequences (color only), which are obtained through the method presented in Chapter 3. Sample eye image sequences are shown in Fig. 7.2b. Since the eye activity that we target can be recognized from eye observations over fixed time windows (Bellet et al., 2019), we adopt a temporal sliding window approach relying on a neural network consisting of convolutional and fully connected layers, rather than using explicit temporal models like recurrent architectures.

Input images

In our model, the label of a frame is estimated by taking as input several images, by processing them first individually to extract relevant and abstract information about the gaze (iris position), concatenate them in a common part to process the sequence of these features and perform the final classification. In the individual part, the feature extraction is the same for all eye images so the weights are shared. In the common part, the sequence of head poses is injected in the network for better results, as eye appearance changes can be due to head pose variations instead of eye movements, like when fixation occur along with head gesture motion.

Architecture

The architecture is presented in Fig. 7.3. In our design, the network takes 9 consecutive frames (from $t-4$ to $t+4$) with dimension 36x60 as input and predicts the label of the frame t. Each eye frame is processed by 4 convolutional layers for feature extraction. Then the extracted features of the 9 frames are stacked and further processed by 2 convolutional layers processing along the temporal dimension. The estimated head pose (rotation angles of yaw, pitch, and roll) is introduced at this stage by concatenating the features from the previous layers with the

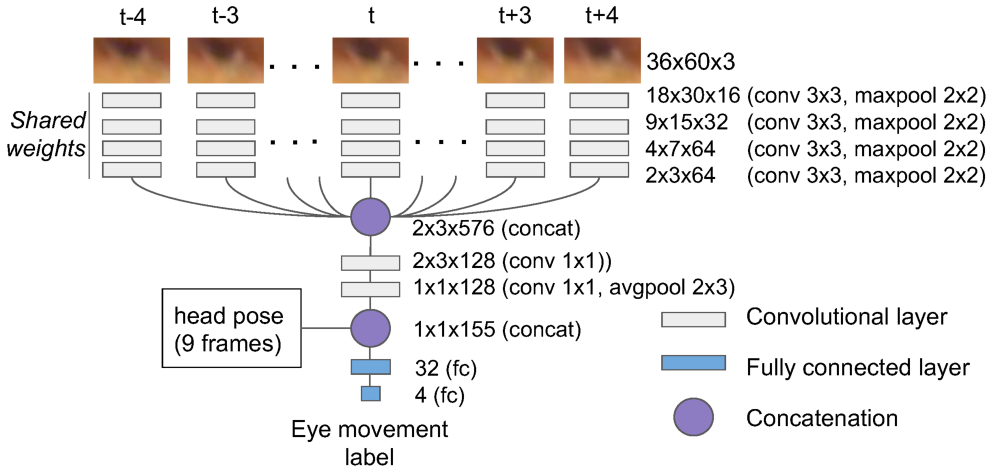


Figure 7.3 – Network architecture of the propsoed eye movement detector. It takes as input 9 consecutive images and outputs the probability for each eye movement type, depending on the task. Eye images are first processed individually, using branches that share parameters. Then, they are concatenated and processed together and the resulting feature vector is processed, together with head pose angles, by two fully connected layers.

rotation angles of the 9 frames. The result is forwarded to 2 fully connected layers for making the final prediction. In this work, eye movement detection is modeled as a classification task and we use a cross-entropy loss for training the network.

This model was developed by investigating and comparing several architectures. Among others, we saw that using past information only decrease the performance and that adding more than 4 frames in the past and future does not improve the results. Our hypothesis is that 4 frames in the future (i.e. 130 ms) are enough for the network to decide if a variation in the eye appearance is due to a blink (recover the same appearance in the future), to a saccade (appearance changes in the future) or to some noise.

7.3 Experimental Setup

7.3.1 Baseline methods

Lacking direct comparison, we used baseline methods which use as input an XY gaze signal instead of eye images. Experiments were made on the same dataset, extracting gaze direction from color eye images using a multi-level Histogram of Oriented Gradient based Support vector Regression (Funes Mora and Odobez, 2016) trained on a separate dataset (Eyediap). This method was shown to deliver state-of-the-art performance on low-resolution images involving large head rotation. Note that it does not explicitly detect blinks, but it usually generates a down-up pattern in pitch.

- *Dispersion-Threshold Identification (I-DT)* (Salvucci and Goldberg, 2000). This classic

method distinguishes fixations and saccades by measuring dispersion in a moving time window. Parameters were trained to maximize the Cohen's kappa.

- *Naive Segmented Linear Regression (NSLR-HMM)* (Pekkanen and Lappi, 2017). First, a Segmented Linear Regression denoises and segments the signal, then an Hidden Markov Model classifies the obtained segments into fixation, saccade, smooth pursuit, and post-saccadic oscillations (PSO) classes. We used the implementation and the trained model provided by the author (https://gitlab.com/nslr/) and considered smooth-pursuit and PSO as fixation to allow comparison with our method.
- *Convolutional Neural Network (FFT-CNN)* (Hoppe and Bulling, 2016). This CNN uses the Fast Fourier Transform of the XY gaze signal as input to classify time windows as fixation, saccade or smooth pursuit. We implemented and trained this model from scratch.

7.3.2 Experimental protocol

Dataset

We used the video recordings from the KTH-Idiap dataset. As mentioned in Sec. 4.2, eye movements were annotated on all 20 videos based on 5 classes: fixation, blink during a fixation (fix-blink), saccade, blink during a saccade (sac-blink) and unknown.

We used the "leave one subject out" protocol, ignoring frames labeled as "unknown". To get balanced classes for training we applied down- and up-sampling on the 20 training videos to obtain 20'000 frames for each class, including frames with interpolated labels. However, we used neither interpolation nor resampling on the video used for testing.

Performance measure

To compare methods' performances, we relied on the Cohen's kappa (Cohen, 1960), which measures agreement for classifications taking into account the probability of random agreements which is especially important in unbalanced datasets:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{7.1}$$

where p_o is the observed agreement probability and p_e the probability of random agreement. Usually, the agreement is considered as weak if $\kappa > 0.2$, as moderate if $\kappa > 0.4$ and as strong if $\kappa > 0.6$. It was computed on each video, to get the variance across subjects.

Tasks

As we compare methods that do not extract all the same eye movements, we need to define a way to combine and/or ignore some labels. We defined the following tasks:

Table 7.1 – Evaluation of methods.

Method	Task	κ mean	κ std
our	4 classes	.536	.093
our	fix-sac-blink	.552	.091
our	blink-others	.671	.104
our	fix-sac	.501	.130
FFT-CNN (Hoppe and Bulling, 2016)	4 classes	.417	.062
FFT-CNN (Hoppe and Bulling, 2016)	fix-sac-blink	.431	.059
FFT-CNN (Hoppe and Bulling, 2016)	blink-others	.297	.064
FFT-CNN (Hoppe and Bulling, 2016)	fix-sac	.480	.122
I-DT (Salvucci and Goldberg, 2000)	fix-sac	.306	.095
NSLR-HMM (Pekkanen and Lappi, 2017)	fix-sac	.369	.065

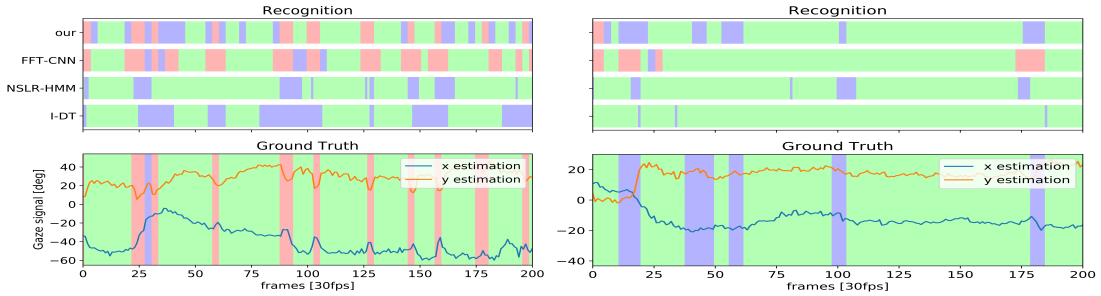


Figure 7.4 – Qualitative comparison over 2 segments. Recognition for four different methods including ours (top), with labels indicated by colors (green: fixation, blue: saccade, pink: blink). and groun truth (bottom), together with the estimated gaze signal (yaw and pitch).

- **4 classes.** 4 classes: fixation, fix-blink, saccade and sac-blink;
- **fix-sac-blink.** 3 classes: fixation, saccade and blink (fix-blink and sac-blink are merged in blink);
- **blink-others.** 2 classes: blink (i.e. fix-blink and sac-blink) and others, which is the combination of fixation and saccade;
- **fix-sac.** 2 classes: fixation and saccade. Fix-blink and sac-blink frames in the ground truth are ignored for evaluation and frames recognized as fix-blink/sac-blink are considered as fixation/saccade respectively.

7.4 Results

Method evaluation

Results are reported in Tab. 7.1. Our method achieves an overall moderate agreement with the ground truth, which is a good result given the difficulty of the tasks. It seems particularly suited to detect blink (blink-others task) but saccade recognition is more challenging (fix-sac

task). Note that the standard deviations show that performances variate across subjects.

Comparison with baselines

Looking at baselines performances in Tab. 7.1, one can notice the overall low scores of all methods, highlighting again the difficulty of the task. The FFT-CNN method reaches lower performance than our method, although it is not significant for the fix-sac one. It struggles to distinguish saccades from blinks, which is consistent with the intuition that blinks are better handled using eye images than the gaze signal. Per-class accuracy validates this result: FFT-CNN reaches 86% accuracy on fixation but only 33% for saccades (versus 81% and 77% for our method), often confusing saccade and blinks. It shows that (1) deep learning seems an appropriate approach as FFT-CNN and our method beat the two other baselines and (2) that using eye images helps to detect blinks, while not decreasing saccade detection performances.

Qualitative results

Examples of recognition are presented in Fig. 7.4. The left side of Fig. 7.4 presents an example in which all classifiers detect most of the blinks, although I-DT and NSLR-HMM are predicting saccades for blinks. I-DT tends to merge successive blinks and deep learning methods mistakenly predict saccades before and after blinks. Here, NSLR-HMM performs well. In the example on the right of Fig. 7.4, we can see that I-DT and FFT-CNN struggle to detect saccades. NSLR-HMM is already better but still misses two events. Those events correspond to small saccadic movements which are difficult to distinguish from noise in the gaze signal. It shows that using eye images helps to recognize subtle saccades that would be mixed with noise in the gaze signal.

7.5 Discussions and limitations

Our method relies on future information, creating a delay of about 130 ms between frame acquisition and eye movement estimation. Some applications will suffer from this, but instant reactivity is not always needed, like for off-line analysis, global statistics computation (blink rate) and low-frequency behavior estimation (attention).

One limitation of our work is the precision of the annotations, as the sampling rate of the sensor used to record the data is relatively slow compared to the events we want to detect. Considering the available data and annotations, chosing a training loss and a metric based on events rather than frames might be more appropriate. For example, we could use a many-to-one recurrent network with the ground truth being the presence or absence of a given eye movement in the input sequence. Regarding metrics, intesection over union might be interesting (Startsev et al., 2019).

Also, we compared our approach with baseline methods that were not designed for the exact

same task, in term of detected eye movements or data quality. That shows an advantage of deep learning methods, which can be retrained on a different set of labels for direct comparison. Moreover, all the proposed baselines rely on the same gaze estimation method, chosen because it tends to react consistently to eye movements. It would be interesting to make experiments with more recent methods to see if those baseline methods can be improved.

Finally, regarding performance, most errors of our method consist in predicting saccades instead of fixations around blinks. It might be interesting to check whether a temporal method, like Hidden Markov Model or Long Short-Term Memory (LSTM) cells, could help to better learn the label transition statistics and feature dynamics. However, the few tests we made using LSTM were not conclusive, showing that it is not straight forward.

7.6 Conclusion

In this chapter, we proposed a method based on computer vision and deep learning to detect fixation, saccade, and blink in natural interaction video recorded with remote sensors. We showed that deep learning approaches outperform classical methods for saccade detection when facing noisy data coming from computer vision methods instead of dedicated IR eye tracker sensors. Also, our method outperforms another deep learning approach on blink detection task, using eye images instead of gaze signal.

Finally, the overall performances obtained on the presented dataset show that detecting eye movements in low-sampling rate data acquired with remote sensors in natural conditions remains a challenging task, although it is of high interest in many fields.

8 Conclusion

In this last chapter, we recall the thesis contributions and discuss the limitations as well as the perspectives of the presented work.

8.1 Summary

The main objective of this thesis was to improve the accuracy and the usability of gaze and VFOA estimation in weakly constrained settings relying on consumer sensors. In this regard, we addressed the following challenges:

- **Remote sensors and dynamic environment.** The accuracy of gaze estimation, VFOA estimation, and eye movement recognition significantly improved in the last year, thanks to new sensors and methods. However, the application of these methods is challenging and the best results are still achieves using restrictive devices, like head-mounted devices or screen-based eye trackers. Also, most gaze datasets consist of recordings of people looking passively at visual stimuli. Similarly, most VFOA datasets consist of static setups, with a fixed number of participants who are not moving in the scene. By working on three datasets where people acting naturally were recorded with remote sensors, we could evaluate the accuracy of gaze estimation, VFOA estimation, and eye movement recognition in this kind of challenging scenario. While we did not investigate VFOA estimation in setups with moving people, the proposed approach demonstrated promising results when the setup is not the same during training and testing.
- **Environments and users diversity.** The ultimate goal of human sensing methods is to train models that can be applied in diverse environments with a minimal amount of adaptation, which is still difficult to achieve today. Also, diversity in appearances and behaviors across people makes it more challenging to learn models with good generalization capabilities. By relying on a background method that was trained in a separate dataset (and thus different people) and by proposing cross-datasets experiments we could see how the presented methods behave under this challenging condition.

8.2 Contributions

In the following, we recall our contributions regarding the three tasks we were interested in.

Gaze estimation. Due to invisible anatomical differences, even the best gaze estimation methods suffer a person-specific bias. To compensate it, it is necessary to collect calibration points, i.e. user's eye images which a known gaze. This collection is usually performed in a dedicated session previous to the interaction with the system, but such an approach is cumbersome in applications where the system encounter short interactions will multiple people, e.g. for information panels or robots in shopping malls, or when making such a calibration session could bias the interaction, e.g. for psychological assessment.

In Chapter 5, we proposed a comprehensive approach relying on context-based priors to get weak VFOA labels, which allow collecting calibration samples. Gaze calibration is then performed using a robust estimator to account for the inaccuracy of the weak VFOA labeling. This approach was evaluated on three datasets, with two different tasks, namely conversation and object manipulation. We show that the proposed method consistently improves the results, given that the chosen calibration model fits the task. After calibration, gaze estimation reached an 8° to 12° angular error, which corresponds to the state-of-the-art results for cross-datasets experiments.

VFOA estimation. There are two main approaches to estimate the VFOA of people in 3D scenes. The first relies on estimating subject's features, typically head pose and gaze, and comparing them to the position of the potential visual targets. The other one consists in modeling the VFOA of all the people involved in the interaction, by taking into account features coming from everyone and context. There is a trade-off between these approaches: the first one is more flexible, as it is not related to a specific setup, but the second one can be more accurate, does not necessarily require the location of people, and is less dependent on the accuracy of the people's features.

In Chapter 6, we proposed a method that takes into account contextual information, but encodes all the features into a fixed-sized representation and can thus be applied to an arbitrary number of people and targets. Moreover, it can handle moving targets, as their position is explicitly given as input. Experiments in cross-datasets settings validated the ability of this method to generalize to a new setup. Also, it was shown better than a strong probabilistic baseline, which, despite being much simple, is comparable in terms of inputs and usability.

Eye movements recognition. Fixations and saccades recognition is usually performed by analyzing the gaze signal, which is usually acquired using precise, but invasive, eye trackers. However, remote sensors provide eye images with a lower resolution and a lower sampling rate, leading to less accurate gaze estimation, which in turn makes it more difficult to recognize eye movements. In Chapter 7, we proposed a method to recognize fixations, saccades, and blinks

directly from eye images. A CNN takes 9 consecutive eye images as input and outputs directly the eye movements probability. Doing so was shown more effective than processing the gaze signal, which suffers the noise inherited from the data quality and the potential inaccuracy of the gaze estimation method.

8.3 Limitations and perspectives

In the following, we recall the main limitations of the presented work and present some perspectives of future work.

Scene monitoring. The presented gaze calibration and VFOA estimation methods both rely on contextual information provided by scene monitoring, but we did not address how to get the position of the objects and people of interest, as well as their speaking status or the action they perform. Indeed, while scene monitoring can be easily set up in some applications, like in the UBImpressed or KTH-Idiap setup examples, other applications may not allow to spread sensors in the scene. It is typically the case with conversational robots, where all sensors are embedded on the robot and only capture part of the scene or in sequence (by moving the head). For this reason, it would be interesting to evaluate the presented methods on data recorded in partially monitored environments, for example by recording a conversation from a single point of view, with an RGB camera and a microphone array. However, extracting 3D positions and recognizing speakers from such sensors are challenging tasks in themselves. One possibility would be to simulate some uncertainties in feature extraction, to allow studying the robustness of the proposed approaches.

Contextual features. We relied on contextual features to select calibration samples and to improve VFOA estimation accuracy. In the conversation case, we relied mainly on the speaking status of people, but other information might improve the results. For example, we could take into account the head pose and/or gaze of the targets. Also, it could be interesting to detect head or hand gestures and to incorporate them in attention priors and 2D feature maps for VFOA estimation.

Manipulation. In this thesis, we presented some experiments in object manipulation setup, which could be improved in several ways. The pick and place gestures could be automatically detected instead of being manually annotated. Also, while the ManiGaze dataset provides a good way to benchmark gaze estimation from embedded sensors, the interaction is limited to conversation. Experiments could be done in other more collaborative activities, like handovers or actual learning by demonstration.

Context-based few-shot learning. In Chapter 5, we focus on the calibration samples collection and propose robust, but rather simple, calibration models. It would be interesting to use the proposed method together with a few-shot learning approach based on deep learning, as it was shown more effective than linear calibration models. However, the question of robustness

remains, as we not only need to estimate good calibration parameters but also to select which calibration sample are true and relevant.

Temporal aspect. The gaze and VFOA estimation methods used in this work process the input frame-by-frame. However, the eye dynamic is an important factor to understand gaze and eye movements, so taking into account time, e.g. by using temporal layers in the different neural networks, could improve the performances.

Eye movements and gaze estimation. In Chapter 7, we recognized fixations, saccades, and blinks with the improvement of gaze estimation as a motivation. Indeed, a possible future direction would be to estimate eye movements and gaze together, e.g. in a multi-task manner, to improve both estimations.

Bibliography

Admoni, H. and Scassellati, B. (2017). Social Eye Gaze in Human-Robot Interaction: A Review. *Journal of Human-Robot Interaction*, 6(1):25–53.

Anantrasirichai, N., Gilchrist, I. D., and Bull, D. R. (2016). Fixation identification for low-sample-rate mobile eye trackers. In *International Conference on Image Processing (ICIP)*, pages 3126–3130. IEEE.

Anas, E. R., Henriquez, P., and Matuszewski, B. J. (2017). Online eye status detection in the wild with convolutional neural networks. In *International Conference on Computer Vision Theory and Applications*, volume 7, pages 88–95. SciTePress.

Andersson, R., Larsson, L., Holmqvist, K., Stridh, M., and Nystrom, M. (2017). One algorithm to rule them all? an evaluation and discussion of ten eye movement event-detection algorithms. *Behavior Research Methods*, 49(2):616–637.

Andrist, S., Zhi Tan, X., Gleicher, M., and Mutlu, B. (2014). Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 25–32, New York, USA. ACM.

Anzalone, S. M., Xavier, J., Boucenna, S., Billeci, L., Narzisi, A., Muratori, F., Cohen, D., and Chetouani, M. (2019). Quantifying patterns of joint attention during human-robot interactions: An application for autism spectrum disorder assessment. *Pattern Recognition Letters*, 118:42–50.

Asteriadis, S., Karpouzis, K., and Kollias, S. (2014). Visual focus of attention in non-calibrated environments using gaze estimation. *International Journal of Computer Vision*, 107(3):293–316.

Azagra, P., Golemo, F., Mollard, Y., Lopes, M., Civera, J., and Murillo, A. C. (2017). A multimodal dataset for object model learning from natural human-robot interaction. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 6134–6141. IEEE.

Ba, S. and Odobez, J.-M. (2008a). Recognizing visual focus of attention from head pose in natural meetings. *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1).

Ba, S. O., Hung, H., and Odobez, J.-M. (2009). Visual activity context for focus of attention estimation in dynamic meetings. In *International Conference on Multimedia and Expo*, pages 1424–1427. IEEE.

Ba, S. O. and Odobez, J.-M. (2008b). Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *International Conference on Acoustics, Speech and Signal Processing*, pages 2221–2224. IEEE.

Ba, S. O. and Odobez, J.-M. (2011). Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):101–116.

Baccour, M. H., Driewer, F., Kasneci, E., and Rosenstiel, W. (2019). Camera-based eye blink detection algorithm for assessing driver drowsiness. In *Intelligent Vehicles Symposium (IV)*, pages 987–993. IEEE.

Bader, T., Vogelgesang, M., and Klaus, E. (2009). Multimodal integration of natural gaze behavior for intention recognition during object manipulation. In *Proceedings of the international conference on Multimodal interfaces*, pages 199–206.

Bai, C., Kumar, S., Leskovec, J., Metzger, M., Nunamaker, J., and Subrahmanian, V. S. (2019). Predicting the visual focus of attention in multi-person discussion videos. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 4504–4510. International Joint Conferences on Artificial Intelligence Organization.

Baio, J. (2018). Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. *MMWR. Surveillance Summaries*, 67.

Batrinca, L. M., Mana, N., Lepri, B., Pianesi, F., and Sebe, N. (2011). Please, tell me about yourself: automatic personality assessment using short self-presentations. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 255–262.

Bellet, M. E., Bellet, J., Nienborg, H., Hafed, Z. M., and Berens, P. (2019). Human-level saccade detection performance using deep neural networks. *Journal of neurophysiology*, 121(2):646–661.

Bennewitz, M., Faber, F., Joho, D., and Behnke, S. (2007). Fritz-a humanoid communication robot. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*, pages 1072–1077. IEEE.

Bohus, D. and Horvitz, E. (2009). Learning to predict engagement with a spoken dialog system in open-world settings. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '09, pages 244–252. Association for Computational Linguistics.

Bohus, D. and Horvitz, E. (2010). Computational models for multiparty turn taking. Technical report, Microsoft Research Technical Report MSR-TR 2010-115.

Boraston, Z. and Blakemore, S.-J. (2007). The application of eye-tracking technology in the study of autism. *The Journal of physiology*, 581(3):893–898.

Borji, A. and Itti, L. (2012). State-of-the-art in visual attention modeling. *Transactions on pattern analysis and machine intelligence*, 35(1):185–207.

Canévet, O., He, W., Motlicek, P., and Odobez, J.-M. (2020). The mummer data set for robot perception in multi-party hri scenarios. In *Proceedings of the 29th IEEE International Conference on Robot & Human Interactive Communication*.

Cao, Z., Martinez, G. H., Simon, T., Wei, S., and Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chau, M. and Betke, M. (2005). Real time eye tracking and blink detection with usb cameras. Technical report, Boston University Computer Science Department.

Chen, C., Heili, A., and Odobez, J.-M. (2011). A joint estimation of head and body orientation cues in surveillance video. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 860–867. IEEE.

Chen, M. C., Anderson, J. R., and Sohn, M. H. (2001). What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. In *CHI'01 extended abstracts on Human factors in computing systems*. ACM.

Chen, Z. and Shi, B. E. (2018). Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer.

Cheng, Y., Lu, F., and Zhang, X. (2018). Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *Proceedings of the European Conference on Computer Vision*, pages 100–115.

Cheng, Y., Wang, H., Bao, Y., and Lu, F. (2021). Appearance-based gaze estimation with deep learning: A review and benchmark.

Chennamma, H. and Yuan, X. (2013). A survey on eye-gaze tracking techniques. *Indian Journal of Computer Science and Engineering*, 4:388–393.

Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A., and Rehg, J. M. (2018). Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 383–398.

Chong, E., Wang, Y., Ruiz, N., and Rehg, J. M. (2020). Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5406.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cornia, M., Baraldi, L., Serra, G., and Cucchiara, R. (2016). A deep multi-level network for saliency prediction. In *23rd International Conference on Pattern Recognition (ICPR)*, pages 3488–3493. IEEE.

Cortacero, K., Fischer, T., and Demiris, Y. (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*.

Divjak, M. and Bischof, H. (2009). Eye blink based fatigue detection for prevention of computer vision syndrome. In *MVA*, pages 350–353.

Dong, L., Di, H., Tao, L., Xu, G., and Oliver, P. (2009). Visual focus of attention recognition in the ambient kitchen. In *Asian Conference on Computer Vision*, pages 548–559. Springer.

Dreißig, M., Baccour, M. H., Schäck, T., and Kasneci, E. (2020). Driver drowsiness classification based on eye blink and head movement features using the k-nn algorithm. In *IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE.

Duffner, S. and Garcia, C. (2016). Visual focus of Attention Estimation With Unsupervised Incremental Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(12):2264–2272.

Duque, A. and Vázquez, C. (2015). Double attention bias for positive and negative emotional faces in clinical depression: Evidence from an eye-tracking study. *Journal of Behavior Therapy and Experimental Psychiatry*, 46:107–114.

Feng, Y., Cheung, G., Tan, W.-t., and Ji, Y. (2011). Hidden markov model for eye gaze prediction in networked video streaming. In *International Conference on Multimedia and Expo*. IEEE.

Fischer, T., Chang, H. J., and Demiris, Y. (2018). Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–352.

Fletcher, L. and Zelinsky, A. (2009). Driver inattention detection based on eye gaze—road event correlation. *The international journal of robotics research*, 28(6).

Foster, M. E., Craenen, B., Deshmukh, A., Lemon, O., Bastianelli, E., Dondrup, C., Papaioannou, I., Vanzo, A., Odobez, J.-M., Canévet, O., et al. (2019). Mummer: Socially intelligent human-robot interaction in public spaces. In *Artificial Intelligence for Human-Robot Interaction Symposium (AI-HRI)*.

Foster, M. E., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., and Petrick, R. P. (2012). Two people walk into a bar: Dynamic multi-party social interaction with a robot agent. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 3–10.

Frauendorfer, D., Mast, M. S., Nguyen, L., and Gatica-Perez, D. (2014). Nonverbal social sensing in action: Unobtrusive recording and extracting of nonverbal behavior in social interactions illustrated with a research example. *Journal of Nonverbal Behavior*, 38(2):231–245.

Funes Mora, K. A., Monay, F., and Odobez, J.-M. (2014). Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proc. of the ACM Symposium on Eye Tracking Research and Applications*.

Funes Mora, K. A., Nguyen, L. S., Gatica-Perez, D., and Odobez, J.-M. (2013). A semi-automated system for accurate gaze coding in natural dyadic interactions. In *15th ACM International Conference on Multimodal Interaction*. ACM.

Funes Mora, K. A. and Odobez, J.-M. (2012). Gaze estimation from multimodal kinect data. In *Conference in Computer Vision and Pattern Recognition, Workshop on Gesture Recognition*. IEEE.

Funes Mora, K. A. and Odobez, J.-M. (2014). 3d gaze tracking and automatic gaze coding from rgb-d cameras. In *IEEE Conference in Computer Vision and Pattern Recognition, Vision Meets Cognition Workshop*.

Funes Mora, K. A. and Odobez, J.-M. (2016). Gaze estimation in the 3d space using rgb-d sensors. *International Journal of Computer Vision*, 118:194–216.

Gatica-Perez, D., Vinciarelli, A., and Odobez, J.-M. (2014). Nonverbal behavior analysis. *Multimodal interactive systems management*, pages 165–187.

Gorga, S. and Otsuka, K. (2010). Conversation scene analysis based on dynamic bayesian network and image-based gaze detection. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 1–8.

Grauman, K., Betke, M., Lombardi, J., Gips, J., and Bradski, G. R. (2003). Communication via eye blinks and eyebrow raises: Video-based human-computer interfaces. *Universal Access in the Information Society*, 2(4):359–373.

Guestrin, E. D. and Eizenman, M. (2006). General theory of remote gaze estimation using the pupil center andcorneal reflections. *Transactions on bio-medical engi-neering*, 53(6).

Hansen, D. W. and Ji, Q. (2009). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500.

Hayhoe, M. and Ballard, D. (2005). Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194.

He, W., Motlicek, P., and Odobez, J.-M. (2018a). Deep neural networks for multiple speaker detection and localization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 74–79.

He, W., Motlicek, P., and Odobez, J.-M. (2018b). Joint localization and classification of multiple sound sources using a multi-task neural network. In *Proceedings of Interspeech*, pages 312–316.

Hessels, R. S., Niehorster, D. C., Kemner, C., and Hooge, I. T. C. (2017). Noise-robust fixation detection in eye movement data: Identification by two-means clustering (i2mc). *Behavior Research Methods*, 49(5):1802–1823.

Holland, C., Garza, A., Kurtova, E., Cruz, J., and Komogortsev, O. (2013). Usability evaluation of eye tracking on an unmodified common tablet. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery.

Hoppe, S. and Bulling, A. (2016). End-to-end eye movement detection using convolutional neural networks.

Hoque, M., Courgeon, M., Martin, J.-C., Mutlu, B., and Picard, R. W. (2013). Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 697–706.

Huang, J.-B., Cai, Q., Liu, Z., Ahuja, N., and Zhang, Z. (2014). Towards accurate and robust cross-ratio based gaze trackers through learning from simulation. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 75–82.

Huang, M. X., Li, J., Ngai, G., and Leong, H. V. (2016). Stressclick: Sensing stress from gaze-click patterns. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1395–1404.

Hung, H., Huang, Y., Friedland, G., and Gatica-Perez, D. (2011). Estimating dominance in multi-party meetings using speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):847–860.

Isaac, L., Vrijsen, J. N., Rinck, M., Speckens, A., and Becker, E. S. (2014). Shorter gaze duration for happy faces in current but not remitted depression: Evidence from eye movements. *Psychiatry Research*, 218(1-2):79–86.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259.

Jayagopi, D. B., Sheiki, S., Klotz, D., Wienke, J., Odobez, J.-M., Wrede, S., Khalidov, V., Nyugen, L., Wrede, B., and Gatica-Perez, D. (2013). The vernissage corpus: A conversational human-robot-interaction dataset. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

Jianfeng, L. and Shigang, L. (2014). Eye-model-based gaze estimation by rgb-d camera. In *Conference on Computer Vision and Pattern Recognition Workshops*, pages 606–610. IEEE.

Johansson, R., Westling, G., Backstrom, A., and Flanagan, R. (2001). Eye-Hand Coordination in Object Manipulation. *Journal of Neuroscience*, 21(17):6917–6932.

Kar, A. and Corcoran, P. (2017). A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access*, 5:16495–16519.

Katsuki, F. and Constantinidis, C. (2014). Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist*, 20(5):509–521.

Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., and Torralba, A. (2019). Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta psychologica*, 26:22–63.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758.

Kleinke, C. L. (1986). Gaze and eye contact: a research review. *Psychological bulletin*, 100(1):78.

Kluttz, N., Mayes, B., West, R., and Kerby, D. (2009). The effect of head turn on the perception of gaze. *Vision Research*, 49(15):1979–1993.

Knapp, M. L., Hall, J. A., and Horgan, T. G. (2013). *Nonverbal communication in human interaction*. Cengage Learning.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., and Torralba, A. (2016). Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184.

Land, M. (2009). Vision, eye movements, and natural behavior. *Visual Neuroscience*, 26(1):51–62.

Langton, S., Honeyman, H., and Tessler, E. (2004). The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception & psychophysics*, 66(5):752–771.

Lappi, O., Lehtonen, E., Pekkanen, J., and Itkonen, T. (2013). Beyond the tangent point: gaze targets in naturalistic driving. *Journal of vision*, 13(13):11–11.

Larsson, L., Nystrom, M., and Stridh, M. (2013). Detection of saccades and postsaccadic oscillations in the presence of smooth pursuit. *IEEE Transactions on Biomedical Engineering*, 60(9):2484–2493.

Lee, W. J., Kim, J. H., Shin, Y. U., Hwang, S., and Lim, H. W. (2019). Differences in eye movement range based on age and gaze direction. *Eye*, 33(7):1145–1151.

Li, G. and Yu, Y. (2016). Visual saliency detection based on multiscale deep cnn features. *Transactions on image processing*, 25(11):5012–5024.

Li, N. and Busso, C. (2014). User independent gaze estimation by exploiting similarity measures in the eye pair appearance eigenspace. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 335–338. ACM.

Linden, E., Sjostrand, J., and Proutiere, A. (2019). Learning to personalize in appearance-based gaze tracking. In *Proceedings of the International Conference on Computer Vision Workshops*. IEEE.

Liu, G., Yu, Y., and Odobez, J.-M. (2020). A differential approach for gaze estimation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*.

Lupu, R. G., Ungureanu, F., and Siriteanu, V. (2013). Eye tracking mouse for human computer interaction. In *E-Health and Bioengineering Conference (EHB)*. IEEE.

Mansouryar, M., Steil, J., Sugano, Y., and Bulling, A. (2016). 3d gaze estimation from 2d pupil positions on monocular head-mounted eye trackers. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 197–200.

Marin-Jimenez, M. J., Kalogeiton, V., Medina-Suarez, P., and Zisserman, A. (2019). Laeo-net: revisiting people looking at each other in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3477–3485.

Martinez, F., Carbone, A., and Pissaloux, E. (2012). Gaze estimation using local features and non-linear regression. In *19th International Conference on Image Processing*, pages 1961–1964. IEEE.

Martínez-González, A., Villamizar, M., Canévet, O., and Odobez, J.-M. (2020a). Efficient convolutional neural networks for depth-based multi-person pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):4207–4221.

Martínez-González, A., Villamizar, M., Canévet, O., and Odobez, J.-M. (2020b). Residual pose: A decoupled approach for depth-based 3d human pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Masame, K. (1990). Perception of where a person is looking: Overestimation and underestimation of gaze direction. *Tohoku Psychologica Folia*, 49:33–41.

Masko, D. (2017). Calibration in eye tracking using transfer learning.

Mayberry, A., Hu, P., Marlin, B., Salthouse, C., and Ganesan, D. (2014). ishadow: design of a wearable, real-time mobile gaze tracker. In *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, pages 82–94.

Mohanakrishnan, J., Nakashima, S., Odagiri, J., and Yu, S. (2013). A novel blink detection system for user monitoring. In *1st Workshop on User-Centered Computer Vision (UCCV)*, pages 37–42. IEEE.

Moon, A., Troniak, D., Gleeson, B., Pan, M., Zheng, M., Blumer, B., MacLean, K., and Croft, E. (2014). Meet me where i'm gazing: How shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 334–341, New York, USA. ACM.

Morimoto, C. and Mimica, M. (2005). Eye gaze tracking techniques for interactive applications. *Computer vision and image understanding*, 98(1):4–24.

Muller, P., Buschek, D., Huang, M. X., and Bulling, A. (2019). Reducing calibration drift in mobile eye trackers by exploiting mobile phone usage. In *Proc. of the ACM Symp. on Eye Tracking Research & Applications*.

Müller, P., Huang, M. X., Zhang, X., and Bulling, A. (2018). Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–10.

Mulvey, F. (2011). Eye anatomy, eye movements and vision. In *Gaze interaction and applications of eye tracking: Advances in assistive technologies*, pages 10–20. IGI Global.

Muralidhar, S., Costa, J. M. R., Nguyen, L. S., and Gatica-Perez, D. (2016). Dites-moi: Wearable feedback on conversational behavior. In *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia*.

Muralidhar, S., Siegfried, R., Odobez, J.-M., and Gatica-Perez, D. (2018). Facing employers and customers: What do gaze and expressions tell about soft skills? In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*. ACM.

Newman, B. A., Aronson, R. M., Srinivasa, S. S., Kitani, K., and Admoni, H. (2018). Harmonic: A multimodal dataset of assistive human-robot collaboration. *arXiv preprint arXiv:1807.11154*.

Nguyen, L. S., Frauendorfer, D., Mast, M. S., and Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *Transactions on multimedia*, 16(4):1018–1031.

Ning, M., Daniels, J., Schwartz, J., Dunlap, K., Washington, P., Kalantarian, H., Du, M., and Wall, D. P. (2019). Identification and Quantification of Gaps in Access to Autism Resources in the United States: An Infodemiological Study. *Journal of Medical Internet Research*, 21(7):e13094.

Nystrom, M., Andersson, R., Holmqvist, K., and van de Weijer, J. (2013). The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods*, 45(1):272–288.

Oertel, C., Funes Mora, K. A., Gustafson, J., and Odobez, J.-M. (2015). Deciphering the silent participant: On the use of audio-visual cues for the classification of listener categories in group discussions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 107–114. ACM.

Oertel, C., Funes Mora, K. A., Sheikhi, S., Odobez, J.-M., and Gustafson, J. (2014). Who will get the grant?: A multimodal corpus for the analysis of conversational behaviours in group interviews. In *Proceedings of the workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, pages 27–32.

Oertel, C., Wlodarczak, M., Edlund, J., Wagner, P., and Gustafson, J. (2013). Gaze patterns in turn-taking. In *Annual Conference of the International Speech Communication Association*.

Otsuka, K., Kasuga, K., and Kohler, M. (2018). Estimating visual focus of attention in multi-party meetings using deep convolutional neural networks. In *29th British Machine Vision Conference*.

Otsuka, K., Yamato, J., Takemae, Y., and Murase, H. (2006). Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, pages 1175–1180.

Palazzi, A., Abati, D., Solera, F., Cucchiara, R., et al. (2018). Predicting the driver's focus of attention: the dr (eye) ve project. *Transactions on pattern analysis and machine intelligence*, 41(7):1720–1733.

Palmero, C., Selva, J., Bagheri, M. A., and Escalera, S. (2018). Recurrent cnn for 3d gaze estimation using appearance and shape cues. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 191–199.

Pan, J., Ferrer, C. C., McGuinness, K., O'Connor, N. E., Torres, J., Sayrol, E., and Giro-i Nieto, X. (2017). Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*.

Panning, A., Al-Hamadi, A., and Michaelis, B. (2011). A color based approach for eye blink detection in image sequences. In *International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 40–45. IEEE.

Park, S., Aksan, E., Zhang, X., and Hilliges, O. (2020). Towards end-to-end video-based eye-tracking. In *European Conference on Computer Vision*, pages 747–763. Springer.

Park, S., Mello, S. D., Molchanov, P., Iqbal, U., Hilliges, O., and Kautz, J. (2019). Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9368–9377.

Park, S., Zhang, X., Bulling, A., and Hilliges, O. (2018). Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proc. of the ACM Symp. on Eye Tracking Research & Applications*.

Patney, A., Salvi, M., Kim, J., Kaplanyan, A., Wyman, C., Benty, N., Luebke, D., and Lefohn, A. (2016). Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics*, 35(6).

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. (2009). A 3d face model for pose and illumination invariant face recognition. In *International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. IEEE.

Pekkanen, J. and Lappi, O. (2017). A new and general approach to signal denoising and eye movement classification based on segmented linear regression. *Scientific Reports*, 7(1):17726.

Pfeuffer, K., Vidal, M., Turner, J., Bulling, A., and Gellersen, H. (2013). Pursuit calibration: Making gaze calibration less tedious and more flexible. In *Proceedings of the Symposium on User Interface Software and Technology*. AM.

Pi, J. and Shi, B. E. (2019). Task-embedded online eye-tracker calibration for improving robustness to head motion. In *Proceedings of the ACM Symposium on Eye Tracking Research & Applications*.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Rehg, J., Abowd, G., Rozga, A., Romero, M., Clements, M., Sclaroff, S., Essa, I., Ousley, O., Li, Y., Kim, C., et al. (2013). Decoding children's social behavior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3414–3421.

Reiter-Palmon, R., Sinha, T., Gevers, J., Odobez, J.-M., and Volpe, G. (2017). Theories and models of teams and groups. *Small Group Research*, 48(5):544–567.

Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880.

Salam, H., Celiktutan, O., Hupont, I., Gunes, H., and Chetouani, M. (2016). Fully automatic analysis of engagement and its relationship to personality in human-robot interactions. *IEEE Access*, 5:705–721.

Salvucci, D. D. and Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78. ACM.

Sanchez-Cortes, D., Aran, O., Jayagopi, D. B., Schmid Mast, M., and Gatica-Perez, D. (2013). Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces*, 7(1-2):39–53.

Santini, T., Fuhl, W., and Kasneci, E. (2017). Calibme: Fast and unsupervised eye tracker calibration for gaze-based pervasive human-computer interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 2594–2605. ACM.

Santini, T., Fuhl, W., Kübler, T., and Kasneci, E. (2016). Bayesian identification of fixations, saccades, and smooth pursuits. In *Proceedings of the 9th Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 163–170. ACM.

Scherer, S., Stratou, G., Lucas, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A. S., and Morency, L.-P. (2014). Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, 32(10):648–658.

Sheikhi, S. and Odobez, J.-M. (2015). Combining dynamic head pose–gaze mapping with the robot conversational state for attention recognition in human–robot interactions. *Pattern Recognition Letters*, 66:81–90.

Sidenmark, L. and Gellersen, H. (2019a). Eye, head and torso coordination during gaze shifts in virtual reality. *Transactions on Computer-Human Interaction*, 27(1).

Sidenmark, L. and Gellersen, H. (2019b). Eye, head and torso coordination during gaze shifts in virtual reality. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(1):1–40.

Sidenmark, L. and Lundström, A. (2019). Gaze behaviour on interacted objects during hand interaction in virtual reality for eye tracking calibration. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 1–9.

Siegfried, R., Aminian, B., and Odobez, J.-M. (2020). Manigaze: a dataset for evaluating remote gaze estimator in object manipulation situations. In *Symposium on Eye Tracking Research and Applications*. ACM.

Siegfried, R. and Odobez, J.-M. (2017). Supervised gaze bias correction for gaze coding in interactions. In *ECEM COGAIN Symposium*.

Siegfried, R. and Odobez, J.-M. (2021a). Context-based unsupervised calibration for gaze estimation. *ACM Transactions on Multimedia Computing, Communications, and Applications*. (under publication process).

Siegfried, R. and Odobez, J.-M. (2021b). 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 Workshops*. IEEE.

Siegfried, R., Yu, Y., and Odobez, J.-M. (2017). Towards the use of social interaction conventions as prior for gaze model adaptation. In *Proceedings of 19th ACM International Conference on Multimodal Interaction*. ACM.

Siegfried, R., Yu, Y., and Odobez, J.-M. (2019). A deep learning approach for robust head pose independent eye movements recognition from videos. In *2019 ACM Symposium on Eye Tracking Research and Applications*. ACM.

Sigari, M.-H., Pourshahabi, M.-R., Soryani, M., and Fathy, M. (2014). A review on driver face monitoring systems for fatigue and distraction detection. *International Journal of Advanced Science and Technology*, 64:73–100.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, B., Yin, Q., Feiner, S., and Nayar, S. (2013). Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the ACM symposium on User interface software and technology*. ACM.

Soukupova, T. and Cech, J. (2016). Eye blink detection using facial landmarks. In *21st computer vision winter workshop, Rimske Toplice, Slovenia*.

Startsev, M., Göb, S., and Dorr, M. (2019). A novel gaze event detection metric that is not fooled by gaze-independent baselines. In *Proceedings of the 11th ACM symposium on eye tracking research & applications*, pages 1–9.

Stiefelhagen, R., Yang, J., and Waibel, A. (2002). Modeling focus of attention for meeting indexing based on multiple cues. *Transactions on Neural Networks*, 13(4):928–938.

Strabala, K., Lee, M. K., Dragan, A., Forlizzi, J., Srinivasa, S. S., Cakmak, M., and Micelli, V. (2013). Toward seamless human-robot handovers. *Journal of Human-Robot Interaction*, 2(1):112–132.

Sugano, Y., Matsushita, Y., and Sato, Y. (2014). Learning-by-synthesis for appearance-based 3d gaze estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 1821–1828. IEEE.

Sugano, Y., Matsushita, Y., Sato, Y., and Koike, H. (2015). Appearance-based gaze estimation with online calibration from mouse operations. *Transactions on Human-Machine Systems*, 45(6):750–760.

Sundaram, D. S. and Webster, C. (2000). The role of nonverbal communication in service encounters. *Journal of Services Marketing*.

Tafaj, E., Kasneci, G., Rosenstiel, W., and Bogdan, M. (2012). Bayesian online clustering of eye movement data. In *Proceedings of the symposium on eye tracking research and applications*, pages 285–288.

Valenti, R., Sebe, N., and Gevers, T. (2011). Combining head pose and eye location information for gaze estimation. *Transactions on Image Processing*, 21(2):802–815.

Veneri, G., Piu, P., Rosini, F., Federighi, P., Federico, A., and Rufa, A. (2011). Automatic eye fixations identification based on analysis of variance and covariance. *Pattern Recognition Letters*, 32(13):1588–1593.

Vidal, M., Turner, J., Bulling, A., and Gellersen, H. (2012). Wearable eye tracking for mental health monitoring. *Computer Communications*, 35(11):1306–1311.

Voit, M. and Stiefelhagen, R. (2008). Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 173–180.

Wang, K., Su, H., and Ji, Q. (2019). Neuro-inspired eye tracking with eye movement dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9831–9840.

Xia, Y., Zhang, D., Kim, J., Nakayama, K., Zipser, K., and Whitney, D. (2018). Predicting driver attention in critical situations. In *Asian conference on computer vision*, pages 658–674. Springer.

Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., Samaras, D., and Hoai, M. (2020). Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 193–202.

Yu, Y., Funes Mora, K. A., and Odobez, J.-M. (2018a). Headfusion: 360 degree head pose tracking combining 3d morphable model and 3d reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11).

Yu, Y., Liu, G., and Odobez, J.-M. (2018b). Deep multitask gaze estimation with a constrained landmark-gaze model. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.

Yu, Y., Liu, G., and Odobez, J.-M. (2019). Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11937–11946.

Yu, Y. and Odobez, J.-M. (2020). Unsupervised representation learning for gaze estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yücel, Z., Salah, A. A., Meriçli, Ç., Meriçli, T., Valenti, R., and Gevers, T. (2013). Joint attention by gaze interpolation and saliency. *Transactions on cybernetics*, 43(3):829–842.

Zemblys, R., Niehorster, D. C., Komogortsev, O., and Holmqvist, K. (2016). Using machine learning to detect events in eye-tracking data. *Behavior Research Methods*, 50(1):160–181.

Zhang, L., Morgan, M., Bhattacharya, I., Foley, M., Braasch, J., Riedl, C., Foucault Welles, B., and Radke, R. J. (2019a). Improved visual focus of attention estimation and prosodic features for analyzing group interactions. In *International Conference on Multimodal Interaction*, pages 385–394.

Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., and Hilliges, O. (2020). Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, pages 365–381. Springer.

Zhang, X., Sugano, Y., and Bulling, A. (2019b). Evaluation of appearance-based methods and implications for gaze-based applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM.

Zhang, X., Sugano, Y., Fritz, M., and Bulling, A. (2015). Appearance-based gaze estimation in the wild. In *Proceedings of the conference on computer vision and pattern recognition*, pages 4511–4520.

Zhang, X., Sugano, Y., Fritz, M., and Bulling, A. (2017a). 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*, pages 51–60.

Zhang, X., Sugano, Y., Fritz, M., and Bulling, A. (2017b). Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *Transactions on Pattern Analysis and Machine Intelligence*, 41(1):162–175.

Zhou, D., Luo, J., Silenzio, V., Zhou, Y., Hu, J., Currier, G., and Kautz, H. (2015). Tackling mental health by integrating unobtrusive multimodal sensing. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 1401–1408. AAAI Press.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

Rémy SIEGFRIED

Engineer in microengineering

Swiss nationality

remysieg@gmail.com

Rue de la Moya 8, 1920 Martigny

079 350 50 43

rsieg.ch

Education

PhD – Electrical engineering

EPFL, 2017 – June 2021

Thesis: ''Infering and modelling attention in human behaviors or for human-robot interactions''
Attended lectures: machine learning for engineers, deep learning, bayesian optimization

Master (MSc) – Microengineering

EPFL, 2014 – 2016

Specialization: Robotic and autonomous systems

Mobile robotic, Artificial intelligence, Data analytics, Mecanics, Electronics, Control, Computer science

Work experience

Research assistant – PhD student

Idiap Research Institute, 2017 – June 2021

Joined the « Perception and activity understanding » group in the frame of the european MuMMER project (see below)

Focus: gaze and attention estimation, eye movements classification, RGB and depth cameras

Soft skills: litterature review, academical writing, presentation

Publications: 6 international conferences (4 as main author), 1 international journal

Research assistant – Intern

EPFL (MOBOTS group), 2016 (6 monthes)

Hired after my master project to enhance robotic teaching using Thymio, an educational robot
Focus: simulator integration for online code evaluation
Soft skills: experiment design and organisation
Publication: 1 international conference (main author)

R&D engineer – Intern

SenseFly SA, 2015 (7 monthes)

Focus: development of drivers for a quadrotor's autopilot, electronical adaptation of a camera for a glider drone
Soft skills: flexibility, adaptation

Other activities

Judo club in Saint-Maurice

Practice Judo for 25 years (2nd Dan)
Technical director (since 2015) and teacher (since 2009)
Member of the regional technical comittee (since 2015)
Soft skills: teaching, coaching, work in comittee

Compulsory swiss military service

Staff officier at the rank of captain (since 2015)
Main activities: planification, organisation, and teaching (1 month per year, end in 2022)
Soft skills: management, work under time pressure

Hobbies and interests

Judo (2nd dan), Iaido (Ikkyu)
Video games, table-top roleplaying games

Technology, robotic
Psychology, learning and teaching techniques

Other research activities

Teaching Assistance

Idiap - UniDistance, 2021

« Multimodal computational sensing of people » lecture
Prepared and assessed one lab

Reviews

ICMI – 2018, 2020, 2021
IROS – 2020
TVCJ – 2020

Last research projects

MultiModal Mall Entertainment Robot (MuMMER) – EU H2020
http://mummer-project.eu
Goal: Development of a perception module for a conversational robot for information and entertainment purposes
My task: Visual focus of attention estimation from RGB data recorded through embedded sensors

Ubiquitous First Impressions and Ubiquitous Awareness (UBImpressed) - SNF
https://www.idiap.ch/project/ubimpressed
Goal: Psychological study focused on the building of first impression and its application in hospitality employees training
My task: Gaze estimation from RGB-D data recorded during dyadic interactions

Most relevant publications

R. Siegfried and J.-M. Odobez, ***Robust Unsupervised Gaze Calibration using Conversation and Manipulation Attention Priors***, accepted for publication in ACM Transactions on Multimedia Computing, Communications, and Applications

R. Siegfried and J.-M. Odobez***, Visual Focus of Attention Estimation in 3D Scene with an Arbitrary Number of Targets***, in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, June 2021

R. Siegfried et al., ***ManiGaze: a Dataset for Evaluating Remote Gaze Estimator in Object Manipulation Situations***, in ACM Symposium on Eye Tracking Research and Applications (ETRA), Stuttgart, June 2020

R. Siegfried et al., ***A Deep Learning Approach for Robust Head Pose Independent Eye Movements Recognition from Videos***, in ACM Symposium on Eye Tracking Research & Applications (ETRA), Denver, June 2019

R. Siegfried et al., ***Towards the Use of Social Interaction Conventions As Prior for Gaze Model Adaptation***, in ACM International Conference on Multimodal Interaction (ICMI), Glasgow, November 2017

R. Siegfried et al., ***Improved mobile robot programming performance through real-time program assessment***, in ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE), Bologna, July 2017

Skills

Programming	**C, C++, Python**
Libraries & frameworks	**ROS, pyTorch, Keras, JupyterHub**
Software	**Microsoft Office, LaTex, Matlab, git**
French	**Mother tongue**
English	**Fluent (worked for 5 years in English)**
German	**B2 (Goethe-Institut)**

REFERENCES

Dr. Jean-Marc ODOBEZ
(Thesis director)
EPFL MER and senior researcher at Idiap
odobez@idiap.ch
+41 27 721 77 26

Prof. Dr. Francesco Mondada
Prof. at EPFL and Head of the MOBOTS group
francesco.mondada@epfl.ch
+41 21 693 73 57